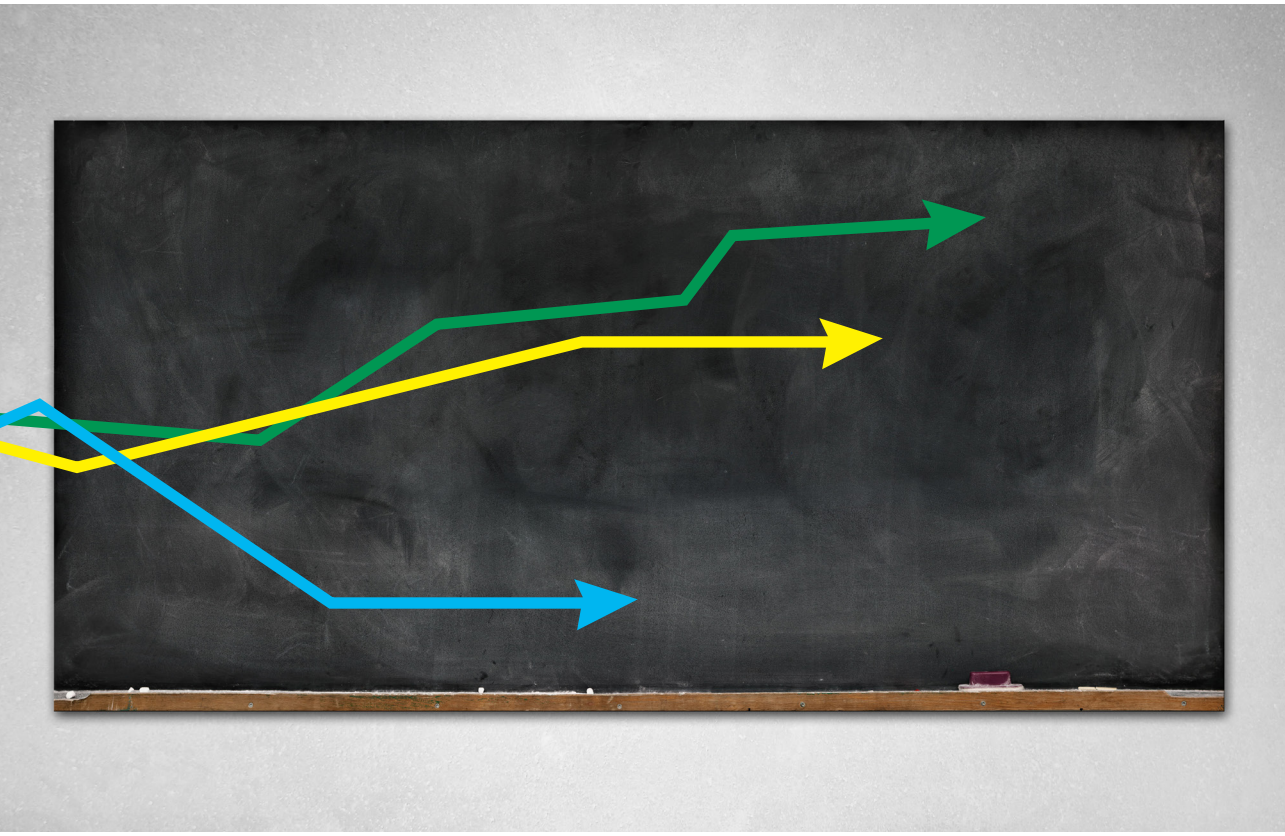
Learning Analytics & Educational Data Mining for Inquiry-Based Learning



Mehrnoosh Vahdat

Learning Analytics and Educational Data Mining for Inquiry-Based Learning

Mehrnoosh Vahdat

Vahdat, Mehrnoosh

Learning analytics and educational data mining for inquiry-based learning
Technische Universiteit Eindhoven, 2017.-Proefschrift-

A catalogue record is available from the Eindhoven University of Technology library
ISBN: 978-90-386-4240-6

Keywords: Learning analytics / Educational data mining / Inquiry-based learning / Machine learning / Rademacher complexity / Algorithmic stability / Process Mining / Cluster analysis / Concept learning / Simulation / Puzzle games

Typeset with LaTeX
Printed by proefschriftmaken.nl, The Netherlands

Learning Analytics and Educational Data Mining for Inquiry-Based Learning

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de
Technische Universiteit Eindhoven, op gezag van de
rector magnificus prof.dr.ir. F.P.T. Frank Baaijens, voor een
commissie aangewezen door het College voor
Promoties, in het openbaar te verdedigen
op maandag 3 april 2017 om 16.00 uur

door

Mehrnoosh Vahdat

geboren te Teheran, Iran

Dit proefschrift is goedgekeurd door de promotoren en de samenstelling van de promotiecommissie is als volgt:

voorzitter:	prof.dr.ir. J.H. Eggen
1^{e} promotor:	prof.dr. D. Anguita (Università degli Studi di Genova)
2^{e} promotor:	prof.dr. G.W.M. Rauterberg
1^{e} co-promotor:	dr. M. Funk
2^{e} co-promotor:	dr. L. Oneto (Università degli Studi di Genova)
leden:	dr. H. Drachsler (Open Universiteit)
	dr. E. Lavoué (Université Jean Moulin Lyon 3)
	prof.dr. M. Pechenizkiy

Het onderzoek of ontwerp dat in dit proefschrift wordt beschreven is uitgevoerd in overeenstemming met de TU/e Gedragscode Wetenschapsbeoefening.

UNIVERSITÀ DEGLI STUDI
DI GENOVA

This dissertation was produced under Erasmus Mundus Joint Doctorate Program in Interactive and Cognitive Environments. The research was conducted towards a joint double PhD degree between the following partner universities:

Università degli Studi di Genova

&

Technische Universiteit Eindhoven

ICE PhD Acknowledgements

This PhD Thesis has been developed in the framework of, and according to, the rules of the Erasmus Mundus Joint Doctorate on Interactive and Cognitive Environments EMJD ICE [FPA n^o 2010-0012] with the cooperation of the following Universities:

According to ICE regulations, the Italian PhD title has also been awarded by the Università degli Studi di Genova.

Acknowledgments

First of all, I would like to specially thank my partner Remi Brochenin for his continuous support and encouragement in this PhD adventure. Thank you for your patience in the never-ending discussions about my research and for teaching and helping me understand various aspects of Math and Computer Science.

This PhD research was carried out in two partner universities, under the guidance of four supervisors and co-supervisors. I would like to express my sincere gratitude to Davide Anguita and Luca Oneto from the University of Genoa for their invaluable support of my PhD research, for their significant insights, motivation, and immense knowledge. Also, my sincere thanks go to Mathias Funk and Matthias Rauterberg from the Eindhoven University of Technology whom guidance and constant feedback were extremely valuable during my entire PhD and writing of this thesis.

Besides my supervisors, I would like to thank all my colleagues in SmartLab (at UNIGE) and the Design Intelligence group (at TU/e) for helping me understand many details of the PhD path. In particular, I am grateful to Jorge Luis Reyes Ortiz and Isah Lawal who were open to my questions at any time. And indeed, Maira Brandao Carvalho, I agree with you! We made a great research team together.

I would like to thank all the professors of the University of Genoa and my fellow labmates for their precious support to the experiments of my research. In particular, I thank Domenico Ponta, Giuliano Donzellini, Emanuele Fumeo, Alessandro Ghio, Ilenia Orlandi, and Marjan Asadi. Also, I thank the students of UNIGE and TU/e for their participation in the experiments which helped me get results of better quality.

Finally, I would like to acknowledge my family and friends for their unconditional support during my professional and life experiences.

Learning Analytics and Educational Data Mining
for Inquiry-Based Learning

Summary

The growing interest in recent years towards Learning Analytics (LA) and Educational Data Mining (EDM) has motivated the development of novel approaches and advancements in educational settings. The wide variety of research and practice in this context has enforced important possibilities and applications from adaptation and personalization of Technology Enhanced Learning (TEL) systems to the improvement of instructional design and pedagogy choices based on students' needs. LA and EDM play an important role in enhancing learning processes by offering innovative applications of analytics methods. This leads to the knowledge discovery about the learning processes, and development and integration of more personalized, adaptive, and interactive educational environments. Inquiry-based learning (IBL) environments are considered as promising TEL environments to increase the knowledge and skills of learners. IBL focuses on contexts where learners are meant to discover knowledge rather than passively memorizing the concepts. LA and EDM are gaining attention in IBL contexts as a way to help facilitate learning and improve learning achievements of the students.

In this thesis, we aim to present novel applications of LA and EDM focused on IBL contexts. In particular, we aim to address what analytics methods can quantify the learning processes in an IBL cycle. We consider a learner-centered inquiry cycle as a structure to explain our objectives regarding three educational contexts. This cycle comprises of three main learning phases: 1 - conceptualization (generating hypothesis and question), 2 - investigation and discovery, 3 - conclusion and reflection. We focus on each phase in a different educational context through the application of LA and EDM. The three educational contexts are concept learning, simulation-based learning, and game-based puzzle-solving as follows.

- In the first part of this thesis, we perform an empirical study in the context of **human concept learning** where we investigate the learners' hypothesis creation (the first phase of IBL cycle). We apply Machine Learning that is usually exploited as a tool for analyzing data coming from experimental studies, but it has been recently applied to humans as if they were algorithms that learn from data. One example is the application of Rademacher Complexity, which measures the capacity of a learning machine, to human learning. In this line of research, we propose a more powerful measure of complexity, the Human Algorithmic Stability, as a tool to better understand the learning process of humans in particular their hypothesis creation. The results from three different experiments, with more than 600 engineering students

from the University of Genoa, shows the advantages of Algorithmic Stability over Rademacher Complexity for a better understanding of the human learning process.

- In the second part, we perform an empirical study in the context of **simulation-based learning** where we study the learner's investigation and discovery behavior (the second phase of IBL cycle). We propose an analytics approach based on Process Mining (PM) and the Cyclomatic Complexity metric (CM) to gain insight into the learning processes of students from their interaction data. We collected data through six laboratory sessions where first-year students of Computer Engineering were using a digital electronics simulator called *Deeds*. This study shows the capabilities of PM in combination with CM in explaining the properties of the learning behavior.

- In the third part, we perform an empirical study in the context of **game-based puzzle-solving** where we investigate the learners' conclusion and reflection (the third phase of IBL cycle). In this study, we investigate the use of LA and EDM in digital puzzle games to explore the way players learn game skills and solve problems in an open-source puzzle game called *Lix*. We performed an experiment with 15 participants, who played one puzzle for a total of 272 times. We applied PM and cluster analysis, given the resulting event log, in a three-step analysis approach. Our results indicate that the discovered process models are representative of players' tactics, as the members of each cluster converged to their cluster reference. This approach can be used as a basis for recommending interventions so as to facilitate the puzzle-solving process of players.

In conclusion, this thesis presents three novel applications of LA and EDM methods in the IBL cycle for three different educational contexts. The findings of each empirical study raise awareness about the IBL phases of learners, from developing inquiries and hypothesis generation to performing experiments for testing the hypothesis and explaining the results of their investigation process. Our findings can be used as the basis for generating feedback and recommendations to the stakeholders. Teachers, learners, TEL designers, and researchers are potential stakeholders of this thesis who can benefit from the knowledge discovered through LA and EDM to improve and facilitate the learners' IBL process.

Contents

1 Introduction

1.1 Motivation 1
1.2 Main Contributions 3
1.3 Thesis Outline 4

2 Learning Analytics and Educational Data Mining

2.1 Introduction 7
2.2 What are LA and EDM? 8
2.3 Inquiry-Based Learning 10
2.3.1 Phase One: Generating Hypothesis and Question 12
2.3.2 Phase Two: Investigation and Discovery 13
2.3.3 Phase Three: Conclusion and Reflection 14
2.4 Learning Contexts 14
2.4.1 Concept Learning 14
2.4.2 Simulation-Based Learning 16
2.4.3 Game-Based Puzzle-Solving 16
2.5 LA and EDM for IBL 17
2.6 Analytics Methods and Data Features 18
2.6.1 Classification 20
2.6.2 Clustering 20
2.6.3 Process Mining 21
2.7 Summary 22

3 State of the Art

3.1 Introduction 23
3.2 Overview on Applications 24
3.2.1 Classification in Education 26
3.2.2 Clustering in Education 28
3.2.3 Process Mining in Education 29
3.3 Evidences from Learning Contexts 30
3.3.1 Concept Learning 31
3.3.2 Simulation-Based Learning 31

3.3.3 Game-Based Puzzle-Solving 32
3.4 Summary 33

4 IBL Phase One: Application of Machine Learning in Concept Learning
4.1 Introduction 35
4.2 Rademacher Complexity and Algorithmic Stability in Machine Learning . 38
4.2.1 Understanding Learning Ability through Rademacher Complexity 41
4.2.2 Understanding Learning Ability through Algorithmic Stability . . 43
4.3 From Machine Learning to Human Learning 45
4.3.1 Human Error 46
4.3.2 Human Rademacher Complexity 47
4.3.3 Human Algorithmic Stability 48
4.4 Experimental Design 49
4.4.1 Experimental Design: EX^1 50
4.4.2 Experimental Design: EX^2 55
4.4.3 Experimental Design: EX^3 56
4.5 Results 57
4.5.1 Results for EX^1 58
4.5.2 Results for EX^2 60
4.5.3 Results for EX^3 63
4.6 Discussion 63
4.7 Summary 64

5 IBL Phase Two: Application of Process Mining in Simulation-Based Learning
5.1 Introduction 67
5.2 Process Mining and Cyclomatic Complexity Metric 68
5.2.1 Process Mining 68
5.2.2 Event Data 69
5.2.3 Fuzzy Miner: A Process Mining Algorithm 70
5.2.4 Cyclomatic Complexity Metric 71
5.3 Simulator Description 73
5.4 Experimental Design and Data Collection 74
5.5 Learning Analytics Approach 76
5.6 Data Set 79
5.6.1 Feature Selection 79
5.6.2 Activity Selection and Mapping 80
5.7 Results 82
5.7.1 Results from Process Mining 82
5.7.2 Results from Cyclomatic Complexity Metric 88

5.7.3 Feedback from Instructors . . . 92
5.8 Discussion . . . 94
5.9 Summary . . . 96

6 IBL Phase Three: Application of Data Mining in Game-Based Puzzle-Solving
6.1 Introduction . . . 99
6.2 Game Description . . . 100
6.3 Experimental Design and Data Collection . . . 102
6.4 Learning Analytics Approach . . . 103
6.4.1 Cluster Analysis of Tactics . . . 104
6.4.2 Process Mining of Clusters . . . 107
6.4.3 Validation . . . 108
6.5 Results . . . 108
6.5.1 Results from Cluster Analysis . . . 108
6.5.2 Results from Process Mining . . . 110
6.5.3 Validation of Results through Convergence . . . 111
6.6 Discussion . . . 112
6.7 Summary . . . 114

7 Discussion and Conclusions
7.1 Proposed Framework for Applying the Analytics Methods . . . 115
7.2 Proposed Analytics Methods for IBL phases . . . 118
7.3 Limitations . . . 120
7.4 Future Work . . . 122

Appendix A EPM Data Set
A.1 Data Set Description . . . 125
A.2 Features . . . 126
A.3 Grades Data . . . 127
A.4 Exercises . . . 127

Appendix B Lix Data Set
B.1 Data Set Description . . . 129
B.2 Features . . . 129
B.3 Actions . . . 130
B.4 Gameplay . . . 131

Bibliography

Glossary

List of Publications

Chapter 1

Introduction

1.1 Motivation

In recent years, there has been increasing interest in Learning Analytics (LA) and Educational Data Mining (EDM) among both researchers and practitioners of the Technology-Enhanced Learning (TEL) field. Due to the advances in the computer-assisted learning systems and automatic analysis of educational data, many efforts have been carried out in order to enhance the learning achievement (Chatti et al., 2012). In 2011, the Horizon report claims for a fruitful future of LA (Johnson et al., 2011) and considers LA as a great help to discover the hidden information and patterns from raw data collected from educational environments (Siemens, 2012). This is one of the reason motivating the raising interest in LA, and strengthening its connections with data-driven research fields like Data Mining (DM) and Machine Learning (ML).

LA is about the collection and analysis of data from learners and their contexts to understand and optimize learning (Ferguson, 2012), and EDM is concerned with developing and applying computerized methods to detect patterns in large amounts of educational data (Romero et al., 2010). The combination of LA, a new research discipline with a high potential to impact the existing models of education (Gašević et al., 2015a; Siemens, 2012), and EDM, a novice growing research area to apply Data Mining methods on educational data (Bousbia and Belamri, 2014; Koedinger et al., 2015), leads to new insights on learners' behavior, interactions, and learning paths, as well as to improving the TEL methods in a data-driven way. In this regard, LA and EDM can offer opportunities and great potentials to increase our understanding about learning processes so as to optimize learning through educational systems. They can inform and support learners, teachers and their institutions, and therefore help them understanding how these powerful tools can lead to huge benefits in learning and success in educational outcomes, through personalization and adaptation of education based on the learner's needs (Greller and Drachsler, 2012; Romero et al., 2010). In this work, we are interested in exploring the novel ML, DM, and Process Mining (PM) methods in the field of education and

investigate the feasibility of using such methods on educational data. Through LA and EDM, we aim to raise the awareness of stakeholders on the students' learning phases and processes.

The opportunities of LA and EDM have been strengthened by a huge shift in the availability of the data resources. The data availability is an inspiring motivation for growing research and can be considered as benchmarks to advance the current methods and algorithms through comparison to other algorithms (Verbert et al., 2012). In this work, we design experiments to collect detailed data of students' learning process through educational systems. Additionally, we generate data sets from the collected raw data and share with the research community.

LA is gaining attention in inquiry-based engineering education as a way to improve learning achievements of the students. For example, in Pratheesh and Devi (2013), LA is used to help the teacher in a software engineering class distinguish the learning style of the students and adapt their teaching method to different groups of learners. Or the application of LA in remote laboratories in engineering education (in Orduna et al. (2014)) provides an analysis of the usage of laboratories in different contexts through LA dashboards. In this work, we are interested in exploring the IBL cycle of learners in various educational contexts via Learning Analytics and Educational Data Mining.

Understanding the learners' behavior in various IBL phases and learning contexts can be a key element for improving the lessons, instructional planning, and educational systems. This can lead to providing individual assistance to the learners and to improving the learning outcome of students. To gain insight on the learners' actions and interactions in an IBL cycle, appropriate LA and EDM methods need to be implemented with the ability to take the specific characteristics of the educational contexts into account.

Despite the increasing interest in LA and EDM, the evidences of successful application of appropriate analytics methods in IBL settings are still limited. It is indeed challenging to choose and tailor methods of analytics for a specific educational context due to the complexity and the special characteristics of learning processes. This brings us to formulating the following problem statement in this thesis:

What analytics approaches can quantify the learning processes in an IBL cycle?

This question aims to study different tools and the extent to which they can be used for a particular educational context. To address this question, we use the IBL cycle as a structure to explain the application of LA and EDM methods on various learning phases and contexts. We focus particularly on three contexts:

- *Concept Learning* where we investigate how students' hypothesis creation in concept learning varies on the size and domains of categories.
- *Simulation-based Learning* where we explore how the learning process of students in investigation and discovery phase varies based on their grades and task difficulty.

- *Game-based Puzzle-Solving* where we investigate how players solve a puzzle by applying the knowledge learned, if their reflection leads to a tactic, and whether we can discover the problem-solving strategies from the players' individual behavior.

1.2 Main Contributions

The main contributions of this thesis are presented as follows:

- We investigate the IBL cycle of learners and focus on the *hypothesis generation* phase by application of LA and EDM in the context of human concept learning. Our main contributions in this context include:

  We propose a new application of ML algorithms for human **concept learning**. For that, we replicate a study on human Rademacher Complexity, and propose a more powerful measure of complexity, the Human Algorithmic Stability, as a tool to better understand the human hypothesis creation. We compare the two algorithms for human concept learning and discuss the effect of domain and size of the problem on hypothesis generation (Chapter 4).

- We investigate the IBL cycle of learners and focus on the *investigation and discovery* phase by application of LA and EDM in the context of simulation-based learning. Our main contributions in this context include:

  We propose a new application of process mining in combination with the Cyclomatic complexity metric in **simulation-based learning**. We analyze the investigation process of students through Process Mining. We show the variation of processes based on the students' grades and task difficulty of various sessions. Finally, we compare the results with the teachers' judgments about the learning paths of students (Chapter 5).

- We investigate the IBL cycle of learners and focus on the *conclusion and reflection* phase by application of LA and EDM in the context of game-based puzzle solving. Our main contributions in this context include:

  We propose a new application of Process Mining in combination with Cluster Analysis in **game-based puzzle-solving**. We discover the players' successful tactics clusters from their IBL conclusion phase. We then construct process models of tactics and show that the puzzle-solving process of a player reflects the identified tactic (Chapter 6).

- We generate and make publicly available an Educational Process Mining (EPM) data set (Vahdat et al., 2015b) composed of data logs from a group of 115 students of first-year, undergraduate Engineering major of the University of Genoa. The data is collected from a study over a simulation environment named Deeds (Digital

Electronics Education and Design Suite) which is used for e-learning in digital electronics. The data set provides a collection of 230318 activity instances with 13 attributes (Chapter 5 and Appendix A). Our data set can be used by the researchers of LA and EDM who aim to test their analytics methods and pedagogical models. So far, our data set has been accessed over 20000 times which shows the growing interest towards LA data sets.

1.3 Thesis Outline

This thesis comprises of 7 chapters. The chapters 2 and 3 provide an introduction to the topics of this thesis and related works. Then, the next three chapters (chapters 4, 5, and 6) include the empirical studies and the obtained results. The 7th chapter provides the conclusions of this work. The thesis chapters are briefly described below:

- Chapter 2 describes the main ideas behind the development of this thesis. It starts from providing a background on Learning Analytics (LA), Educational Data Mining (EDM), and Inquiry-Based Learning (IBL) concepts. Methods of learning and analytics are explained as far as needed in this chapter.

- Chapter 3 examines the current state of the art in the fields of LA and EDM. It starts with a general introduction regarding the applications of LA and EDM, then it is narrowed down to the specific applications of the three methods of classification, clustering, and process mining in educational contexts. This chapter also highlights the related work regarding the contexts / methods of learning.

- Chapter 4 presents our first empirical study focusing on the first IBL phase of hypothesis generation. This chapter addresses several issues of Human Learning (HL) in hypothesis generation through Machine Learning (ML) methods. It starts with introducing the ML methods, and how we implement these methods in HL. This study is comprehensibly described by providing the details on experimental design of three experiments and the obtained results.

- Chapter 5 presents our second empirical study focusing on the second IBL phase of investigation and discovery. This chapter shows how Process Mining (PM) and Cyclomatic Complexity (CM) can be applied to gain insight into the behavior of students while performing experiments with an educational simulator. It starts with introducing the PM and CM methods, and how we implement these methods in simulation-based learning. The experimental design, analytics approach, and data set of this study are described in detail along with the obtained results.

- Chapter 6 presents our third empirical study focusing on the third IBL phase of conclusion and reflection. This chapter presents an analytics approach which combines PM and cluster analysis to analyze the puzzle-solving behavior of players

while playing a puzzle game. It starts with the game description and the components of players' puzzle-solving. Then, it explains our analytics approach and how we apply it in the context of game-based puzzle-solving. The experimental design and the results of this study are described in detail.

- Chapter 7 provides a discussion over the results of the three empirical studies, summarizes the results and conclusions of this thesis, and proposes future research directions.

Chapter 2

Learning Analytics and Educational Data Mining

2.1 Introduction

This chapter aims to illustrate the fundamental concepts of this thesis[1]. It provides a background on concepts and methodologies adopted in our work. As a starting point, it is necessary to introduce LA and EDM as two emerging fields in technology assisted education. LA is a multi-disciplinary field that is tightly related to Educational Data Mining, and involves recommender systems and personalized adaptive learning (Chatti et al., 2012). LA in combination with novel approaches in EDM lead to improve the learning outcome, since applying such methods can inform and support the stakeholders about the learning processes.

Within the educational contexts and theories, we focus on IBL as a structure to explain our empirical studies. In addition to these fundamental concepts, it is important to provide an introduction on the methods of learning and analytics, as the backbones of our main contributions.

Additionally, the automatic collection and availability of data resources has been inspiring the researchers of LA and EDM to test and improve analytics methods, and contribute to the data sharing community. There has been many efforts to provide simplified access of LA data for researchers and practitioners (Taibi and Dietze, 2013). Such examples are PSLC DataShop and educational enriched data from MOOC (Baker and Yacef, 2009), and data mining and Machine Learning repositories like the UCI Machine Learning repository (Murphy and Aha, 1995). The accessibility to anatomized data sets for research on teaching and learning in LA field, facilitate researchers to test their model or theory on a data set or to analyze the students' performance and learning by application of new analytics methods. Also various efforts such as encouraging researchers to apply LA methods on available data sets from different learning environments or linked

[1]This chapter is written partly based on Vahdat et al. (2015a).

data, in the forms of competitions, shows the growing interest in collection and use of LA data (Drachsler et al., 2014).

The structure of this chapter is as follows. First, the concepts of LA and EDM as well as their benefits and challenges are described in Section 2.2. We explain the concept of IBL in Section 2.3. Then, a review of several learning methods (Section 2.4), and how we orient our work toward the IBL cycle and phases (Section 2.5) are provided. The educational data features and the common applied methods of analytics in current research are explained in Section 2.6. Finally, the chapter is summarized in Section 2.7 .

2.2 What are LA and EDM?

LA and EDM are both two emerging fields that have a lot in common, although they have differences in their origins and applications. LA is a multi-disciplinary field that involves ML, artificial intelligence, information retrieval, statistics, and visualization. Additionally, it is related to the Technology Enhanced Learning (TEL) areas of research such as EDM, recommender systems, and personalized adaptive learning (Chatti et al., 2012).

EDM is concerned with: "developing, researching, and applying computerized methods to detect patterns in large collections of educational data that would otherwise be hard or impossible to analyze due to the enormous volume of data within which they exist" (Romero et al., 2010). While, LA initially was defined as: "the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs" (Ferguson, 2012), one can easily remark that these fields talk about the same area of research and follow similar aims of improving education and data analysis to support research and practice in education (Siemens and Baker, 2012). Therefore, we use them interchangeably in the entire thesis.

Differences of LA and EDM have been highlighted in various studies such that in LA, human judgment has an important role, while in EDM, the automation tools are influential on the final decision. Thus, EDM follows a bottom-up approach and looks for new patterns in data, and investigates for developing new models, whereas LA has a top-down approach and applies the existing methods to assess the learning theories about how students learn (Baker and Yacef, 2009; Bienkowski et al., 2012; Siemens and Baker, 2012).

Benefits

The benefits of LA and EDM are explained further in many studies. For instance, the UNESCO policy brief explains the LA benefits in micro (individual user actions), meso (institution-wide), and macro (regional, national, or international) levels covering various

stakeholders (Buckingham Shum, 2012). These stakeholders are considered in three main groups: educators, learners, and administrators.

Educators are responsible to design and plan the TEL systems, and they are most aware of learning process of the students, their needs, and common problems. LA and EDM can increase the instructors' awareness about the performance of learners, identify struggling or disconnected students, and empower instructors with pedagogically meaningful information (Baker and Inventado, 2014; Papamitsiou and Economides, 2014). Such information can be a great help for educators to monitor the learning process and adapt their teaching activities to the needs of students. The second group is the learners who can benefit from recommendation and more personalized feedback on their learning activities, resources, and paths. In this context, LA and EDM can gain a better understanding of the learner through student modeling to detect the learning needs and adapt the teaching methods and content to the individuals (Peña-Ayala, 2014). Finally, administrators are dealing with decision-making and budget allowance, and can influence the process of improving the systems and learning resources (Romero and Ventura, 2007; Siemens and Long, 2011).

In general, in both fields, improving learning and gaining insights into learning processes is the ultimate goal: LA and EDM are valuable concerning the prediction of the future learning behavior in order to provide feedbacks and adapt recommender systems based on learners' attitudes. Additionally, they are helpful to discover and enhance the learning domain models and to evaluate learning materials and courseware. Also they can advance the scientific knowledge about learners, detect their abnormal behavior and problems, as well as improve the pedagogical support by learning software (Bienkowski et al., 2012; He, 2013). In fact, these two research areas are considered complementary due to the holistic framework of LA and reductionistic viewpoint of EDM in gaining insights into learning processes (Papamitsiou and Economides, 2014). Figure 2.1 shows the LA/ EDM process concerning the data collection, processing, and giving feedback to the learners (or teachers) as a purpose of intervention and optimizing the learning outcome.

Challenges

Although LA and EDM have beneficial advantages, their drawbacks and challenges need to be considered by the researchers and practitioners of the field as well. Since LA and EDM emerged from various fields of analytics and data mining, it is challenging for them to obtain the connections with cognition, metacognition, and pedagogy, which are indeed mandatory in understanding the learning processes. Researchers need to pay attention to learning sciences to ensure effective pedagogy and improved learning design (Ferguson, 2012).

Other factors, mentioned in many studies, are the high costs of applications and techniques, as well as the issues regarding the data interoperability and reliability. There have been efforts to standardize the educational data and enhance its mobility such as

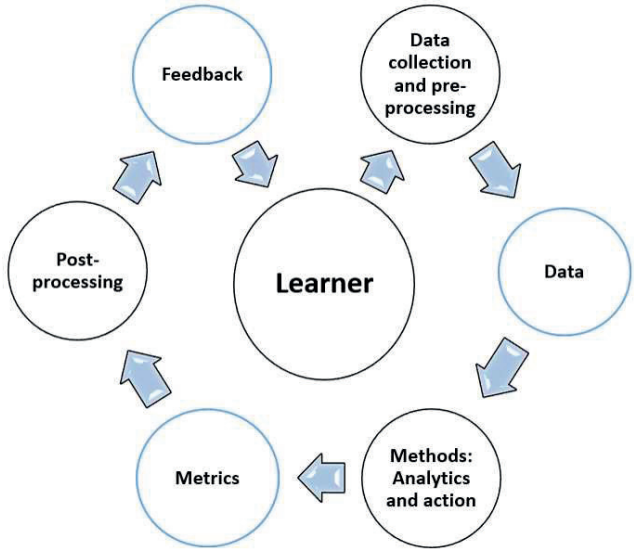


Figure 2.1: A learner-centric LA/ EDM process starts with learner whose data is collected and analyzed, and after post processing, feedback and interventions are made in order to improve learning (based on Chatti et al. (2012) and Clow (2012)).

IEEE standard for learning technology (IEEE SLT) and Experience API. However, the current state of interoperability is not effective enough to bring all data levels together. As for reliability, there are many challenges on the way of understanding the role of user in activity data and making sense of its context through unorganized information. Furthermore, ethical obligations such as privacy and anonymity is a growing difficulty due to the increase of data resources and powerful tools that need to be taken care of (Bienkowski et al., 2012; Del Blanco et al., 2013; Ferguson, 2012; Gyllstrom, 2009).

2.3 Inquiry-Based Learning

Inquiry-Based Learning (IBL) focuses on contexts where learners are meant to discover knowledge rather than memorizing concepts (Kruse and Pongsajapan, 2012; Prince and Felder, 2006). In this way, the learning session begins with a set of observations to interpret, and the learner tries to analyze data or solve a problem by the help of guiding principles (Lee, 2011), resulting in a more effective educational approach (De Jong et al., 2014). In other words, the learner tries to formulate hypotheses and test them to discover new causal relations (Pedaste et al., 2012).

IBL has gained a lot of attention, and many researchers have worked on the IBL process and cycle (Pedaste et al., 2015). The growing interest in promoting IBL at

schools for various ages shows the importance of IBL in instruction. For instance, the PRIMAS[2] and mascil[3] European projects focus on teaching and learning of mathematics and science in the context of IBL and support teachers to implement IBL methods in day-to-day teaching. According to PRIMAS, in today's dynamic society, attainment of facts alone is not sufficient and students need to develop further competencies to be able to apply the knowledge learned in the real problem-solving situations.

IBL has a long history in teaching science at schools. The idea of developing critical thinking rather than memorizing facts is not new, however, with the emergence and popularity of technology in education, IBL has received more attention (Pedaste et al., 2015). In the environments where IBL is applied rather than reproductive learning (Montuori, 2012), TEL systems can play an important role in providing guidance and principles to optimize learning. Inquiry refers to teaching and learning practices that are more student-driven, explorative, self-directed, and accompanied with guidance of the instructor (Justice et al., 2009) thus, technology-based instruction can facilitate the process of active learning. In such a process, students are encouraged to use curiosity to explore and understand topics through asking questions, performing research and experiments, and reflect over what they have concluded.

According to a literature analysis (Pedaste et al., 2015), IBL is organized into several inquiry phases that form a cycle together. This study provides a comprehensive literature review on IBL and claims that the phases and the form of the IBL cycle differ among researchers due to the various learning contexts. In this work, an inquiry-based learning framework presents the inquiry phases and sub-phases aggregated from 32 reviewed studies. Pedaste et al. (2015) collected the core features of IBL and proposes a synthesized inquiry cycle that combines the existing IBL frameworks. They identify five inquiry phases: orientation, conceptualization, investigation, conclusion, and discussion. Every general phase consists of more sub-phases. A brief description of IBL phases is provided as follows.

Orientation: The IBL cycle starts with 'orientation' where the learners are introduced to a topic or a problem to be solved. This phase engages students in the learning task, makes them focus on the learning problem, and makes connections with their background knowledge and experience (Bybee et al., 2006).

Conceptualization: The 'orientation' phase is followed by the 'conceptualization' phase where the learners form the key concepts to be studies in either a hypothesis-driven approach or a question-driven approach. In the question-driven approach, the students explore a phenomenon by forming open questions while in the hypothesis-driven approach, they follow a theory-based approach on what to investigate.

Investigation: Experimentation and data interpretation comes afterward as part of the 'investigation' process. In this phase, the learner tries to respond to the formed

[2]http://www.primas-project.eu, last accessed 27 September 2016.
[3]http://www.mascil-project.eu, last accessed 27 September 2016.

research questions or test the hypotheses by performing experiments, collecting data, and interpreting them.

Conclusion: Finally, the learners address the research questions or hypotheses in the phase of 'conclusion'. In this phase, the learners might evaluate and reflect over their results which can lead to the formulation of new questions or hypotheses, therefore, the learners restart the IBL cycle.

Discussion: Throughout the whole process, the 'discussion' phase can be connected with all the phases. The learners can communicate their ideas and discuss them at any moment. Depending on the context, this phase can lead to a discussion with others or happens as a form of self-reflection. This phase can be considered as optional in IBL cycle (Bybee et al., 2006; Pedaste et al., 2015; Scanlon et al., 2011; White and Frederiksen, 1998; White et al., 1999).

Additionally, there can be a flexibility in the way of linking between these phases. I.e, the learner can reflect on the result of the work and goes back to the questioning or investigation phase. Also, there can be a transition from investigation to the questioning phase.

In the context of our study, we only consider the phases of IBL engaging the role of the learner. In Figure 2.2, the inquiry cycle is shown based on the work of Pedaste et al. (2015) in which the phases are presented in a way to form a learner-centered cycle. The communication and discussion phase is shown on the right as a separate phase, as it can be optionally considered at any moment in the cycle. Also, we chose to show the orientation phase on the top of the diagram and separately from the cycle since, introduction of a topic or a learning challenge to the learner happens before the learner's role begins (usually by the instructor). Thus, we consider the learner-centered inquiry cycle as: conceptualization (generating hypothesis and question)- investigation and discovery - conclusion and reflection. We chose to add reflection to the phase of conclusion since a reflection over the results can lead to begin the inquiry cycle again.

2.3.1 Phase One: Generating Hypothesis and Question

In this phase, the learners perform conceptualization of the topics they are exposed to during orientation. They try to either understand the questions provided by the instructor or ask questions themselves. In this context, they identify the problem and set hypotheses to be tested later. In a proposed goal structure for inquiry, White et al. (1999) explains that the students start by formulating a research question, then they generate predictions and alternative hypotheses related to their question. This issue was indicated by the Heiss, Obourn, and Hoffman Learning Cycle that was based on Dewey's Instructional Model (Heiss et al., 1950). This cycle indicates a phase of 'exploring the unit' in which the learners observe demonstrations to come up with research questions and propose a hypothesis to address the raised question, then they plan for testing the hypothesis. In

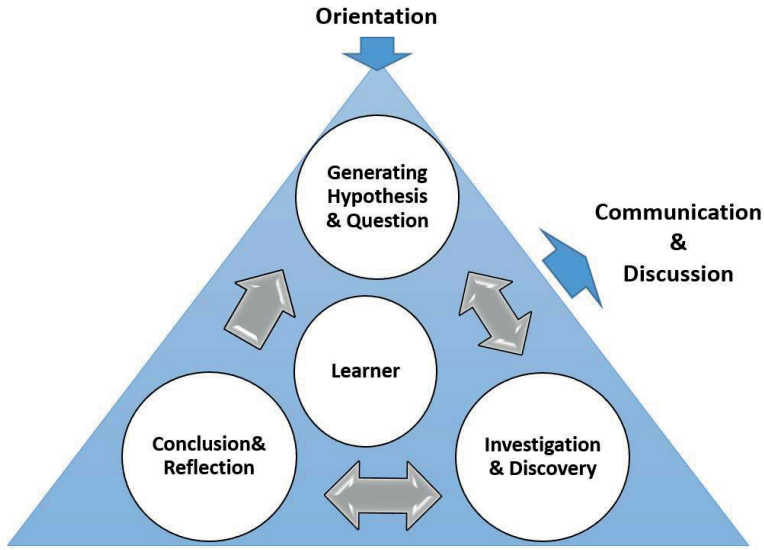


Figure 2.2: The IBL cycle (based on Pedaste et al. (2015)). Our focus is on the three phases in the blue triangle.

this step, the learner tries to establish a relationship between the perplexing situation and previous experiences (Heiss et al., 1950).

This phase is called 'conceptualization' by Pedaste et al. (2015) and divided into two subsequent sub-phases: questioning and hypothesis generation. Depending on the nature of the inquiry learning, inductive (data-driven) or deductive (hypothesis-driven) approach can be considered in this phase. For instance, Klahr and Dunbar (1988) suggests that both inductive and deductive approaches can exist side-by-side in a scientific reasoning process. The selection and arrangement of inquiry phases can be affected by the balance between the inductive and deductive approaches.

2.3.2 Phase Two: Investigation and Discovery

After defining the problem and hypotheses, the learner performs various types of experiments in order to test the hypotheses (Progressive education association, 1937). In this phase, the learners collect and interpret data of their performed experiments or investigations (Heiss et al., 1950) to form a conclusion. Through these experimental investigations, the learners try to determine the accuracy of their hypotheses. Or following a data-driven approach, the learners analyze their data to see if there are any patterns (White et al., 1999).

In this phase, the learners can investigate, observe, or explore depending on whether

they have a specific plan based on a clear hypothesis (deductive approach), or they are flexible in making discoveries related to their questions (inductive approach) (Pedaste et al., 2015). In this context, performing laboratory activities can be helpful for exploration experiences and conducting investigations. These experiences can lead to formulating concepts, processes, and skills. In this phase, educational software can facilitate the formulation of concepts in a scientific way (Bybee et al., 2006).

2.3.3 Phase Three: Conclusion and Reflection

Conclusions are drawn based on the investigation and analysis phase. In the 'conclusion' phase, the learners try to explain their findings and the obtained scientific knowledge by preparing the results and summaries. They try to demonstrate their conceptual understanding of a concept, process, or skill in the form of an explanation, a presentation, or a test. Additionally, reflection, reasoning with evidence, and comparing the obtained results with previous knowledge can be considered in the conclusion phase. They may also evaluate their findings by applying them to a different situation and determine the limitations of their investigations. Afterward, their understandings can be evaluated and they may receive feedback. Finally, this phase can lead to raise new research questions or formulate new hypotheses and the learner starts the IBL cycle again (Bybee et al., 2006; Heiss et al., 1950; Pedaste et al., 2015).

In the next section, we provide examples of learning methods in which the IBL cycle is implemented.

2.4 Learning Contexts

In this section, we provide three examples of learning methods / contexts in which IBL is implemented. For each, first we describe the method / learning context and then, the specific IBL cycle of each method.

2.4.1 Concept Learning

Concept Learning (CL) is a field of Cognitive Science (COGS) that explores how concepts are attained in humans. Various approaches exist in categorising concepts and how they are attained (Goodman et al., 2008; Vats et al., 2013). One approach is to consider concepts as mental representations which help identify and separate objects, events, and relationships. Another approach considers that concepts are learned inductively even from sparse and noisy evidences. In addition, concepts can be formed by combining other simpler concepts, and their meanings are derived from the ones of their constituents.

Various theories integrate different approaches of CL: for instance, Exemplar Theory (Kruschke, 1992) suggests that the categorization takes place by the proximity of the

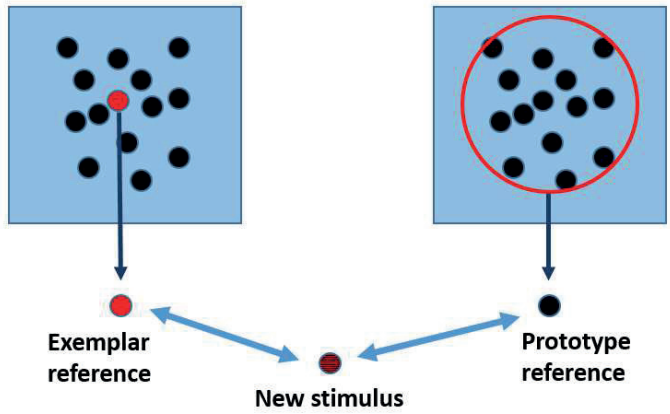


Figure 2.3: Exemplar theory vs. Prototype theory. The new stimulus is compared to the exemplar reference on the left, and to the prototype reference on the right.

new stimulus to the members of the category that one has observed, and by comparison of similarities, the label is assigned to the stimulus. Another theory, called Prototype Theory (Medin and Schaffer, 1978), explains that categorization takes place in a similar way as Exemplar Theory, while the comparison is carried out to the average of category members not to a specific member. In this case, at first, the attributes of members of a category are derived (named prototypes), then categorization is done by considering the similarity to the generated prototypes. Figure 2.3 visualizes the mechanisms of the Exemplar and Prototype theories.

In addition to these theories, researchers discovered that rule-based theories are important in the initial formation of categories (Bruner and Austin, 1986; Nosofsky and Palmeri, 1998): first the distinguishing attributes of new items are extracted from the category, then Examplar or Prototype theory, for categorizing distinct items, is applied. In this approach, concepts are constructed by combination (Murphy, 2002). In particular, a concept is represented by some rule that determines whether a stimulus belongs to a category (Goldstone and Kersten, 2003). Thus, humans try to find a rule (learn a model) when being confronted with a new example.

CL derived from rule-based theories can be considered in the context of IBL, due to the exploration, formulation, and validation of category rules. In this way, the IBL cycle can be presented as: orientation - hypothesis generation - hypothesis testing - conclusion and reflection. The IBL cycle starts with observation of a set of items in the orientation phase. By observing the set, the learner tries to distinguish the attributes of the set items and generate hypotheses (second phase) to explain the rule of distinct categories. In the next phase, by presenting a new stimulus, the learner tests the hypothesis by determining

whether the rule can be applied on the new stimulus. The result of hypothesis testing can be determined as conclusion and an evaluation can lead to generation of a new hypothesis. We further investigate this concept in Chapter 4.

2.4.2 Simulation-Based Learning

Simulation-based environments are efficient tools to improve learning and have been used in many educational contexts especially when dealing with inquiry-based problems (Donzellini and Ponta, 2007; Van Joolingen and De Jong, 2003). Learning with computer simulators is investigated in many studies (De Jong and Van Joolingen, 1998; Njoo and De Jong, 1993; Reid et al., 2003). For instance in Njoo and De Jong (1993), computer simulations are considered for exploratory learning (a technique of inquiry-based learning) since they can provide learning environments that facilitate interaction and discovery. Simulators can increase the instructional value of a system since they provide more control over various kinds of events and their frequencies compared to a real-life situation, e.g., dealing with accidents in a driving simulation can be practiced as many times as needed (Van Joolingen and De Jong, 2003). However, it is worth mentioning that simulators are limited by their underlying model thus they can impose limitations on the exploration process of learners.

Simulation environments are developed specifically to facilitate IBL. For instance, Co-Lab simulation environments and SimQuest facilitate instructors to tune simulations to their own needs (Van Joolingen and De Jong, 2003). Or in the Go-Lab project, inquiry learning spaces can structure the process of learning and guide students through an inquiry cycle (Gillet et al., 2013; Govaerts et al., 2013). In the context of this project, an experiment on the application of LA in online simulation labs shows the increase of awareness of teachers about the progress and the difficulties of students. In this study, several contextual LA apps are implemented and helped the teachers to monitor the progress of students (Vozniuk et al., 2015).

Simulation-Based Learning can be depicted in an IBL cycle. In this way, the IBL cycle can be presented as: orientation - question and hypothesis generation - performing experiments via simulations - conclusion and reflection. The IBL cycle starts with presenting a problem in the orientation phase. After understanding the problem, the learner plans and designs experiments to obtain supporting evidence for answering the question. The learner investigate the research questions by using computer simulations. The result of investigation can be determined as conclusion or an evaluation can lead to further experiments.

2.4.3 Game-Based Puzzle-Solving

There is a growing field of investigation about the application of games as learning tools. Research indicates that video game playing can enhance performance in a wide variety

of tasks and cognitive skills, e.g., allocation of attentional resources (Green et al., 2010), task-switching (Oei and Patterson, 2014), creativity (Jackson et al., 2012), problem solving (Shute et al., 2015) and spatial skills (Green and Bavelier, 2007; Shute et al., 2015), in addition to other non-cognitive skills such as persistence (Shute et al., 2015) and openness to experiences (Ventura et al., 2012). Games are increasingly being seen as an alternative to complement or enhance traditional education, particularly with games produced with the specific purpose of achieving certain learning goals (Bellotti et al., 2010; Erhel and Jamet, 2013; Kebritchi et al., 2010).

One specific class of games is the class of digital puzzle games. This type of game is commonly used for educational purposes (Liu and Lin, 2009), possibly given its typical reliance on problem-solving and on logical and mathematical intelligence (Becker, 2005). Puzzles can facilitate motivation, logical thinking, and problem-solving strategies for students. Puzzle-solving can help the students to apply their mathematics and logic skills to a different situation and reinforce their problem-solving abilities. Also, puzzles can be considered as problem-solving games since the players are usually engaged in the process of inquiry, reasoning, and discovery (Huang et al., 2007).

Game-Based Puzzle-Solving can be explained by an IBL cycle. In this way, the IBL cycle can be presented as: orientation - question and hypothesis generation - exploration and formulation of strategy - conclusion and reflection. The IBL cycle starts with presenting the puzzle in the orientation phase. After comprehending the aim of the puzzle, the player explores and tries to formulate strategies to reach the goal. In computer puzzle games, the player has the opportunity to test the strategy and play the game over and over again to reach a more efficient strategy. Finally, the result of puzzle-solving can be determined as execution of the competing strategy.

2.5 LA and EDM for IBL

Here, we explain a framework in the context of this thesis, to apply the LA / EDM process (shown in Section 2.2) to the IBL cycle (presented in Section 2.3). Note that, this framework is context-specific and only proposed as a structure in which we explain our empirical studies. As explained in Section 2.3, among the various IBL phases, we choose the student-centered phases that are essential in the IBL cycle. These phases are: conceptualization (generating hypothesis and question), investigation and discovery, and conclusion and reflection on which we perform LA and EDM.

In this framework (Figure 2.4), LA and EDM are performed on three main phases of the IBL cycle. The LA / EDM process starts with collecting the learner's data from the three IBL phases, the data is processed and analyzed through appropriate methods and metrics. Then, the feedback is given to the stakeholders informing them about the learner's IBL process. Depending on the research question, the data from one or all of the IBL phases can be taken into account.

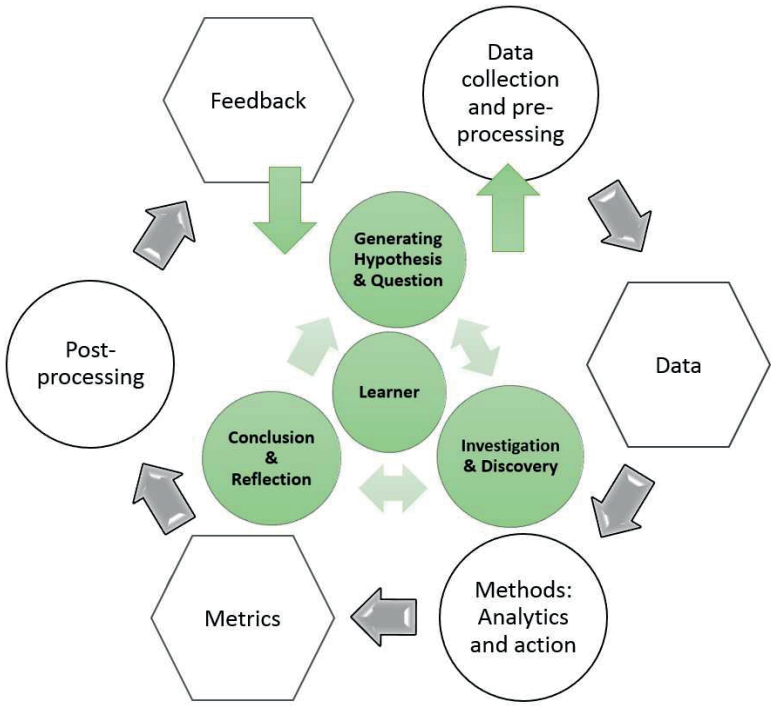


Figure 2.4: LA and EDM process for the IBL cycle. The inner green diagram shows the learner-centric IBL cycle. The LA and EDM process is applied on three IBL phases.

Note that, we discard the orientation and communication shown in Figure 2.2 since in this LA / EDM process, we focus on the student-performed activities and results, and not the orientation given by the instructor or the communication that might happen during the learning process. For instance during orientation, the instructor presents the theoretical content or the challenge to the students. Therefore, orientation is not included in the LA / EDM process (for detailed information about LA and EDM, see Section 2.2). The same applies to the discussion phase since we do not collect any communication data from learners. However, we include the reflection phase in conclusion since it encompasses the reasoning of the learner about the outcome of experiment and this phase plays an important role in returning to the IBL cycle.

2.6 Analytics Methods and Data Features

We survey in this section the most common analytics approaches, adopted by LA and EDM. A propaedeutic step is represented by reviewing data features typically used in the field, as well as their categories and important features.

Normally, data is collected from a wide variety of environments in educational settings. These environments can be the Student Information Systems (SIS), Learning Management Systems (LMS), Intelligent Tutoring Systems (ITS), Personal Learning environments (PLE), game-based and simulation-based learning environments (e.g., in Vahdat et al. (2014)). However, research has not been limited only to these settings, involving data gathering from other disparate sources, like social media, where applicable. Web-based courses and open data sets are also rich data sources for LA and EDM research (Chatti et al., 2012). Data gathered from these environments include the following types:

- Students-performed actions which normally consist of learning time span and sequence of concepts and practices. This kind of data should be heterogeneous and hierarchical so as to cover the various data levels nested inside one another. Additionally, the context of gathered data is crucial in order to explain the results and learning models.

- Data can be also collected through questionnaires assessing the learners' knowledge or other features. For instance, data categories are concerned with the student profile such as age, background, and interests in many LA and EDM cases (Bienkowski et al., 2012; Del Blanco et al., 2013).

Exploring Data Mining (DM) methods adopted in educational contexts, and introducing the common methods and applications over the gathered data is an important aspect of the LA and EDM review. A recent study shows that the most popular DM method used in the field is classification. Clustering, regression, and discovery with models are followed respectively in its ranking (Papamitsiou and Economides, 2014). In some state-of-the-art studies over EDM, the adopted methods are categorized into their data mining roles such as prediction, clustering, relationship mining, and discovery with models. Statistics and visualization have been mentioned in various studies as well. For information on the applications of these methods and their purpose, see Section 3.2.

These methods have been explained in further detail below to show the precise methodology and algorithms adopted in various LA and EDM studies. As an example, a prediction instance contains the classification of students' achievements from their activity data. Or exploring the sequential events to find occurring students' mistakes is an example of association rule mining (Baker and Yacef, 2009; Bienkowski et al., 2012; Romero and Ventura, 2007). Similarly, LA technical methods which are drawn from EDM, are mainly related to academic analytics, action analytics, and predictive analytics. As an example, in social network analysis, the data drown from students' tacit actions helps to identify disconnected students (Bienkowski et al., 2012).

While an exhaustive treatment of Analytics methods is beyond our scope, we still provide a general explanation over several fundamental concepts that are the foundations of the methods we employed in this thesis.

2.6.1 Classification

Classification is a Machine Learning (ML) method of supervised learning which looks for algorithms that can learn from data. These algorithms produce general hypotheses (infer functions) by learning a set of rules from the labeled instances (training set). In other words, a classifier is built for prediction of the class labels of future instances (test set). This method is called supervised learning since the labels or categories of the given instances are known beforehand in contrast to unsupervised learning where the labels are unknown (Bishop, 2006; Kotsiantis et al., 2007).

In addition to proposing new algorithms and tools, ML develops different methods for measuring the effectiveness of a learning process. In particular, ML studies the learning ability of an algorithm in order to avoid data memorization and to improve its generalization performance, which is the ability to learn the targeted concept effectively (Vapnik, 1998).

Examples of techniques for assessing the performance of a learning algorithm are: Hypothesis Space-based methods (Anguita et al., 2012) (based on the VC-Dimension (Vapnik, 1998), Rademacher Complexity (RC) (Anguita et al., 2011b; Bartlett et al., 2005; Koltchinskii, 2001), and PAC Bayes Theory (Lever et al., 2013; McAllester, 1998)) and Algorithm-based methods (Oneto et al., 2015a) (based on Compression Bounds (Floyd and Warmuth, 1995), and Algorithmic Stability (AS) Theory (Bousquet and Elisseeff, 2002; Poggio et al., 2004)). Thanks to these approaches, many valuable parameters that describe how a particular machine learns can be quantified. For example, it is possible to rigorously measure the generalization performance of a learned model.

In Chapter 4, we explain how we adopt the concepts of RC and AS in the context of classification and concept learning.

2.6.2 Clustering

Cluster analysis has broad applications in various disciplines and can be applied in a variety of contexts. It is beneficial in exploratory data analysis (Jain et al., 1999) and is considered as an iterative procedure of knowledge discovery rather than an automatic process that leads to directly usable results. The process of cluster analysis requires human knowledge for benchmarking algorithms and parameters (Bauckhage et al., 2015).

Clustering is an unsupervised classification of data items into groups (clusters), according to some item similarity measures. In this context, the items of interest are represented by a set of features, and the proximity containing pairwise similarities between the items are measured. The clusters are discovered in a way that the items in a cluster are more similar to each other than to the ones in other clusters. Depending on algorithms and parameters considered, the result of cluster analysis can lead to different groupings thus the process of clustering requires care and effort of experts (Bauckhage et al., 2015; Castro et al., 2007; Jain et al., 1999).

A clustering task comprises of several components. The task starts with the pattern representation concerning the extraction and selection of features, and identification of the number of classes. The next step is to define a pattern proximity measure that should be adapted to the data domain. Pattern proximity (or dissimilarity) is measured by a distance function over the pairs of patterns. Finally, clustering is performed and the groupings are discovered. The outcome of clustering needs to be validated through appropriate criteria and techniques (Halkidi et al., 2001; Jain and Dubes, 1988; Jain et al., 1999).

There are various clustering algorithms and validation techniques based on the context of study and the data domain. For instance, Jain et al. (1999) classifies the clustering algorithms into four categories of Partitional, Hierarchical, Density-based, and Grid-based clustering. We do not provide a review over the various techniques, however, we will explain the details of the method we employ in the context of our study (see Section 6.4.1).

In Chapter 6, we explain how we adopt clustering in the context of game-based puzzle-solving.

2.6.3 Process Mining

Process Mining is a method that has emerged from the fields of Business Project Management and Business Intelligence to obtain valuable knowledge from a process. By adopting various PM techniques such as process model discovery and conformance checking, process-related knowledge can be extracted from event logs. The aim of PM is to understand better what has happened in a time-ordered series of events and to help improve our insight about the (learning) process. Moreover, it can be used to understand why people deviate from a process or to predict and avoid such potential problems in the future (Van Der Aalst, 2011).

In PM, there are various algorithms used for the purpose of process discovery. To discover process models, these algorithms have different qualities in terms of fitness, precision, generalization and simplicity of the obtained models (Van Der Aalst, 2011). The fitness criteria ensures that the behavior in the event log matches with the model. Precision avoids the irrelevant behavior to be represented by the model and generalization allows to avoid overfitting of the process model. Also, process models are preferred to be as simple as possible to be understandable and to allow visual exploration of processes. Depending on the approach, some or all of these qualities are considered in the discovered model.

There are many PM approaches developed to address various aspects of a process, although most of them are applied to well-structured processes. Examples of such algorithms include 'α miner' that is a basic miner that creates a Petri net from an event log; this algorithm is sensitive to noise (Van der Aalst et al., 2004). 'Heuristic miner' is

very similar to 'α miner' but deals better with noise and incompleteness of the logs (Van Der Aalst, 2011). More recent approaches aim at better generalization and fitness of the process models, such as genetic algorithms and theory of regions (De Medeiros et al., 2006; van Dongen et al., 2007). In all of these cases, processes must be well-structured and explicitly designed to verify the compliant execution. However, the natural processes that are created over time in real life do not follow such requirements. For instance, in education, we need PM approaches that are able to draw hidden knowledge from complex processes that are unknown to us. One of these algorithms is 'fuzzy miner' that is able to deal with the complex and unstructured processes (Günther and Van Der Aalst, 2007).

In Chapter 5, we explain how we adopt the concepts of PM and 'fuzzy miner' algorithm in the context of simulation-based learning.

2.7 Summary

In this chapter, we provided a background over the fundamental concepts of this thesis. The chapter started with describing the LA and EDM areas, their known definitions in the field, their connections and differences, as well as their benefits and challenges. Then, we explained IBL as a learning cycle where the educational contexts of this thesis are based on. We showed that the IBL cycle comprises of: conceptualization (generating hypothesis and question) - investigation and discovery - conclusion and reflection. The learning and analytics methods are described afterwards to shed light on the educational contexts of this thesis and the methods employed to quantify the learning processes of students. The most important analytics methods of this thesis are classification, clustering, and process mining. The educational contexts of our empirical studies include: concept learning, simulation-based learning, and game-based puzzle-solving.

In the following chapter, applications of LA and EDM will be explained in further detail to illustrate the common use of approaches from these fields. We will provide examples from the literature concerning the analytics methods as well as the methods of learning.

Chapter 3

State of the Art

3.1 Introduction

The growing interest in recent years towards Learning Analytics (LA) and Educational Data Mining (EDM) has enabled novel approaches and advancements in educational settings. The wide variety of research and practice in this context has enforced important possibilities and applications from adaptation and personalization of Technology Enhanced Learning (TEL) systems to improvements of instructional design and pedagogy choices based on students needs. LA and EDM play an important role in enhancing learning processes by offering innovative methods of development and integration of more personalized, adaptive, and interactive educational environments. Additionally, there has been a rise in development and use of TEL systems and in the easy access to vast amounts of data about learners and learning contexts. This has been the motivation for many efforts to develop new methods or apply the existing ones in the field of education to make sense of learners data. Here, a review of research and practice in LA and EDM is presented accompanied by their most central methods and applications.

In this chapter[1], we provide an overview on the applications of LA and EDM by the most relevant works in the literature. This chapter is organized in two main parts: 'Overview on applications' in Section 3.2, where the state of the art of LA and EDM is covered with a particular attention on the application of methods. Also, the future trends of this field from the perspective of current literature is discussed. We discuss the applications of three particular methods of analytics namely classification, clustering, and process mining in educational contexts in Sections 3.2.1, 3.2.2, and 3.2.3 respectively.

The second part of this chapter includes 'Learning contexts' in Section 3.3, where the focus alters from analytics methods to the methods of learning. Namely, the most relevant works in the contexts of concept learning, simulation-based learning, and game-based puzzle-solving are provided in Sections 3.3.1, 3.3.2, and 3.3.3 respectively. Finally, we summarize the state of the art in Section 3.4.

[1]This chapter is written partly based on Vahdat et al. (2015a).

In every part, we start with a general overview on the topic accompanied by the relevant works in the field, then we narrow down the relevant examples by providing the closest literature to our empirical studies.

3.2 Overview on Applications

The strong points of LA and EDM have been shown through their applications in various studies. These applications address a wide verity of goals like those explained in Section 2.2. Among LA and EDM applications, an analysis of recent cases shows that student modeling is the most addressed approach. In this context, the aim is to gain a better understanding of the learner to detect the learning needs and adapt the teaching methods and content to the individuals (Peña-Ayala, 2014). Student modeling is applied to detect the students' states such as motivation, satisfaction, experience, and progress. Detecting the problems in the learning processes is a way to improve the educational systems and provide interventions at the right time and situation (Bienkowski et al., 2012; Romero et al., 2010).

Another common application is to predict the students' grades or to classify their behavior based on their learning outcome (Bienkowski et al., 2012). Prediction of learner's performance can lead to the increase of student and teacher awareness of learners' situation and enhancement of provided feedback and assessment services (Papamitsiou and Economides, 2014). Additionally, it can be used for analyzing the domain structure (topics and learning components) and measuring its quality. Also, LA and EDM can be applied for improving the quality of courses by giving hints to the educators and administrators. By analyzing the students' learning behavior in the process of problem solving, recommendations can be generated to guide students about the content and tasks.

A literature analysis on a wide variety of studies and applications of LA and EDM shows that there is a growing need and effort from research into practice (Vahdat et al., 2015a). There have been many LA cases that started as a research initiative and then transformed into an applied system for everyday educational practices. For instance, there exist several cases of developed applications for learning mathematics, which present the practical use of LA in improving the didactic support in primary schools. In such studies, LA is applied to gather and interpret students' data for increasing the reflection and deep learning (Ebner et al., 2014; Schön et al., 2012). Another attempt was done to employ LA research for improving the instructional planning. In this approach, instructional experts are provided with the visualizations of learners' interaction with TEL systems. These summaries of activities increased the insights into the learning processes and consequently, form instructional interventions. This case shows that LA can effectively influence on the selections of pedagogy based on seeing data of students' interactions (Brooks, 2013).

Although applications of LA and EDM in various educational contexts are numerous, it is challenging to find successful cases indicating the positive effect of LA and EDM.

A case is considered successful if the application of LA and EDM actually leads to the improvement of educational outcome. Studies like Gašević et al. (2015b); Romero and Ventura (2015) have recently collected some success case studies. For instance, Course Signals developed at Purdue University collected the students' traces from Blackboard LMS and SIS, and tried to identify the students at risk and in need of more attention by applying data mining methods. The study shows that this application actually led to a rise in the retention of the students who used the Course Signals tool (Arnold and Pistilli, 2012). In the context of LA and EDM, there are many attempts to develop dashboards in order to raise awareness of learners and teachers, and to give feedback about the learning process. However, there have been a few research studies to measure the effectiveness of these dashboards on learning. An evaluation study was done on an LA dashboard named StepUp! and shows that the evaluation process is a complex task and critical in LA research (Santos et al., 2013). eLAT is another example of an LA Toolkit that enables teachers to explore the students' learning behavior, and allow them monitor the students through visualizations (Dyckhoff et al., 2012).

The increasing awareness toward LA and EDM benefits has led to the increasing funds given to institutions worldwide. Particularly, European Commission recently granted opportunities from research projects to implementations and community building in the field. Examples of such projects are LACE[2], PELARS[3], LEAs BOX[4], Next-Tell[5], WeSpot[6], and WatchMe[7]. The presence of the success cases of LA and EDM are strong in these initiatives. For instance in LACE FP7 European project which aims to gather and integrate the active communities in LA/EDM research and practice, created a knowledge base of evidence that captures effective evidences of the field and assign them particular evidence types and sectors, highlighting works like Sao Pedro et al. (2013) as a positive evidence for improving teaching. Figure 3.1[8] is the evidence flow map taken from LACE Evidence Hub showing the LA cases separated by types on the left and sectors on the right. The interactive map shows if the results of such studies have been positive or negative. For instance Sao Pedro et al. (2013) is presented as a positive evidence for improving teaching.

Various studies that explored the LA and EDM opportunities (Bienkowski et al., 2012; Chatti et al., 2012; Romero and Ventura, 2007; Romero et al., 2010; Siemens and Baker, 2012) have framed future trends of research for the experts of the field. For instance, they claim that there is still a lack of user-friendly mining tools that can be used by

[2] `http://www.laceproject.eu`
[3] `http://www.pelars-project.eu`
[4] `http://www.leas-box.eu`
[5] `http://next-tell.eu`
[6] `http://wespot.net`
[7] `http://www.project-watchme.eu`
[8] The figure is taken from LACE Project Evidence Hub: `http://evidence.laceproject.eu`, last accessed 9 June 2016.

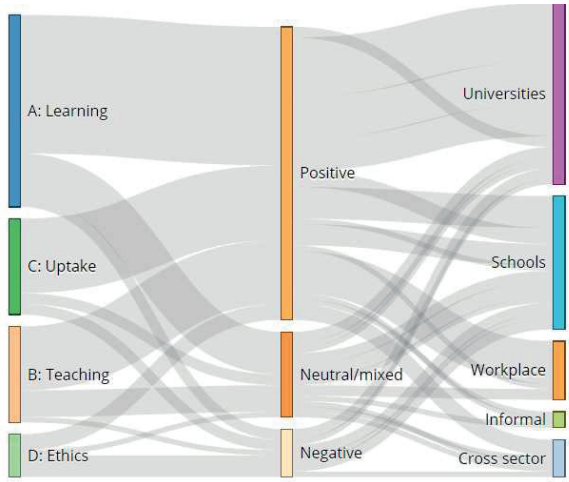


Figure 3.1: Evidence flow diagram by LACE Project Evidence Hub.

the educators, since currently the majority of tools require expertise in data mining to be adopted. Further integration and development of the recommendation engines into e-learning systems is suggested by these studies to improve the instruction in a more effective way. Also, there is no doubt that standardization of methods and data is an important challenge in the fields of LA and EDM. This will help to re-use tools and resources from one context to another, and save a lot of time and costs. Due to the ethical concerns, advancing the anonymization of data is suggested as well in order to protect the individual privacy across various educational applications.

The literature in LA and EDM covers a vast range of techniques, ranging from descriptive approaches to predictive methods. Here, we provide examples in three categories of methods including predictive approaches like classification, and descriptive ones such as clustering and process mining.

3.2.1 Classification in Education

Classification is one of the most common tasks applied in LA and EDM (Papamitsiou and Economides, 2014). In classification, the aim is to develop a model that predicts a single (predicted) variable from a set of predictor variables. Such a predictive method can have a remarkable importance in supporting education and learning. For instance, by predicting whether a student will succeed or fail, and informing the educator about the at-risk student (Baker and Yacef, 2009; Papamitsiou and Economides, 2014).

This task is used in a wide variety of educational contexts to predict the future learning behavior by identifying the variables that are most associated with the learning

behavior. Some studies have compared different classification methods and techniques for classifying students (Romero et al., 2008). For example, Kotsiantis and Pintelas (2005) uses ML techniques to predict the students' marks in the context of distance education. In this study, the demographic characteristics and students' marks in written assignments are used as predictor variables, and various regression algorithms are compared to predict the students' performance. The study shows that the education domain can offer a wide variety of challenging applications for ML and DM tasks.

Among the most recent literature, an interesting application of predictive methodologies is related to pupils, who do not finish their secondary education. Studies indicate that ML can be used to predict high-school dropouts, which allows for early interventions. Şara et al. (2015) present the first large-scale study of high-school dropouts prediction. It considers pupils who finished at least their first six months of Danish high-school education, and the goal was to predict a dropout in the next three months.

The importance of predicting the performance of students based on their behavior and their characteristics is highlighted by Nakayama et al. (2015). In a blended learning course, in particular, participant's note-taking activity correlates with their learning performance. The possibility of predicting performance in final exams is examined using metrics of participant's characteristics and features of the contents of notes taken during the course by Nakayama et al. (2015). According to the results of this prediction performance, features of note-taking activities are a significant source of information to predict the score of final exams.

Classification is used not only in education as a prediction method, but also as a methodology to measure human learning capabilities. Zhu et al. (2010) argues that concept learning or categorization in humans is similar to the classification task in ML. Similarly to ML, first the labeled training data (a set of stimulus) is given to the learner. Then a classifier is built (a rule determining the reason behind such labeling is drawn in human mind). And finally in the prediction phase, the classifier (rule) is applied to the test set (new stimulus).

For instance in several studies on concept learning, the idea of binary classification is adopted to design an experiment on humans, and quantify their learning performance (Castro et al., 2009; Zhu et al., 2009). In this context, Zhu et al. (2009) presents parallelisms between ML and Human Learning (HL): in order to maximize knowledge when dealing with an unknown phenomenon, humans, as algorithms, should grasp what originated experimental evidences, rather than simply memorizing them. In this work, the authors perform exploratory experiments based on binary classification towards studying human learning using ML complexity measures. In ML, the learning process of an algorithm given a set of evidences is studied via complexity measures. The way towards using ML complexity measures in the HL domain has been paved by Zhu et al. (2009), which introduced Human Rademacher Complexity (HRC). In Chapter 4, we investigate another powerful complexity measure to study the concept learning of humans.

3.2.2 Clustering in Education

Descriptive analytics methods are numerous and can be used for different purposes. Some of the most used descriptive approaches belong to the domain of clustering methods, which have shown their effectiveness in several heterogeneous domains, and various educational contexts. Clustering is used in education to discover the groups / clusters of similar behavior, educational materials and contents, or learning sessions. The grouping can be done based on various characteristics such as the interaction patterns of the learners, the amount of access to the educational contents, or the type of problems the learners are dealing with (Baker and Yacef, 2009; Romero and Ventura, 2007).

An example in the context of web learning is the work of Tang and McCalla (2002) in which the students are clustered based on the similarity of their access to educational contents such as the sequence and the contents of the pages they visited. The authors indicate the advantages of such method as promoting an effective group-based collaborative learning, and diagnosis of learners while browsing the content. In this way, the system can provide appropriate feedback to the learners and deliver them a personalized course content.

In a recent study (Saarela and Kärkkäinen, 2015), the authors present an efficient version of a robust clustering algorithm for sparse educational data that considers weights, allowing to enable generalization and tuning of a sample with respect to the corresponding population, into account. The algorithm is utilized to divide the Finnish student population of PISA 2012 (namely, the latest data from the Programme for International Student Assessment) into groups, according to their attitudes and perceptions towards mathematics, for which one third of the data is missing.

Clustering methods have been used in a wide variety of interactive systems due to their strength in finding similar behavioral patterns. One well-known area is computer games where clustering can be applied for mining game behavioral data and profiling the characteristics of players (Bauckhage et al., 2015).

When it comes to the identification of playing styles and capabilities of players in dealing with challenges during the game, few research has used unsupervised learning approaches (Charles and Black, 2004). The literature gets even more sparse when the aim is to model the players' strategies. The strategy refers to the overall plan of actions to achieve an ultimate goal (Bakkes et al., 2012). In an educational game, the strategy can be referred as a problem-solving plan of action designed to reach a learning goal. We initially aimed at finding successful cases of clustering in educational games where the problem-solving strategies of players are discovered. It was quite challenging to find such cases of cluster analysis in educational games, thus we provide several examples where clustering is applied for mining of game behavioral data.

For instance, Drachen et al. (2009) applied clustering on high-level behavioral data to construct the models of players. This method is suggested for automation of user testing of the game design, as well as to tune the game in real-time for serving the

players' needs during the game. In another study, cluster analysis is applied to analyze the affective reactions of players to the game events in the context of an educational game called PrimeClimb (Amershi et al., 2006). In this study, the patterns of affective behavior are clustered in order to understand the influence of the emotions of players on learning. Recognizing such groups is suggested for adaptation of the educational game to the learners' needs. As another example, Ramirez-Cano et al. (2010) used a meta-clustering scheme including three levels of clustering based on actions, preferences, and social network metrics of players. In this study, such combined clustering analysis determining different levels of similarity between players was suggested to improve the social interaction among players (Ramirez-Cano et al., 2010). In Chapter 6, we further investigate the application of clustering for mining the game behavioral data.

3.2.3 Process Mining in Education

In the context of LA and EDM, very few process-oriented approaches focus on the process as a whole (Trcka et al., 2010). Educational PM can be valuable for obtaining a better insight on the underlying educational processes (Trcka and Pechenizkiy, 2009). PM in education can be applied to construct educational process models through model discovery of the observed behavior. Later, the obtained model can be used to project the information extracted from the logs and to check whether the observed behavior is reflected by the model, leading to different kinds of decision support (Trcka and Pechenizkiy, 2009). There are various studies showing the potential of PM in the educational contexts.

For example, Pechenizkiy et al. (2009) show some applications of the PM methods and apply them to online assessment data for analyzing the process of answering the questions or requesting feedback by students. The authors demonstrate a PM tool called ProM for the analysis of data from two case studies where in one, the students followed a strict order in answering questions and feedback requests, and in the other, the students were flexible to answer the questions in any order and to revisit their answers. The descriptive visualizations give insight about the answering behavior of the students and the time spent on various tasks.

In Trcka and Pechenizkiy (2009), a framework is introduced for integrating the domain knowledge in educational PM in order to facilitate interactive PM and help educators analyze educational processes based on resulting process models. In this work, the authors introduce and formalize patterns in an academic curriculum describing the possibilities and constraints of students. Having a process model can help students be aware of what they need to do, help the educators to check whether the curriculum was respected, and facilitate real-time detection of curriculum violations.

So far, the presented studies focused on a single process for the purpose of process model discovery from the learners' interaction data and conformance checking of the logs

with the obtained model. Nevertheless, there are few efforts on comparing the process models and analyzing the differences from an educational point of view. For instance, Bannert et al. (2014) analyzes student processes of self-regulated learning and relates the discovered process models with the theories of self-regulated learning and metacognition. In this study, the differences of learning behavior between successful students and less successful ones are highlighted via their temporal patterns. The authors explain that the successful students perform regulatory activities with a higher frequency and in a different order than the less successful students. Also, the types of self-regulated learning activities of the two groups differ. In this study, the comparison of the two groups of students is mostly performed in a descriptive manner.

A recent study has proposed a numerical approach on comparative process mining using process cubes where events and process models are arranged in various dimensions (van der Aalst et al., 2013). In this study, the attributes are referred as the dimensions of a cube where the process models can be produced and compared. For instance, the authors compared the process of passed versus failed students of a course from their data of watching video lectures. The results indicate that the average trace fitness for all students that passed was significantly higher than the one of the failed students. However, the results are not presented as a generalization but rather a starting point for a better analysis of learning processes. In Chapter 5, we investigate another method to quantify the learning processes of students for the purpose of comparison. Quantification of process models helps a better comparison of processes and identification of their relation with other parameters such as students' grades.

PM can be used in combination with other data mining methods to improve the discovery of the models with a better performance and comprehensibility. In Bogarín et al. (2014), PM is applied in combination with clustering to improve the obtained educational process models in a study where the students followed an online course using Moodle (a learning management system). In this work, first the cluster analysis is performed to group the students based on their Moodle usage data and their grades, then PM is applied on the clusters separately to obtain the process models, resulting in a better performance / fitness and comprehensibility / size. In Chapter 6, we further investigate application of PM in combination with clustering to analyze the puzzle-solving behavior of learners.

3.3 Evidences from Learning Contexts

So far, we presented the various applications of LA and EDM in education. Here, we present the relevant literature in various learning contexts of this thesis. By learning contexts, we mean the learning methods where IBL cycle is implemented. In particular, we provide the related work in three contexts: concept learning, simulation-based learning, and game-based puzzle-solving. These three learning contexts are the basis of our

empirical studies presented in Chapters 4, 5, and 6 respectively.

3.3.1 Concept Learning

The latest approach towards human rule-based learning (explained in Section 2.4.1) has been a motivation for CL to benefit from the research of other fields like Artificial Intelligence, Information Theory, and ML. In this context, the cross between HL and ML (Deák et al., 2007; Madani and Sabourin, 2011; Matsuka et al., 2008) leads to development of sophisticated formal models of CL (Goldstone and Kersten, 2003; Joachims, 2015; Lan et al., 2014; Piech et al., 2012).

For instance, Griffiths et al. (2008) measures the ability of humans in category learning by applying Bayesian approaches in iterative learning. In this context, a human learns a concept and produces a hypothesis on the given data, then, another human learns the previously developed hypothesis and generates a new one. This method was adopted for identifying the inductive biases in humans. In another study (Feldman, 2000), the difficulty of concepts in relation with HL is exploited. In this context, the subjective difficulty of boolean concepts for humans is measured, and it is shown that the subjective difficulty is proportional to the complexity of boolean statements (length of the statement). Thus, by knowing the complexity of the logical structure of concepts, it is possible to predict how difficult that concept is for humans.

Other examples are Zhu et al. (2009) and Vahdat et al. (2016b, 2015d) where ML Theory, which helps to understand the learning ability of ML algorithms, has been used to explore HL. As discussed in Section 3.2.1, Zhu et al. (2009) proposed the application of Rademacher Complexity (RC) approaches (Bartlett and Mendelson, 2002) to estimate human capability of extracting knowledge. We discuss the limitations of this method later in Chapter 4, and propose a better method to investigate CL.

3.3.2 Simulation-Based Learning

Many studies have been done on learning with computer simulators, and explored their advantages and difficulties (De Jong and Van Joolingen, 1998; Njoo and De Jong, 1993; Reid et al., 2003). A meta-analysis of recent literature on the use of simulations for learning shows that students who learned with simulation were more successful than those who followed a non-simulation-based instruction (D'Angelo et al., 2014). Such examples (see Section 2.4.1) include Co-Lab simulation environments and SimQuest that allows instructors to author existing simulations to their own needs (Van Joolingen and De Jong, 2003). Co-Lab is a collaborative learning environment in which learners experiment through simulations and remote laboratories (van Joolingen et al., 2005).

LA and EDM have been used in IBL simulation environments to facilitate learning. For example in the Go-Lab project (Gillet et al., 2013; Govaerts et al., 2013), application of LA in online labs was investigated. In this study, several contextual LA apps

are implemented for the teachers to monitor the progress of the students and for the students to receive assistance through scaffolding based on LA and the teacher's configurations (Vozniuk et al., 2015). This study shows the valuable effect of LA to increase the awareness of teachers about the progress and difficulties of the students.

In another example, Sao Pedro et al. (2012) applied LA and EDM on a physical science simulation environment that fosters scientific inquiry in middle-school science classes. This study developed predictive models to assess the inquiry skills of students in designing experiments and hypothesis testing. In this work, predictive modeling was helpful to identify the students who had lack of skills or needed scaffolding, and to intervene and support the students as quickly as possible. EvoRoom is another simulation environment that is designed for high school biology curriculum (Slotta et al., 2013). This environment simulates a rainforest in a dynamic and interactive way to engage students in collaborative inquiry activities where they make observations and perform tasks. A real-time LA is applied to provide ambient feedback and dynamic visualizations through capturing and aggregating the students' behavior and progress.

Another simulation environment is 'Deeds' (Digital Electronics Education and Design Suite) (Donzellini and Ponta, 2003, 2007; Ponta et al., 1998) which is used for e-learning in digital electronics. The environment provides learning materials through specialized browsers for the students and ask them to solve the varied-level problems. It has been shown effective in the students' learning results for over ten years since it provides a general-purpose simulator with a highly interactive e-learning environment. Analysis of data derived from simulation based environments is worthwhile since it provides in-depth insight into simulation-based learning development. The Deeds design team persuade an ongoing effort to improve its functionalities by getting feedback from the students. We will show in Chapter 5 that application of LA and EDM with Deeds is a valuable opportunity to gain insight on the learning process of students while interacting with Deeds learning tools.

3.3.3 Game-Based Puzzle-Solving

Computer games have been attracting the ever-growing attention of researchers due to their active, experiential, situated, and problem-based features leading to motivation, learning, and developing useful skills (Boyle et al., 2011; Subrahmanyam and Greenfield, 1994). Many studies have investigated the advantages of educational computer games and many efforts are done for developing effective learning games / serious games where the players develop problem-solving strategies. For instance in Hwang et al. (2013), concept mapping is integrated for developing educational computer games, where students use concept maps to organize what they have learned. Additionally, an experiment is performed implementing the concept map-embedded gaming approach in a role-playing game which improved the student's learning achievements.

In another study, the usefulness of online games in vocabulary learning is presented (Yip and Kwan, 2006). During an experiment, a group of students learned vocabularies through online games and a control group was presented the same vocabularies through activity-based lessons. the result showed that the learners preferred the online games to the conventional method. Also, the authors indicate that the learners who played the game, learned and retained the vocabularies better than the control group.

Among various game genres, puzzle games are reported as the second most popular after simulations in studies of games for learning (Connolly et al., 2012). Puzzles are considered as problem-solving games that encourage logical thinking and facilitate problem-solving strategies. There have been studies on design of puzzle games and analyzing the players behaviors and strategies / tactics when solving a puzzle. For example, Huang et al. (2007) designed a number puzzle game, called Number Jigsaw Puzzle (NJP), and investigate the players' problem-solving strategies. The results from conducting a pilot study showed that five different strategies had been found, and most of the participants were able to use their previous knowledge to construct the strategies. In Chapter 6, we further investigate the puzzle-solving strategies of players through LA and EDM methods.

3.4 Summary

In this chapter, we presented essential applications of LA and EDM in the existing research literature. It included relevant up-to-date research regarding both the methods of analytics applied in the educational contexts and the relevant methods of learning highlighted in the existing literature. In every section, the methods were presented starting from a more generic point of view, then the closest examples to our empirical studies were presented.

In the three following chapters, we employ many aspects of the presented literature as the basis of our work. Also in some cases, they will be compared against the methods we propose in our empirical studies. For instance, in Chapter 4, we start from the concept of Human Rademacher Complexity (Zhu et al., 2009) presented at the end of the sections 3.2.1 and 3.3.1. We will compare this work to another method we propose to gain insight on human concept learning. In Chapter 5, we adopt the concept of Process Mining in simulation-based learning covered in Sections 3.2.3 and 3.3.2 respectively, to uncover the underlying educational processes. Finally, in Chapter 6, we adopt the concept of clustering in educational games covered in Sections 3.2.2 and 3.3.3 to gain insight on the players' strategies.

Chapter 4

IBL Phase One: Application of Machine Learning in Concept Learning

4.1 Introduction

Since the emergence of Technology-Enhanced Learning (TEL) systems and automatic analysis of educational data, many efforts have been carried out to enhance the learning experience (Chatti et al., 2012; Lee and Brunskill, 2012). For this reason, Learning Analytics (LA) and Educational Data Mining have recently gained a lot of attention as one of their major interests is to explore the way humans learn (Brown, 2012; Papamitsiou and Economides, 2014) (for more information on applications of LA and EDM, see Section 3.2). New advances in LA enable measuring, collecting and analyzing data about learners and their contexts and allow exploring the behavior of people while learning (e.g. through Machine Learning models), opening the door towards optimized and personalized education (Bienkowski et al., 2012; Koedinger et al., 2013; Piech et al., 2013; Siemens and Long, 2011). LA is a multi-disciplinary field which is tightly connected to Statistics and ML on the one side and to Cognitive Science (COGS) and Pedagogy on the other side (Ferguson, 2012; Polk and Seifert, 2002). For more background on LA and EDM, see Section 2.2.

Machine Learning (ML) is a field of Computer Science which develops and studies algorithms that can learn from and make predictions on data (Bishop, 2006). Such algorithms, used as tools in LA, build models from data in order to make data-driven decisions or predictions. For solving many real world problems, ML offers various tools (Lillo-Castellano et al., 2015; Tian et al., 2015; Yuan et al., 2015; Zhang and Li, 2015): classification, regression, clustering, online learning, semi-supervised learning, reinforcement learning, etc (Bishop, 2006; Hastie et al., 2009; MacKay, 2003; Shawe-Taylor and Cristianini, 2004). According to Baker (2000) there are three ways of using models of

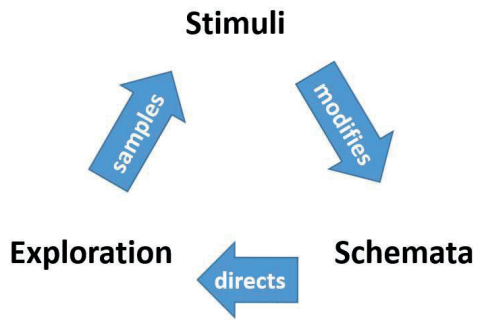


Figure 4.1: Neisser's perceptual cycle (Neisser, 1976). Learning can be described by the perception of stimuli as mental schemata that directs exploratory behavior.

educational processes: the first one includes models used as scientific tools to understand an educational situation, such as using models to predict the student academic success (Kotsiantis et al., 2004). In the second one, models are used as a component of educational artefacts such as student modeling and its application in a TEL environment (Brusilovsky et al., 2005; Rauterberg et al., 1997) or integrating the model of student problem-solving into a TEL system with the aim to personalize and adapt educational materials to their needs (Arnold and Pistilli, 2012). Finally, the third one includes models used as basis for design of TEL systems (Triantafillou et al., 2003).

As explained in Section 2.6.1, ML is not limited to proposing algorithms. It also can develop various methods for measuring the effectiveness of a learning process (Vapnik, 1998). This is done through assessing the performance of an algorithm in order to avoid data memorization and to improve its generalization performance. Rademacher Complexity (RC) and Algorithmic Stability (AS) Theory are examples of powerful ML techniques to measure the generalization performance of a learned model.

While ML studies learning algorithms, COGS studies and analyses how learning takes place in humans (Bruner and Austin, 1986; Pashler and Mozer, 2013; Watanabe, 1985). In this context, humans can be considered as information processing systems (as suggested in Rauterberg (1995)) with a high learning potential and learning is a permanent process that is regulated by optimizing the complexity of the learning context, based on actions and mental schemata of humans (Rauterberg and Ulich, 1996). This process is illustrated in the Neisser's perceptual cycle (Figure 4.1) where learning implies an abstraction of environment (stimuli, information) in the form of mental schemata which guide exploratory behavior. This process continues till the knowledge of the environment is obtained or a task is completed (Neisser, 1976).

In the context of COGS, Concept Learning (CL) investigates how concepts are attained in humans (Human Learning - HL). As highlighted in Section 2.4.1, there are

various approaches in categorizing and attainment of concepts by humans. We focus here on the latest approach towards human rule-based learning. This approach has been a motivation for CL to benefit from the research done in the field of ML. In this context, the cross between HL and ML are advantageous for development of sophisticated formal models of CL (See the examples in Section 3.3.1), and for investigating how people tackle new problems and extract knowledge from observations. For example in the context of inquiry-based CL, the learners discover knowledge through observations and hypothesis testing (Lee, 2012) (for more information about Inquiry-Based Learning (IBL), see Section 2.3). The IBL cycle begins with a set of observations to interpret, and the learner tries generate hypotheses, analyze the data, and test the hypotheses (Pedaste et al., 2015), resulting in a generalizable rule and a more effective CL (see Section 2.4).

In this chapter[1], our main contribution is to build a connection between ML and HL. In particular, we focus on the 'first phase of IBL' (see Section 2.3.1) and apply ML methods to measure the capacity of students in **hypothesis generation** and finding meaningful rules given various inquiry-based problems. Measuring the ability of a human to capture information rather than simply memorizing can be the key to optimize and improve HL. In this sense, the parallelism with ML is straightforward: for example, several approaches in the last decades dealt with the development of measures to assess the generalization ability of learning algorithms in order to minimize risks of overfitting (memorization). As a consequence, merging ML studies on the generalization ability estimation and HL has been proposed by some researchers.

In particular, Zhu et al. (2009) proposes the application of ML approaches Bartlett and Mendelson (2002) to estimate the human capability of extracting knowledge (Human Rademacher Complexity - HRC). Unfortunately, (H)RC requires a set of models to be defined a priori, which includes the models to be explored by the learner (being either an algorithm or a human) (Oneto et al., 2015a), while this hypothesis is not always satisfied by ML methods. For example, k-Nearest Neighbors (k-NN) algorithm, that groups similar objects into the same class, is developed through heuristic training procedures. In this way, the hypothesis space (a set of functions / models from which the best one fitting the data is selected) is not defined in advance and requires data to be defined (Klesk and Korzen, 2011). Defining a priori a list of alternative models for humans is an even tougher task (Schacter et al., 2010; Zhu, 2015). This leads to formulating further assumptions (Zhu et al., 2009), which do not often hold in practice. We will explain these assumptions in Section 4.3.2.

As an alternative, we propose to exploit AS (Bousquet and Elisseeff, 2002; Oneto et al., 2015a) in order to compute the Human Algorithmic Stability (HAS), which does not rely on the definition of a set of models and does not require any additional assumptions. In this study, we comparatively benchmark HRC and HAS, by designing experiments

[1]This chapter is written based on Vahdat et al. (2016b) which completes and extends the preliminary results reported in Vahdat et al. (2015d).

to analyze the way a group of students learns the tasks with different difficulties, and we compare the two approaches to verify which one is the most informative for getting more insights into HL. To reach this purpose, three different experiments were performed from October 2014 till May 2015 with 606 students of various engineering majors from the University of Genoa, Italy. We generated unique questionnaires for every student to measure HRC and HAS over 7 groups of students, as described in the next sections. Filled questionnaires were collected, anonymized, digitized, and analyzed. Our results show that HAS is influenced by the nature and the complexity of the problem to learn. Moreover, contrarily to HRC, HAS is also able to capture the fast-learning ability of a human when dealing with simple problems: this allows providing new perspectives with reference to the human tendency to overfit training data depending on the nature of the problem faced. These results can thus function as a bridge between ML and HL, for the measure of the propensity of the learner towards CL versus simple memorization.

This Chapter is structured as follows: Section 4.2 presents the theoretical ML framework, Section 4.3 relates the ML framework to HL, Section 4.4 describes our experimental design, Section 4.5 reports the results of our study, the conclusions are discussed in Section 4.6, and finally a summary is provided in Section 4.7.

4.2 Rademacher Complexity and Algorithmic Stability in Machine Learning

Let us consider the classical binary classification framework (Vapnik, 1998). Let $\mathcal{X}$ and $\mathcal{Y} = \{\pm 1\}$ be, respectively, an input and an output space. We consider a set of labeled independent and identically distributed (i.i.d.) data $\mathcal{S}_n : \{Z_1, \ldots, Z_n\}$ of size n, where $Z_{i \in \{1,\ldots,n\}} = (X_i, Y_i)$, with $X_i \in \mathcal{X}$ and $Y_i \in \mathcal{Y}$, sampled from an unknown distribution μ over $\mathcal{X} \times \mathcal{Y}$. We also define two modified training sets: $\mathcal{S}_n^{\setminus i}$, where the i-th element is removed and $\mathcal{S}_n^i$, where the i-th element is replaced with Z_i', which is another i.i.d. pattern sampled from μ:

$$\mathcal{S}_n^{\setminus i} : \{Z_1, \ldots, Z_{i-1}, Z_{i+1}, \ldots, Z_n\}, \quad \mathcal{S}_n^i : \{Z_1, \ldots, Z_{i-1}, Z_i', Z_{i+1}, \ldots, Z_n\}. \tag{4.1}$$

A learning algorithm $\mathscr{A}$ maps $\mathcal{S}_n$ into a function $f : \mathscr{A}_{\mathcal{S}_n}$ from $\mathcal{X}$ to $\mathcal{Y}$. In particular, $\mathscr{A}$ allows designing $f \in \mathcal{F}$ and defining the hypothesis space $\mathcal{F}$, which is generally unknown.

Even if often not specified (Bousquet and Elisseeff, 2002; Oneto et al., 2015a), there are some properties that the algorithm $\mathscr{A}$ must satisfy in order to ensure the validity of the results of the next sections. In particular, we consider only deterministic algorithms. It is also assumed that the algorithm $\mathscr{A}$ is symmetric with respect to $\mathcal{S}_n$, i.e. it does not depend on the order of the elements in the training set.

The accuracy of $\mathscr{A}_{\mathcal{S}_n}$ in representing the hidden relationship μ is measured with reference to $\ell(\mathscr{A}_{\mathcal{S}_n}, Z) : \mathcal{F} \times (\mathcal{X} \times \mathcal{Y}) \to \mathbb{R}$.

$\ell(\mathscr{A}_{\mathcal{S}_n}, Z)$ is called the *loss* function and measures the loss or discrepancy between $\mathscr{A}_{\mathcal{S}_n}$ and Z. Since we are dealing with binary classification problems, we use the hard loss function:

$$\ell(\mathscr{A}_{\mathcal{S}_n}, Z) = \frac{1 - Yf(X)}{2} \in \{0, 1\}, \tag{4.2}$$

which counts the number of misclassified examples.

The expected value of the *loss* is given by the risk functional R. The quantity of interest is defined as the generalization error, namely the error that a model will perform on new data generated by μ and previously unseen:

$$R(\mathscr{A}_{\mathcal{S}_n}) = \mathbb{E}_Z \ell(\mathscr{A}_{\mathcal{S}_n}, Z). \tag{4.3}$$

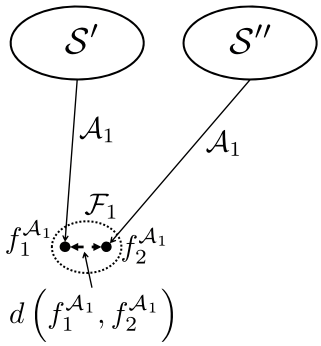
Unfortunately since μ is unknown, the random variable $R(\mathscr{A}_{\mathcal{S}_n})$ cannot be computed and, consequently, must be estimated. Two common empirical estimators are the Empirical ($\widehat{R}_{\text{EMP}}(\mathscr{A}_{\mathcal{S}_n})$) and Leave-One-Out ($\widehat{R}_{\text{LOO}}(\mathscr{A}_{\mathcal{S}_n})$) errors:

$$\widehat{R}_{\text{EMP}}(\mathscr{A}_{\mathcal{S}_n}, \mathcal{S}_n) = \frac{1}{n} \sum_{Z \in \mathcal{S}_n} \ell(\mathscr{A}_{\mathcal{S}_n}, Z), \quad \widehat{R}_{\text{LOO}}(\mathscr{A}_{\mathcal{S}_n}, \mathcal{S}_n) = \frac{1}{n} \sum_{i=1}^{n} \ell(\mathscr{A}_{\mathcal{S}_n^{\setminus i}}, Z_i). \tag{4.4}$$

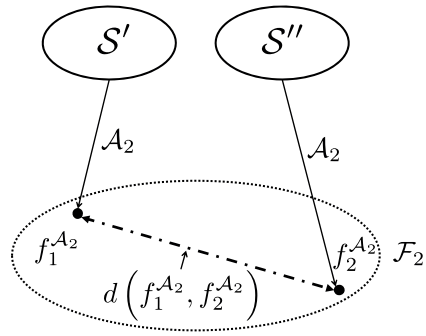
In order to estimate the generalization error, given one of its empirical estimators, we make use of two powerful statistical measures: RC (Anguita et al., 2012; Bartlett et al., 2002; Bartlett and Mendelson, 2002; Koltchinskii, 2001; Oneto et al., 2015b,c) and AS (Bousquet and Elisseeff, 2002; Mukherjee et al., 2006; Oneto et al., 2015a; Poggio et al., 2004). The difference between the two approaches can be clarified through the graphical representation, proposed in Figure 4.2.

RC introduction: Basically, RC allows estimating the size of a class of functions. Let $\mathscr{A}_1$ and $\mathscr{A}_2$ be two algorithms that chooses the model, respectively, in the classes $\mathcal{F}_1$ and $\mathcal{F}_2$, and $\mathcal{S}'$ and $\mathcal{S}''$ be two different training sets, originated from the same distribution μ. The learning phase consists of finding a model from the selected hypothesis space, say $\mathcal{F}_1$, which best fits the data: if $\mathcal{S}'$ is used, then $f_1^{\mathscr{A}_1}$ is obtained, while $f_2^{\mathscr{A}_1}$ is selected if we opt to learn the dataset $\mathcal{S}''$, instead. Since, in this case, the hypothesis space $\mathcal{F}_1$ is small (namely simple) enough, $f_1^{\mathcal{F}_1}$ will be forced to be "close" (with respect to some distance $d(f_1^{\mathscr{A}_1}, f_2^{\mathscr{A}_1})$) to $f_2^{\mathcal{F}_1}$. In other words, the final outcome of the learning phase will not be heavily influenced by the randomness of the process generating the data, so the risk of learning noise, i.e. of overfitting the data, will be small (see Figure 4.2a). If, instead, we use $\mathscr{A}_2$ which chooses functions from a larger hypothesis set $\mathcal{F}_2$, the model $f_1^{\mathcal{F}_2}$ could end up "far"from $f_2^{\mathcal{F}_2}$ (with respect to the distance $d(f_1^{\mathscr{A}_2}, f_2^{\mathscr{A}_2})$): this means that the hypothesis class is too large for a particular learning task and the risk of overfitting the data is high (see Figure 4.2b). In practice, RC tries to estimate the largest $d(f_1^{\mathscr{A}}, f_2^{\mathscr{A}})$ given the hypothesis space $\mathcal{F}$ from which the algorithm chooses the model.

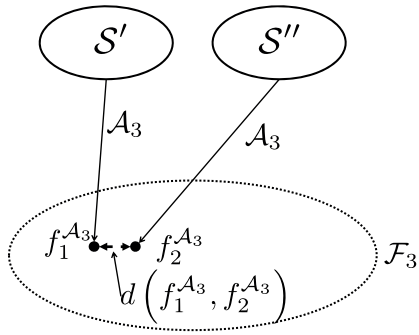
AS introduction: The advantage of using AS is that the hypothesis space $\mathcal{F}$ does not need to be known in advance. This means that, even if the algorithm chooses from



(a) $\mathcal{F}_1$: RC and AS are small (considering the size of $\mathcal{F}_1$ and $d(f_1^{\mathscr{A}_1}, f_2^{\mathscr{A}_1})$).



(b) $\mathcal{F}_2$: RC and AS are large (considering the size of $\mathcal{F}_2$ and $d(f_1^{\mathscr{A}_2}, f_2^{\mathscr{A}_2})$).



(c) $\mathcal{F}_3$: RC is large but AS is small (considering the size of $\mathcal{F}_3$ and $d(f_1^{\mathscr{A}_3}, f_2^{\mathscr{A}_3})$).

Figure 4.2: Different approaches to learning. $\mathscr{A}_1$ and $\mathscr{A}_2$ are two algorithms that chooses the model from the classes $\mathcal{F}_1$ and $\mathcal{F}_2$. $\mathcal{S}'$ and $\mathcal{S}''$ are two different training sets, originated from the same distribution μ. $d(f_1^{\mathscr{A}_{1,2,3}}, f_2^{\mathscr{A}_{1,2,3}})$ shows the distance between the chosen models during the two different learning processes.

a large hypothesis space, given a particular μ, the algorithm will choose models that are close to each other. A graphical description is shown in Figure 4.2: let us consider, again, algorithms $\mathscr{A}_1$ (see Figure 4.2a) and $\mathscr{A}_2$(see Figure 4.2b), and the two training sets $\mathcal{S}'$ and $\mathcal{S}''$. In this case, $d(f_1^{\mathscr{A}_2}, f_2^{\mathscr{A}_2}) \gg d(f_1^{\mathscr{A}_1}, f_2^{\mathscr{A}_1})$ consequently, we derive the same conclusion of the analysis performed with RC-based approach. Instead, if we apply a stable algorithm $\mathscr{A}_3$ (see Figure 4.2c) on $\mathcal{S}'$ and $\mathcal{S}''$, we obtain that $d(f_1^{\mathscr{A}_3}, f_2^{\mathscr{A}_3})$ is small even if $\mathcal{F}_3$ is a large hypothesis space. Note that, since we are just looking at $d(f_1^{\mathscr{A}}, f_2^{\mathscr{A}})$, we do not need to explicitly know $\mathcal{F}$: the stability of a particular learning algorithm is simply computed by measuring $d(f_1^{\mathscr{A}}, f_2^{\mathscr{A}})$.

In the next sections, we recall some results that are useful for the remaining part of this work. Some of the theoretical results can be retrieved in previous studies (Anguita et al., 2012; Bartlett and Mendelson, 2002; Bousquet and Elisseeff, 2002; Oneto et al., 2015a; Poggio et al., 2004), but the adaptation of these approaches to the context of this work is entirely new.

4.2.1 Understanding Learning Ability through Rademacher Complexity

As detailed in the previous section, the quantity of interest in any learning procedure is the generalization error $R(\mathscr{A}_{\mathcal{S}_n})$ which is not measurable since μ is unknown. We can however measure its empirical estimators which are optimistically biased estimators of $R(\mathscr{A}_{\mathcal{S}_n})$ (Bousquet and Elisseeff, 2002; Vapnik, 1998). In order to estimate this bias, we can use RC to bound Uniform Deviation which is the maximum distance between the generalization error and the empirical error (Anguita et al., 2012; Bartlett and Mendelson, 2002):

$$\widehat{U}_n(\mathcal{F}) = \sup_{f \in \mathcal{F}} \left[R(f) + \widehat{R}_{\text{EMP}}(f, \mathcal{S}_n) \right], \tag{4.5}$$

since

$$R(\mathscr{A}_{\mathcal{S}_n}) \leq \widehat{R}_{\text{EMP}}(\mathscr{A}_{\mathcal{S}_n}, \mathcal{S}_n) + \widehat{U}_n(\mathcal{F}). \tag{4.6}$$

The expected value of $\widehat{U}_n(\mathcal{F})$ can be defined as:

$$U_n(\mathcal{F}) = \mathbb{E}_{\mathcal{S}_n} \widehat{U}_n(\mathcal{F}). \tag{4.7}$$

The RC of a class of functions $\mathcal{F}$ and its expected value are defined as:

$$\widehat{C}_n(\mathcal{F}) = \mathbb{E}_{\boldsymbol{\sigma}} \sup_{f \in \mathcal{F}} \frac{2}{n} \sum_{i=1}^{n} \sigma_i \ell(f, Z_i), \quad C_n(\mathcal{F}) = \mathbb{E}_{\mathcal{S}_n} \widehat{C}_n(\mathcal{F}), \tag{4.8}$$

where $\sigma_1, \ldots, \sigma_n$ are n independent random variables for which $\mathbb{P}(\sigma_i = +1) = \mathbb{P}(\sigma_i = -1) = 1/2$. Note that, since we use the hard loss function, it is possible to prove that:

$$\widehat{C}_n(\mathcal{F}) = \mathbb{E}_{\boldsymbol{\sigma}} \sup_{f \in \mathcal{F}} \frac{2}{n} \sum_{i=1}^{n} \sigma_i \ell(f, Z_i) = 1 - 2\mathbb{E}_{\boldsymbol{\sigma}} \inf_{f \in \mathcal{F}} \widehat{R}_{\text{EMP}}(f, \mathcal{S}_n^{\boldsymbol{\sigma}}), \tag{4.9}$$

where $\mathcal{S}_n^\sigma : \{Z_1^\sigma, \ldots, Z_n^\sigma\}$ of size n, and $Z_{i\in\{1,\ldots,n\}}^\sigma = (X_i, \sigma_i)$. An upper bound of $R(\mathscr{A}_{\mathcal{S}_n})$ in terms of $\widehat{C}(\mathcal{F})$ was proposed in Bartlett and Mendelson (2002) and the proof consists mainly of an application of the McDiarmid's inequality (McDiarmid, 1989). Since both RC and Uniform Deviation are bounded difference functions (Anguita et al., 2012; Bartlett and Mendelson, 2002) then:

$$C_n(\mathcal{F}) \leq \widehat{C}_n(\mathcal{F}) + \sqrt{\frac{2\log(\frac{1}{\delta})}{n}}, \quad \widehat{U}_n(\mathcal{L}) \leq U_n(\mathcal{F}) + \sqrt{\frac{\log(\frac{1}{\delta})}{2n}}. \tag{4.10}$$

The bounds hold with probability $(1-\delta)$, where δ is a user-defined level of confidence. Using these results, it is possible to bound $\widehat{U}_n(\mathcal{F})$, by noting that (Bartlett and Mendelson, 2002):

$$U_n(\mathcal{F}) \leq C_n(\mathcal{F}). \tag{4.11}$$

By exploiting Eqns. (4.10) and (4.11), we obtain the following relation which holds with probability $(1-\delta)$ (Anguita et al., 2012; Bartlett and Mendelson, 2002):

$$R(\mathscr{A}_{\mathcal{S}_n}) \leq \widehat{R}_{\text{EMP}}(\mathscr{A}_{\mathcal{S}_n}, \mathcal{S}_n) + \widehat{C}_n(\mathcal{F}) + 3\sqrt{\frac{\log(\frac{2}{\delta})}{2n}}. \tag{4.12}$$

All the quantities involved in the bound of Eq. (4.12) can be empirically computed by exploiting the available observations $(\mathcal{S}_n)$. Note that $\widehat{C}_n(\mathcal{F})$ requires that an expectation over all σ is performed. This is obviously infeasible in practice, and a Monte Carlo estimation of the quantity can be computed instead (Anguita et al., 2012; Bartlett and Mendelson, 2002). A more refined alternative is to exploit a recent result (Bartlett and Mendelson, 2002; Klesk and Korzen, 2011), which shows that the following quantity can be used:

$$\widehat{\widehat{C}}_n(\mathcal{F}) = 1 - 2\inf_{f\in\mathcal{F}} \widehat{R}_{\text{EMP}}(f, \mathcal{S}_n^\sigma). \tag{4.13}$$

that is RC computed with just a single draw of $\boldsymbol{\sigma}$. In fact, $\widehat{\widehat{C}}_n(\mathcal{F})$ is also a bounded difference function and consequently, the following bound holds with probability $(1-\delta)$:

$$C_n(\mathcal{F}) \leq \widehat{\widehat{C}}_n(\mathcal{F}) + \sqrt{\frac{8\log(\frac{1}{\delta})}{n}}. \tag{4.14}$$

By exploiting Eqns. (4.10), (4.11), and (4.14), we can prove that the following bound holds with probability $(1-\delta)$:

$$R(\mathscr{A}_{\mathcal{S}_n}) \leq \widehat{R}_{\text{EMP}}(\mathscr{A}_{\mathcal{S}_n}, \mathcal{S}_n) + \widehat{\widehat{C}}_n(\mathcal{F}) + 5\sqrt{\frac{\log(\frac{2}{\delta})}{2n}}. \tag{4.15}$$

The bound of Eq. (4.15) is slightly looser but it is more computational tractable respect to the one of Eq. (4.12). It is worth noting that all the constants in the bound can be improved (Oneto et al., 2015b) but this is out of the scope of our work.

Unfortunately, RC requires the hypothesis space to be fixed before seeing the data and even functions that will never be picked up by the learning procedure are taken into account when estimating the Uniform Deviation. This could compromise our ability to understand the learning properties of our algorithm: if an effective algorithm chooses functions that belong to a complex class, the bound of Eq. (4.12) is loose (Oneto et al., 2015a; Poggio et al., 2004) and so we cannot guarantee the performance of the algorithm.

4.2.2 Understanding Learning Ability through Algorithmic Stability

Stability does not require a class of functions $\mathcal{F}$ to be defined a priori, since the set of models is implicitly derived by the algorithm $\mathscr{A}$ itself. For this reason, the simple deviation $\widehat{D}(\mathscr{A}_{\mathcal{S}_n}, \mathcal{S}_n)$ of the generalization error from $\widehat{R}_{\text{EMP}}(\mathscr{A}_{\mathcal{S}_n}, \mathcal{S}_n)$ or $\widehat{R}_{\text{LOO}}(\mathscr{A}_{\mathcal{S}_n}, \mathcal{S}_n)$ is analyzed (Bousquet and Elisseeff, 2002; Mukherjee et al., 2006; Oneto et al., 2015a; Poggio et al., 2004):

$$\widehat{D}_{\text{EMP}}(\mathscr{A}_{\mathcal{S}_n}, \mathcal{S}_n) = \left| R(\mathscr{A}_{\mathcal{S}_n}) - \widehat{R}_{\text{EMP}}(\mathscr{A}_{\mathcal{S}_n}, \mathcal{S}_n) \right|, \tag{4.16}$$

$$\widehat{D}_{\text{LOO}}(\mathscr{A}_{\mathcal{S}_n}, \mathcal{S}_n) = \left| R(\mathscr{A}_{\mathcal{S}_n}) - \widehat{R}_{\text{LOO}}(\mathscr{A}_{\mathcal{S}_n}, \mathcal{S}_n) \right|. \tag{4.17}$$

Consider that, the deterministic squared counterpart of $\widehat{D}_{\text{EMP}}(\mathscr{A}_{\mathcal{S}_n}, \mathcal{S}_n)$ and $\widehat{D}_{\text{LOO}}(\mathscr{A}_{\mathcal{S}_n}, \mathcal{S}_n)$ can be defined as:

$$D^2_{\text{EMP}}(\mathscr{A}, n) = \mathbb{E}_{\mathcal{S}_n}\left[\widehat{D}_{\text{EMP}}(\mathscr{A}_{\mathcal{S}_n}, \mathcal{S}_n)\right]^2, \tag{4.18}$$

$$D^2_{\text{LOO}}(\mathscr{A}, n) = \mathbb{E}_{\mathcal{S}_n}\left[\widehat{D}_{\text{LOO}}(\mathscr{A}_{\mathcal{S}_n}, \mathcal{S}_n)\right]^2. \tag{4.19}$$

In order to study $\widehat{D}(\mathscr{A}_{\mathcal{S}_n}, \mathcal{S}_n)$, we can adopt different approaches. The first one consists of using Hypothesis Stability $H(\mathscr{A}, n)$:

$$H_{\text{EMP}}(\mathscr{A}, n) = \mathbb{E}_{\mathcal{S}_n, Z'_i}\left| \ell(\mathscr{A}_{\mathcal{S}_n}, Z_i) - \ell(\mathscr{A}_{\mathcal{S}_n^i}, Z_i) \right| \leq \beta_{\text{EMP}}, \tag{4.20}$$

$$H_{\text{LOO}}(\mathscr{A}, n) = \mathbb{E}_{\mathcal{S}_n, Z}\left| \ell(\mathscr{A}_{\mathcal{S}_n}, Z) - \ell(\mathscr{A}_{\mathcal{S}_n^{\setminus i}}, Z) \right| \leq \beta_{\text{LOO}}. \tag{4.21}$$

Lemma 3 in Bousquet and Elisseeff (2002) proves that:

$$D^2_{\text{EMP}}(\mathscr{A}, n) \leq \frac{1}{2n} + 3H_{\text{EMP}}(\mathscr{A}, n), \quad D^2_{\text{LOO}}(\mathscr{A}, n) \leq \frac{1}{2n} + 3H_{\text{LOO}}(\mathscr{A}, n). \tag{4.22}$$

By exploiting the Chebyshev inequality (Casella and Berger, 2002), for a random variable a, with probability $(1-\delta)$ we obtain:

$$\mathbb{P}[a > t] \leq \mathbb{E}[a^2]/t^2, \quad a < \sqrt{\mathbb{E}[a^2]/\delta}. \tag{4.23}$$

Then, by combining Eqns. (4.16), (4.20), and (4.22) (or analogously, Eqns. (4.17), (4.21), and (4.22)) with probability $(1-\delta)$ we obtain that:

$$R(\mathscr{A}_{\mathcal{S}_n}) \leq \widehat{R}_{\mathrm{EMP}}(\mathscr{A}_{\mathcal{S}_n}, \mathcal{S}_n) + \sqrt{\frac{1}{2n\delta} + \frac{3\beta_{\mathrm{EMP}}}{\delta}}, \tag{4.24}$$

$$R(\mathscr{A}_{\mathcal{S}_n}) \leq \widehat{R}_{\mathrm{LOO}}(\mathscr{A}_{\mathcal{S}_n}, \mathcal{S}_n) + \sqrt{\frac{1}{2n\delta} + \frac{3\beta_{\mathrm{LOO}}}{\delta}}. \tag{4.25}$$

In Theorem 11 of Bousquet and Elisseeff (2002), it is proved that $H_{\mathrm{EMP}}(\mathscr{A}, n) \leq 2H_{\mathrm{LOO}}(\mathscr{A}, n)$ but unfortunately, as remarked in Mukherjee et al. (2006), this proof contains an error and consequently Leave-One-Out Hypothesis Stability does not imply $H_{\mathrm{EMP}}(\mathscr{A}, n)$ Stability. Moreover, in Oneto et al. (2015a) it is proved that $H_{\mathrm{LOO}}(\mathscr{A}, n)$ can be estimated from the data, while this is not possible for $H_{\mathrm{EMP}}(\mathscr{A}, n)$. Consequently, from now on, we only deal with Leave-One-Out Hypothesis Stability since, as described in the rest of this section, it is the only one which leads to a fully empirical stability-based bound.

In order to estimate $H_{\mathrm{LOO}}(\mathscr{A}, n)$ from the data, we need to suppose that for the algorithm under exam ($\mathscr{A}$), Hypothesis Stability does not increase with the cardinality of the training set:

$$H_{\mathrm{LOO}}(\mathscr{A}, n) \leq H_{\mathrm{LOO}}(\mathscr{A}, {}^{\sqrt{n}}/_{2}). \tag{4.26}$$

We point out that Property (4.26) is a desirable requirement for any learning algorithm: in fact, the impact on the learning procedure of removing samples from $\mathcal{S}_n$ should decrease on average, as n grows. Note that, Property (4.26) has already been studied by many researchers in the past. In particular, Property (4.26) is related to the notion of consistency (Devroye et al., 1996; Steinwart, 2005) and connections can also be identified with the trend of the learning curves of an algorithm (Dietrich et al., 1999; Mukherjee et al., 2003; Opper, 1995; Opper et al., 1990). Moreover, such quantities are strictly linked to the concept of Smart Rule (Devroye et al., 1996). It is worth underlining that, in the above-referenced works, Property (4.26) is proved to be satisfied by many well-known algorithms (Support Vector Machines, Kernelized Regularized Least Squares, k–Local Rules with $k > 1$).

In order to derive a fully empirical bound we have to study $H_{\mathrm{LOO}}(\mathscr{A}, {}^{\sqrt{n}}/_{2})$. For this purpose, the following empirical quantity can be introduced (Oneto et al., 2015a):

$$\widehat{H}_{\mathrm{LOO}}(\mathscr{A}, {}^{\sqrt{n}}/_{2}, \mathcal{S}_n) = \frac{8}{n\sqrt{n}} \sum_{k=1}^{{}^{\sqrt{n}}/_{2}} \sum_{j=1}^{{}^{\sqrt{n}}/_{2}} \sum_{i=1}^{{}^{\sqrt{n}}/_{2}} \left| \ell(\mathscr{A}_{\breve{\mathcal{S}}^k_{{}^{\sqrt{n}}/_{2}}}, \breve{Z}^k_j) - \ell(\mathscr{A}_{(\breve{\mathcal{S}}^k_{{}^{\sqrt{n}}/_{2}})^{\setminus i}}, \breve{Z}^k_j) \right|, \tag{4.27}$$

where $\forall k \in \{1, \ldots, {}^{\sqrt{n}}/_{2}\}$:

$$\breve{\mathcal{S}}^k_{{}^{\sqrt{n}}/_{2}} : \left\{ Z_{(k-1)\sqrt{n}+1}, \ldots, Z_{(k-1)\sqrt{n}+{}^{\sqrt{n}}/_{2}} \right\}, \quad \breve{Z}^k_j : Z_{(k-1)\sqrt{n}+{}^{\sqrt{n}}/_{2}+j}. \tag{4.28}$$

Note that, the quantity of Eq. (4.27) is the empirical unbiased estimator of $H_{\mathrm{LOO}}(\mathscr{A}, \sqrt{n}/2)$ and then:

$$H_{\mathrm{LOO}}(\mathscr{A}, \sqrt{n}/2) = \mathbb{E}_{\mathcal{S}_n} \widehat{H}_{\mathrm{LOO}}(\mathscr{A}, \sqrt{n}/2, \mathcal{S}_n). \tag{4.29}$$

As a consequence, with probability $(1-\delta)$, the difference between $H_{\mathrm{LOO}}(\mathscr{A}, \sqrt{n}/2)$ and $\widehat{H}_{\mathrm{LOO}}(\mathscr{A}, \sqrt{n}/2, \mathcal{S}_n)$ can be bounded by exploiting, for example, the Hoeffding inequality (Hoeffding, 1963; Oneto et al., 2015a):

$$H_{\mathrm{LOO}}(\mathscr{A}, \sqrt{n}/2) \leq \widehat{H}_{\mathrm{LOO}}(\mathscr{A}, \sqrt{n}/2, \mathcal{S}_n) + \sqrt{\frac{\log(\frac{1}{\delta})}{\sqrt{n}}}. \tag{4.30}$$

Combining Eqns. (4.25) and (4.30), the following stability bound, holding with probability $(1-\delta)$, can be derived:

$$R(\mathscr{A}_{\mathcal{S}_n}) \leq \widehat{R}_{\mathrm{LOO}}(\mathscr{A}_{\mathcal{S}_n}, \mathcal{S}_n) + \sqrt{\frac{2}{\delta}\left[\frac{1}{2n} + 3\left(\widehat{H}_{\mathrm{LOO}}(\mathscr{A}, \sqrt{n}/2, \mathcal{S}_n) + \sqrt{\frac{\log(\frac{2}{\delta})}{\sqrt{n}}}\right)\right]}. \tag{4.31}$$

The bound of Eq. (4.31) takes into account only empirical quantities and, in order to compute them, we do not need to know $\mathcal{F}$ but we need to apply the algorithm $\mathscr{A}$ on a series of modified training sets that are all built using $\mathcal{S}_n$. Moreover, only the functions that, given a set of data, could be actually learned by $\mathscr{A}$ are contemplated, differently from RC. When stability bounds are used, we only need to prove that Property (4.26) holds for the chosen algorithm.

Up to now, the preliminaries of this work were presented. In the next section, we show how the concepts of RC and AS are implemented to study HL.

4.3 From Machine Learning to Human Learning

In this section we show how to measure the three different quantities as described in Section 4.2 in human learning: the generalization error (that we call Human Error), HRC, and HAS. We describe how these quantities should behave from a ML point of view, and through ad-hoc experiments on humans, we check whether these behaviours can be observed for HL as well.

In order to measure the three above-mentioned quantities, it is important to consider that a human is fundamentally different from a ML algorithm in the sense that, humans have memory while ML algorithms do not. In other words, we cannot state that humans are deterministic in the sense that, if we give the same problem P to a human at time t and at time $t+\Delta$, the selected models, respectively M^t and $M^{t+\Delta}$, could be different ($M^t \neq M^{t+\Delta}$). Moreover, if a human is given a problem $P1$ at time t, and another problem $P2$ at time $t+\Delta$, the selected models respectively M^t_{P1} and $M^{t+\Delta}_{P2}$, could be different from the one selected by the same person given the problem $P2$ at time t and problem $P1$ at

time $t + \Delta$, respectively M_{P2}^{t} and $M_{P1}^{t+\Delta}$. In other words, $M_{P2}^{t} \neq M_{P2}^{t+\Delta}$ and $M_{P1}^{t} \neq M_{P1}^{t+\Delta}$. Note that, this problem is also underlined in Zhu et al. (2009). These issues, as are shown in the next section, prevent us to measure, for example, $\widehat{C}_n(\mathcal{F})$, $\widehat{R}_{\text{LOO}}(\mathscr{A}_{\mathcal{S}_n}, \mathcal{S}_n)$, and $\widehat{H}_{\text{LOO}}(\mathscr{A}, \sqrt{n}/2, \mathcal{S}_n)$ for a human. This is due to the fact that, contrary to a ML algorithm, a human cannot be questioned as many times as we need, because we risk to falsify the results. Unfortunately, for measuring the quantity of interest for a single human we should ask her to solve several slightly modified instances of the problem as described in Section 4.2. For this reason, we propose to average these quantities over different humans as in Zhu et al. (2009).

In order to show how to measure Human Error, HRC, and HAS, first we need to introduce some additional quantities. In particular, we consider m sets $\mathcal{D}_m^{\text{LEARN}} : \{\mathcal{L}_{1,n}, \ldots, \mathcal{L}_{m,n}\}$ of n labeled i.i.d. data and other m sets $\mathcal{D}_m^{\text{TEST}} : \{\mathcal{T}_{1,p}, \ldots, \mathcal{T}_{m,p}\}$ of p labeled i.i.d. data all sampled from μ. Moreover, we consider a group of $2m$ humans $\mathcal{G}^{2m} : \{\mathscr{H}^1, \ldots, \mathscr{H}^{2m}\}$ and each $\mathscr{H}^{i \in \{1,\ldots,2m\}}$ chooses functions in the unknown space $\mathcal{F}^{i \in \{1,\ldots,2m\}}$.

4.3.1 Human Error

The first quantity of interest in learning, as described in Section 4.2, is the generalization error of an algorithm $\mathscr{A}$. In this case, the algorithm $\mathscr{A}$ is the human $\mathscr{H}$ that takes some sample of the distribution μ, which basically is a task to learn, and returns a model f in an unknown $\mathcal{F}$. Obviously we cannot explicitly access f but we can ask the human $\mathscr{H}$ to label some previously unseen unlabeled samples, and check whether her answer is in agreement with the true label of the sample. Here, we are interested in measuring the expected Human Error:

$$R_n^{\mu} = \mathbb{E}_{\mathscr{H}, \mathcal{S}_n, Z} \ell(\mathscr{H}_{\mathcal{S}_n}, Z). \tag{4.32}$$

In other words, R_n^{μ} is the average ability of the human to learn a binary classification task μ, given n i.i.d. samples $\mathcal{S}_n$ sampled from μ. Since we do not have all the humans but just a finite group of them $\mathcal{G}^{2m}$, and since we cannot give them many times different samples to learn coming from the same μ (they would remember the samples and this leads to compromising the results), we can estimate R_n^{μ} thanks to $\mathcal{D}_m^{\text{LEARN}}$ and $\mathcal{D}_m^{\text{TEST}}$:

$$\widehat{R}_n^{\mu} = \frac{1}{m} \sum_{i=1}^{m} \frac{1}{p} \sum_{Z \in \mathcal{T}_{i,p}} \ell(\mathscr{H}^i_{\mathcal{L}_{i,n}}, Z). \tag{4.33}$$

Note that, $\widehat{R}_n^{\mu}$ is an unbiased estimator of R_n^{μ} and in particular we can use the Hoeffding inequality to state that the following bound holds, with probability $(1 - \delta)$:

$$R_n^{\mu} \leq \widehat{R}_n^{\mu} + \sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{2m}}. \tag{4.34}$$

Consider that, the accuracy in estimating R_n^μ through $\widehat{R}_n^\mu$ depends only on m. We decided to add the parameter p (more data to label for each human) since, as can be seen later in the experimental section, we discovered that this addition leads to a more stable estimate.

Based on these definitions, one can investigate how R_n^μ varies by changing task (μ) or the number of samples n available for learning. From a ML perspective, it is expected that R_n^μ decreases with n for consistency reasons: the more samples we use for learning, the smaller error we will obtain on previously unseen data. If this behaviour is not observed, it would show that the samples have been memorized rather than learned (Devroye et al., 1996; Mukherjee et al., 2003; Opper, 1995; Steinwart, 2005; Vapnik, 1998).

4.3.2 Human Rademacher Complexity

For what concerns HRC, the average RC of a human needs to be estimated:

$$.C_n(\mathcal{F}) = \mathbb{E}_{\mathcal{F},\mathcal{S}_n,\boldsymbol{\sigma}}\left[1 - 2\inf_{f\in\mathcal{F}} \widehat{R}_{\mathrm{EMP}}(f, \mathcal{S}_n^{\boldsymbol{\sigma}})\right]. \tag{4.35}$$

Note that in this case, $C_n(\mathcal{F})$ cannot be computed since the space $\mathcal{F}$ is unknown. For this reason, an assumption is made: every human is always able to choose the best model of which she/he is capable. Based on this assumption, it is possible to measure the infimum in Eq. (4.35). Unfortunately, this assumption does not hold in practice because the process of learning of a human is not equivalent to a simple minimization process. Instead, learning may be viewed as a complex process, rather than a collection of factual and procedural knowledge (Schacter et al., 2010).

Based on this assumption, we can reformulate $C_n(\mathcal{F})$ in the following way:

$$C_n^\mu = \mathbb{E}_{\mathscr{H},\mathcal{S}_n,\boldsymbol{\sigma}}\left[1 - 2\widehat{R}_{\mathrm{EMP}}(\mathscr{H}_{\mathcal{S}_n^{\boldsymbol{\sigma}}}, \mathcal{S}_n^{\boldsymbol{\sigma}})\right]. \tag{4.36}$$

Consider that, C_n^μ (as $C_n(\mathcal{F})$) does not depend on $\mathbb{P}\{Y|X\}$ but just on $\mathbb{P}\{X\}$, since when computing C_n^μ the labels are disregarded (see Section 4.2.1 for details) (Anguita et al., 2011a). In other words, many μ with the same $\mathbb{P}\{X\}$, but different $\mathbb{P}\{Y|X\}$, have the same C_n^μ.

Unfortunately, C_n^μ cannot be computed for the same reason that R_n^μ cannot be computed (see Section 4.3.1), so we can estimate C_n^μ by using $\mathcal{G}^{2m}$ and $\mathcal{D}_m^{\mathrm{LEARN}}$ with the following empirical estimator:

$$\widehat{C}_n^\mu = \frac{1}{m}\sum_{i=1}^{m}\left[1 - 2\widehat{R}_{\mathrm{EMP}}(\mathscr{H}^i_{\mathcal{L}_{i,n}^{\boldsymbol{\sigma}^i}}, \mathcal{L}_{i,n}^{\boldsymbol{\sigma}^i})\right]. \tag{4.37}$$

Note that $\widehat{C}_n^\mu$ is an unbiased estimator of C_n^μ and in particular, we can use the Hoeffding inequality to state that the following bound holds, with probability $(1-\delta)$:

$$C_n^\mu \leq \widehat{C}_n^\mu + \sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{2m}}. \tag{4.38}$$

Based on these definitions, one can investigate how C_n^μ varies by changing task (μ) or the number of samples n available for learning. From the ML perspective, we expect that RC decreases when n increases. The more data we need to learn, the less $\mathcal{F}$ is able to fit random noise as $\mathbb{P}\{Y|X\}$. If C_n^μ does not decrease with n, it shows that the human is not learning but memorizing the information. In other words, $\mathcal{F}$ is too large to be able to learn a particular task (Anguita et al., 2011a; Bartlett and Mendelson, 2002; Koltchinskii, 2001; Zhu et al., 2009). Note that C_n^μ is just a possible indicator of the ability of the algorithm to learn the task. There are cases where, even if it is not possible to prove through RC that an algorithm is learning the task, indeed the algorithm is effectively learning (Poggio et al., 2004) or it is learning at much faster rate than we can prove (Bartlett et al., 2005; Oneto et al., 2015d).

4.3.3 Human Algorithmic Stability

The last quantity to estimate is the average Human Leave-One-Out Hypothesis Stability. In particular, we are interested in the quantity of Eq. (4.21). As explained in Section 4.2.2, stability does not require the knowledge of the $\mathcal{F}$ from which the humans will choose the model, but it requires that the humans solve many instances of a particular problem. Consequently, the average Human Leave-One-Out Hypothesis Stability can be expressed as:

$$\begin{aligned} H_n^\mu &= \mathbb{E}_{\mathscr{H},\mathcal{S}_n,Z}\left|\ell(\mathscr{H}_{\mathcal{S}_n},Z)-\ell(\mathscr{H}_{\mathcal{S}_n^{\setminus i\in\{1,\dots,n\}}},Z)\right| \\ &= \mathbb{E}_{\mathscr{H},\mathcal{S}_n,Z}\left|\ell(\mathscr{H}_{\mathcal{S}_n},Z)-\ell(\mathscr{H}_{\mathcal{S}_n^{\setminus 1}},Z)\right|, \end{aligned} \tag{4.39}$$

Differently from to RC, H_n^μ depends both on $\mathbb{P}\{X\}$ and $\mathbb{P}\{Y|X\}$. In other words, each μ has, in general, a different H_n^μ even if $\mathbb{P}\{X\}$ is the same.

H_n^μ cannot be computed for the same reason we cannot compute R_n^μ and C_n^μ, so we can estimate it by using $\mathcal{G}^{2m}$, $\mathcal{D}_m^{\text{LEARN}}$, and $\mathcal{D}_m^{\text{TEST}}$, and the following unbiased empirical estimator:

$$\widehat{H}_n^\mu = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{p}\sum_{Z\in\mathcal{T}_{i,p}}\left|\ell(\mathscr{H}^i_{\mathcal{L}_{i,n}},Z)-\ell(\mathscr{H}^i_{\mathcal{L}_{i,n}^{\setminus i}},Z)\right| \tag{4.40}$$

By using the Hoeffding inequality again, we can state that the following bound holds, with probability $(1-\delta)$:

$$H_n^\mu \leq \widehat{H}_n^\mu + \sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{2m}}. \tag{4.41}$$

Note that, as for $\widehat{R}_n^\mu$, even if the accuracy of our estimate depends only on m, we introduce p for making $\widehat{H}_n^\mu$ a more stable estimator of H_n^μ as can be seen later in the experimental section.

A problem in computing H_n^μ and $\widehat{H}_n^\mu$ is that they do not take into account the ability of humans to memorize. In particular, it is difficult that people will change their mind once they learned a concept (Schacter et al., 2010). This is a problem when measuring $|\ell(\mathscr{H}^i_{\mathcal{L}_{i,n}}, Z) - \ell(\mathscr{H}^i_{\mathcal{L}^{\setminus i}_{i,n}}, Z)|$, since we need to provide the same data-set to the human with one sample removed in two different moments in time. This produces a bias in the estimation, as described in the beginning of Section 4.3: in fact, our experiments (see Section 4.5) will show that this bias is noticeable. A solution to this problem is to measure the average cross-human stability instead of the average self-human stability. In other words, we can consider the stability as a property that connects different humans instead of the property of a single human (analogously to the size of the space of function for RC). Consequently, we can measure how stable two humans are, one with respect to the other, averaged over a group of people: this interpretation of stability measures the stability of a group of people while learning a task μ. Based on these considerations, we reformulate H_n^μ in order to measure the following notion of stability:

$$\begin{aligned}\overline{H}_n^\mu &= \mathbb{E}_{\mathscr{H}',\mathscr{H}'',\mathcal{S}_n,Z}\left|\ell(\mathscr{H}'_{\mathcal{S}_n}, Z) - \ell(\mathscr{H}''_{\mathcal{S}_n^{\setminus i\in\{1,\dots,n\}}}, Z)\right| \\ &= \mathbb{E}_{\mathscr{H}',\mathscr{H}'',\mathcal{S}_n,Z}\left|\ell(\mathscr{H}'_{\mathcal{S}_n}, Z) - \ell(\mathscr{H}''_{\mathcal{S}_n^{\setminus 1}}, Z)\right|,\end{aligned} \tag{4.42}$$

Analogously to H_n^μ, $\overline{H}_n^\mu$ can be estimated by using $\mathcal{G}^{2m}$, $\mathcal{D}_m^{\text{LEARN}}$, and $\mathcal{D}_m^{\text{TEST}}$ with the following unbiased empirical estimator:

$$\widehat{\overline{H}}_n^\mu = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{p}\sum_{Z\in\mathcal{T}_{i,p}}\left|\ell(\mathscr{H}^i_{\mathcal{L}_{i,n}}, Z) - \ell(\mathscr{H}^{i+m}_{\mathcal{L}^{\setminus i}_{i,n}}, Z)\right| \tag{4.43}$$

Knowing that, with probability $(1-\delta)$:

$$\overline{H}_n^\mu \leq \widehat{\overline{H}}_n^\mu + \sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{2m}}. \tag{4.44}$$

Based on these definitions, we investigate how $\overline{H}_n^\mu$ and H_n^μ vary by changing task (μ) or the number of samples (n) available for learning. From a ML perspective, we expect that stability will decreases as n increases: the more data is available for learning, the less impact there is in removing one single example from the data. If $\overline{H}_n^\mu$ and H_n^μ do not decrease with n, the reason would be that the machine is not learning and there are not enough examples available to retrieve the information hidden in one example from the others. In other words, $\mathscr{H}$ is not able to find a stable solution or is not able to learn the particular task effectively (Bousquet and Elisseeff, 2002; Mukherjee et al., 2006; Oneto et al., 2015a; Poggio et al., 2004).

4.4 Experimental Design

A previous work (Zhu et al., 2009) reports on an experiment targeted towards measuring HRC. Given the drawbacks of RC with respect to AS, as highlighted in the previous

sections, we built on the previous efforts and experiments to design and carry out a new experiment, which aims at estimating the average HRC and HAS. Our objective is to compare these two quantities and identify which one is the most informative for getting more insights into HL.

We performed three different experiments EX^1 (a pilot experiment), EX^2, and EX^3. EX^1 was performed in October 2014, EX^2 in November 2014, and EX^3 in April/May 2015. EX^2 was designed based on the observed results of EX^1 while EX^3 extends EX^2. The experiments were carried out with a total of 606 students from the University of Genoa (Italy): 70 students of Bioinformatics Engineering participated in EX^1; EX^2 was carried out with 307 undergraduate students of Electronic Engineering (68 students), Bioinformatics Engineering (136 students), and Computer Engineering (103 students); finally EX^3 was carried out with 229 students of Bioinformatics Engineering (63 students), Electrical Engineering (53 students), Computer Engineering (33 students), and Mechanical Engineering (80 students). Each student participated only one time in the study and was given a unique automatically generated questionnaire. Experiments were carried out with 7 groups of students carefully controlled by the professors and researchers involved in the study together with the help of university employees. Filled questionnaires were collected, digitized, anonymized, and analyzed. We describe the details of the questionnaire design in the next sections.

4.4.1 Experimental Design: EX^1

In this phase, our experiment involved two types of statistical measures: HRC and HAS. For HRC in particular, we followed the approach designed in Zhu et al. (2009). Then we designed a new experiment in order to measure HAS. We use the word "task" (or "problem") and "domain" to refer, respectively, to $\mathbb{P}\{Y|X\}$ (or μ) and $\mathbb{P}\{X\}$. The first step consisted of defining the task to be learned by students. Two domains were defined: Shape and Word. Two problems were defined for each domain: simple (linear) and difficult (non-linear)[2]. We report the detailed description of these problems as follows.

Shape Domain

The Shape domain consists of 321 computer-generated 3D shapes, parametrized by $\alpha \in [-8, +8]$, such that a small value of α leads to spiky shapes, while a large α allows to obtain smooth ones. A label was assigned to each shape, and two problems were defined in accordance with ad-hoc rules to depict tasks of increasing complexity:

- Shape Simple (SS), where $Y = +1$ if $\alpha \leq 0$ and $Y = -1$ otherwise.
- Shape Difficult (SD), where $Y = +1$ if $-4 \leq \alpha \leq 4$ and $Y = -1$ otherwise.

In Figure 4.3, the samples from Shape domain are shown, we note that for small changes of α recognizing the label for a human can be difficult since the shapes look similar. The

[2]We thank the authors of Zhu et al. (2009) and Medler et al. (2005) for providing their dataset.

(a) $\alpha = -8$ (b) $\alpha = -7.95$ (c) α=0 (d) α=0.05 (e) α=7.95 (f) α=8

Figure 4.3: Samples from Shape domain (Zhu et al., 2009). The computer-generated shapes are parametrized by $\alpha \in [-8, +8]$ from spiky shapes to smooth ones.

Word	e.v.	Word	e.v.	Word	e.v.
rape	-5.60	jeer	-2.20	smile	4.76
killer	-5.55	snub	-2.18	fun	4.91
funeral	-5.47	meal	2.54	laughter	4.95
slavery	-5.41	bunny	2.55	joy	5.19

Table 4.1: Samples from the Word domain with their emotional valence (e.v. $\in [-5.60, +5.19]$): negative emotion vs. positive emotion (Medler et al., 2005).

probability distribution over the shapes is uniform.

Word Domain

The Word domain consists of 321 words[3], sampled from the Wisconsin Perceptual Attribute Ratings Database (Medler et al., 2005), which includes words rated by 350 undergraduates based on their emotional valence. Two rules were defined for labeling data, analogously to Zhu et al. (2009) and to what is done above:

- Word Simple (WS): words were sorted by their length and the 161 longest ones were labeled with $Y = +1$, while the others were labeled with $Y = -1$.
- Word Difficult (WD): words were sorted by their emotional valence and the 161 most positive ones were labeled with $Y = +1$, while the others were labeled with $Y = -1$.

The probability distribution over the words is uniform. In Table 4.1, the samples from the Word domain are shown with their emotional valence.

Human Rademacher Complexity Experimental Design

In order to compute HRC, the same procedure of Zhu et al. (2009) has been adopted. RC does not depend on the problem ($\mathbb{P}\{Y|X\}$) but only on the domain ($\mathbb{P}\{X\}$), consequently we measured HRC for the two above-mentioned domains: Shape and Word.

[3]Having to deal with Italian students only, words have been translated into Italian.

Thus, for each domain and for a fixed value of n, Eq. (4.37) was computed. In particular, for each student $\mathscr{H}^i$ with $i \in \{1, \dots, m\}$, we measured $\widehat{R}_{\text{EMP}}(\mathscr{H}^i_{\mathcal{L}^{\sigma^i}_{i,n}}, \mathcal{L}^{\sigma^i}_{i,n})$. Consequently, we needed to build $\mathcal{L}^{\sigma^i}_{i,n}$ with $i \in \{1, \dots, m\}$ for the desired domain as follows:

1. sample randomly n samples from the desired domain;
2. discard the labels;
3. assign random labels to the samples.

After creating $\mathcal{L}^{\sigma^i}_{i,n}$, $\widehat{R}_{\text{EMP}}(\mathscr{H}^i_{\mathcal{L}^{\sigma^i}_{i,n}}, \mathcal{L}^{\sigma^i}_{i,n})$ was measured as follows:

1. each student $\mathscr{H}^i$ was asked to learn the rule by exploiting the dataset $\mathcal{L}^{\sigma^i}_{i,n}$;
2. each student was asked to label the same samples $\mathcal{L}^{\sigma^i}_{i,n}$ where the labels had been removed.

We designed the questionnaire for measuring $\widehat{R}_{\text{EMP}}(\mathscr{H}^i_{\mathcal{L}^{\sigma^i}_{i,n}}, \mathcal{L}^{\sigma^i}_{i,n})$, for each n, student, and domain. The questionnaire consists of 4 steps (each student was asked to complete the step within the time given in parentheses):

I students were asked to learn the underlying rule from $\mathcal{L}^{\sigma^i}_{i,n}$, with a short time limit (3 minutes);

II students were asked to perform a filler task consisting of some two-digit addition/subtraction questions, to reduce risks of memorization: this forced the students to learn a rule rather than memorizing the examples. In other words, it forced them to think about an irrelevant concept and reset the short-term memory (Schacter et al., 2010; Zhu et al., 2009) (1 minute);

III students were asked to classify the same samples $\mathcal{L}^{\sigma^i}_{i,n}$ where the labels had been removed, and the order in which the samples were presented was different from Step I. Students were not aware of this fact, and they were encouraged to guess if necessary (without time limit);

IV students were asked to describe the rule they identified, and to estimate the confidence of their decision (1 minute);

Then, for each student $\mathscr{H}^i$ with $i \in \{1, \dots, m\}$, a unique questionnaire was given, the answers were collected, $\widehat{R}_{\text{EMP}}(\mathscr{H}^i_{\mathcal{L}^{\sigma^i}_{i,n}}, \mathcal{L}^{\sigma^i}_{i,n})$ was computed, and finally HRC ($\widehat{C}^{\mu}_n$) was derived by averaging the results over the m students with the same n and domain according to Eq. (4.37).

This procedure is also detailed in Zhu et al. (2009) but we varied $n \in \{3, 5, 7, 10, 15, 20, 25\}$ instead of $n \in \{5, 10, 20, 40\}$ because, as reported in the experimental result, this range allows to better interpret the behaviour of the measured quantities. Later, we compare our results with Zhu et al. (2009). We report in $\widehat{C}^{\text{SHAPE}}_n$ and $\widehat{C}^{\text{WORD}}_n$ the result for the two domains (Shape and Word) and different values of n.

Human Algorithmic Stability Experimental Design

A new experimental protocol was designed, differently from what was done for HRC, in order to measure the average HAS. Since AS changes not only based on the domain but also based on the problem, as it depends on $\mathbb{P}\{Y|X\}$, we measured it over four different problems: SS, SD, WS, and WD.

Thus, for each problem and for a fixed value of n, the value of Eq. (4.40) was computed. In particular, for each student $\mathscr{H}^i$ with $i \in \{1, \ldots, m\}$ we measured $|\ell(\mathscr{H}^i_{\mathcal{L}_{i,n}}, Z) - \ell(\mathscr{H}^i_{\mathcal{L}^{\setminus i}_{i,n}}, Z)|$ with $Z \in \mathcal{T}_{i,p}$. Since we use the hard loss function:

$$\left|\ell(\mathscr{H}^i_{\mathcal{L}_{i,n}}, Z) - \ell(\mathscr{H}^i_{\mathcal{L}^{\setminus i}_{i,n}}, Z)\right| = \left|\mathscr{H}^i_{\mathcal{L}_{i,n}}(X) - \mathscr{H}^i_{\mathcal{L}^{\setminus i}_{i,n}}(X)\right|, \tag{4.45}$$

which means that in order to compute $|\ell(\mathscr{H}^i_{\mathcal{L}_{i,n}}, Z) - \ell(\mathscr{H}^i_{\mathcal{L}^{\setminus i}_{i,n}}, Z)|$, we simply needed to check if the sample Z had been labeled by the students in the same way after learning $\mathcal{L}^{\setminus i}_{i,n}$ or $\mathcal{L}_{i,n}$. In order to measure this quantity, the following steps were performed:

1. each student $\mathscr{H}^i$ was asked to learn the rule by exploiting the dataset $\mathcal{L}^{\setminus i}_{i,n}$, which was the dataset $\mathcal{L}_{i,n}$ where a random sample $i \in \{1, \ldots, n\}$ had been removed;
2. each student was asked to label the samples $Z \in \mathcal{T}_{i,p}$ where the labels had been removed.
3. each student $\mathscr{H}^i$ was asked to learn the rule by exploiting the whole dataset $\mathcal{L}_{i,n}$;
4. each student was asked to label the samples $Z \in \mathcal{T}_{i,p}$ again where the labels had been removed.

Once all these data were collected, one could measure $|\ell(\mathscr{H}^i_{\mathcal{L}_{i,n}}, Z) - \ell(\mathscr{H}^i_{\mathcal{L}^{\setminus i}_{i,n}}, Z)|$. As in the previous section we built the questionnaire, consisting of 8 steps:

I students were asked to learn the underlying rule from $\mathcal{L}^{\setminus i}_{i,n}$ with a short time limit (3 minutes);

II students were asked to perform a filler task, as in the HRC experiment (1 minute);

III students were asked to classify the samples in $\mathcal{T}_{i,p}$ where the labels had been removed (without time limit);

IV students were asked to describe the rule they identified, and to estimate the confidence of their decision (1 minute).

V students were asked to learn the underlying rule from $\mathcal{L}_{i,n}$ (these samples were presented in a random order respect to Step I) with a short time limit (3 minutes). They were not aware that $n - 1$ training instances were the same as in Step I;

VI students were asked to perform a filler task (1 minute);

VII students were asked to classify the samples in $\mathcal{T}_{i,p}$ (these samples were presented in a random order respect to Step III) where the labels had been removed (without time limit);

VIII students were asked to describe the rule they identified, and to estimate the confidence of their decision (1 minute).

Note that, first $\mathcal{L}_{i,n}^{\setminus i}$ and then $\mathcal{L}_{i,n}$ were provided in order to avoid the risk of memorization.

Then, for each student $\mathscr{H}^i$ with $i \in \{1, \ldots, m\}$, a unique questionnaire was given, the answers were collected, $|\ell(\mathscr{H}^i_{\mathcal{L}_{i,n}}, Z) - \ell(\mathscr{H}^i_{\mathcal{L}^{\setminus i}_{i,n}}, Z)|$ was computed, and finally HAS ($\widehat{H}^{\mu}_n$) was derived by averaging the results over the m students with the same n and problem according to Eq. (4.40).

The results for two domains and two levels of difficulty were collected, by varying $n \in \{3, 5, 7, 10, 15, 20, 25\}$, in $\widehat{H}^{\text{SS}}_n$, $\widehat{H}^{\text{SD}}_n$, $\widehat{H}^{\text{WS}}_n$, and $\widehat{H}^{\text{WD}}_n$ respectively.

Human Error Experimental Design

The experiment designed for measuring HAS (see Section 4.4.1) allowed us to additionally measure Human Error.

In particular, for each student $\mathscr{H}^i$ with $i \in \{1, \ldots, m\}$, n, and problem, we measured $\ell(\mathscr{H}^i_{\mathcal{L}_{i,n}}, Z)$ with $Z \in \mathcal{T}_{i,p}$. In other words:

1. each student $\mathscr{H}^i$ was asked to learn the rule by exploiting the dataset $\mathcal{L}_{i,n}$;
2. each student was asked to label the samples $Z \in \mathcal{T}_{i,p}$ where the labels had been removed.

The questionnaire was designed in the same way as the one of Section 4.4.1 and in particular, Steps V, VI, VII, and VIII: this was advantageous for us since all the quantities were already available and there was no need to build an additional experiment. Moreover, from Steps I, II, III, and IV the data for measuring Human Error on $n - 1$ was also available.

Then, for each student $\mathscr{H}^i$ with $i \in \{1, \ldots, m\}$, a unique questionnaire was given, the answers were collected, $\ell(\mathscr{H}^i_{\mathcal{L}_{i,n}}, Z)$ with $Z \in \mathcal{T}_{i,p}$ was computed, and finally Human Error ($\widehat{R}^{\mu}_n$) was derived by averaging the results over the m students with the same n and problem according to Eq. (4.40) (since Human Error depends on $\mathbb{P}\{Y|X\}$).

Based on the designed HAS experiment, it was possible to collect data for $n \in \{3, 5, 7, 10, 15, 20, 25\}$ and $n - 1$. The results for two domains and two levels of difficulty were collected, by varying n, in $\widehat{R}^{\text{SS}}_n$, $\widehat{R}^{\text{SD}}_n$, $\widehat{R}^{\text{WS}}_n$, and $\widehat{R}^{\text{WD}}_n$ respectively.

Aggregated Questionnaire Structure

The aggregated questionnaire consists of 3 sub-questionnaires, 2 for HAS and 1 for HRC:

- A questionnaire of Section 4.4.1 for SS or SD
- A questionnaire of Section 4.4.1 for WD or WS
- A questionnaire of Section 4.4.1 for Word or Shape

We avoided including two simple or two difficult problems in one questionnaire. Note that with two students, there was a complete set of problems: HAS for SS, SD, WS, and WD, as well as HRC for Word and Shape. Human Error was computed for SS, SD, WS, and WD thanks to the HAS questionnaires.

Before giving the questionnaires to the students, we devoted approximately 10 minutes to explain the experiment procedure and our study goals to the students. In the early trials of this study, we noticed that this kind of explanation increases the motivation of the students to try their best in answering the questions.

The size $n \in \{3, 5, 7, 10, 15, 20, 25\}$ of the sets were randomly chosen. In EX^1, since estimation accuracy of Human Error, AS, and RC depends on m (see Section 4.3), we set $p = 1$. Each experiment took about 30 minutes in total.

A first trial, conducted on 70 volunteers, showed that students were able to mentally link Steps I and III with Steps V and VII due to the problems described in Section 4.3.3. This result confirmed our hypothesis that humans easily memorize and they cannot easily forget a previously memorized rule. The detailed results are reported in the experimental result sections. In the next section, we explain a modified experiment which allowed us to successfully measure HAS.

4.4.2 Experimental Design: EX^2

As described above, we had to modify the experiment in order to measure HAS; but for HRC, we kept the design of EX^1.

Human Algorithmic Stability Experiment Design

Instead of measuring $\widehat{H}^{\mu}_{n}$, we needed to measure $\widehat{\widehat{H}}^{\mu}_{n}$ for each value of n and problem (see Eq. (4.43)). In particular, for a pair of students $\mathscr{H}^i$ and $\mathscr{H}^{i+m}$ with $i \in \{1, \ldots, m\}$, we measured $|\ell(\mathscr{H}^i_{\mathcal{L}_{i,n}}, Z) - \ell(\mathscr{H}^{i+m}_{\mathcal{L}^{\setminus i}_{i,n}}, Z)|$ with $Z \in \mathcal{T}_{i,p}$. In order to compute $|\ell(\mathscr{H}^i_{\mathcal{L}_{i,n}}, Z) - \ell(\mathscr{H}^{i+m}_{\mathcal{L}^{\setminus i}_{i,n}}, Z)|$, we simply needed to check if the same label had been assigned to Z by both students when one learned the rule through $\mathcal{L}^{\setminus i}_{i,n}$, and the other one through $\mathcal{L}_{i,n}$. In order to measure $|\ell(\mathscr{H}^i_{\mathcal{L}_{i,n}}, Z) - \ell(\mathscr{H}^{i+m}_{\mathcal{L}^{\setminus i}_{i,n}}, Z)|$, the following steps were performed:

1. each student $\mathscr{H}^i$ was asked to learn the rule by exploiting the dataset $\mathcal{L}^{\setminus i}_{i,n}$, while student $\mathscr{H}^{i+m}$ was asked to learn the rule by exploiting the whole dataset $\mathcal{L}_{i,n}$;
2. student $\mathscr{H}^i$ and student $\mathscr{H}^{i+m}$ were asked to label the samples $Z \in \mathcal{T}_{i,p}$ where the labels had been removed.

The aggregated questionnaire was the same as Section 4.4.1 but split over a pair of students: phases I, II, III, and IV were assigned to one of the students and phases V, VI, VII, and VIII to the other. Thanks to this procedure, all quantities, necessary to compute HAS, could be derived: for every pair of students $\mathscr{H}^i$ and $\mathscr{H}^{i+m}$ with $i \in \{1, \ldots, m\}$, a unique questionnaire was given, the answers were collected, $|\ell(\mathscr{H}^i_{\mathcal{L}_{i,n}}, Z) - \ell(\mathscr{H}^{i+m}_{\mathcal{L}^{\setminus i}_{i,n}}, Z)|$ was computed, and finally HAS $\widehat{\widehat{H}}^{\mu}_{n}$ was derived by averaging the results over m pair of students with same n and problem according to Eq. (4.43).

As in Section 4.4.1, The results for two domains and two levels of difficulty were

collected, by varying n, in $\widehat{\overline{H}}_n^{\mathrm{SS}}$, $\widehat{\overline{H}}_n^{\mathrm{SD}}$, $\widehat{\overline{H}}_n^{\mathrm{WS}}$, and $\widehat{\overline{H}}_n^{\mathrm{WD}}$ respectively.

Human Error

As for EX^1, the AS questionnaire allows to measure Human Error as well, although, the data had to be collected from a pair of students instead of a single student. Note that the available data is exactly the same as Section 4.4.1.

Aggregated Questionnaire Structure

The aggregated questionnaire consists of 5 sub-questionnaires, 4 for HAS and 1 for HRC but split over two students:

- four questionnaires of Section 4.4.2 for SS, SD, WS, and WD: phases I, II, III, and IV were given to one of the students, while phases V, VI, VII, and VIII were given to the other;
- one questionnaire of Section 4.4.1 for Word or Shape: Word was given to one of the students, while Shape was given to the other.

Analogously to EX^1, for EX^2, n was randomly chosen in $\{3, 5, 7, 10, 15, 20, 25\}$, we set $p = 1$ and approximately 10 minutes were devoted before the experiment, to explain the experiment procedure and our study goals to the students.

We obtained interesting results by providing these questionnaires to 307 volunteers. We report here the complete set of results including the extension of this experiment to more students and an additional domain. In the next section we explain how the experiment was extended. In particular, we address a problem raised in Vahdat et al. (2015d) showing that it is possible to reduce the oscillation of the results with a better estimate of the quantity of interest (Human Error and HAS).

4.4.3 Experimental Design: EX^3

In EX^3, we decided not to measure HRC for two reasons: firstly, as can be seen later in the experimental results, we managed to replicate and obtain satisfactory results with EX^1 and EX^2 which are consistent with the one reported in Zhu et al. (2009). Consequently, our fist goal of being able to reproduce and verify the results of Zhu et al. (2009) was already achieved. Secondly, by excluding RC, we had the possibility to insert another domain into the experiment without increasing the length of the experiment, and compromising the level of attention of the students, which could produce artefacts in the final results (Schacter et al., 2010).

In EX^3, we made two major changes respect to EX^2 (apart from the above-mentioned changes). Firstly, instead of setting $p = 1$, we decided to set $p = 15$ in order to improve the accuracy and statistical robustness of the estimator (see Section 4.3). In other words,

we computed a more accurate estimate of the difference between the rules found by the students in each pair. Secondly, we added a new domain: the Math domain.

In the next sections, these modifications are described in detail.

Math Domain

The Math domain consists of 321 numbers $x \in \{0, 1, \ldots, 320\}$. A label was assigned to each number, and two problems were defined in accordance with ad-hoc rules to depict tasks of increasing complexity:

- Math Simple (MS), where $Y = +1$ if $x \leq 159$ and $Y = -1$ otherwise.
- Math Difficult (MD), where $Y = +1$ if $108 \leq x \leq 214$ and $Y = -1$ otherwise.

The probability distribution over the numbers is uniform.

Aggregated Questionnaire Structure

The aggregated questionnaire consists of 6 HAS sub-questionnaires of Section 4.4.2, one for each problem among SS, SD, WS, WD, MS, and MD. As in EX^2, phases I, II, III, and IV were given to a student of a pair, while phases V, VI, VII, and VIII were given to the other. The order in which the problems were presented was randomized since, in case of crowded classes, it was impossible for students to copy the labels from each other.

Analogously to EX^2, in EX^3, n was randomly chosen in $\{3, 5, 7, 10, 15, 20, 25\}$ but, differently from EX^2, we set $p = 15$. As was the case with EX^1 and EX^2, we devoted approximately 10 minutes before the experiment, to explain the experiment procedure and our study goals to the students. The experiment itself took about 30 minutes, the same as EX^1 and EX^2.

4.5 Results

In the following sections, the results of the three experiments (EX^1, EX^2, and EX^3) are presented[4]. In particular, for EX^1, the following quantities are reported:

- Human Error, $\widehat{R}_n^\mu$
- Human Rademacher Complexity, $\widehat{C}_n^\mu$ (where we set $p = 1$)
- Human Algorithmic Stability, $\widehat{H}_n^\mu$ ($p = 1$)

In the first experiment $\mu \in \{\text{SS}, \text{SD}, \text{WS}, \text{WD}\}$. Note that, since $\widehat{C}_n^\mu$ depends just on the domain, $\widehat{C}_n^{\text{SHAPE}} = \widehat{C}_n^{\text{SS}} = \widehat{C}_n^{\text{SD}}$ and $\widehat{C}_n^{\text{WORD}} = \widehat{C}_n^{\text{WS}} = \widehat{C}_n^{\text{WD}}$ (see Section 4.4.1).

For EX^2 instead, we report the following quantities:

- Human Error, $\widehat{R}_n^\mu$
- Human Rademacher Complexity, $\widehat{C}_n^\mu$ ($p = 1$)
- Human Algorithmic Stability, $\widehat{\widetilde{H}}_n^\mu$ ($p = 1$)

[4]Some examples of the filled and anonimized questionnaires can be retrieved from `www.la.smartlab.ws/filled-questionnaires.zip`

Note that in the second experiment, $\mu \in \{\text{SS, SD, WS, WD}\}$ as in the first one, but we measure $\widehat{\widehat{H}}_n^{\mu}$ instead of $\widehat{H}_n^{\mu}$ for the reason that $\widehat{H}_n^{\mu}$ cannot be effectively measured (see Section 4.4.2).

Finally, for EX^3, we report the following quantities:

- Human Error, $\widehat{R}_n^{\mu}$ (where we set $p = 15$)
- Human Algorithmic Stability, $\widehat{\widehat{H}}_n^{\mu}$ ($p = 15$)

In the third experiment, $\mu \in \{\text{SS, SD, WS, WD, MS, MD}\}$ as we added another domain (Math) and two problems (simple and difficult). Note that in this case HRC was not measured.

The results are measured for $n \in \{3, 5, 7, 10, 15, 20, 25\}$ in all the experiments. Thanks to the designed experiments, we can also measure $\widehat{R}_n^{\mu}$ for $n - 1$ (see Section 4.4.1).

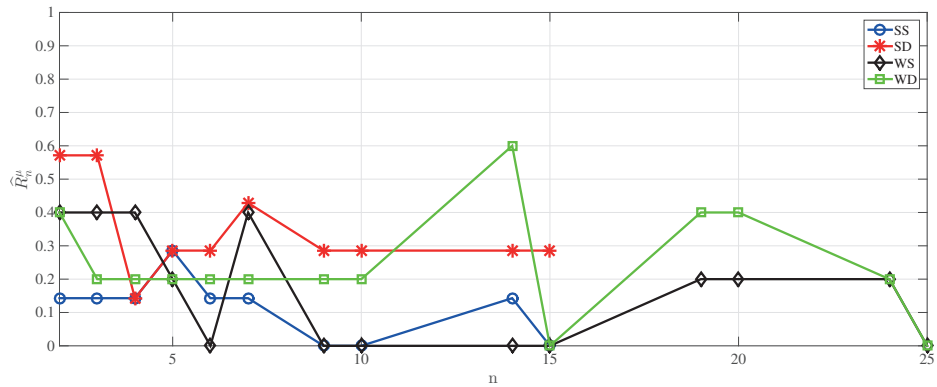
4.5.1 Results for EX^1

The first experiment was carried out as a pilot test on a small population, a group of 70 students. Therefore, we lack data for some problems and for some values of n. For example, in Figure 4.4c, $\widehat{C}_{20}^{\text{SHAPE}}$ and $\widehat{C}_{25}^{\text{SHAPE}}$ are missing.

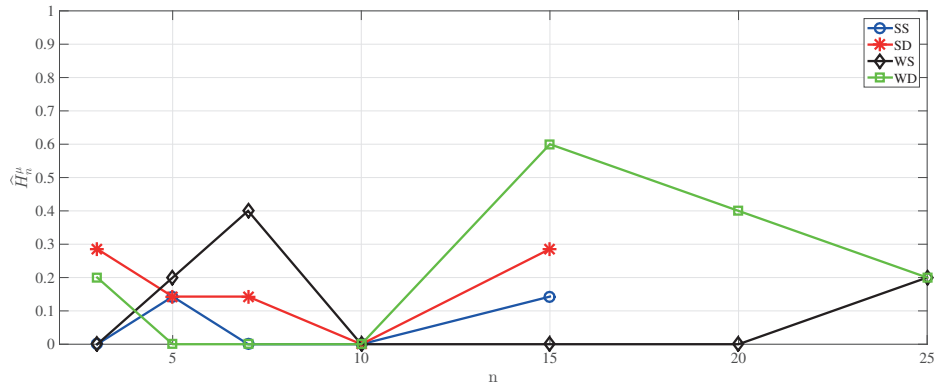
Figure 4.4a shows the trend of $\widehat{R}_n^{\mu}$, for different values of n and different problems μ: as expected from ML Theory, $\widehat{R}_n^{\mu}$ is generally smaller for simple tasks than for difficult ones and this is confirmed in HL as well. However, analogies end here. While the error of ML models usually decreases with n, results on HL are characterized by oscillations, even for small variations of n. This is due to the fact that a small sample is considered.

Figure 4.4c shows the trend for $\widehat{C}_n^{\mu}$. Contrarily to HAS, HRC is not able to discriminate between the task complexities, since labels are neglected when computing $\widehat{C}_n^{\mu}$. HRC decreases with n (as in ML), and this trend is substantially uncorrelated with the errors for the considered domains. Note that, due to the fact that a small sample of students is considered, HRC is characterized by oscillations as well.

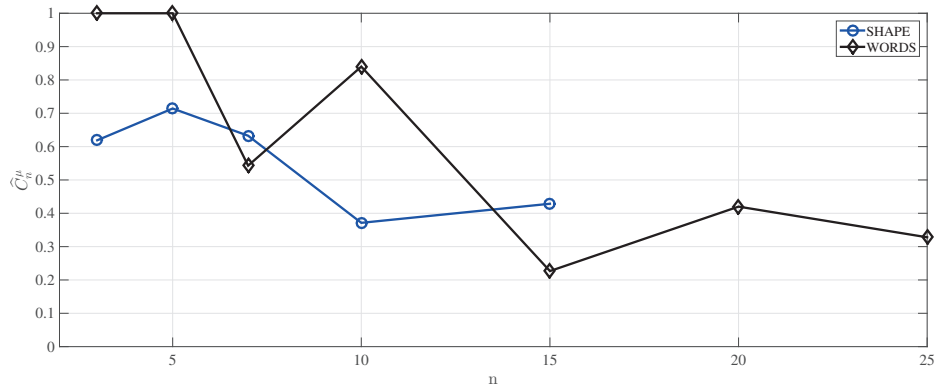
Finally, Figure 4.4b presents the obtained results when computing $\widehat{H}_n^{\mu}$ (with $p = 1$) as n varies. Despite being designed in the ML framework, it is worth highlighting how HAS is able to grasp the nature and peculiarities of HL. As a matter of fact, simple tasks are characterized by smaller values of $\widehat{H}_n^{\mu}$, and HAS for the Shape domain is generally smaller than for the Word domain. Both results are in accordance with the trend of the error, registered in HL, and the nature of the analyzed phenomenon: in this sense, HAS offers interesting insights on HL, because it raises questions about the ability of humans to learn in and across different domains. By analyzing the results and in particular the answers to the questions IV and VIII of the HAS questionnaire (see Section 4.4.1), in a majority of cases, the students did not change the rule when learned during the step VI respect to the step I (see Section 4.4.1). Therefore, a lot of zeros in Figure 4.4b can be seen. After observing this phenomenon, we decided to ask the students about the reason



(a) Human Error for different values of n and different problems μ (Shape Simple (SS), Shape Difficult (SD), Word Simple (WS), and Word Difficult (WD)).



(b) Human Algorithmic Stability (HAS) for different values of n and different problems μ (Shape Simple (SS), Shape Difficult (SD), Word Simple (WS), and Word Difficult (WD)).



(c) Human Rademacher Complexity (HRC) for different values of n and different domains (Shape and Word).

Figure 4.4: Results of EX[1].

behind their choice. Their answers showed that they recognized that the problems of steps I and V are the same thus, instead of learning a new rule, they just memorized their previous answer to the problem. This is a quite interesting result since it confirms our initial idea about the learning behaviour of humans. Again, the oscillations in HAS results are due to the fact that a small sample is considered.

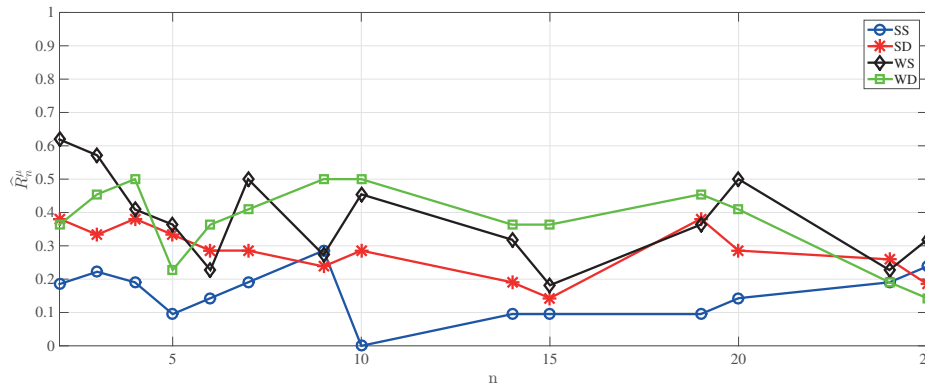
4.5.2 Results for EX^2

In EX^2 the experiment was carried out over 306 students in order to address the issues of EX^1 related to the lack of data. Moreover, we measure HAS in a different way: we use $\widehat{\widehat{H}}_n^{\mu}$ (with $p = 1$) instead of $\widehat{H}_n^{\mu}$ in order to fix the memorization problem encountered during EX^1.

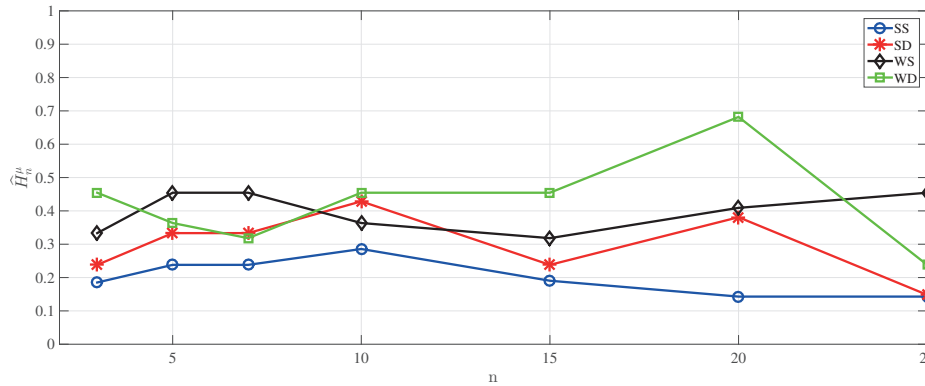
Figure 4.5a represents Human Error for different problems with varied number of samples available for learning: the results are quite similar to the ones of EX^1 where the oscillations are still a problem. At this point we had the impression that only a subset of students were willing to perform at their best when completing the questionnaire (due to the lack of motivation and concentration) and this could compromise the result. We explored this hypothesis by analyzing the filler tasks, in order to verify the students' level of attention. We discarded the students with low level of attention (the ones with many errors or incomplete filler tasks) but the results remained almost unchanged. Consequently, in EX^3 we set $p = 15$ in order to get a more stable estimate. Note that, in any case, the curve is generally higher for difficult problems respect to the simple ones and for the Word domain with respect to the Shape domain. Unfortunately, Human Error still does not decrease with n.

Figure 4.5c represents HRC by varying n in different domains. In this case, we significantly reduced the oscillations respect to EX^1 (Figure 4.4b). It is worth highlighting that the curve is very similar to the one obtained by Zhu et al. (2009) for the measure of HRC. Note that, RC is able to grasp the difference between the domains (Shape domain is a simpler task compared to the Word one). Moreover, it has the same drawbacks depicted in the results of EX^1.

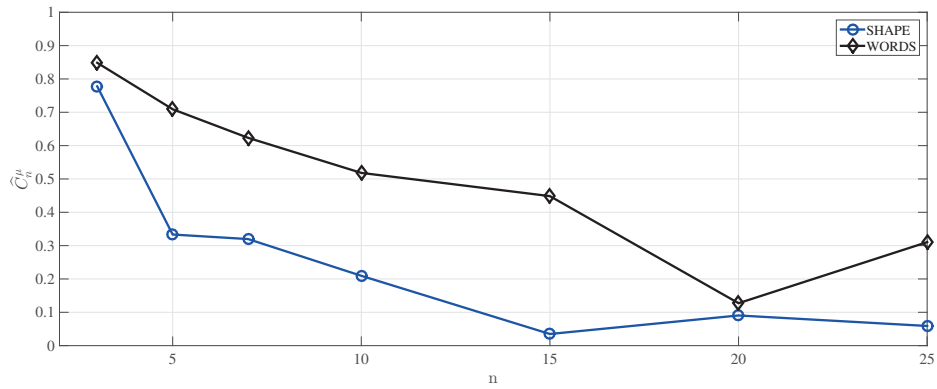
Finally, Figure 4.5b shows the trend of HAS by varying n. Results are, again, similar to the one of EX^1 (the curve is generally higher for difficult problems respect to the simpler ones and for the Word domain with respect to the Shape domain, analogously to what happens with Human Error) although, the oscillations are mostly reduced compared to EX^1. Also in this case, we tried to discard the students with low level of attention but, again, the results did not noticeably change and for this reason we did not report it. We address this issue in EX^3 by adding the parameter $p = 15$ in measuring both Human Error and HAS.



(a) Human Error for different values of n and different problems μ (Shape Simple (SS), Shape Difficult (SD), Word Simple (WS), and Word Difficult (WD)).

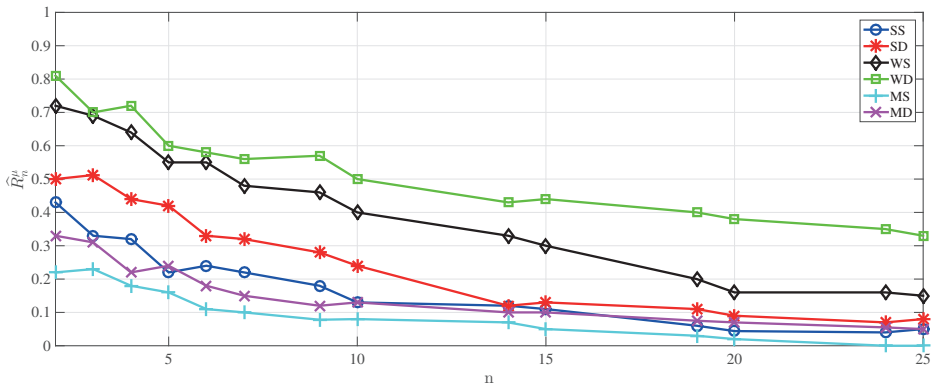


(b) Human Algorithmic Stability (HAS) for different values of n and different problems μ (Shape Simple (SS), Shape Difficult (SD), Word Simple (WS), and Word Difficult (WD)).

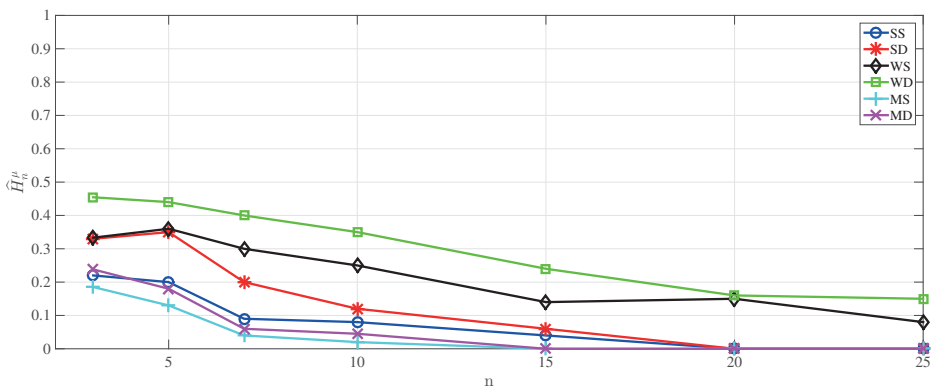


(c) Human Rademacher Complexity (HRC) for different values of n and different domains (Shape and Word).

Figure 4.5: Results of EX^2.



(a) Human Error for different values of n and different problems μ (Shape Simple (SS), Shape Difficult (SD), Word Simple (WS), Word Difficult (WD), Math Simple (MS), and Math Difficult (MD)).



(b) Human Algorithmic Stability (HAS) for different values of n and different problems μ (Shape Simple (SS), Shape Difficult (SD), Word Simple (WS), Word Difficult (WD), Math Simple (MS), and Math Difficult (MD)).

Figure 4.6: Results of EX^3.

4.5.3 Results for EX^3

In EX^3 the experiment was carried out with 229 participants in order to address the issues of EX^2 related to stability of the estimators. Moreover, we did not measure HRC, and we use $\widehat{R}_n^{\mu}$ and $\widehat{\widehat{H}}_n^{\mu}$ with $p = 15$ in order to obtain a more stable estimate.

In both Figures 4.6a and 4.6b, it can be seen that the oscillations are substantially reduced. Both Human Error and HAS decreases with n (as expected from ML Theory). Results show that the Math domain is the simplest while the Word domain is the most difficult for the students. This can be due to the fact that all students are from engineering majors, so they are more used to the problems related to Mathematics. Moreover, Human Error and HAS are larger for difficult problems compared to easier ones. The ranking of problems (from difficult to simple) results to be: WD, WS, SD, SS, MD, and MS. While there is a clear distinction between the Word problems (difficult or simple) and the other domains, the distinction between Shape and Math problems is not as clear. This effect is more visible as n increases. This can be due to the fact that the Math and the Shape domains leave less space for imagination to find meaningful rules. Instead, the Word domain might open a whole world of possible interpretations and rules which can be discovered and selected. For instance, we observed that some students related the words as part of a common story or image for discovering the rule. This observation is also supported by the HRC experiment which shows how the capacity of HL in case of the Word domain, is larger respect the one of Shape domain.

As a matter of fact, our study results are derived from and limited to the students of engineering majors. These results can be dependent on the background knowledge and the major of the students. For example, students of humanities or literature majors might have better results for the Word domain as opposed to the Math or Shape domain.

4.6 Discussion

Recently ML tools have been exploited in COGS to understand HL. In addition to providing algorithms for extracting information from the data, ML provides tools to analyze the learning capacity of algorithms.

AS is an effective ML tool for understanding the learning ability of algorithms. In this study, we propose to exploit this tool for obtaining insight into HL and we show that HAS is more informative towards HL than HRC which was previously studied in Zhu et al. (2009). We conducted our experiments with 606 students of various engineering majors from the University of Genoa. Our results showed that: both HRC and HAS are able to detect the difficulty level of different domains for a group of students. However, contrary to HAS, HRC requires some additional assumptions to be measured that are seldom or never satisfied in HL (Schacter et al., 2010). Additionally, HAS extends these results by detecting the difficulty level of different problems in the same domain. In particular, the

difficulty scale of the problems (from difficult to simple) for the students is: WD, WS, SD, SS, MD, and MS.

Our results suggest that ML can offer new opportunities in the study of HL in the fields of LA and COGS. In particular, CL can be enhanced and studied further by application of ML methods and integrated into instructional design in classrooms or TEL systems to improve education. In this context, educational reform has been mentioned as the most important and relevant application of CL (Goldstone and Kersten, 2003) such that approaches in manipulating category labels, presentation order, learning strategies, and category variability can optimize CL. Consequently, ML opens the doors toward improving CL in educational settings.

Recent works in CL (Chater and Vitányi, 2003; Feldman, 2000; Griffiths et al., 2008; Zhu et al., 2009) highlight how cross fertilization between ML and HL can be extended to better understand how people tackle new problems and extract knowledge from observations. CL can be seen as a method of learning where IBL cycle is implemented. As explained in Section 2.4.1, the IBL cycle in the context of CL starts with the phases of observing a set of items by the learner and generating a hypothesis / a rule to explain the distinction among concepts / categories. Then the hypothesis can be tested when the learner is exposed to a new stimulus, and the IBL cycle can restart by generating of a new hypothesis. In this study, we showed that ML methods provide more insight into the hypothesis generation of humans, and as such they can give more insight on the IBL cycle of learners. Such insight can help teachers improve their instruction and adapt it to the expected capacity of concept attainments for various domains, with various levels of difficulty, and various groups of learners. For instance, HAS can improve the ability of educators to detect when learners are passively memorizing the concepts rather than discovering new knowledge. Note that HAS, unlike HRC, is not only able to explain the difficulty level of the particular domain for learners, but also is able to detect the difficulty of a problem in a domain. Consequently, this ability can lead to better personalization and adaptation of education to the needs of students.

4.7 Summary

In this chapter we presented our first empirical study to investigate the IBL cycle in particular the phase of hypothesis generation in the context of CL. For this purpose, we adopted the concepts of LA and EDM where strong links between COGS and ML help us to understand the learning process of humans.

We showed that ML can be applied to humans as if they were algorithms that learn from data. We successfully replicated a recent study (Zhu et al., 2009) on application of Rademacher Complexity, as an ML method, to human learning. Additionally, we proposed another powerful ML tool called Algorithmic Stability to gain more insight

on hypothesis generation of humans. We showed the advantages of AS over RC in HL through three experiments over a total of 606 students. Our results are concluded as firstly, HAS can be measured without the assumptions required by HRC. Secondly, HAS depends not only on the knowledge domain, as HRC, but also on the complexity of the problem. Finally, HAS and HRC can be exploited for better understanding of the process of hypothesis generation in humans.

In the next chapter, we move from hypothesis generation and focus more on the second phase of IBL: investigation and discovery, in an educational context where again IBL cycle is implemented.

Chapter 5

IBL Phase Two: Application of Process Mining in Simulation-Based Learning

5.1 Introduction

LA and EDM aim at gaining insight into the behavior of learners by building models based on data collected from learning tools (Siemens, 2012). As a result, LA and EDM applied in educational settings can improve the understanding of various stakeholders (students, teachers, administrators, etc.) about the way people learn, in a data-driven way. Additionally, Technology Enhanced Learning (TEL) systems can be improved to be more personalized and adaptive through LA and EDM, and learning can be optimized (Chatti et al., 2012; Romero and Ventura, 2007; Siemens and Long, 2011).

Inquiry-based simulation environments are considered as promising TEL environments in higher education to increase the knowledge and skills of students in engineering education (D'Angelo et al., 2014; Gillet et al., 2013; Njoo and De Jong, 1993). In this context, the learners discover knowledge through investigation and performing experiments (for more information about Inquiry-Based Learning (IBL), see Section 2.3). The IBL cycle begins with presenting a problem, and the learner tries to study the task, perform experiments by using computer simulations, and address the problem (Pedaste et al., 2015), resulting in a conclusion or further experiments (see Section 2.4).

To enhance this learning experience, and to increase the awareness of teachers about the learning processes in laboratory sessions, application of LA and EDM is proposed (Vahdat et al., 2015c, 2014). In this regard, LA and EDM methods can help discover the hidden information and patterns from raw data collected from educational environments (Johnson et al., 2011; Siemens, 2012). In particular, the use of Process Mining (PM) can be effective to obtain knowledge from educational processes (Bannert et al., 2014; Pechenizkiy et al., 2009; Trcka and Pechenizkiy, 2009).

In this work[1], we focus on the 'second phase of IBL' (see Section 2.3.2) and apply PM methods in combination with the Cyclomatic Complexity (CM) metric to increase our understanding about the **investigation and discovery** behavior of students in the context of inquiry-based simulation environments. We propose to apply the CM metric, normally applied to study the complexity of computer programs, to study and quantify the learning processes of students. We apply fuzzy miner to obtain the process models of various students' clusters and use the CM metric to quantify and compare the process models.

For addressing these issues, we carried out our study over a simulation environment named Deeds (Digital Electronics Education and Design Suite) (Ponta et al., 1998; Vahdat et al., 2014) which is used for the university-level e-learning in digital electronics. Additionally, as a contribution to the data sharing community, we made our data set publicly available. We provide an overview of our data set built from client side data logs.

This study shows that many properties of learning behavior of students can be explained by PM and academic success of students can be explained based on CM. Additionally, our data set has been viewed thousands of times, which can be an indicator of growing need for data sets in the LA and EDM research fields.

This chapter is structured as follows. In Section 5.2, preliminaries of PM and CM are described. We explain fuzzy miner and the CM metric for measuring the complexity of process models. In Section 5.3, the simulator and the learning process of students are presented. In Sections 5.4, 5.5, and 5.6, we describe the experimental design, analytics approach, and the characteristics of our data set respectively. In Sections 5.7 and Section 5.8, we report the results and discuss them. Finally, a summary is provided in Section 5.9.

5.2 Process Mining and Cyclomatic Complexity Metric

In this section, we provide a detailed description of the applied methods in this work.

5.2.1 Process Mining

Learning can be considered as an ongoing process which happens over time (Fautley and Savage, 2008). The data gathered from learning environments include concrete actions performed by students associated with the duration and outcome of students' activities (Bienkowski et al., 2012; Del Blanco et al., 2013). Therefore, to analyze these data, it can be valuable to integrate methods that focus on learning as a temporal process rather

[1]This chapter is written based on Vahdat et al. (2015c, 2014).

than only a set of features with an outcome. Studying the learning behavior based on the temporal processes can lead to the development of formative assessments, and to help the teachers get alerted as early as possible about the students' difficulties.

Among process related methods, Process Mining (PM) is emerging in TEL since it can enhance our understanding by discovering models of student actions and learning processes (Trcka et al., 2010).

There are various challenges for PM in education: Reaching an expectation model of behavior is very challenging due to the wide variety of behavior. Also, the behavior of learners is usually very complex and unordered. Therefore, the process discovery needs particular attention and PM methods need to be selected carefully. Another challenge is that the event log contains just a fraction of all possible behaviors of students. The students do not limit their learning to the time we log their data. For example, they might learn by discussing with their peers at breaks or studying at home when we cannot record such information. This limitation makes it impossible to gather all the learning behavior and represent it in a process model (Van Der Aalst, 2011). We try to minimize the information loss by imposing restrictions over the work time and submission of the students. Additionally, we explain how we try to overcome the complexity of the models by choosing appropriate methods for discovering the models of students' behavior.

5.2.2 Event Data

Here, we explain some specific PM terms as they are used throughout the chapter. A process obtained from an event log that is recorded sequentially consists of cases, events, activities, timestamps, and various attributes. In Figure 5.1, an extract of a log of student behavior indicating the case, activity, and timestamps is shown. The 'case' can be allocated to the session, student, or exercise. These concepts are explained as follows (Van der Aalst et al., 2004):

- 'Case': a process consists of cases. A case consist of a set of events. Allocating the case Id in order to group events plays an important role in how to discover a process model. As an example in our educational context, allocating a case per student allows us to obtain the common processes among the students while by allocating a case per exercise, we can obtain the processes occurred during various exercises. Similarly, a case can be assigned to the sessions of a course to compare the processes among various sessions. A session refers to a laboratory session of the course in which students work on various exercises on the same topic.

- 'Event': a log consists of a set of events (rows in Figure 5.1) and each particular event is associated with a case. For each event, there is a timestamp, and other attributes if applicable.

Session	Student	Exercise	Activity	Start-time	End-time
1	10	Es_1_3	TextEditor_Es_1_3	2.10.2014 12:30:33	2.10.2014 12:31:6
1	10	Es_1_3	Deeds_Es_1_3	2.10.2014 12:31:7	2.10.2014 12:31:14
1	10	Es_1_3	Diagram	2.10.2014 12:31:15	2.10.2014 12:31:17
1	10	Es_1_3	Deeds_Es_1_3	2.10.2014 12:31:18	2.10.2014 12:31:20
1	10	Es_1_3	Diagram	2.10.2014 12:31:21	2.10.2014 12:32:18
1	10	Es_1_3	TextEditor_Es_1_3	2.10.2014 12:32:19	2.10.2014 12:32:23

Figure 5.1: An extract of an event log in which 'Case' can be allocated to session, student, or exercise.

- 'Attribute': the main attributes of an event are activity, timestamps, resource, and cost. The two most important attributes are described below.
  - 'Activity': 'event' classes found in the log are shown as 'activity' nodes in the process map.
  - 'Timestamps': each activity is associated with timestamps. For instance, start time and end time of an activity can be taken into account.

5.2.3 Fuzzy Miner: A Process Mining Algorithm

We try to summarize the fuzzy miner concept based on Günther and Van Der Aalst (2007) for explaining how our process models are obtained. Fuzzy miner is built over the metaphor of a road map. Similarly to a road map that the most important roads are visualized and showing all the unnecessary details is avoided, fuzzy miner provides a high-level view on the process. In this way, fuzzy miner tries to consider the most important activities and relations between activities and abstract unnecessary details.

This miner is effective when the event log is unstructured, and the process model is not known beforehand. This kind of modeling is required when dealing with so-called *spaghetti-like* processes since understanding them is very challenging (van der Aalst and Günth, 2007). For instance in educational contexts, the fuzzy miner algorithm is applied as an effective tool to obtain the process models from the logs of students (Bannert et al., 2014) and improved the abstraction of complex student processes.

This process is done in three steps (for the formal description of fuzzy miner and its properties refer to Günther and Van Der Aalst (2007)): to discover the process model from an event log, first a partial order of the activities (nodes) is created based on how they follow one another, then the transitions (edges) that seem to be less important are removed according to some metrics, and finally the less important nodes are aggregated to reach a higher level of importance while compared with a predefined metric, and those that do not reach this level are removed.

To measure the level of importance and create abstract views of the process model, two metrics are defined in fuzzy miner: significance and correlation.

Significance is determined for both activities and the transitions between them (nodes and edges) and calculates the relative importance of behavior. For this measure, the frequency of events, the count of distinct predecessors and successors, and the frequency of binary relations are considered.

Correlation measures the importance of the relation between two events considering additional attributes. For this measurement, the proximity of two successive events based on their average execution distance, the overlap of the two events' attributes, and similarity of event names are taken into account.

The fuzzy miner algorithm preserves the events with high significance and high correlation, and removes the ones with low significance and low correlation. Less significant but highly correlated events are aggregated. Through this procedure, an abstract and simplified process model is obtained. The process model is normally visualized by a colored graph in which the intensity of colors in nodes and edges is proportional to the significance of behavior.

In order to obtain the process models, the Disco tool (Van Der Aalst, 2011) can be used for discovering models of complex processes through abstraction and generalization facilitated by the fuzzy miner algorithm. Figure 5.2 shows the fuzzy miner capability in aggregating and abstracting a student process model, based on the above-mentioned metrics, during one of the sessions of the course. The relative darkness of the nodes and edges shows that the activity A9 is the most significant activity of the process model followed by A5. Also, the transition from A9 to A10 is the most significant. We can also observe that A9 and A10 are more correlated and can be aggregated into one cluster (if more abstraction is needed) compared to the relation between A9 and A8.

On the top of Figure 5.2, the process is shown with no aggregation, and on the bottom, the fuzzy miner reduces the model to the most significant and correlated activities and paths by applying the node and edge cutoff parameters (Günther and Van Der Aalst, 2007) to 20% and 0% respectively (0% is the indicator of minimum number of paths throughout the process). These parameters are set in order to show a simple and yet meaningful process structure. Note that, the Disco tool only allows to abstract through the high-level cutoff parameters and not through the significance and correlation metrics.

5.2.4 Cyclomatic Complexity Metric

The understandability measure of a process model can indicate how the model is easy to read and understand. In Figl and Laue (2011) it is shown that, although process models help people get a better understanding of processes, understanding complex models faces cognitive limits. For instance, it can be difficult for teachers to comprehend the complex learning processes of students. To read and understand the complex processes, we need metrics to explain the properties of the process models. In this context, we adopt the concept of measuring the understandability of process models, and transfer it as a tool

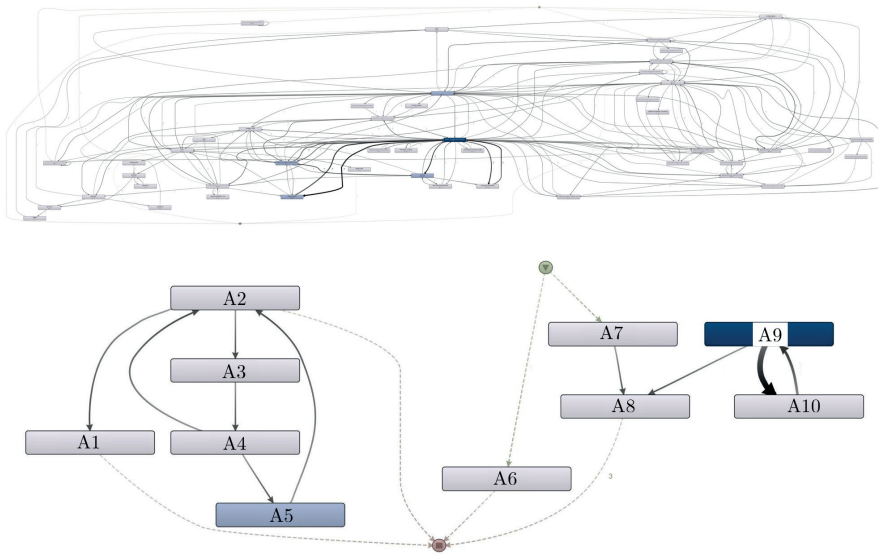


Figure 5.2: An example of a student process model obtained by Disco tool through applying the fuzzy miner. Every A refers to an activity (light to dark blue: activities with low to high frequency).

to assist us comprehend better the properties of learning processes.

Complexity metrics can measure understandability and maintainability of a workflow net or a business process model by extending the metrics applied to measure the software complexity (Gruhn and Laue, 2006; Štuikys and Damaševičius, 2013). These metrics in PM have been introduced for petri nets or workflow nets but can be extended to other formalisms as well (Lassen and Van Der Aalst, 2009).

For instance, Cyclomatic Complexity of McCabe (CM) has been used as a metric to measure the complexity of a control-flow graph of a program (Gruhn and Laue, 2006; Lassen and Van Der Aalst, 2009; McCabe, 1976), and it is usually applied to software complexity in the meaning of *the difficulty to maintain, change and understand software* (Zuse, 1991). CM is measured by calculating the number of linearly-independent paths through a graph (Gruhn and Laue, 2006).

Based on McCabe (1976) for computing CM we need to define $G = (V, E)$ as a control-flow graph where E is the number of independent paths and V is the number of activities. The Cyclomatic metric CM is defined as:

$$\text{CM} = |E| - |V| + P$$

where $|x|$ is the cardinality of the set and P is the number of connected components. P in our case is a constant with value of 1 since the process per session is one connected

graph with no interruption.

Since complexity has strong relation with various concepts such as simplicity, ease of use, uncertainty and the context of application, it has been a topic of interest to the designers of Human-Computer Interaction (HCI) (Ham et al., 2011; Holzinger et al., 2012; Rauterberg, 1992; Thomas and Richards, 2009). For instance in Rauterberg (1996), this metric is applied into the context of HCI to measure the complexity of the observable behavior by analyzing the recorded task solving process, and obtained the behavioral complexity. In the educational context, CM is used in various studies to measure the complexity of the programming assignments of students while learning (Mohamed et al., 2013; Sapounidis et al., 2015). For instance, Sapounidis et al. (2015) compares CM of children programing tasks when using two interfaces among different age groups. CM is the most widely used metric for measuring the complexity of structured programs (Štuikys and Damaševičius, 2013). However, to our knowledge, CM has never been used in the context of educational PM. By applying this metric along with PM, we hope to explain more properties of students' learning processes. In the next section, we explain how CM is adopted in the context of our study.

5.3 Simulator Description

Deeds (Digital Electronics Education and Design Suite) (Ponta et al., 1998) is a simulation environment for e-learning in digital electronics. The environment provides learning materials through specialized browsers for the students, and asks them to solve various problems with different levels of difficulty. It has been effective, for over ten years, in teaching and improving the learning outcome of students since it provides a general-purpose simulator with a highly interactive e-learning environment. The course of digital electronics is organized in separate theoretical and laboratory sessions where students work with the Deeds simulator. Here, we explain the learning process of students in the laboratory sessions and explain how IBL is implemented in the context of education with the Deeds simulator.

The laboratory sessions are designed, to enable the students to be engaged in an inquiry-based learning process, as follows. The sessions are assisted by facilitators and in the beginning of each session a set of problem-solving exercises are given to the students. For each exercise, the students follow a learning process that involves understanding the given problem and dividing it into various tasks, making observations in the simulation environment, doing experiments to find the answer, and finally explaining and justifying their solution and method used. This follows the second level of open inquiry outlined by Schwab and Brandwein (1966), which states that students are given the questions while the methods to solve the problem are developed by the students. We distinguish three main phases in their learning process that are aligned with the characteristics of inquiry-based learning described in Section 2.3. Note that the orientation is done

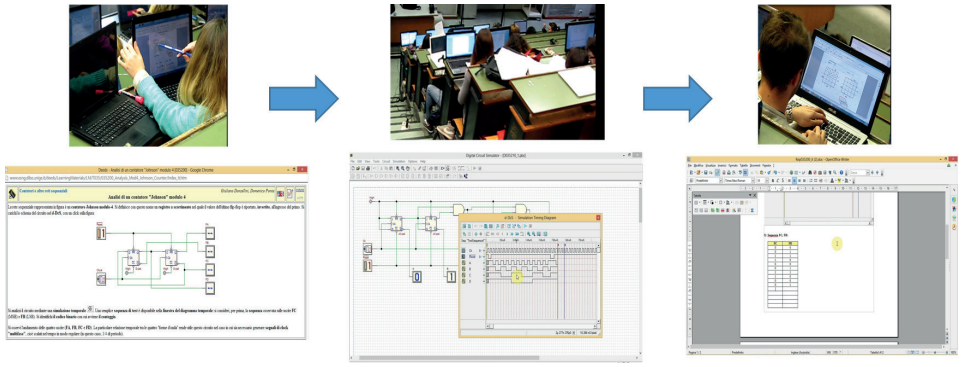


Figure 5.3: The three phases of Inquiry-Based Learning process performed by students during a laboratory session. From left to right: students conceptualize when studying the exercise, investigate and perform experiments in simulator, and conclude by explaining the results in a text editor.

through theoretical sessions separately from laboratory sessions. Here, we only focus on the student-centered IBL cycle as shown in Figure 2.2. The following learning process with the Deeds simulator can be seen in Figure 5.3:

1. **Conceptualization**: the students study the question and understand what is expected. In this phase, they can divide the question in smaller parts to better understand the problem. They try to identify the problem and set hypotheses to be tested later.

2. **Investigation and discovery** with simulation: the students open the question in the simulator and perform experiments to obtain supporting evidence for answering the question and test their hypotheses. In this phase, students are asked to find a solution while experiencing with the simulator and observing its behavior. Depending on the type of problem, they might need to work with various components of the simulator and timing diagram (will be explained further in Section 5.6).

3. **Conclusion and reflection**: the students interpret and explain the result and evidence collected based on their work with simulator. In this phase, they open the exercise in a text editor and explain the knowledge they obtained from the investigation and discovery phase. In some cases, they need to provide justifications for their choice.

5.4 Experimental Design and Data Collection

This study was carried out on two groups of BSc students during the laboratory sessions of a course of digital electronics at the University of Genoa. We performed our experiment

in two rounds: the first round, during the spring semester (March to June 2014) as a pilot experiment, and the second round, during the autumn semester (October to December 2014). 125 students participated in the pilot study and 100 students participated in the real experiment. The experiment design was improved after the pilot to limit the amount of noise in the collected data. For instance, during the first round (pilot), students were free to continue working at home and submit their exercises with a delay; while in the second round, we set more strict time limits to the process (the result of pilot study helped us improve our experiment and analytics approach). In this way, the students were obliged to work and submit their results during the time limits of each session (about three hours).

LADC

Since we deal with a simulation-based environment, having access to the client side data plays an important role in understanding of how the user interacts with the simulator for problem-solving. Therefore, data is gathered through a logging application called LA Data Collector (LADC), custom implemented for the study purposes. LADC is developed to log data from the client system, and currently can log the interaction data with Deeds browsers and external applications (we tried to collect as much data as possible without re-engineering the simulator code). Log data were collected from 100 students by LADC from six laboratory sessions. Additionally, two people were assigned to help the students during the data collection process in all the sessions. In this way, we ensured the accuracy of data collection by LADC and avoidance of data loss.

In addition to data collection by LADC, we performed the experiment through the following steps: semi-structured interviews with the instructors of the course; a demographic questionnaire designed to acquire general information, motivation, and background knowledge from the learners; and a feedback questionnaire at the end of the course.

The feedback questionnaire captured the satisfaction of students about the teaching method in the laboratory sessions, working with the Deeds simulator, and their feedback on our experiment. The final exam was designed to address the topics of laboratory sessions separately. In this way, intermediate evaluations of exercises, in addition to the final grade, can be related to the interaction data of each session.

The process of data collection was performed carefully to maintain the data accuracy and validity. About 86% of students in average were present and submitted their logs per session collected by LADC (100 out of 115 logs were considered for analysis). After each session, testing algorithms were applied on the logs for detecting problems and informing the students about their mistakes. In particular, we tested if the logs contained a minimum amount of data, and if data was relevant to the session exercises.

Additionally, five students volunteered for recording videos of their screen. The videos were collected from all the sessions in order to have an explicable source for our logs and

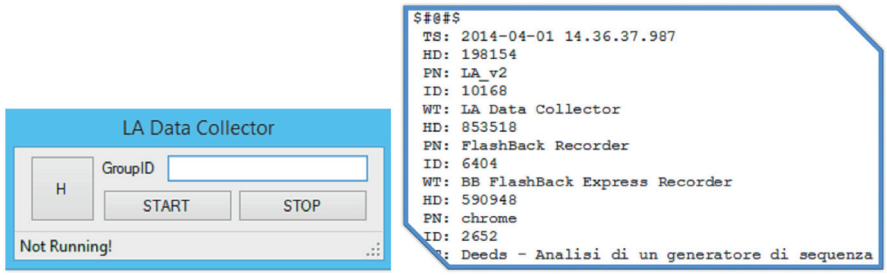


Figure 5.4: Data collection by LADC in a laboratory session. From left to right: LADC interface and raw data obtained from LADC.

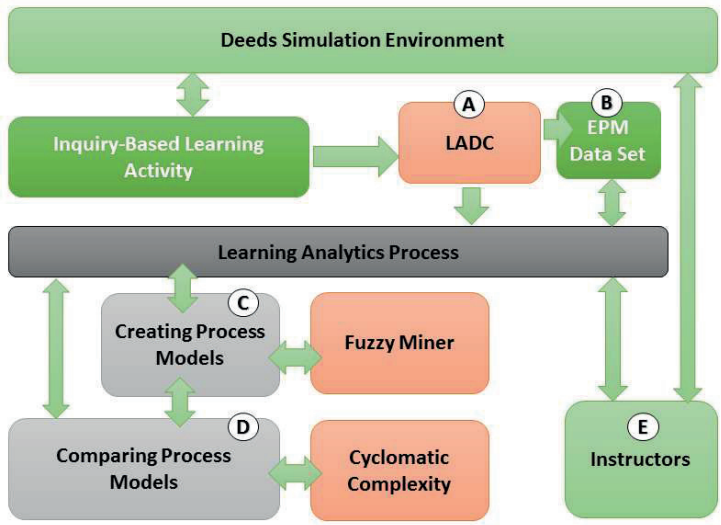


Figure 5.5: Learning Analytics approach based on PM and CM analysis

to understand better the raw data and the reasons behind meaningless or irrelevant data in the log. Due to the selection bias from these videos, they were only used to resolve the events and the bias was not transferred to results. Figure 5.4 shows the LADC interface and an extract of raw data collected from a student.

5.5 Learning Analytics Approach

The LA approach developed in our study (steps A-E in Figure 5.5) includes the data collection (A), pre-processing and data set generation (B), analytics (C and D) by applying PM and CM methods, and post-processing (E) by feedback to and from instructors. We explain each of these steps (A-E) as follows.

- Step A: The data collection phase through LADC was explained in Section 5.4.
- Step B: In this step, we generate a data set from raw data logs by data cleaning, noise detection, activity selection and mapping, and identifying the levels of granularity in our data. These steps helped us build a data set ready for further analysis. Also to understand better the data logs, we apply the fuzzy miner algorithm, as a tool to abstract complex or noisy processes. We use this method in order to detect the irrelevant activities to the process of learning (for instance, when students setup programs, or visit web pages that are irrelevant to the course) and identify the various levels of granularity. Level of granularity indicates the way we group the student performed-events in the PM process. The data set generation resulted from this phase is explained in detail in Section 5.6.

The analytics phase combines PM and CM in the context of inquiry-based learning. As explained in Section 5.2.1, it is challenging to obtain knowledge from educational processes since they are usually very complex and unordered. To overcome this issue, we propose to apply the fuzzy miner algorithm in combination with the CM metric to understand the differences between the learning paths of various students' clusters. In this context, the data analytics process includes creating the process models using fuzzy miner and comparing them by application of CM metric. These steps are explained further as follows.

- Step C: We aim to understand the properties of students' learning behavior through the discovery of process models. This phase can help us make sense of sequences of students' actions and identify the type of process that is followed by the students. The main questions to address are:

  1. What are the most significant and time-consuming activities throughout the whole course?
  2. What are the differences between exercises in terms of the students' effort?
  3. What are the differences of students' behavior based on their grades?

  To address these questions, we create the process models for sessions, exercises, and students. This is referred to the allocation of 'case' in PM as explained in Section 5.2.1. We present our findings in Section 5.7.1.

- Step D: We try to quantify the learning processes in order to compare them. For that, we measure CM of the process models of students in all the levels of granularity. As explained in Section 5.2.4, metrics like CM have been applied in PM for measuring the understandability of a process model (Lassen and Van Der Aalst, 2009). We further propose to apply CM to quantify the complex learning processes and compare them in terms of the academic achievements of students in particular,

in the context of inquiry-based learning applications. When the students experiment with the simulator and its various components, the inquiry process helps them develop methods to solve the problem. Therefore, we form the following set of hypotheses:

1. The more difficult exercises are for the students, the more they perform intensive activities and their CM increases. We use 'intensive' when referring to high CM to give an intuition on what CM measures. High CM means that the difference between the number of independent transitions and the number of distinct activities is high (see Section 5.2.4). In our context, we rely on the large difference between the number of nodes and the number of transitions. Given that the size of the set of activities / nodes is rather small compare to transitions and the students have equal amount of time per session, we can assume: if CM is high, the number of transitions compared to the size of the set of activities would be high. It means that the student is very active moving between the activities.
2. The more the students perform intensive IBL activities, the more they will be successful in learning the concepts.

We will verify these hypotheses in Section 5.7.2.

- Step E: We perform a semi-structured interview with the instructors of the digital electronics course to compare their expectations with our findings. Concretely, we aim to receive their expectations on the difficulty of course sessions for the students, and the process models of the students in terms of frequency and order of IBL activities. This phase can show whether the results of this study can add value to the insight of the instructors about the learning process of students. The opinion of instructors on the difficulty of sessions is an important factor to organize their instruction and the workload of given exercises in every laboratory session (e.g., they can avoid allocating too many difficult exercises to a single session). Our results can help the instructors improve their planning. Also, the expectation of instructors about the way students perform the IBL activities in terms of time, frequency, and order might differ from the reality. Our results can increase the instructors' awareness about the learning processes of students and warn them about the problems or outliers (e.g, whether a student spends too much time on an exercise) in the process.

  Four instructors of the course are interviewed, each from 30 minutes to 1 hour. The instructors do not have the same level of experience. The responsible professor has over ten years experience of teaching with this particular course, and the experience of the others varies from several years to several months. The results of this phase are presented in Section 5.7.3.

5.6 Data Set

In this section, we explain the attributes of our data set[2] built from the event logs of students which are characterized by multivariate, sequential, and time-series data. Since in building the data set the properties of PM are considered, and activities are chosen in a way to obtain the most abstract and descriptive process models, it is important to give an overview on the generation of our data set. As one of the contributions of this study, the data set is published[3] on the UCI machine learning repository with 230318 instances, and 13 attributes, and no missing values. The number of web hits for the first four months (October 2015 to January 2016) exceeded 10000 views. A comparison with educational data sets on the same repository displays our data set among the most viewed ones. Also, the successful access rate shows the growing need and request towards LA and EDM data sets.

5.6.1 Feature Selection

The raw data was collected from the students' interactions by LADC (see Section 5.4) in a text file. Each text file contains data of a session per student and records the interactions for almost every second in the following order: date and time, name of the application in use on the system in combination with the name of its window title, followed by title of the window in focus. Finally, data present start and end time of the process, idle time, mouse clicks, mouse movements, and keystrokes. The raw data gives us valuable information on the activities per session, per student, and per exercise since each exercise is designed to contain a code in the title which appears in the browser, the simulator title, as well as in the editor used for conclusion (further details are provided in Section 5.6). Based on the titles in use, the activities and cases are withdrawn sequentially in order to discover the process model of each student through PM. The extracted events are later aggregated into three levels of granularity and CM is measured on each one.

As the features/ attributes are selected and presented in a suitable format for Process Mining, activities can be linked together in a process instance or case. In this context, the potential cases in our data set are: session, student_Id, and exercise[4]. As explained in Section 5.2.1, 'case' can be allocated based on the needs of researchers. For example, if the aim is to create the process model of students and exploit the common processes among the students, the case needs to be set to 'student_Id' while if one needs to compare the activities from one exercise to another, the case needs to be allocated to 'exercise'. In

[2]For detailed information on data set, see appendix A.1

[3]The data set can be retrieved from:
www.la.smartlab.ws
The data set on UCI can be retrieved from:
https://goo.gl/Xo0TCq

[4]For accessing the content of exercises, see appendix A.4

the same way, sessions can be targeted if an overview on the total activities per session is needed[5].

5.6.2 Activity Selection and Mapping

In this section, we provide the details about the activities and their descriptions. The way we grouped activities lies on the inquiry-learning process of students explained in Section 5.3. We assigned labels in a way to get better insight on the various levels of the learning process while paying less attention to non-IBL activities of students. The activities are labeled based on the title of web pages that are on focus/ in the view of the student. In this context, for coding the activity names, we followed an order (described below) to ensure the accuracy of activity selection process. The reason of such order lies in the nested characteristics of recorded text. After finding all the keywords of our interest relevant to the course and educational activities of the students, if nothing was found, we assigned 'Other' as the activity name. Additionally, the way we built our data set is based on the second level of granularity (as explained in Section 5.5). In this way, we extracted all activities relevant to the topics of the course, and categorized the rest as 'Other'. Moreover, labeling of activities is done carefully since, it has an effect on the functionality of the fuzzy miner for correlation measurements (refer to Section 5.2.3). For instance, we named those activities of students performed with with a text editor in a way that starts with 'TextEditor' so that the algorithm considers them relevant to each other. The order and description of activities are as follows:

1. Study_Es_# of session_# of exercise: indicates that a student is studying or viewing the content of a specific exercise (e.g., Study_Es_6_1).

2. Deeds_Es_# of session_# of exercise: indicates that the student is working on a specific exercise inside the Deeds simulator (Digital Circuit Simulator) (e.g., Deeds_Es_6_1).

3. Deeds_Es: shows when the student is in the Deeds simulator but it is not clear what exercise the student is working on. To assign the number of exercise, it is possible to look for the previous activity, or the 'exercise' feature.

4. Deeds: contains other activities related to Deeds, e.g., when the students save a circuit image or export VHDL.

5. TextEditor_Es_# of session_# of exercise: when the student is writing the results of his work to submit later to the instructor. The students use a text editor (Word, Office, etc.) to answer to the questions and explain the solution they found with the Deeds simulator (e.g., TextEditor_Es_6_1).

[5]For further details on feature selection, see appendix A.2

6. TextEditor_Es: indicates that the student is working on an exercise in the text editor but it is not clear which exercise it is. This happens due to change of file names by the student, so we cannot recognize automatically which exercise he works on.

7. TextEditor: shows that the student is using the text editor but not for explaining the exercise results, this can contain other activities, e.g., when they open or save their works.

8. Diagram: when the students use 'Simulation Timing Diagram' to test the timing simulation of the logic networks, while using the Deeds simulator. It also contains these components: "Input Test Sequence" and "Timing Diagram View Manager ToolBar".

9. Properties: the Deeds simulator, the Simulation Timing diagram, and FSM (Finite State Machine Simulator) contain the properties window, which allows to set all the required parameters of the component under construction. For instance, the Properties can contain: "Switch Input", "Push-Button", "Clock properties", "Output properties", "textbox properties". We label all as 'Properties'.

10. Study_Materials: when the student is viewing some materials relevant to the course (provided by the instructor).

11. FSM_Es_# of session_# of exercise: when the student is working on a specific exercise on 'Finite State Machine Simulator' (e.g., FSM_Es_6_1).

12. FSM_Related: when the student is handling the components of Finite State Machine Simulator.

13. Aulaweb: students are using Aulaweb (a learning management system based on Moodle) which is used for the course of digital electronics at the University of Genoa. In Aulaweb, the students might access and download the exercises, upload their work, and check the forum news.

14. Blank: when the title of a visited page is not recorded.

15. Other: when the student is not viewing any pages described above, 'Other' is assigned to the activity. This includes, for majority of the cases, the student non-IBL activity (e.g., if the student is on Facebook).

Mapping scheme for analyzing the IBL activities

Here we explain how we map and group the coded above activities to the IBL phases described in Section 5.3. We use the IBL framework and map the learning activities to the three main phases of the IBL cycle as follows.

Conceptualization contains all the activities relevant to studying the exercises (activity 1 in 5.6.2), accessing the exercises (activity 13), and reviewing the topics relevant to the exercise (activity 10). *Investigation and discovery with simulation* contains all the activities relevant to experimenting with Deeds simulators and their components (activities 2, 3, 4, 8, 9, 11, and 12). *Conclusion and reflection* contains all the activities relevant to writing the results (activities 5, 6, and 7).

Note that we include the preparation and organization tasks (i.e. activities 4, 7, and 13) in the corresponding IBL phases since they occur naturally when the students are involved in the learning process. The rest of the activities that are not relevant to any of the IBL phases are mapped to non-IBL activities (activity 15). Activity 14 will be filtered as noise. The coding and mapping of activities are done through a comprehensive log analysis. All the titles of the visited pages by the students are extracted from the logs, compared to the screen records of the students to draw their meaning, coded into the activities, and associated with the IBL phases. PM is also used to detect the coding problems as will be explained in the next section.

In the following section, we describe the results of this study by applying PM and the CM metric over the presented data set and the above-mentioned activities.

5.7 Results

We present the results of this study based on our previous efforts (Vahdat et al., 2015c, 2014) in three steps: the results obtained by applying PM on the time-series data of the students behavior, the results obtained by applying the CM metric for comparison and further analysis of the process models, as well as the results of interviews with the instructors and comparison of their opinions with our results. In this study, the results are presented based on the second level of granularity, since from our analysis, this seems to be the best trade-off between the accuracy and compression of the process models. The findings of this study are explained in more depth as follows:

5.7.1 Results from Process Mining

The results of our work through Process Mining and application of fuzzy miner (addressing the steps B and C of Section 5.5) can be summarized into four main contributions:

- Identification of granularity levels and noise.
- Comparison of sessions.
- Comparison of exercises.
- Comparison of behavior of the most successful and the least successful students based on their final grades.

Identification of granularity levels and noise

Identifying granularity levels: applying fuzzy miner for process discovery facilitated the detection of less significant behaviors as well as less related events. By applying this miner to all of the events, we discovered that the non-IBL events (when students are not using the main applications relevant to the course) are the least significant student behavior in the process of problem solving. For instance, in Figure 5.6, the process model of a student is shown, considering the 20% most significant activities. On the right, the process shows the most frequent activities (from light to dark blue) while on the left, the most timely activities of student performance are shown (from light to dark red). Here, we do not show the titles of activities in the process model since they are very long and not yet anonymized.

We argued conceptually in Section 5.3 that the most important activities in the learning process are the IBL related activities. Here, the process models obtained from every session showed analytically as well that the significant activities of the models present the IBL related activities.

Three levels of granularity are obtained from analysis of the created models. In the first level, we consider all the events carried out by a student including both IBL and non-IBL events (relevant or not to the inquiry-based learning process). In the second level, we consider only the IBL events. In the third level, we group and map the events based on meta-cognitive activities such as: studying the exercise content, model building, simulating, and reasoning for each specific exercise.

Detecting noise: a comparison of the two process models (left and right models in Figure 5.6) showed that there is a lot of noise in the data which needs to be reduced. For instance, we noticed that activity A4 that is represented with high frequency and low performance time in the model, contains a blank page that occurs as noise when students switch between the applications. We found the reason behind the occurrence of such activities by comparing the models to the screen records. In this work, exploring the data logs in the form of process models and comparing them with the screen records of the students was a great help for reducing the noise. This phase also helped us in the process of activity selection for generating our data set explained in Section 5.6.2.

Comparison of sessions

As described in Section 5.2.1, case allocation in PM is important in discovering the learning processes. Here, we focus on the case 'session' and obtain a process model of the entire laboratory sessions to identify the types of activities that were the most performed by students. This approach helps us address the first question of step C in Section 5.5.

The result (Figure 5.7) shows that the most frequent activities for the students in all the sessions are 'Diagram' and its 'Properties' of which relative frequencies are much higher than the following activity 'Aulaweb' (7% difference in their relative frequencies).

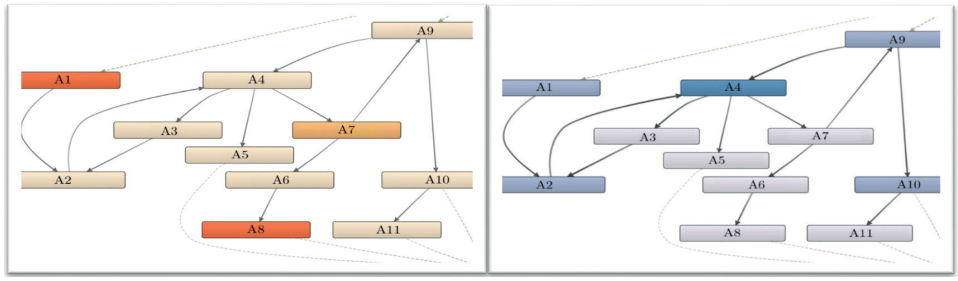


Figure 5.6: Comparison of frequency versus performance of an extract of student process model, obtained by Disco tool and applying the fuzzy miner. Every 'A' refers to an activity (On the left, light to dark red: activities with low to high performance time. On the right, light to dark blue: activities with low to high frequency).

'Diagram' is the most time-consuming activity among all. This result presents that the students spent more time and effort in IBL phase of 'Investigation and discovery with simulation'. This result can be explained as follows: Since the students learn by experimenting with the simulator, the most effort is performed on the simulator components. These results can help the instructors provide more assistance with those components that require more time and effort from the students.

Comparison of exercises

Here, we address the second question of step C in Section 5.5 by allocating the 'case' to 'student_Id' and 'activity' to 'exercise'. This approach allows us to compare the exercises in terms of the frequency and performance of the students' actions. In other words, we consider the activities per exercise to understand better which exercises take more time and effort. For that, we aggregated the activities relevant to an exercise and labeled them similarly (e.g., all the activities were performed by a particular student while solving the 4th exercise of the session 5 were aggregated into 'Es_5_4'). Note that, we tested this approach over the first and the second levels of granularity and obtained similar results in terms of relative frequency and performance time of exercises.

The results show that sessions 4 and 6 contain exercises with the most frequent activities while sessions 3, 4, and 6 contain exercises with the most time-consuming activities. Figure 5.8 shows an extract of frequency and performance process models by applying the node and edge cutoff parameters of 0% and 100%. As an example, 'Es_3_1' is indicated as the most time-consuming exercise while its frequency (of activities) is not high relative to the other exercises. A possible explanation is that during this exercise, students have a high proportion of waiting time.

In another example, the first three exercises of the 6th session are performed with high frequency and amount of time, and the three following exercises are characterized

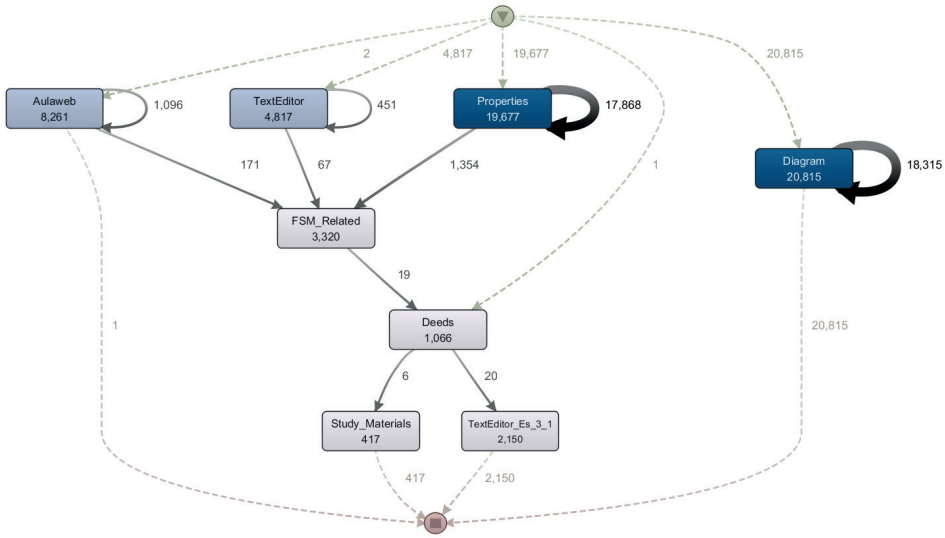


Figure 5.7: A process model showing the most frequent activities throughout all the sessions, by setting the node and edge cutoff parameters to 0%. Here, case is allocated to 'Session'.

by the least effort while their contents are as difficult. A possible explanation can be that the students spent most of their effort on the first exercises, and gave up on the last ones. Our results can increase the awareness of teachers on the relative level of effort performed by students for various exercises and help them balance the difficulty level of sessions.

Comparison of behavior of the most successful and the least successful students

Here, we address the third question of step C in Section 5.5. We compare the common activities of successful versus non-successful students to understand the difference of behavior of students based on their grades. Similarly to the previous result, the 'case' in PM is allocated to 'student' and 'activity' to 'exercise'. We only considered the students who attended in all 6 sessions of the course and clustered them into the most and the least successful students based on their final grades (12 students per cluster). Note that, we assume that the grades are valid and reliable, since the course of digital electronics with the Deeds simulator are held in the respective university for over ten years and the tests are established and have the standard quality[6].

A comparison between the process models of the two clusters (Figure 5.9) shows that the most and the least successful students spent their time and effort differently on the

[6]For more information on the grades, see appendix A.3

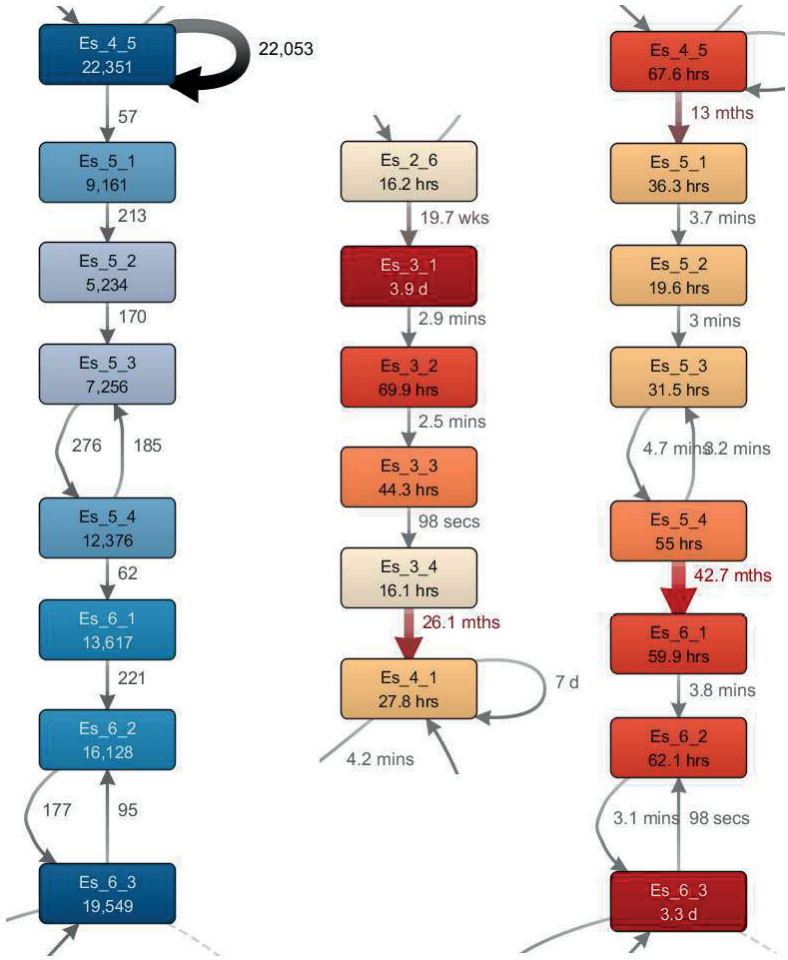


Figure 5.8: An extract of the most frequent activities in exercise level (in blue on the left) and the most time-consuming activities (in red on the right) throughout exercises, by applying the node and edge cutoff parameters to 0% and 100%. Here, case is allocated to 'student_Id' and activity to 'exercise'.

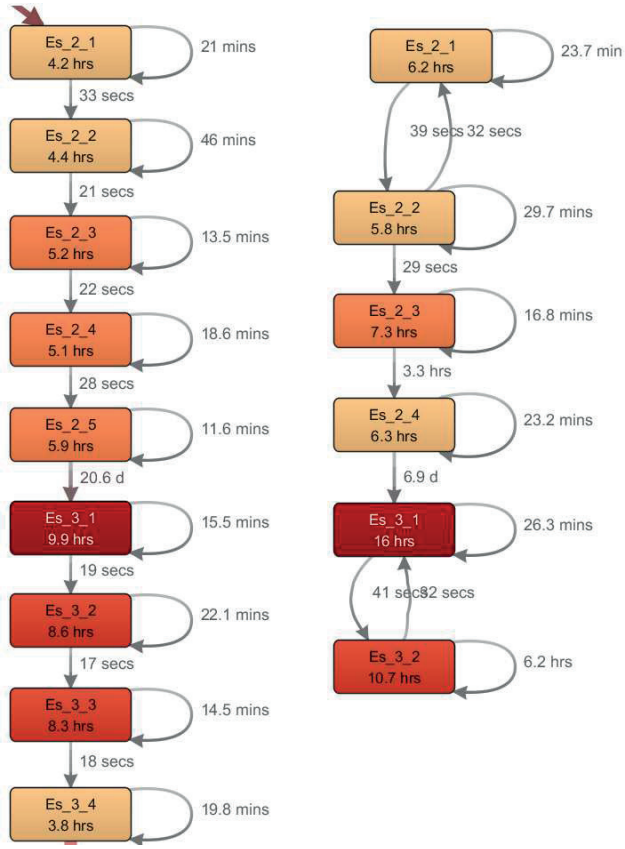


Figure 5.9: An extract of the performance model of most successful students (left), and performance model of less successful students (right).

exercises. For instance, the least successful ones spent less effort or totally skipped the last exercises of the second and third sessions (Es_2_5, Es_2_6, Es_3_3, Es_3_4) compared to the successful group. Also, a comparison of the time spent on the exercises shows that the most successful students are persistent on spending as much time or even more when the exercises become more difficult even if they are approaching the end of the session. Contrarily, the less successful ones tend to spend less time when they solve the ensuing exercises. We tested this approach on the first and second levels of granularity and obtained similar results.

Up to this point, the results are drawn based on the visualizations and statistics of frequencies and performances drawn by fuzzy miner in Disco tool. In the following section, a more parametric method to compare the process models of the students is presented. In this context, application of CM metric helps compare the process models of the students quantitatively.

5.7.2 Results from Cyclomatic Complexity Metric

Our results from application of CM on process models address the hypotheses described in step D of Section 5.5, and can be summarized into three main contributions:

- About the difficulty level of sessions: the average CM per session has a negative correlation with the average of intermediate grades of students.
- About the interactive behavior of students: the average CM per session of most successful students is higher than the one of least successful ones.
- About the similarity of student behavior depending on the level of granularity, and their behavior on IBL versus non-IBL events: successful students have more similar behavior in terms of intensive IBL activities while the less successful ones perform more non-IBL activities.

An indicator of session difficulty

Correlation: we compared the average CM of various laboratory sessions, as measured from the student processes, with the difficulty level of sessions. The results show that CM in average is related to the difficulty of the assignments given to the students during a session. We interpret the 'difficulty level of session' based on the average of intermediate grades of students on assignments related to a particular session of the course. An intermediate grade of a student is the sub-grade of her final exam's grade that assesses the student knowledge about the topics of a particular session.

A comparison of the average of intermediate grades ($\overline{IG}$) with the average CM of total students ($\overline{CM}$) is performed. The result shows that $\overline{CM}$ per session has a strong negative correlation with $\overline{IG}$ (r = -0.93). In Figure 5.10, when $\overline{CM}$ increases, $\overline{IG}$ decreases. This relationship is shown for the second level of granularity, however a similar result is obtained for the two other levels of granularity. These results address the first hypothesis of step D in Section 5.5. The more difficult a session is for the students, the more they perform intensive activities, therefore, their $\overline{CM}$ increases.

Note that for computing the correlation, the first session is not considered in the analysis since the complexity metric in the first session of the course is an indicator of trial and error and the students mainly explore the environment. This is due to the fact that the simulator is new for all of the students. Based on the demographic questionnaire, 74% of students did not have any experience with the simulator, and the rest had very little experience. So the background knowledge of students about the simulator is considered homogeneous. Additionally, during the first session of the course, $\overline{CM}$ is relatively high while $\overline{IG}$ is the highest. This can be explained as follows: interacting with the system was challenging for them in the beginning, while the assigned exercises were very simple. From the second session of the course, the average complexity is an indicator of difficulty level of exercises and effort of students when confronting with various exercises.

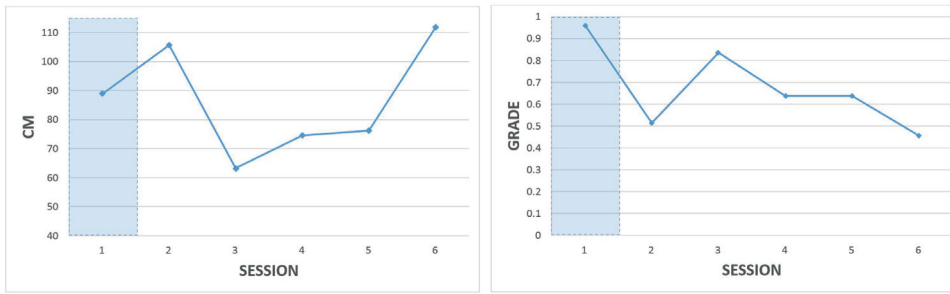


Figure 5.10: Average of Cyclomatic Complexity (CM) per session (left), average of intermediate grades per session (right).

Variance: additionally, we considered the variance of CM throughout all sessions. Figure 5.11 shows the variance of complexity of behavior per session. It is worthwhile to mention that when a session is more difficult for the students (lower grade), not only their $\overline{CM}$ increases but also the variance of their CM is higher. In other terms, the CM variance has a negative correlation with the intermediate grades (r = -0.8, excluding the first session from the analysis). Also, there are more outliers in the difficult sessions (e.g., sessions 2 and 6). This can be explained as follows: when a session is very difficult, some students try much harder than average while some of them might give up.

Our results can be used for predicting the perceived task difficulty by the students based on their interaction with the Deeds simulator. In this way, instructors can spend more time and effort on the topics that are perceived more complicated by the students. Additionally, this course is offered to many engineering majors in University of Genoa, and the students of various intakes have different backgrounds. Therefore, the complexity of the tasks and sessions can be different for diverse classes and majors.

In this regard, $\overline{CM}$ shows how difficult or easy a session is for a group of students. Teachers might be interested in comparing various groups of students based on their background knowledge and the difficulty level of exercises perceived by students. Later, we compare our results with the interviews of the instructors, and show that the findings can be valuable for the teachers in order to organize their lessons and to balance the difficulty of the sessions for the students.

An indicator of difference of behavior between the least and the most successful students

Another contribution of this study shows the difference of behavior of students when they are clustered into two groups based on their final grades: high-graded versus low-graded students. Our results show that $\overline{CM}$ of high-graded students is higher than the one of low-graded students. This difference grows when the distance of the student clusters increases. We incrementally re-clustered the students by increasing the difference of their

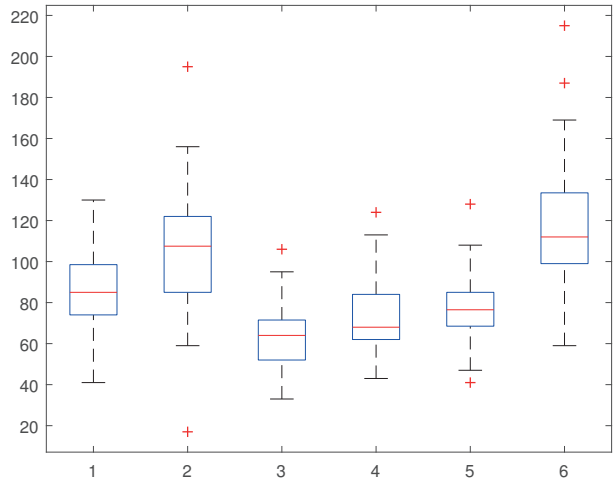


Figure 5.11: Variance of complexity per session.

grades: high-graded are the one with grades over the mean value of the grades (μ) plus n times the standard deviation of the grades sigma ($\mu + n\sigma$) while low-graded students are the ones below $\mu - n\sigma$ where n varies in $\{0, 2\}$. Figure 5.12 shows this difference when n is 0 (image on the top) versus 2 (the lower image) for the second level of granularity. It means that when n=2, the clusters are the furthest from each other where only the most and the least graded students are considered. The results from the second and third level of granularity show a better separation in $\overline{CM}$ compared to the first level. This can be due to the non-IBL events in the first level of granularity that contributes as noise to the results.

Our findings confirm the result of PM shown earlier (Section 5.7.1) where the process models of the most and the least successful students are compared. As explained, comparing the process models of students at the exercise level showed that the less successful ones tend to quit the session earlier or skip some exercises, therefore the complexity decreases. Additionally, these results address the second hypothesis of step D in Section 5.5. The better the students are, the more they perform intensive activities, therefore, their $\overline{CM}$ increases.

We performed a correlation analysis between the average of $\overline{CM}$ throughout all sessions for clusters ($\overline{\overline{CM}}$) with the average of the final grades. The analysis showed a strong positive correlation between $\overline{\overline{CM}}$ and the average of final grades (r = 0.99). This result validates our previous analysis: the students who got better grades, had a larger $\overline{CM}$ per session in average.

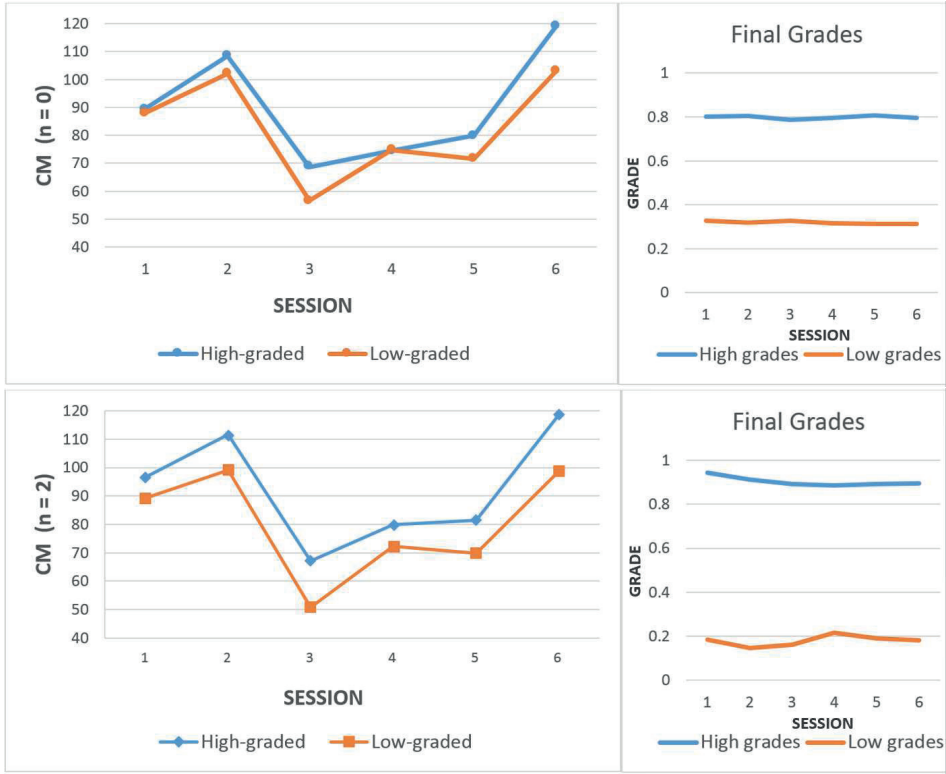


Figure 5.12: Comparison of student clusters via the difference of average Cyclomatic Complexity (CM) for the second level of granularity. The average final grades are shown on the right (the grades are normalized to $[0, 1]$ due to privacy issues). $n = 2$ and $n = 0$ indicate that the clusters are the furthest from and closest to each other respectively.

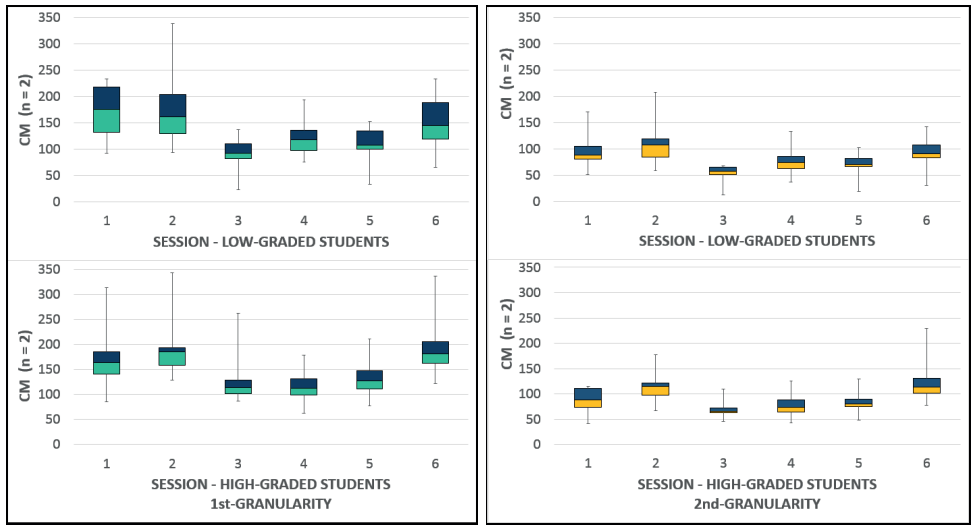


Figure 5.13: The variance of Cyclomatic Complexity (CM) for: Low-graded students (on the top) versus high-graded students (on the bottom), and first granularity level (on the left) versus second granularity level (on the right).

An indicator of similarity of the behavior of students based on the granularity levels

We compared the variance of CM between the most and the least successful students for the first and the second level of granularity. In this way, we could compare the behavior of students via IBL versus all events (including non-IBL events). The results show that when considering all the events (Figure 5.13 on the left), the behaviors of high-graded students are closer to each other than the low-graded ones for the majority of sessions (i.e. there are smaller variances on the bottom left diagram).

When we exclude the non-IBL events in the second level (Figure 5.13 on the right), there is more contrast in the variance between the two levels of granularity for the least successful students than for the most successful ones (i.e. there is a bigger shift moving from the upper left diagram to the upper right one compared to the lower diagrams in Figure 5.13). This comparison supports the second hypothesis of step D in Section 5.5 showing that the low-graded students perform more non-IBL activities than the better-graded students during their learning process.

5.7.3 Feedback from Instructors

The results obtained from semi-structured interviews with the instructors of the digital electronics course (step E in Section 5.5) are valuable for the interpretation of our findings. The interview questions were about the difficulty level of each course session for the

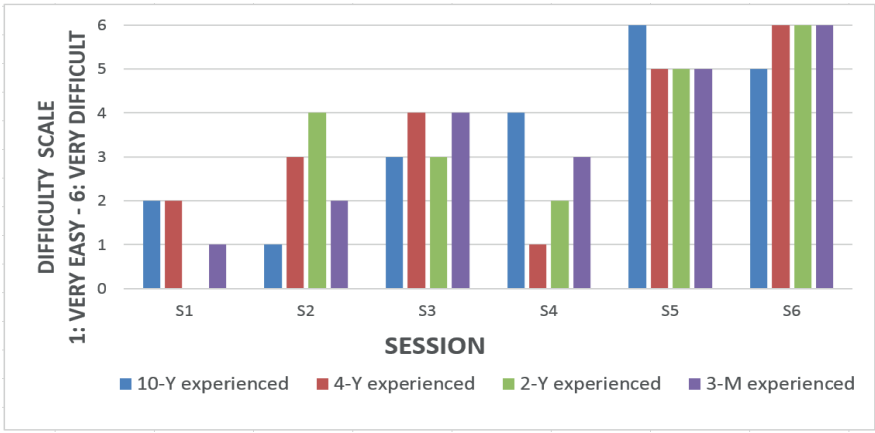


Figure 5.14: The difficulty of sessions based on the instructors own teaching experience (from 10 years to 3 months).

students as well as their experience and feedback during the laboratory sessions. They were asked to sort the sessions based on the difficulty level from very difficult to very easy (Figure 5.14). Also, the results of this study were shown to the instructors in order to get their interpretation and feedback.

Rating the difficulty: the results show that all of the instructors agree on which sessions were the most difficult for the students (sessions 6 and 5) but their opinions varied about the difficulty level of the rest of the sessions (see Figure 5.14). For example, two teachers indicated session 4 as medium difficult and the two other indicated it as an easy session. The feedback of the instructors helped us interpret the results better. For instance, we showed in Section 5.7.1 that there is waiting time in an exercise of the third session while the instructors reported the third session with medium difficulty during the interviews. As the reason behind such result, one of the instructors suggested that the students might have not understood the question well and hesitated to perform experiments.

Describing the processes: also, we asked the instructors to draw an expected process model of a student, given the main IBL events. We compared the models drawn by instructors and observed that their process models varied in size of events and linear paths. The expected process models of instructors were much less complex in number of activities and transitions in comparison with the real process models of the students. Figure 5.15 shows, on the left, a simplified process model of a student for the second session of the course. On the right, a process model sample drawn by an instructor presenting the expected graph of activities and transitions for a normal student. A comparison of these two shows that even a very simplified student model is more complex than the expectation of instructors.

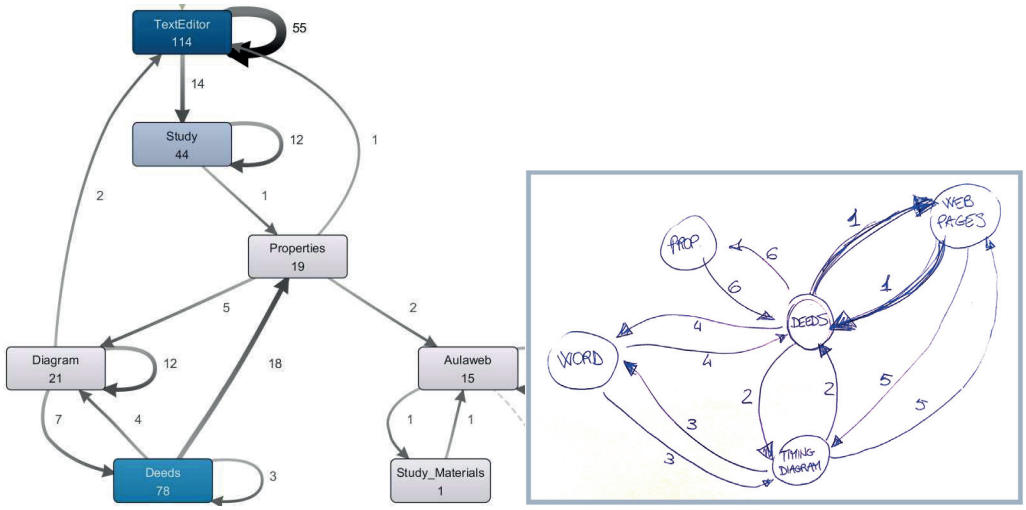


Figure 5.15: Comparing an instructor's expected model with a simplified student process model. A sample of an expected process model drawn by an instructor is shown on the right, presenting activities and transitions for a normal student. The labels on the arrows are suggesting the number of transitions.

Also, the opinions of instructors varied about the significance of relationship between the events. For instance, two indicated that increasing amount of transitions is an indicator of less successful behavior and the other two stated that it is not an indicator of learning.

These results confirm that the average CM of students and the obtained process models can give valuable insights to the instructors of the course, in particular to those with less experience. The fact that the instructors have varied opinions about the difficulty levels of exercises and performance level of activities shows that identifying where the students might face difficulties in the course is not trivial. Our results can help the instructors distinguish better the difficulties of the students and provide a more focused guidance.

5.8 Discussion

PM methods help improve our understanding of the behavior of learners through abstraction and discovery of learning processes. Applying these methods can lead to better detection of problems in particular, in the contexts like simulation-based learning where we can collect data through a logging system.

In this study, we applied PM methods to discover, understand, and compare the learning processes of students. We obtained the data from logging interactions of stu-

dents while they were working with an educational simulator named Deeds during six sessions of a course of digital electronics. A data set of temporal and sequential activities per student and per course session is collected by the LADC custom tool developed for our research. We used the complexity metric CM to quantify the processes of student learning and compare the behavior of various clusters of students based on their academic achievements. We also interviewed the four instructors of the course for their interpretation and feedback.

As described in Section 3.2.3, there are few efforts on applying PM in the educational contexts in particular, for comparing the process models and analyzing the differences of learning behavior. Some examples are Bannert et al. (2014) and van der Aalst et al. (2013) where the process models of different groups of students are compared through the methods of conformance checking. Contrarily, we quantified the student processes through CM which allowed us to relate the processes with other parameters such as students' grades. Quantifying the students' processes through CM allowed to consider a student learning process as a whole and to compare it with the average CM of successful students.

Our findings show that process discovery and CM of the student models can give insight about the learning process of students, and give valuable information about how students act in the process of problem solving. We applied fuzzy miner, a PM algorithm, to abstract and simplify the process models, to detect and remove the irrelevant activities of the learning process, and to help define the granularity levels. By application of PM, various sessions of the course could be compared based on the activities of the students. Similarly, the exercises given to the students could be compared based on their behavior. In this way, we compared the behavior of the most successful and the least successful students. Note that, we presented some extracts of students' process models to address the raised questions about the learning behavior of students. Depending on the needs of instructors, and by allocation of cases and activities differently, we can obtain more process models, and compare the activities in a higher or lower level.

Additionally, our results indicate that: CM in average is positively correlated with the average final grades of students and inversely correlated with the difficulty level of laboratory sessions. As a result, the average CM of students can be used as an indicator of the difficulty level of laboratory sessions for a group of students (from various engineering majors or intakes). These results are in conformity with inquiry-based learning theory explained in Section 3.3. The more difficult a session is for the students, the more they are engaged in the inquiry-based process and perform more intensive activities, therefore their average CM increases. Similarly, the better the students are, the more they are focused on the inquiry-based process and perform more intensive IBL activities, therefore their average CM increases. We also contributed to the community by sharing data, providing free access to our data set built from client side data logs.

The advantage of our study can be described as follows. It provides tangible visu-

alizations through PM for analysis of students learning behavior. These visualizations can give insight about the learning process of students in particular: where the students are stuck and spend a lot of time, which sessions and exercises take more effort from the students, whether the difficulty level of exercises throughout the course sessions is balanced, and which students have given up the exercises and need more help. Furthermore, applying CM can be helpful for the instructors to know how involved the students are in the IBL process and help the ones who are less involved.

In the future, we aim to develop user-friendly tools for the instructors and use our results as a basis for generating feedback and recommendations. For instance, CM can be used to show the abnormal behavior of students during a session. When CM of a student is far from a threshold, teachers can be notified to intervene. Also, CM of a group of students that reflects on the difficulty level can help teachers balance the difficulty of exercises based on the needs of the students and their different academic backgrounds.

5.9 Summary

In this chapter, we presented our second empirical study to investigate the IBL cycle in particular the phase of 'investigation and discovery' in the context of simulation-based learning. Note that, we considered the three IBL phases in this study, however, most of the efforts of the students were carried out in the phase of 'investigation and discovery' with simulation.

To analyze the behavior of students while interacting with the simulator, we adopted PM methods in combination with CM to quantify the learning processes of the students and compare them.

Our results concluded that PM can be applied to discover the process models of students' behavior while learning with an educational simulator. PM was valuable to obtain the general students' behavior throughout all the sessions, and their specific behavior for various exercises. We also could compare the process models of the most successful and the least successful students based on their final grades. Additionally, we proposed CM to quantify the learning processes and showed that the average CM per session has a negative correlation with the average of intermediate grades of students. The average CM per session of most successful students is higher than the one of least successful ones. And successful students have more similar behavior in terms of intensive IBL activities while the less successful ones perform more non-IBL activities.

We propose to implement PM methods to the simulation environment by developing dashboards which provide reports based on the student interactions in different phases of IBL. Such visualizations can raise awareness of teachers about the learning process of students and help them detect the students who deviate from the appropriate learning path and provide help on time. The instructors can be informed about the students who need more attention on a particular IBL phase and the details of the problem they face.

For instance, if a student has problem in understanding the content of an assignment or has difficulty performing the experiments with the simulator, the teacher can be alerted about the student current state.

More details on the limitations and future perspectives of this study will be discussed in chapter 7. In the next chapter, we focus on the third phase of IBL: Conclusion and reflection, in a puzzle game where again IBL cycle is implemented.

Chapter 6

IBL Phase Three: Application of Data Mining in Game-Based Puzzle-Solving

6.1 Introduction

Learning Analytics (LA) and Educational Data Mining (EDM) that focus on the learning experience can be applied in combination with game analytics that focus on the gameplay experience, to support the achievement of learning goals (Bohannon, 2010; Serrano-Laguna et al., 2014). In particular, LA and EDM aim at collecting and analyzing data about learners and their learning context by applying methods to extract non-trivial patterns from data (Romero et al., 2010; Siemens, 2012).

LA and EDM can be particularly effective in IBL environments (De Jong et al., 2014), where the learner is involved in active discovery of knowledge through exploration and experiments rather than reproductive approaches (Kruse and Pongsajapan, 2012). One such learning environment is the category of learning games as described in Section 2.4. In this context, LA and EDM methods can enhance understanding about the learning behavior in a data-driven way, allowing for the development of personalized and adaptive computer-assisted learning systems (Siemens and Long, 2011).

Various methods of analytics in e-learning and game analytics help researchers make sense of data collected from user behavior, particularly through the use of modeling techniques (Siemens, 2012). As explained in Section 2.6.3, one such technique is Process Mining (PM), used for modeling students' sequences of activities and behaviors, with the objective of understanding the underlying processes of the learning behavior (Trcka and Pechenizkiy, 2009; Trcka et al., 2010). Another technique is cluster analysis (see Section 2.6.2), which is used in game behavioral data analytics for pattern discovery and player profiling (Bauckhage et al., 2015).

The application of LA and EDM methods is proposed in various studies on game-

based learning. Baalsrud Hauge et al. (2014) propose an LA framework for games for learning, contemplating both real time and after-game analysis. Serrano-Laguna et al. (2014) define a scalable LA model based on the identification of generic traces that can be the basis for game-based assessment rules. In another example, a game is designed to teach XML basics to the students. LA is used to simplify the tasks of teachers by providing real-time visualizations of students behavior. Therefore, the teacher could obtain a more thorough view of the students' performance during a lesson (Serrano-Laguna and Fernández-Manjón, 2014). In our work, we are interested in applying LA and EDM differently, to investigate the learning behavior of players considering the characteristics of their IBL cycle.

In this work[1], we aim to investigate the use of LA methods in one specific class of games: digital puzzle games. This type of game is commonly used for educational purposes (Liu and Lin, 2009), possibly given its typical reliance on problem-solving and on logical and mathematical intelligence (Becker, 2005). Since this type of game highly constrains the interaction and has clear success criteria, LA can be applied effectively.

We describe our approach to explore the way players learn game skills and solve problems in an open-source puzzle game called *Lix* (Naarmann, 2011). We performed a study with 15 participants, who played the game for 272 times in total. We focused on the IBL phase of 'conclusion and reflection' (see Section 2.3.3) where the players define and execute tactics to succeed the game. We applied PM and cluster analysis to automatically extract players' tactics as process models. These models are built in order to detect the most significant components of a tactic, and represent the puzzle-solving behavior of the players of a cluster. The components are used later as references for validation of clustering results.

This chapter is organized as follows. In Section 6.2, we describe the open-source puzzle game that we extended to collect data about detailed player behavior. We then describe our experiment to collect data of players in Section 6.3. In Section 6.4, we present the LA approach adopted in this study. The results of our study are presented in Section 6.5. In Section 6.6, we discuss our results and indicate pointers for future work. Finally in Section 6.7, we provide a summary of this chapter and conclusions.

6.2 Game Description

We extended an existing open-source puzzle game called *Lix*, which is inspired by *Lemmings*, a 1991 game by DMA Design. In *Lix*, the objective is to guide a group of lixes (simple characters in the game) to a designated exit by assigning a limited number of skills to specific lixes, which allows the selected lix to alter the landscape, to affect the behavior of other lixes, or to clear obstacles to create a safe passage for the rest of the

[1]This chapter is written partly based on Vahdat et al. (2016a).

lixes. The player assigns one skill at a time to a lix, so that it performs the desired task (e.g. assign 'JUMPER' to jump). Figure 6.1 shows the interface of the game.

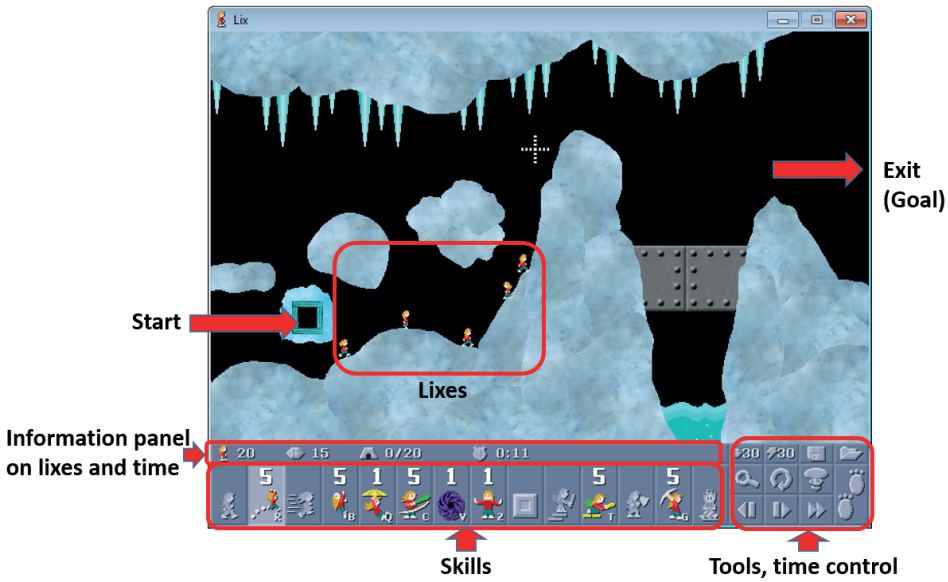


Figure 6.1: *Lix* game interface. The player guides a group of lixes to the exit by using the skills and tools in the game.

In its original version, *Lix* has a wide selection of puzzles with different levels of difficulty. We selected one intermediate-level puzzle, which could typically be solved within one hour by players with no prior experience in this type of game. The selected puzzle contains a *spacial insight problem* (Dow and Mayer, 2004) – it requires a shift in perspective to overcome the familiar way of looking at the problem, and implies the need to remove a self-imposed spatial constraint.

There are three main components of the process of solving the selected puzzle: task comprehension, definition of a tactic, and precision of execution.

Task comprehension: First of all, players need to understand what is expected (Bottino et al., 2007). In the case of the selected puzzle, this involves understanding the nature of the goal itself (i.e. finding the exit) and understanding the effects of their activities in the game environment (i.e. learning how to use the available skills that can be assigned to a lix).

Definition of a tactic: Once players have understood the task itself, they need to define a tactic, that is, devise the correct order and frequency of activities to achieve the goal (Bottino et al., 2007). Here, players face the spatial insight problem in which a shift in perspective is required; they need to find the self-imposed spatial constraint and remove it (Dow and Mayer, 2004).

Precision of execution: When the tactic is defined, players still need to execute it with precision, i.e. clicking in the right place at the right time, which depends on the current dynamics of lixes and their environment. The game interface has keyboard shortcuts and secondary controls that can be learned and used to facilitate execution of in-game actions.

In this work, we are interested in the two components of *definition of a tactic* and *precision of execution* and we relate them to the third IBL phase of 'conclusion and reflection' (see Section 2.4.3). The first and second phases of IBL occur when the players try to understand the task and learn the skills by exploring and experimenting in the game. When the players define a tactic and execute it precisely, they actually reach a conclusion.

6.3 Experimental Design and Data Collection

The study was carried out with 15 adult participants. All participants declared familiarity with computers. Each participant played the game individually, one at a time. Participants were given a brief explanation of the experiment and the goal of the game. No explanations about the game user interface were given. Participants were then given a pre-test questionnaire to collect demographic data, gaming experience, and familiarity with the games *Lix* or *Lemmings*.

Participants were asked to play the puzzle as many times as they wanted, with freedom to quit at any time[2]. They played the game 272 times in total. They were asked to think aloud, so that we could take notes on their reactions, tactics, reasoning, and persistence in dead-ends. The observation notes and creation of the replays helped us form expectations about the players' tactics. These expectations are used for further evaluation of our analytics approach.

The expected duration of the experiment was 30–40 minutes. However, we did not set an upper limit to the game duration. The actual duration depended on persistence and problem-solving skills of each player. When the participants solved the puzzle, they were asked to play once more to try to shorten their game time and improve the precision of execution. In this way, we could obtain the last used successful tactic of a player while reducing noise in the data.

To collect game data, we altered the game to perform network calls to a web service that listens to relevant game events, and records them in a database[3]. The events recorded are of two types: *game* traces and *meaningful variable* traces (Serrano-Laguna et al., 2014). The *game* traces we collect indicate timestamps of when the player started the game, started or restarted a puzzle, paused the game and returned to the menu. The *meaningful variable* traces consist of a simple record of a timestamp, a short code de-

[2]For more information on the gameplay, see appendix B.4.

[3]For detailed information on data set, see appendix B.

scribing the skill assigned to a Lix, an internal identifier of the Lix to which the skill was assigned and an internal measure of game time (called "update"). Table 6.1 illustrates the format of the events recorded[4].

timestamp	**activity**	**update**	**which lix**	**lix saved**	**skills used**
2015-12-10T14:29:19Z	STARTLEVEL				
2015-12-10T14:30:20Z	ASSIGN=MINER	949	1		
2015-12-10T14:31:15Z	ASSIGN=CLIMBER	1766	9		
2015-12-10T14:31:23Z	ASSIGN=PLATFORMER	1890	13		
2015-12-10T14:31:41Z	RESULT			20	13

Table 6.1: An extract of game events recorded and sent to the listening web service.

Every attempt at solving the puzzle is recorded (one attempt is represented in Table 6.1 as the lines between 'STARTLEVEL' to 'RESULT'). At the end of each attempt, an entry is recorded with the result, indicating the duration (in "updates" and in seconds), the number of assigned skills, and the number of saved lixes. We do not record *input* traces (i.e. mouse clicks or keyboard entries); however, since the game relies on a completely deterministic physics engine, it is possible to replay the attempts using the data we collect. The replays from log data are used as valuable sources for our observation.

Currently, collected data is used for offline analysis (after the game is ended). Future development will include real-time suggestions of adaptations in the game to conform to the player's skills. The source code of the game is available online (Carvalho, 2015).

6.4 Learning Analytics Approach

Our goal is to understand how participants solve puzzles in games, through a data-driven approach. In particular, we aim to identify the clusters of the tactics applied by the players and identify a reference sequence for each cluster. We obtain the reference sequences by building process models of clustered tactics, which identify the most significant activities and transitions. These references will play a central role in validation of the results. By comparing a player's process to previously established successful references, we can detect whether the player behaves similarly. The outcome of our approach can serve as a basis for recommending interventions, e.g. providing hints when a player is likely to fail.

Our analytics approach comprises of three main steps. A preliminary step is collecting the data from the game ('A' in Figure 6.2). The first step of our approach is to identify the tactics adopted in the game by players through cluster analysis ('B'). In the second step, we obtain the process models of the identified tactics through PM and fuzzy miner. These

[4]For more information on features, see appendix B.2.

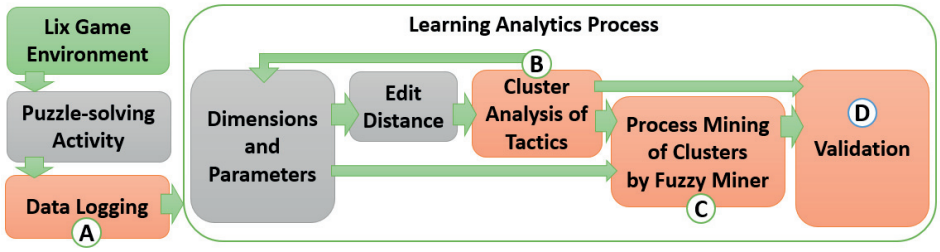


Figure 6.2: Learning Analytics approach

models represent the most significant components of the tactics (refer to the description of fuzzy miner in Section 5.2.3) which yield references that are central in validation of our results ('C'). Finally, we validate the results of cluster analysis and PM by measuring how the elements of a tactic cluster converge to their reference ('D'). Details of each main step of the approach ('B'–'D') are given in the following sections.

6.4.1 Cluster Analysis of Tactics

In the context of our puzzle game, there can be several possible tactics to succeed in the puzzle. The optimal tactic can be suggested by the game designer. In our approach, we aim to identify successful tactics from the collected data, as there can be more solutions that the designer has not anticipated. The advantage of this method is that it is generalizable, and it can be used even when possible success tactics are not known in advance. To identify success tactics, we extract the sequences of activities by encoding various dimensions, then perform cluster analysis to obtain the tactic clusters ('B' in Figure 6.2).

As explained in Section 2.6.2, cluster analysis is an unsupervised learning method that has been used successfully for the analysis of game behavioral data (Bauckhage et al., 2015; Drachen et al., 2009; Thurau et al., 2003). Hierarchical clustering is a method that looks for existing hierarchy of clusters in data (Aggarwal and Reddy, 2013; Jain and Dubes, 1988). This method can take a dissimilarity matrix of data pairs as input and identify clusters using a linkage method; the validity of a linkage method can be reflected by a high value of the cophenetic correlation coefficient.

A linkage method specifies how to measure the distance between clusters, of which some popular examples are: Ward, complete, single, centroid, and median (Rasmussen, 1992; Ward Jr, 1963). The resulting clusters can be shown as a dendrogram (a tree diagram with U-shaped lines showing how clusters merge or split into other clusters) (Fraley and Raftery, 1998). If the dendrogram generated by the assigned linkage method is valid, the cophenetic correlation coefficient would show a strong correlation. This coefficient is the correlation between the distance values obtained from the linkage method and the ones from the original dissimilarity matrix.

Additionally, the consistency of dendrogram links can show the quality of the clustering results. The consistency can be evaluated by comparison of the heights of each dendrogram link with the heights of neighboring links below it in a specific level of hierarchy (Zahn, 1971). The closer the heights are below a cluster, the smaller is the distance between its members, implying high *intracluster consistency*. Instead, more distinct division between the clusters leads to low *inter-cluster consistency*.

In our approach, in order to obtain the clusters of players' tactics: first, we determine dimensions to obtain gameplay sequences; then, we obtain a distance matrix from the dissimilarity between the sequences; finally, we apply hierarchical clustering. We explain the process in detail below.

Determining dimensions of gameplay sequences

To analyze the puzzle-solving processes of players and perform clustering, we encode the event data of players as strings with dimensional mappings.

An *attempt string* is the sequence of characters which represents one attempt of solving the game. For this purpose, each activity[5] is encoded as a character as follows: 'CLIMBER' (C), 'MINER' (M), 'PLATFORMER' (P), 'BLOCKER' (K), 'BATTER' (B), and 'JUMPER' ('J'). Note that, an activity can be a skill, a start or an end of an attempt, or the usage of various tools in the game. A *gameplay sequence* is the concatenation of the selected attempt strings of one player in chronological order. For more information on string selection, see the next section (Distance matrix based on Edit Distance).

In each of the dimensions defined below, we create the gameplay sequences, perform clustering, and benchmark the obtained clusters. The benchmarking is done against the best one matching our observation as well as some specific clustering metrics that are explained later in this section. We consider four dimensions which are used in creating the gameplay sequences incrementally:

1. Skills: only the skills used by players are considered in clustering. E.g., {CCMPK}.

2. Sessions: the start of each attempt is added to the gameplay sequences. The character '∧' is added at the beginning of every attempt string, e.g. {∧CCMPK}.

3. Feedback about success: the result of each attempt is added to the gameplay sequences. This is done in two alternative ways, feedback of success and failure and feedback in scale. In the first one, the character '+' (resp. '?') is added at the end of the attempt string if the result of the attempt was a success (resp. failure); e.g., {∧CCMPK+}. In the second one, four different characters are added, one to each attempt string; these characters represent success (100%), complete failure (0%), or two intermediate possibilities (1-50% and 51-99%).

[5]For detailed information on the meaning of each activity, see appendix B.3.

4. Time: we added the time to the gameplay sequences in seconds. Between two consecutive characters of a string, we inserted a character or a sequence of characters representing the time elapsed between two activities. This is done in two ways: the first is to insert '.' for each 5 seconds (or another fixed duration) elapsed; e.g., {C..C} means that 10 seconds were elapsed between two C activities. The second is to insert an individual character in a set of six characters which each encodes an interval of elapsed time, such as 10 to 19 seconds; e.g., {C,M;P} means that 10 to 19 seconds passed between C and M, and 20 to 29 seconds passed between M and P. Table 6.2 shows the mapping scheme of the time dimension in the second way.

Elapsed time in seconds	**Character**
5-9	.
10-19	,
20-29	;
30-39	:
40-59	/
≥ 60	%

Table 6.2: The mapping scheme of the time dimension. The elapsed time in seconds is added as a character to the gameplay sequences.

Encoding the sequences in the approach mentioned above and benchmarking the results against our expectation can be a way to deal with the practical difficulties caused by increases in dimensionality (Donoho, 2000). We aim for finding the important dimensions of the puzzle-solving process in an iterative way while avoiding the problems of high dimensionality. This approach can be advantageous in our case since we deal with a limited number of samples. Also, this approach can show the potential of iterative sequence mining with simple encoding.

Gameplay sequences are classified according to two more criteria: *Sequences with vs. without short attempt strings*: sometimes a player starts an attempt but does not perform many activities, and restarts immediately. We test clustering both with and excluding those attempts (less than 4 activities). *Sequences of all players vs. only successful ones*: we test whether the players who did not succeed have a bias effect over clustering the success tactics.

Distance matrix based on Edit Distance

For every gameplay sequence obtained from the previously explained process, we compute the distance matrix (the matrix of dissimilarities between sequences) based on Edit Distance. The Edit Distance (or Levenshtein distance (Levenshtein, 1966)) of two strings

is the minimum number of character substitutions, deletions, and insertions necessary for transforming of one string into the other. We compute this distance in two ways: first, we compute edit distance with no modification which gives an equal weight to all the edit operations, and second, we compute the distance by adding double weight for character substitutions. As a result, we can penalize less when two gameplay sequences differ by their length.

In order to avoid the bias from long sequences, we partition gameplay sequences of different lengths and evaluate which length gives an appropriate measure of distance. In this way, first, all of the attempt strings of a player are appended to form one single gameplay sequence. Second, only the last and most successful attempt string of each player is selected, and third, the last few attempt strings are appended so that sequences of equal length are obtained. Note that, in successful cases, we expect the players would elaborate their tactic during the last attempts, thus such gameplay sequences would contain less random behavior.

Hierarchical clustering of gameplay sequences

We obtain the clusters by grouping data points in a binary hierarchical cluster tree using agglomerative hierarchical clustering and evaluating over various linkage methods (Ward, complete, single, centroid, and median). Since our data points are strings of characters, the Edit Distance is used as a measure of dissimilarity between pairs of gameplay sequences. The best clustering result is when the clusters show a high cophenetic correlation coefficient, and high intracluster yet low inter-cluster consistency. Afterward, we visualize a dendrogram and detect the best natural groupings in clusters (by setting the maximum number of clusters). The results are explained in detail in Section 6.5.1. After the clusters are obtained, we try to discover the process models of tactics by applying PM. This method is explained in the next section.

6.4.2 Process Mining of Clusters

We employ PM methods (see Section 2.6.3) in combination with cluster analysis, in order to obtain process models that represent the puzzle-solving process of players belonging to each cluster ('C' in Figure 6.2). Such process models are valuable since they can show us the common components of each tactic cluster, by detecting the most significant activities and transitions performed by players. For this purpose, we apply fuzzy miner and the Disco tool (see Section 5.2.3), so as to identify a generalized and abstract model from the gameplay sequences. Fuzzy miner includes the most frequent order of activities and excludes the ones not significant in the puzzle-solving process. Therefore, the process model of a cluster yields a *cluster reference* (a reference sequence of activities) that is representative of the members of a cluster. After identifying the cluster reference, we validate the result by testing the convergence of players to the reference of their cluster.

This method is explained in detail in the next section. The process models of clusters are reported in Section 6.5.2.

6.4.3 Validation

After the discovery of clusters of tactics and their cluster reference through PM, we compare the game behavior with respect to these references. This is a validation method ('D' in Figure 6.2) for the clustering results, and helps us understand better the nature of clusters. This method will allow us to determine whether the obtained clusters are meaningful and reflect the different tactic choices of players.

We measure the distance of all of the attempt strings of players with the identified references and evaluate whether this distance decreases by the end of the game. We expect that the process of all the players, who are members a particular cluster, would converge to a small distance from their cluster reference. Also, those who do not converge to any successful reference would be unsuccessful. We consider a reference cluster as a successful reference when all the members of the cluster succeed in the game. The results of validation are reported in Section 6.5.3.

6.5 Results

6.5.1 Results from Cluster Analysis

As explained in Section 6.4.1, we performed agglomerative hierarchical clustering on the gameplay sequences obtained from assigned dimensions. We obtained several candidates for clusters and selected the candidate with high cophenetic correlation coefficient, high intracluster consistency, and low inter-cluster consistency. Two distinct clusters including all of the successful players were obtained with a relatively high cophenetic correlation coefficient ($c = 0.9162$), high intracluster consistency and inter-cluster division. Unsuccessful players were not all joined into one single cluster when all the players were considered in clustering.

In Figure 6.3, the dendrogram of clusters considering all players is shown. The maximum number of clusters is set to 3, as it could cut off the hierarchy into balanced natural groupings. The links between players are represented as upside-down U-shaped lines and the height of the U indicates the distance between the players' sequences calculated by linkage method of Ward, which performed better than other linkage methods, as showed a higher intracluster consistency.

We identify two clusters as *optimal successful* and *non-optimal successful* tactics, and the rest are categorized as *unsuccessful* players. Note that, the obtained clusters match well with our expectations and they are named accordingly. Our expectations were initially formed based on our observation notes and creation of the replays. In the

optimal successful tactic, the execution is straightforward and less complex while in the non-optimal tactic, more effort on the precision of execution is required to reach the goal. The dendrogram shows the labels of data points we assigned based on our observation (S: optimal successful, N: non-optimal successful, F: unsuccessful).

We also obtained the dendrogram of clusters considering only the successful players to avoid the effect of unsuccessful tactics on clustering the success tactics. The dendrogram in Figure 6.4 confirms the previous result and shows the same clusters.

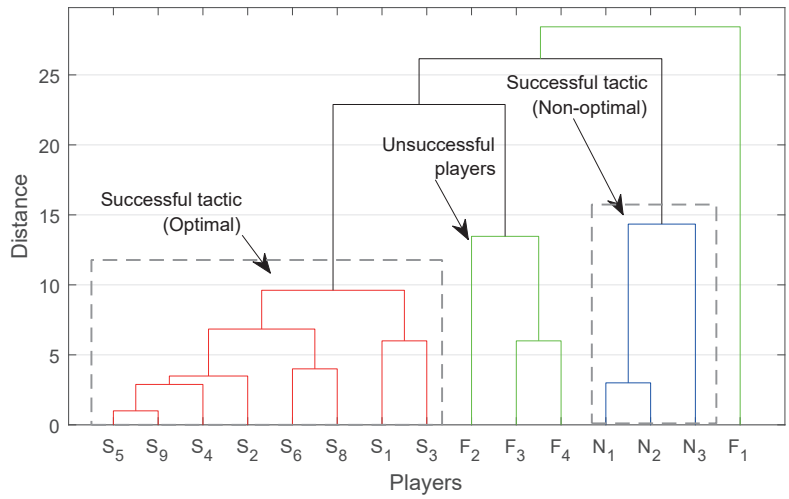


Figure 6.3: Dendrogram of agglomerative hierarchical clustering showing the clusters of tactics performed by all of the players at the Lix game. The labels along the horizontal axis represent the individual players (S: optimal successful tactic in red, N: non-optimal successful tactic in blue, F: unsuccessful players in green). The unsuccessful players are not all clustered, e.g., F_1 is separate from the rest of the unsuccessful players.

The evaluation from applying various dimensions shows that the feedback dimension performs better in clustering and corresponds better to our expectation. This holds for both feedback as success and failure and feedback in scale. The results from dimensions 'skills' and 'sessions' were the second best after feedback. Also, when elapsed time is added as duration with various characters, the results are better than when it is added as repetition of a single character. The edit distance with double weight led to a better division between the obtained clusters. Also, when we considered the last and most successful attempt of each player to measure the edit distance, the result corresponds to our observation the best. Using the last and most successful attempt of each player to measure the edit distance performs a better clustering because the bias towards the length decreased while the gameplay sequences containing the elaborate tactics were obtained.

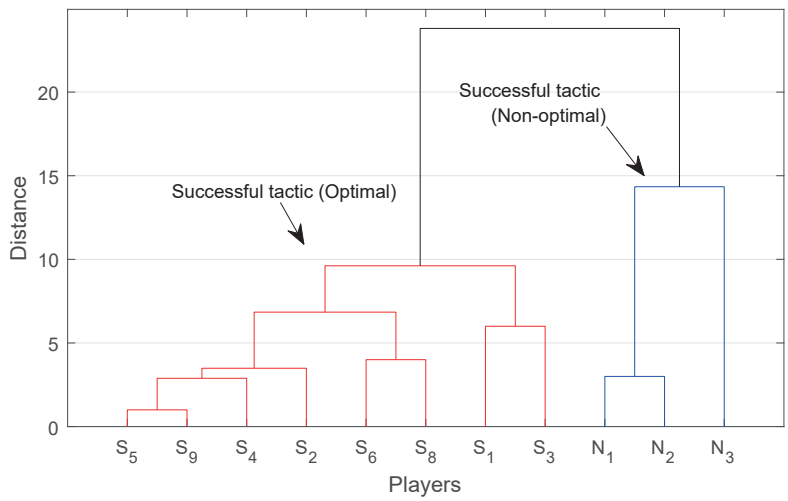


Figure 6.4: Dendrogram of agglomerative hierarchical clustering showing only the clusters of tactics performed by successful players during Lix game. The labels along the horizontal axis represent the individual players (S: optimal successful tactic in red, N: non-optimal successful tactic in blue).

In summary, during the iterative process of cluster analysis, our result is improved when these considerations are taken into account: the feedback dimension in scale, edit distance with double weight for substitutions, linkage method of Ward, when only successful players and last attempt strings are considered, and short attempt strings are discarded.

6.5.2 Results from Process Mining

After obtaining the two clusters of tactics, we performed PM by applying the fuzzy miner algorithm to obtain the process model of each tactic. We built the process models with the relevant events of the gameplay sequences of each cluster. Through the node and edge cutoff parameters in the Disco tool, we obtain an abstract yet informative process model, which can be represented as a reference of our tactic clusters. In Figure 6.5, simplified process models of the two successful tactics are shown. The intensity of colors in nodes and edges is proportional to the frequency of activities and transitions between them. The fuzzy miner algorithm reduced the models to the most significant activities and paths, by applying the node and edge cutoff parameters to 70% and 100% respectively (100% is the indicator of the maximum number of paths throughout the process). These parameters are set in order to show a simple and meaningful process structure.

Figure 6.5 shows the type of activities we observed from the attempt replays. The

process model on the left shows the optimal successful tactic, while the process model on the right shows the non-optimal successful tactic. A comparison between the two models shows the difference of tactics, in terms of the most frequent activities and transitions. E.g., in the optimal successful tactic the sequence {'CLIMBER','MINER'}, and in the non-optimal tactic the sequence {'MINER','PLATFORMER'}, play the most important role in the process.

Using the order and frequency of activities and transitions, we obtained the cluster references. The reference sequence for the optimal successful tactic is {∧CCKMCMBM+}, while the reference sequence for the non-optimal successful tactic sequence is {∧PPMPMP BM+}. Note that 'JUMPER' ('J') appears in the process model of the non-optimal successful tactic, but is filtered out from the reference sequence, since this activity was used to accelerate the end of the game and it was not part of the required tactic.

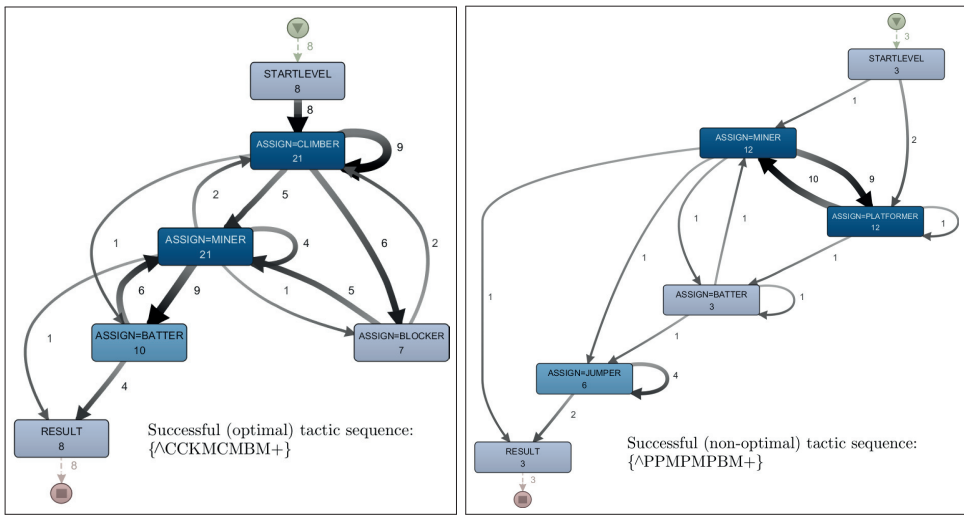


Figure 6.5: The process models of successful tactics showing the abstract and frequent activities of successful attempts of players. From left to right, the optimal successful and non-optimal successful tactics are shown by applying the node and edge cutoff parameters to 70% and 100% using Disco tool and fuzzy miner.

6.5.3 Validation of Results through Convergence

After obtaining the references of clusters, we evaluated whether the members of each cluster converge to their reference obtained by PM. Note that we did not consider a failure tactic for validation since we expect that those who do not converge to any of the success tactics would be unsuccessful. Also, the results of clustering showed that all of the unsuccessful sequences could not be aggregated into a single cluster. By convergence,

we mean that the distance of each member of a cluster with the cluster reference for the last several attempts decreases to a local minimum when finishing the game.

To show the convergence, we simply measured the edit distance (with double weight for substitutions) between each attempt string of a player (from the beginning to the end of the game) and the reference sequences. The distances are averaged over several consecutive attempts for those with a large number of attempts (> 15). This way, we can decrease the effect of bias of edit distance towards the length. Note that for overall convergence, we did not only consider the last successful attempts but all the attempts of the players in order to show their behavior with respect to each tactic throughout the whole game. We observed that the members (players) of each tactic cluster indeed converged to their cluster reference; however, the ones who did not succeed did not converge to any of the cluster references by the end of the game. In figure 6.6, we show three examples each from a cluster. From left to right, the examples are shown from optimal successful tactic, non-optimal successful tactic, and non-successful players.

Additionally, we performed the convergence by assigning the reference to the gameplay sequence of a single player and compared the result with our previous method (obtaining the reference from the process model of a cluster). The result shows that the other members of the same cluster did not converge to this reference since the gameplay sequence of a single player was not abstract and generalized enough.

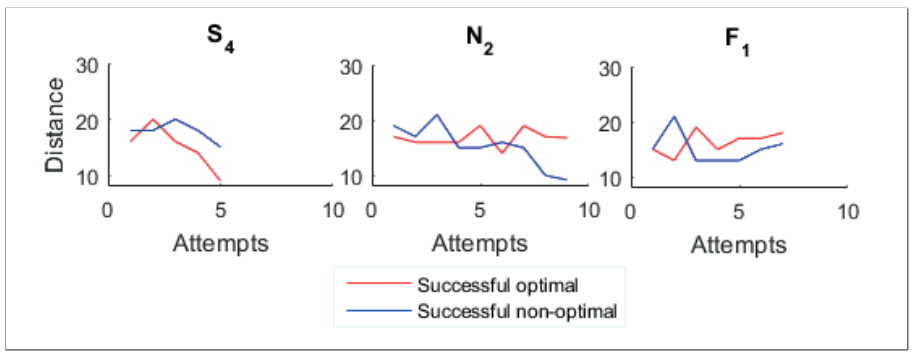


Figure 6.6: Examples of convergence from each cluster: S_4 from optimal successful tactic, N_2 from non-optimal successful tactic, and F_1 as an unsuccessful player.

6.6 Discussion

In this study, we presented a three-step analytics approach that uses clustering, process mining, and validation to extract the puzzle-solving tactics from data.

The clustering result matches our observations: players who succeeded could come up with one of the two success tactics to solve the puzzle during the game, but players

who did not succeed could not approach these tactics sufficiently and performed many random activities. There is indeed a higher distance between unsuccessful players than between the players in the other two clusters, as can be seen in Figure 6.3.

We validated our results by measuring how the elements of a tactic cluster converge to the cluster references. Our validation result shows that the clusters are meaningful and represent a tactic; indeed, all of the players in one cluster converged to their reference. Additionally, we benchmarked various dimensions in the process of cluster analysis and model discovery against our expectation. We show that encoding the sequences taking into account various dimensions can add value in sequence mining even with simple measures like Edit Distance. In every step, we analyzed the results of clustering with respect to our observation of replays, the value of cophenetic correlation coefficient, and the consistency of obtained dendrograms. Those verified through the discovery process are valuable sources to generate a generalized model of tactics discovery.

We can relate the results of the cluster analysis to the three stages of the puzzle-solving process (see Section 6.2 for the definition of stages) and as such, to the respective IBL phases. Clustering shows better results when considering only the last sequences of successful players. We assume that, at that point, players are already in the third IBL phase of 'conclusion and reflection' (*definition of a tactic* stage), in which the processes start to show convergence to a tactic. Conversely, in the first attempts, players are still in the earlier IBL phases (*task comprehension* stage), which are characterized by random exploration of the environment; consequently, those early attempts do not seem to converge to any established tactic. It is possible, however, that a player starts to define a tactic that the algorithm still does not know. In this case, the analytics approach should be able to incorporate this new tactic as a reference for future cluster analysis.

As explained in Section 3.2.2, clustering has been applied in several works to identify the playing styles (Charles and Black, 2004), for user testing of the game design (Drachen et al., 2009), to analyze the affective reactions of players (Amershi et al., 2006), or in combination with PM to improve the obtained educational process models (Bogarín et al., 2014). However, applying clustering in combination with PM to discover the puzzle-solving tactics of players is a new approach. The advantage of our approach is to automatically extract players' tactics as process models. These models represent the puzzle-solving behavior of the players of a cluster through detecting the most significant components of a tactic.

In the future, we will verify the results obtained from our approach by applying the same methodology to more data, in order to cross-validate the obtained process models. Additionally, we aim to extend our approach to other skill-based puzzle games, and use it as input to automatically recognize the different stages of puzzle-solving, and detect which players need assistance, prefer to have more challenges, or prefer to play a different game level. This approach could also be used as the basis for recommending interventions that could allow the game to provide the player/learner with help on time. An effective

intervention strategy would require different types of intervention for each stage. For example, during the *task comprehension* stage, a player would benefit most from help in discovering functionalities with tutorial-like hints. In the *definition of a tactic* stage, however, a player could be given hints when he or she seems to be heading to a dead-end. In the *precision of execution stage*, the game could instead offer tips on keyboard shortcuts and controls to help users improve execution of their chosen tactic.

6.7 Summary

In this chapter, we presented a novel approach to implement LA methods on interaction data in an open-source puzzle game called *Lix*. This game is used in our study because of the value of puzzle games as IBL environments, for educational purposes Becker (2005); Liu and Lin (2009) and, as such, developing ways to automatically analyze players' problem-solving processes can be a valuable tool for educators and game designers alike. In particular, we focused on the IBL phase of 'conclusion and reflection' where the players define and execute tactics to succeed the game. Note that the players' processes of puzzle-solving include the three phases of IBL. However, we only focused on the conclusion part where the players reach a puzzle-solving tactic and refine it.

In this work, we aimed to automatically identifying the tactics adopted in the game by players and the process models of these tactics in order to obtain a reference for the gameplay sequences of similar tactics. We presented a three-step analytics approach that uses clustering, process mining, and validation to extract the relevant puzzle-solving tactics from data and to determine the used tactic for individual players. The results obtained from our LA approach confirm our observations: two main success tactics were discovered through cluster analysis, and the unsuccessful behaviors did not all join into a single cluster. A possible explanation is that these players behaved more randomly and could not refine their tactic towards success. The process models of the success tactics were built, and yielded references for validation. Finally, the validation results indicate that the obtained clusters are relevant clusters of different tactics, as the members of each cluster converged to their cluster reference.

In the next chapter, we conclude the outcome of the three empirical studies, discuss, and indicate pointers for future work of this thesis.

Chapter 7

Discussion and Conclusions

In this thesis, we investigated the methods of Learning Analytics (LA) and Educational Data Mining (EDM) in the context of Inquiry-based Learning (IBL) and applied novel approaches of analytics methods to analyze and quantify the learning processes of students. LA and EDM are foundations for the development of personalized and adaptive educational tools enhancing the support and in-time help provided to the learners (Greller and Drachsler, 2012; Romero et al., 2010). They may also provide a sound basis for modifying the practices in education and training specially when the role of the learner is central. One of the pedagogical approaches that is effective by focusing on the role of the learner is IBL (De Jong et al., 2014). IBL consists of inter-related phases. In IBL environments, the learner is involved in active discovery of knowledge through exploration and experiments rather than reproductive approaches (Kruse and Pongsajapan, 2012). IBL learning processes can be analyzed through the application of LA and EDM.

In this chapter, first we provide a discussion of our general approach towards the application of LA and EDM to the IBL phases in Section 7.1. Then, we summarize our contributions with respect to related work, and discuss our proposed analytics methods for IBL phases in Section 7.2. In Section 7.3, we describe the limitations of this work. Finally, we present the future directions of this research in Section 7.4.

7.1 Proposed Framework for Applying the Analytics Methods

Here, we discuss the general framework of this thesis in the application of LA and EDM methods. We also discuss several issues related to the selection of appropriate analytics methods.

Our general approach towards applying LA and EDM for IBL

In this thesis, we presented a novel view over the application of LA and EDM for IBL. We described a phase-based approach to apply LA and EDM on the learner-centric IBL cycle. This approach, which is rooted in the works of Pedaste et al. (2015), Chatti et al. (2012), and Clow (2012), is described in Section 2.5.

The learner-centric IBL cycle (see Section 2.3) is comprised of three phases: generating hypothesis and question - investigation and discovery - conclusion and reflection. We applied this approach to three educational contexts where in each, we performed an empirical study focusing on an IBL phase. In individual learning contexts, that are usually different in the balance of IBL phases (Pedaste et al., 2015), we focused on a phase that is dominant in the learning process of students. For instance, in Chapter 4 we assumed that in Concept learning (CL), the weight of 'generating hypothesis and question' phase was more than the phase of 'conclusion and reflection'. Or in Chapter 5 on simulation-based learning, the proportion of the 'investigation and discovery' phase was higher than the other two phases.

Results of our approach: the results of our work through the application of our proposed approach are summarized as follows.

- The results of our study on the IBL phase of 'generating hypothesis and question' showed that the LA and EDM process increased our insight into the hypothesis generation of learners. The application of appropriate Machine Learning methods led to discovering the trends of learning by the size and domain of the learning concepts.

- Our results from studying the IBL phase of 'investigation and discovery' showed that appropriate LA and EDM methods can explain the properties of the learning behavior in the IBL cycle. In this context, applying Process Mining in combination with the complexity metric showed which activities were the most significant in the learning process, where the efforts of the students were concentrated, and where the students were stuck.

- By isolating the focus on the IBL phase of 'conclusion and reflection' in the LA and EDM process, the outcome of learning process could be investigated. In this context, application of cluster analysis on the data generated from this phase discovered the puzzle-solving tactics of learners. Moreover, the representatives of players' tactics were discovered through Process Mining.

Benefits of our approach: in our framework, the analytics methods are tuned to the specific features of each phase and as such, they can help balance the learning activities, tailor instruction, and enhance personalization, to meet the individual needs of learners in each phase. The advantages of this approach and its implementation can be explained as follows.

- *More focused, better analytics through tailoring the methods:* this approach helps focus on individual IBL phases in isolation and tailor the analytics methods to address the features of a particular phase and their context.

- *Balancing the learning activities:* viewing the learning traces and providing phase-based feedback can lead to balancing the learning activities between various phases, and bring the learning experience of all the phases to the right level.

- *More efficient use of teacher resources:* the implementation of this approach could help teachers understand better the learning phases of students and direct their time and effort to the parts that need more attention. For instance, teachers can detect if the students do not understand a task or they have difficulties with experimental tools.

- *Engaging learners through enhancing personalization:* the implementation of this approach could provide personalized and adapted assistance based on the learners' needs. Such focused feedback can engage better the students and help them to be more aware of the situation.

Choosing appropriate analytics methods

Quantifying learning processes and interpreting them are very challenging tasks, since learning processes are usually very complex and unordered (Bogarín et al., 2014). LA and EDM through adopting mixed method approaches such as the combination of Data Mining (DM) and Machine Learning (ML) with qualitative methods try to overcome this challenge (Chatti et al., 2012). Furthermore, choosing the appropriate methods needs expertise in DM methods as well as a good understanding about the educational contexts, pedagogical framework of the educational settings, and the knowledge domains. Therefore, the methods that are used in various fields might not satisfy the needs of educational stakeholders and need to be tuned for the specific contexts and pedagogical approaches.

In this thesis, we presented a collection of contributions related to the application of LA and EDM in educational contexts where IBL cycle is implemented. We exploited and chose existing ML, PM, and DM methods, introduced novel applications of these methods, and combined them with qualitative approaches, to study the three IBL phases. In particular, we adopted metrics (such as CM, HRC, and HAS) that are normally used to study the properties of systems, algorithms, or programs, to quantify complex human learning processes.

7.2 Proposed Analytics Methods for IBL phases

In this section, we summarize our contributions and show how our proposed analytics methods are tuned to the particular IBL phases.

Application of Machine Learning in the First IBL Phase

In our first empirical study, we focused on the first phase of the IBL cycle, 'hypothesis generation'. We investigated this IBL phase in the context of human Concept Learning (CL) through the application of Machine Learning (ML). ML is a fast evolving field of research that studies how algorithms learn from data (Bishop, 2006). However in the CL context, ML is now starting to be able to model simple human processes and the long term vision is to imitate more and more complex human activities (Mar et al., 2015).

Our work, through applying ML, is one way to progress in the direction of theoretically characterizing the human ability of learning. We applied ML in our study to humans as if they were algorithms that learn from data (Vapnik, 1998). In particular, we exploited Rademacher Complexity and Algorithmic Stability that are two powerful ML methods for measuring the ability of algorithms to learn.

Human Rademacher Complexity (HRC) and Human Algorithmic Stability (HAS) both increased our insight into 'hypothesis generation' phase of IBL in different ways through presenting the trends of learning by the size and domain of the concepts to learn. We investigated this phase of IBL in three domains (Shape, Word, and Math) which were not referring to any of the study subjects of students. However, HRC and HAS can be tuned through the experimental design to study all IBL phases in a variety of subjects. For instance, these methods can be applied to the subjects of a biology course and study how students hypothesize about the categories of living organisms, and how they test or reflect on their formed hypotheses.

HRC and HAS are not only flexible enough to be applied to many study subjects, but also they can be used directly as educational components for the students of ML and Data Mining. One of the fundamental topics in ML is classification. Humans perform the classification tasks naturally without analyzing the rules behind the category labels. Therefore, it can be challenging for the students to intuitively learn how classification systems work. Our experiments can provide a tangible view on how learning systems actually behave by letting the students experience the same process.

Application of Process Mining in the Second IBL Phase

To investigate further the IBL cycle of learners, we proceed with a second empirical study that relied on the phase of 'investigation and discovery' in the IBL cycle. We focused on simulation-based learning since simulators play an important role in enabling learners to perform experiments in IBL contexts (Van Joolingen and De Jong, 2003).

As explained in Section 3.3.2, LA and EDM have been applied in the context of simulation-based learning mainly to monitor the progress of students (Gillet et al., 2013; Govaerts et al., 2013). In our study, we proposed a different analytics approach that focuses on the analysis of learning processes of students, interacting with the simulation environment and performing IBL tasks. We exploited Process Mining (PM) methods to discover and investigate the learning processes of students, then we compared the processes of the most successful and the least successful students. For this purpose, we adopted a complexity metric called Cyclomatic Complexity, which was originally applied to measure the understandability of source code and later process models. Using a complexity metric in combination with PM to quantify and compare the learning processes is a new approach compared to many studies in the field that have used the classical PM methods (examples are described in Section 3.2.3). As an example, Bannert et al. (2014) analyzed and compared student processes of self-regulated learning through the methods of process discovery and conformance checking.

Our study showed that PM can help explain many properties of the learning behavior of students and the measured complexity can be used as an indicator of academic success of students. PM and CM can be applied to any learning context where temporal data of learners while interacting with a learning tool can be recorded. These methods are particularly useful when the learning processes are complex and unordered. The learning processes appear to be complex when the learners act naturally in the learning environment and their behavior is varied. This situation can be observed in many informal learning contexts where the learners are not obliged to follow strict rules. In this context, PM can abstract the learning processes to obtain the most significant components of the learning behavior. And CM can be used as a summative quantity for the learning processes to be tested for relevance with other variables such as grade, motivation, etc. In this context, the most important resources of a topic, the most time-consuming activities, or the least engaged students can be detected.

Application of Data Mining in the Third IBL Phase

In our third empirical study, we focused on the third phase of the IBL cycle, 'conclusion and reflection'. We focused on game-based puzzle-solving since games are effective learning tools, proved to enhance players' performance in a wide variety of problem-solving tasks (Shute et al., 2015).

LA and EDM have been used to improve game quality and to support the achievement of learning goals. For instance, clustering methods are applied to analyze the behavior of players and to adapt the educational games to the needs of players (Amershi et al., 2006; Drachen et al., 2009). In our study, we investigated the use of LA and EDM in a different way in the context of digital puzzle games, which are commonly used for educational purposes (Connolly et al., 2012). Our approach was to explore the way players learn game skills and solve problems through combining the clustering methods with PM. This

analytics approach allowed us to discover representative process models of tactics from the interaction data of players.

Our proposed approach can be applied to any puzzle-based game environment in which problem-solving happens over repeated attempts. It can be used to automatically extract, from multiple players' interactions with the game, the different tactics that they use when solving the puzzle. These puzzle-solving tactics can be used in several applications, such as monitoring the player's progress, providing feedback, and displaying timely help.

Moreover, this approach can be used as a basis for recommending interventions for different learning tools as far as the learner needs to define and execute a tactic to succeed. For instance, in an interactive platform that teaches coding, there can be different algorithms to a coding challenge. In this case, with the process models of correct algorithms (successful tactics) as successful references of behavior, it would be possible to detect if a learner is converging to or diverging from this reference model. In this way, hints can be generated at the moment a learner is getting far from the solution.

7.3 Limitations

The results of our empirical studies are characterized by and limited to the data collection process, the addressed IBL context, the settings of the experimental design, and the targeted population. Extension of these studies over varied population and educational contexts can help improve the generalizability of our results. The limitations of this work are explained in detail as follows.

Data collection process

We could not investigate all of the IBL phases in a single context, since the data collection was not always possible from all phases. Moreover, it was difficult to distinguish the different IBL phases from observing the learners' behavior or from their data, since in some cases, the learners went back and forth from one phase to another in an unordered manner.

For instance, in Section 5.4 in the context of simulation-based learning, the data from the phase of 'generating hypothesis and question' was not available, as we could not intervene and change the instruction for our experiment. Later, this study can be extended to collect and analyze the learning processes of the three IBL phases. For instance, it would be possible to ask the students what they hypothesize before they perform the experiments. Or in Section 4.4, the results of our work are limited the first phase of IBL, and needs to be extended and investigated further on other IBL phases to reach a generalized statement.

Furthermore, the instruction style of the course where the experiments are performed can have an effect over the data collection process. For instance, in Section 5.4, the

order of the tasks given to the students and the workload of exercises per session could influence the data collection process. When the exercises of a session were easy, most of the students completed the tasks on time and their data was fully collected. While for a difficult session, the students might have continued working after the session was ended, so their data was not fully collected.

In all of our empirical studies, there were limitations regarding the privacy and anonymity for sharing our data sets. Indeed, the contextual data streams could not be shared in full details with the research community. This is a drawback, since our results cannot be completely compared with the related works which use our published data set and therefore, it can influence validity of our results.

Experimental settings

The experiments of this work were performed in two types of settings. The first type is when we performed the experiments as interventions to the educational settings (they were not part of the course instruction). The second type is when the experiments are performed in the real educational setting. Each of these settings has advantages, but drawbacks as well.

For the first type, it is necessary to pay particular attention to control variables and make sure that the results are not biased. Also, the results need to be considered along with the specifications of the experimental design. For instance, in Section 4.4 we did the experiments as interventions. We tried to minimize the bias of intervention through randomization of subjects and the tasks given to the students. For instance, every individual was given a unique set of tasks through the random generation of task content and size. This method was quite time-consuming and required a complex procedure for the experimental design and data analysis.

For the second type, the experiments do not affect the process of learning however, it is not possible to control all variables completely. For instance, in Section 5.4, the study is performed in the real educational setting where the students were using a simulator for learning, and not as an intervention. We collected the students' data as much as possible, without any interventions or controls over the investigation and discovery process of students. Therefore, the bias caused by our experimental design was minimal.

However, we were limited in controlling the experiments since it was not possible to change the regular learning conditions of students. Therefore, the data collection process was challenging and prone to noise. We tried to minimize the information loss by improving our experiments and seting time limits to the course sessions. However, as we were not allowed to intervene the learning process, we could have lost some information of the learning behavior. For instance, students could continue learning at any time and location and not only during the specific sessions of the course. Or, they could meet and practice together before the final exam. Also, the learning method was not limited to the interaction with the simulator. The students were allowed to discuss with their peers and

learn by observing how others interact with the simulator. An extension of this study can be helpful to verify how the results are affected by these limitations.

Population

The number of participants in a study have an influence over the results. When dealing with a limited number of participants, we tried to overcome this limitation partly through the experimental design. For instance, in Section 6.3, we had a limited number of participants. However, we tried to minimize the bias of this limitation by increasing the time of the experiment. The participants were free to play the game as many time as they wanted with no time limits. Therefore, we could collect sufficient data to address the raised research question.

Also, the type of perticipants and their habits can have an effect on the results. In Section 6.3, most of the participants were students with learning habits on a regular basis. Or in Section 4.4, the participants included specific groups of learners (engineering students of University of Genoa). An extension of these studies over varied population can help generalize the results.

The motivation of learners can play an important role in their learning behavior. For instance, in Section 6.3, some participants could be very interested in playing games and enjoyed being confronted with a challenge. In this case, they tried their best in finding the solution. While for some other players, the game might have been very boring. Some players could be accustomed to play video games in general or specifically puzzle games. All of these variables might have influenced the way the players solved the puzzle. Moreover, the mood or time constraints of participants could have an effect on their behavior. For instance, a participant who had a hectic day could be less persistent in finding the solution.

7.4 Future Work

There are many related questions left in the context of LA, EDM, and IBL research to be addressed in the future. We hope that our work serves as a starting point for the researchers of LA and EDM who aim to focus on the IBL cycle of learners through the novel methods of analytics. Our work can be improved by generalization of our framework, further implementation of analytics methods, as well as implementation of learning methods. We describe the perspectives of our research as follows.

Generalization of our framework

Our approach, towards applying LA and EDM for IBL, can be related to any study where similar learner-centric IBL phases are implemented. Through improving the generalizability and interoperability of our approach, it would be possible to integrate the

proposed methods of the empirical studies in any IBL context. And through a comprehensive framework, that offers flexible experimental design and data collection techniques, it would be possible to tune the methods and apply them to all of the phases of an IBL context. For instance, in some contexts, the phase of 'generating hypothesis and question' might be implicit and cannot be recorded by a logging system. Through proper experiments, it is possible to ask the students explicitly about their hypotheses, and exploit the entire IBL cycle. This way, the relations between the phases and how the transitions occur from phase to phase can be investigated.

Implementation of analytics methods

The analytics approaches we presented here are initial efforts in the context of IBL. In order to implement these methods to the entirety of IBL and in a wider context, they need to be standardized. As explained in Section 3.2, standardization of methods and data enhances the mobility and interoperability and allows to re-use tools and resources from one context to another (Chatti et al., 2012).

Also, the application of methods we proposed requires the expertise in Machine Learning and Data Mining. It is necessary to develop user-friendly tools that implement these methods and provide feedback on the learning processes to be directly useful for stakeholders. For instance, interactive dashboards and visualizations can be a way to let the stakeholders explore and understand the output of complex analytics methods.

For all of the empirical studies presented in this thesis, further research is necessary to integrate recommendation systems (which implement our proposes analytics methods) into the respective learning tools for the purpose of improving the learning outcome. Through correct implementation of our proposed methods, personalization and adaptivity of educational environments can be improved, and in-time feedback can be generated to the stakeholders to raise their awareness about the learning processes.

Implementation of learning methods

Our studies are performed with more focus on the technology and analytics side than on the pedagogy and learning theories. Indeed, this thesis does not include the variety of learning theories and pedagogical frameworks applicable to IBL and beyond. It is worthwhile to test various theories of learning through comparative applications of LA and EDM, and to improve and incorporate them into the means of instruction effectively.

In this thesis, we focus on the IBL cycle of learners, however, we only investigate a part of the cycle per study. For future research, we propose to study all phases of IBL which requires the improvement of experimental design and data collection techniques, and choosing a proper context. For instance in the first empirical study, we showed the potentials of two effective ML tools for understanding the hypothesis generation of students in CL. This study can be extended by exploiting the entire IBL cycle of CL

through ML methods, for instance, by modifying the experiments (described in Section 4.4) as follows.

During the experiments, we presented a set of labeled instances to the learners to grasp the underlying rule (for the case of HAS) or to come up with a rule (for the case of HRC). Then, a set of new stimuli was given to students to be labeled. For an extension of this study, a feedback on labels can be presented to learners in an intermediate step, to investigate how consistent the learners are to their previous hypothesis. As a result, it would be possible to study the phase of hypothesis testing and reflection in human CL, and examine, analogously to algorithms, the effect of feedback over learning while being updated over time.

Conclusion

This thesis is one of the steps forward in the application of novel LA and EDM methods and approaches. Our work can open the way to more deep insights in understanding the learning ability of humans and their learning processes. In this line of research, future directions can boost the implementation of LA and EDM in real educational settings to support decision-making of stakeholders, and to enhance the learning experiences of students.

Appendix A

EPM Data Set

A.1 Data Set Description

Our data set[1] contains time-ordered series of activities performed by students during six sessions of laboratory sessions of the course of digital design. There are 6 folders containing the data of students per session. Each 'session' folder contains up to 99 CSV files each dedicated to a specific student log during that session. The number of files in each folder changes due to the number of students present in each session. Each file contains 13 features. Due to ethical reasons and to ensure the anonymity of our users, we did not share the original log files, instead, we shared the data transformed and cleaned in the appropriate format for process mining. Data is presented in separate folders and files as follows:

- 'processes': contains the data files from session 1 to 6.
- 'logs': shows information about the log data per student Id. It shows whether a student has a log in each session.
- 'grades': two files containing the results of the final exam and the grades for assignments of students per session.
- 'exam': two files containing the content of final exam (original in Italian and its translation in English).

The data set metadata includes separate files explaining the attributes. For example, 'features.txt' contains the list of all features, and 'features_info.txt' contains information about the variables used on the feature vector. In the similar way, 'activities_info.txt' contains information about the variable 'activity'. We also include the description of exercises given to the students during each session and the information about the grade data.

[1]The data set on UCI can be retrieved from:
https://goo.gl/Xo0TCq

```
1, 1, Es_1_1, Deeds_Es_1_1, 2.10.2014 11:41:18, 2.10.2014 11:41:29, 421, 0, 0, 9, 0, 879, 0
1, 1, Es_1_1, Diagram, 2.10.2014 11:41:30, 2.10.2014 11:44:46, 4016006, 0, 0, 71, 0, 6904, 0
1, 1, Es_1_1, Deeds_Es_1_1, 2.10.2014 11:44:47, 2.10.2014 11:45:0, 19071, 0, 0, 2, 0, 466, 0
1, 1, Es_1_1, Diagram, 2.10.2014 11:45:1, 2.10.2014 11:49:23, 20223390, 0, 0, 60, 6, 6005, 0
1, 1, Es_1_1, Properties, 2.10.2014 11:49:24, 2.10.2014 11:49:30, 688, 0, 0, 3, 0, 187, 0
1, 1, Es_1_1, Diagram, 2.10.2014 11:49:31, 2.10.2014 11:50:13, 41107, 0, 0, 37, 4, 2550, 0
1, 1, Es_1_1, Properties, 2.10.2014 11:50:14, 2.10.2014 11:50:20, 327, 0, 0, 4, 0, 372, 0
```

Figure A.1: An extract of data set from the data of the first session, presenting the data of a student working on an exercise.

The features selected for this data set come from pre-processing of data collected through LADC. The original logs contain the logging data of client system per approximately a second, while the features are calculated in order to be allocated to a particular activity. In the following sections, we explain further the features of the data set and the detailed descriptions of activities.

A.2 Features

Each CSV file belongs to a specific session and a specific student (named by an anonymous student_Id) and each file contains several exercises of that session presented in 'exercise' feature. Each 'exercise' contains activities, which start-time, end-time, and other features are allocated to that. Figure A.1 shows an extract of data set for one session and one student and the features are shown in the following order:

1) session
2) student_Id
3) exercise
4) activity
5) start_time
6) end_time
7) idle_time
8) mouse_wheel
9) mouse_wheel_click
10) mouse_click_left
11) mouse_click_right
12) mouse_movement
13) keystroke

The first attribute 'session' shows the number of the laboratory session from 1 to 6. 'student_Id' is the Id of student from 1 to 115. Then 'exercise' shows the Id of the exercise the student is working on. Each session contains 4 to 6 exercises, shown as 'Es_# of the session_# of the exercise' ('Es' with no number means the student has not started the exercise yet, and by # we mean the number of the exercise). The 'activity' is labeled based on the title of web pages that are on focus or in the view of the student. 'start_time' and 'end_time' show the start end date and time of a specific activity. And 'idle_time' is the duration of idle time between the start and end time of an activity in milliseconds. The rest of the attributes show the amount of mouse clicks, movements, and keystrokes during a particular activity. To ensure anonymity, we did not publish the

exact name of visited pages by the students thus renamed and augmented the pages into 'activity' names.

A.3 Grades Data

The grades data contain:

- Final grades obtained by the final exam of the course.
- Intermediate grades obtained from the evaluation of students' assignments per session (exercises).

Description of grades data are as follows.

Final grades: The grades are given per student Id. Some students who attended the course did not take the final exam[2], therefore, some Ids are missing in final grades. The exam was held in two times and some students took the exam two times. For those students, we considered their first attempt. The questions of the final exam addressed the concepts of sessions of the course. So, we considered the grades per question based on their reference to the sessions topics in addition to the total final grade.

Intermediate grades: The grades given to the students' assignments from session 2 to 6. Note that there is no grade for Session 1. The students were required to work during each laboratory session and submit their work afterward. The intermediate grades were given after reviewing their works. Note that the students were free to discuss and ask for help during the sessions in order to complete their assignments.

A.4 Exercises

The content of exercises for the laboratory sessions can be retrieved from the official Deeds website[3]. To compare the content of the exercise with exercise feature values, consider the code assigned to it on the website. Each code contains a zip file containing the files to be used in a text editor as well as the format to be used in the Deeds simulator:

[2]The content of the final exam can be retrieved from:
https://goo.gl/Xo0TCq

[3]The content of the exercises can be retrieved from:
http://www.esng.dibe.unige.it/deeds/LearningMaterials/IndexByTopic.htm

Es_1_1 = "001002"	Es_3_1 = "020045"	Es_5_1 = "030180"
Es_1_2 = "005030"	Es_3_2 = "020120"	Es_5_2 = "035200"
Es_1_3 = "005040"	Es_3_3 = "020055"	Es_5_3 = "035210"
Es_1_4 = "005050"	Es_3_4 = "025130"	Es_5_4 = "035230"
Es_2_1 = "015090"	Es_4_1 = "030140"	Es_6_1 = "045270"
Es_2_2 = "015095"	Es_4_2 = "030144"	Es_6_2 = "045280"
Es_2_3 = "015065"	Es_4_3 = "030160"	Es_6_3 = "045290"
Es_2_4 = "015100"	Es_4_4 = "030164"	Es_6_4 = "050300"
Es_2_5 = "015070"	Es_4_5 = "035220"	Es_6_5 = "050310"
Es_2_6 = "015080"		Es_6_6 = "050425"

Appendix B

Lix Data Set

B.1 Data Set Description

Our data set[1] contains the players' time series of activities during gameplay. There are 15 csv files containing the players' data per gameplay. Each file contains 11 features.

We asked the players to play a short minigame of a game called *Lix*[2]. It is a puzzle game, based on *Lemmings*, which was a game released in 1990. The name of the minigame / level is 'Put Your Lix on Ice'.

Data is presented in separate folders and files as follows:

- 'Data' : contains 15 csv files. Each is dedicated to a player.
- 'features_info.txt': contains information about the variables used in the feature vector.
- 'features.txt': list of all features.
- 'actions_info.txt': contains information about the variable 'activity'.

B.2 Features

Each CSV file belongs to a specific player. The features are shown in the following order:

[1]The data set can be retrieved from:
https://www.la.smartlab.ws

[2]Lix game is available here:
https://github.com/SimonN/Lix

1) player
2) timestamp
3) action
4) level
5) update
6) which_lix
7) lix_required
8) lix_saved
9) seconds_used
10) seconds_required
11) seconds_used

The first attribute 'player' shows the Id of the player from 1 to 15. 'timestamp' shows the time of a specific action with the format: yyyy-mm-ddThh:mm:ssZ. 'action' is labelled based on the type of activities the players perform during the game. It describes the skills assigned to a Lix (a game character) and some other activities in the game. The information about actions is provided in 'actions-info.txt'. 'level' shows which minigame of the Lix game is being played. Lix game consists of many minigames. 'update' indicates the internal measure of game time. 'which_lix' shows an internal identifier of the Lix to which the skill was assigned. 'lix_required' shows the number of Lix required to be saved in a specific level. 'lix_saved' shows the number of Lix saved in a try (one time playing the level). 'skills_used' shows the number of skills used in a try. 'seconds_required' shows the time required to reach the goal of the game (in case there is a time limit). 'seconds_used' shows the time spent in a try excluding the time if the game is paused.

B.3 Actions

The details about the actions (some examples) and their meanings are as follows:

- STARTGAME : start of the game.
- STARTLEVEL : start of the level.
- ASSIGN=JUMPER / PLATFORMER / BLOCKER / EXPLODER / etc. : examples of the skills assigned to a Lix
- ASSIGN_LEFT=MINER / ASSIGN_RIGHT=MINER : the player assigns the skills to the left or right of the Lix.
- EMPTYSKILL: nothing is assigned to a Lix while clicked on it.
- FASTSPEED / NORMALSPEED / TURBOSPEED : the player changes the speed of Lixes.
- PAUSE and RESUME: the game is paused and resumed.
- NUKE: the player forces the end of the level.

- PLUS1FRAME / MINUS1FRAME : Lixes are moved ahead or back frame by frame.
- ZOOM : zoom in/out the interface.
- RESULT : the result when a level ends (one try).
- ENDLEVEL: end of the level.
- ENDGAME: exit from the game.

B.4 Gameplay

In this minigame, the player has to guide a group of Lixes through some obstacles to a designated exit. To save the required number of Lixes to solve the puzzle, the player must determine how to assign a limited number of skills to specific Lixes. These skills allow the selected Lix to alter the landscape, to affect the behaviour of other Lixes, or to clear obstacles in order to create a safe passage for the rest of the Lixes.

The player is given only one minigame to play. The goal is to solve the puzzle in the most efficient way - that is, the one that uses the least amount of resources, saves as many Lixes as possible, and takes as little game time as possible. The player can repeat the level as many times as he/ she needs. The goal is not to solve it faster, however, to find the best solution. When the player is satisfied with the tactic he / she reaches, we asked to play it one more time, so that the player has the chance to shorten the game time.

Learning the game controls is also part of this experiment. For this reason, once the minigame starts, we did not help the player, not even with questions about the game user interface. The player is also not allowed to look up for solutions online or to search for hints. The player is asked to explore the game to figure out how to solve the puzzle.

Bibliography

Aggarwal, C. C. and Reddy, C. K. (2013). *Data clustering: algorithms and applications*. Chapman & Hall/CRC.

Amershi, S., Conati, C., and McLaren, H. (2006). Using feature selection and unsupervised clustering to identify affective expressions in educational games. In *Motivational and Affective Issues in ITS*.

Anguita, D., Ghio, A., Oneto, L., and Ridella, S. (2011a). Maximal discrepancy vs. rademacher complexity for error estimation. In *Artificial Neural Networks, Computational Intelligence and Machine Learning*.

Anguita, D., Ghio, A., Oneto, L., and Ridella, S. (2012). In-sample and out-of-sample model selection and error estimation for support vector machines. *IEEE Transactions on Neural Networks and Learning Systems*, 23(9):1390–1406.

Anguita, D., Ghio, A., and Ridella, S. (2011b). Maximal discrepancy for support vector machines. *Neurocomputing*, 74(9):1436–1443.

Arnold, K. E. and Pistilli, M. D. (2012). Course signals at purdue: using learning analytics to increase student success. In *Learning Analytics and Knowledge*.

Baalsrud Hauge, J., Berta, R., Fiucci, G., Fernandez Manjon, B., Padrón-Nápoles, C., Westra, W., and Nadolski, R. (2014). Implications of learning analytics for serious game design. In *ICALT*.

Baker, M. J. (2000). The roles of models in artificial intelligence and education research: a prospective view. *Journal of Artificial Intelligence and Education*, 11:122–143.

Baker, R. and Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining*, 1(1):3–17.

Baker, R. S. and Inventado, P. S. (2014). Educational data mining and learning analytics. In *Learning Analytics: From Research to Practice*.

Bakkes, S. C. J., Spronck, P. H. M., and van Lankveld, G. (2012). Player behavioural modelling for video games. *Entertainment Computing*, 3(3):71–79.

Bannert, M., Reimann, P., and Sonnenberg, C. (2014). Process mining techniques for analysing patterns and strategies in students self-regulated learning. *Metacognition and Learning*, 9(2):161–185.

Bartlett, P. L., Boucheron, S., and Lugosi, G. (2002). Model selection and error estimation. *Machine Learning*, 48(1-3):85–113.

Bartlett, P. L., Bousquet, O., and Mendelson, S. (2005). Local rademacher complexities. *Annals of Statistics*, 33(4):1497–1537.

Bartlett, P. L. and Mendelson, S. (2002). Rademacher and gaussian complexities: Risk bounds and structural results. *The Journal of Machine Learning Research*, 3:463–482.

Bauckhage, C., Drachen, A., and Sifa, R. (2015). Clustering game behavior data. *IEEE Transactions on Computational Intelligence and AI in Games*, 7(3):266–278.

Becker, K. (2005). How are games educational? learning theories embodied in games. In *DiGRA: Changing Views – Worlds in Play*.

Bellotti, F., Berta, R., and De Gloria, A. (2010). Designing effective serious games: Opportunities and challenges for research. *International Journal of Emerging Technologies in Learning*, 5:22–35.

Bienkowski, M., Feng, M., and Means, B. (2012). Enhancing teaching and learning through educational data mining and learning analytics: An issue brief. In *US Department of Education, Office of Educational Technology*.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Bogarín, A., Romero, C., Cerezo, R., and Sánchez-Santillán, M. (2014). Clustering for improving educational process mining. In *Learning Analytics And Knowledge*.

Bohannon, J. (2010). Game-miners grapple with massive data. *Science*, 330(6000):30–31.

Bottino, R. M., Ferlino, L., Ott, M., and Tavella, M. (2007). Developing strategic and reasoning abilities with computer games at primary school level. *Computers & Education*, 49(4):1272–1286.

Bousbia, N. and Belamri, I. (2014). Which contribution does edm provide to computer–based learning environments? In *Educational Data Mining*.

Bousquet, O. and Elisseeff, A. (2002). Stability and generalization. *The Journal of Machine Learning Research*, 2:499–526.

Boyle, E., Connolly, T. M., and Hainey, T. (2011). The role of psychology in understanding the impact of computer games. *Entertainment Computing*, 2(2):69–74.

Brooks, C. (2013). *A data-assisted approach to supporting instructional interventions in technology enhanced learning environments*. PhD thesis, University of Saskatchewan, Saskatoon, Canada.

Brown, M. (2012). Learning analytics: moving from concept to practice. *Educause Learning Initiative Brief*, pages 1 – 5.

Bruner, J. S. and Austin, G. A. (1986). *A study of thinking*. Transaction publishers.

Brusilovsky, P., Sosnovsky, S., and Shcherbinina, O. (2005). User modeling in a distributed e-learning architecture. In *User Modeling*.

Buckingham Shum, S. (2012). Learning analytics. UNESCO Policy Brief.

Bybee, R. W., Taylor, J. A., Gardner, A., Van Scotter, P., Powell, J. C., Westbrook, A., and Landes, N. (2006). The bscs 5e instructional model: Origins and effectiveness. *Colorado Springs, CO: BSCS*, 5:88–98.

Carvalho, M. B. (2015). Adaptivelix. `https://github.com/carvalhomb/AdaptiveLix`.

Casella, G. and Berger, R. L. (2002). *Statistical inference*. Duxbury.

Castro, F., Vellido, A., Nebot, A., and Mugica, F. (2007). Applying data mining techniques to e-learning problems. *Evolution of Teaching and Learning Paradigms in Intelligent Environment*, 62:183–221.

Castro, R. M., Kalish, C., Nowak, R., Qian, R., Rogers, T., and Zhu, X. (2009). Human active learning. In *Advances in Neural Information Processing Systems*.

Charles, D. and Black, M. (2004). Dynamic player modeling: A framework for player-centered digital games. In *Computer Games: Artificial Intelligence, Design and Education*.

Chater, N. and Vitányi, P. (2003). Simplicity: A unifying principle in cognitive science? *Trends in Cognitive Sciences*, 7(1):19–22.

Chatti, M. A., Dyckhoff, A. L., Schroeder, U., and Thüs, H. (2012). A reference model for learning analytics. *International Journal of Technology Enhanced Learning*, 4(5):318–331.

Clow, D. (2012). The learning analytics cycle: Closing the loop effectively. In *Learning Analytics and Knowledge*.

Connolly, T. M., Boyle, E. A., MacArthur, E., Hainey, T., and Boyle, J. M. (2012). A systematic literature review of empirical evidence on computer games and serious games. *Computers & Education*, 59(2):661–686.

D'Angelo, C., Rutstein, D., Harris, C., Bernard, R., Borokhovski, E., and Haertel, G. (2014). Simulations for stem learning: Systematic review and meta-analysis. *SRI International*.

De Jong, T., Sotiriou, S., and Gillet, D. (2014). Innovations in stem education: the go-lab federation of online labs. *Smart Learning Environments*, 1(1):1–16.

De Jong, T. and Van Joolingen, W. R. (1998). Scientific discovery learning with computer simulations of conceptual domains. *Review of Educational Research*, 68(2):179–201.

De Medeiros, A. A., Weijters, A., and Van der Aalst, W. M. P. (2006). Genetic process mining: a basic approach and its challenges. In *Business Process Management Workshops*.

Deák, G. O., Bartlett, M. S., and Jebara, T. (2007). New trends in cognitive science: Integrative approaches to learning and development. *Neurocomputing*, 70(13):2139–2147.

Del Blanco, A., Serrano, A., Freire, M., and Martínez-Ortiz, I.and Fernández-Manjón, B. (2013). E-learning standards and learning analytics. can data collection be improved by using standard data models? In *IEEE EDUCON*.

Devroye, L., Györfi, L., and Lugosi, G. (1996). *A probabilistic theory of pattern recognition*. Springer.

Dietrich, R., Opper, M., and Sompolinsky, H. (1999). Statistical mechanics of support vector networks. *Physical Review Letters*, 82(14):2975.

Donoho, D. L. (2000). High-dimensional data analysis: The curses and blessings of dimensionality. *AMS Math Challenges Lecture*, pages 1–32.

Donzellini, G. and Ponta, D. (2003). Deeds: an e-learning environment for digital design. In *EUNITE*.

Donzellini, G. and Ponta, D. (2007). A simulation environment for e-learning in digital design. *IEEE Transactions on Industrial Electronics*, 54(6):3078–3085.

Dow, G. T. and Mayer, R. E. (2004). Teaching students to solve insight problems: Evidence for domain specificity in creativity training. *Creativity Research Journal*, 16(4):389–398.

Drachen, A., Canossa, A., and Yannakakis, G. N. (2009). Player modeling using self-organization in tomb raider: Underworld. In *Computational Intelligence and Games*.

Drachsler, H., Dietze, S., Herder, E., d'Aquin, M., and Taibi, D. (2014). The learning analytics & knowledge (lak) data challenge 2014. In *Learning Analytics And Knowledge*.

Dyckhoff, A. L., Zielke, D., Bültmann, M., Chatti, M. A., and Schroeder, U. (2012). Design and implementation of a learning analytics toolkit for teachers. *Educational Technology & Society*, 15(3):58–76.

Ebner, M., Schön, M., and Neuhold, B. (2014). Learning analytics in basic math education–first results from the field. *eLearning Papers*, 36:24–27.

Erhel, S. and Jamet, E. (2013). Digital game-based learning: Impact of instructions and feedback on motivation and learning effectiveness. *Computers & Education*, 67:156–167.

Fautley, M. and Savage, J. (2008). *Assessment for learning and teaching in secondary schools*. Learning Matters.

Feldman, J. (2000). Minimization of boolean complexity in human concept learning. *Nature*, 407(6804):630–633.

Ferguson, R. (2012). Learning analytics: drivers, developments and challenges. *International Journal of Technology Enhanced Learning*, 4(5):304–317.

Figl, K. and Laue, R. (2011). Cognitive complexity in business process modeling. In *Advanced Information Systems Engineering*.

Floyd, S. and Warmuth, M. (1995). Sample compression, learnability, and the vapnik-chervonenkis dimension. *Machine Learning*, 21(3):269–304.

Fraley, C. and Raftery, A. E. (1998). How many clusters? which clustering method? answers via model-based cluster analysis. *The Computer Journal*, 41(8):578–588.

Gašević, D., Dawson, S., Mirriahi, N., and Long, P. D. (2015a). Learning analytics–a growing field and community engagement. *Journal of Learning Analytics*, 2(1):1–6.

Gašević, D., Dawson, S., and Siemens, G. (2015b). Let's not forget: Learning analytics are about learning. *TechTrends*, 59(1):64–71.

Gillet, D., De Jong, T., Sotirou, S., and Salzmann, C. (2013). Personalised learning spaces and federated online labs for stem education at school. In *IEEE EDUCON*.

Goldstone, R. L. and Kersten, A. (2003). Concepts and categorization. *Handbook of Psychology*, 4:599–621.

Goodman, N. D., Tenenbaum, J. B., Feldman, J., and Griffiths, T. L. (2008). A rational analysis of rule-based concept learning. *Cognitive Science*, 32(1):108–154.

Govaerts, S., Cao, Y., Vozniuk, A., Holzer, A., Zutin, D. G., Ruiz, E. S., Bollen, L., Manske, S., Faltin, N., Salzmann, C., et al. (2013). Towards an online lab portal for inquiry-based stem learning at school. In *Advances in Web-Based Learning*.

Green, C. S. and Bavelier, D. (2007). Action-video-game experience alters the spatial resolution of vision. *Psychological Science*, 18(1):88–94.

Green, C. S., Li, R., and Bavelier, D. (2010). Perceptual learning during action video game playing. *Topics in Cognitive Science*, 2(2):202–216.

Greller, W. and Drachsler, H. (2012). Translating learning into numbers: A generic framework for learning analytics. *Educational Technology & Society*, 13(3):42–57.

Griffiths, T. L., Christian, B. R., and Kalish, M. L. (2008). Using category structures to test iterated learning as a method for identifying inductive biases. *Cognitive Science*, 32(1):68–107.

Gruhn, V. and Laue, R. (2006). Complexity metrics for business process models. In *Business Information Systems*.

Günther, C. W. and Van Der Aalst, W. M. P. (2007). Fuzzy mining–adaptive process simplification based on multi-perspective metrics. In *Business Process Management*.

Gyllstrom, K. (2009). *Enriching personal information management with document interaction histories*. PhD thesis, The University of North Carolina at Chapel Hill.

Halkidi, M., Batistakis, Y., and Vazirgiannis, M. (2001). On clustering validation techniques. *Journal of Intelligent Information Systems*, 17(2):107–145.

Ham, D. H., Park, J., and Jung, W. (2011). A framework-based approach to identifying and organizing the complexity factors of human-system interaction. *IEEE Systems Journal*, 5(2):213–222.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *Unsupervised learning*. Springer.

He, W. (2013). Examining students' online interaction in a live video streaming environment using data mining and text mining. *Computers in Human Behavior*, 29(1):90–102.

Heiss, E. D., Obourn, S., and Hoffman, C. W. (1950). Modern science teaching. *Journal of Chemical Education*, 27(10):588.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30.

Holzinger, A., Popova, E., Peischl, B., and Ziefle, M. (2012). On complexity reduction of user interfaces for safety-critical systems. In *Multidisciplinary Research and Practice for Information Systems*.

Huang, O. W. S., Cheng, H. N. H., and Chan, T. (2007). Number jigsaw puzzle: A mathematical puzzle game for facilitating players' problem-solving strategies. In *Digital Game and Intelligent Toy Enhanced Learning*.

Hwang, G., Yang, L., and Wang, S. (2013). A concept map-embedded educational computer game for improving students' learning performance in natural science courses. *Computers & Education*, 69:121–130.

Jackson, L. A., Witt, E. A., Games, A. I., Fitzgerald, H. E., von Eye, A., and Zhao, Y. (2012). Information technology use and creativity: Findings from the children and technology project. *Computers in Human Behavior*, 28(2):370–376.

Jain, A. K. and Dubes, R. C. (1988). *Algorithms for clustering data*. Prentice-Hall, Inc.

Jain, A. K., Murty, M. N., and Flynn, P. J. (1999). Data clustering: a review. *ACM Computing Surveys*, 31(3):264–323.

Joachims, T. (2015). Learning representations of student knowledge and educational content. In *Machine Learning*.

Johnson, L., Smith, R., Willis, H., Levine, A., and Haywood, K. (2011). The 2011 horizon report. The New Media Consortium.

Justice, C., Rice, J., Roy, D., Hudspith, B., and Jenkins, H. (2009). Inquiry-based learning in higher education: administrators' perspectives on integrating inquiry pedagogy into the curriculum. *Higher Education*, 58(6):841–855.

Kebritchi, M., Hirumi, A., and Bai, H. (2010). The effects of modern mathematics computer games on mathematics achievement and class motivation. *Computers & Education*, 55(2):427–443.

Klahr, D. and Dunbar, K. (1988). Dual space search during scientific reasoning. *Cognitive science*, 12(1):1–48.

Klesk, P. and Korzen, M. (2011). Sets of approximating functions with finite vapnik–chervonenkis dimension for nearest-neighbors algorithms. *Pattern Recognition Letters*, 32(14):1882–1893.

Koedinger, K. R., Brunskill, E., Baker, S. S., McLaughlin, E. A., and Stamper, J. (2013). New potentials for data-driven intelligent tutoring system development and optimization. *AI Magazine*, 34(3):27–41.

Koedinger, K. R., D'Mello, S., McLaughlin, E. A., Pardos, Z. A., and Rosé, C. P. (2015). Data mining and education. *WIREs Cognitive Science*, 6(4):333–353.

Koltchinskii, V. (2001). Rademacher penalties and structural risk minimization. *IEEE Transactions on Information Theory*, 47(5):1902–1914.

Kotsiantis, S., Pierrakeas, C., and Pintelas, P. (2004). Predicting students' performance in distance learning using machine learning techniques. *Applied Artificial Intelligence*, 18(5):411–426.

Kotsiantis, S. B. and Pintelas, P. E. (2005). Predicting students marks in hellenic open university. In *Advanced Learning Technologies*.

Kotsiantis, S. B., Zaharakis, I., and Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. *Informatica*, 31:249–268.

Kruschke, J. K. (1992). Alcove: an exemplar-based connectionist model of category learning. *Psychological Review*, 99(1):22.

Kruse, A. and Pongsajapan, R. (2012). Student-centered learning analytics. In *CNDLS Thought Papers*.

Lan, A. S., Waters, A. E., Studer, C., and Baraniuk, R. G. (2014). Sparse factor analysis for learning and content analytics. *The Journal of Machine Learning Research*, 15(1):1959–2008.

Lassen, K. B. and Van Der Aalst, W. M. P. (2009). Complexity metrics for workflow nets. *Information and Software Technology*, 51(3):610–626.

Lee, J. I. and Brunskill, E. (2012). The impact on individualizing student models on necessary practice opportunities. In *Educational Data Mining*.

Lee, V. S. (2011). The power of inquiry as a way of learning. *Innovative Higher Education*, 36(3):149–160.

Lee, V. S. (2012). What is inquiry-guided learning? *New Directions for Teaching and Learning*, 2012(129):5–14.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8):707–710.

Lever, G., Laviolette, F., and Shawe-Taylor, J. (2013). Tighter pac-bayes bounds through distribution-dependent priors. *Theoretical Computer Science*, 473:4–28.

Lillo-Castellano, J. M., Mora-Jiménez, I., Figuera-Pozuelo, C., and Rojo-Álvarez, J. L. (2015). Traffic sign segmentation and classification using statistical learning methods. *Neurocomputing*, 153:286–299.

Liu, E. Z. F. and Lin, C. H. (2009). Developing evaluative indicators for educational computer games. *British Journal of Educational Technology*, 40(1):174–178.

MacKay, D. J. C. (2003). *Information theory, inference and learning algorithms*. Cambridge university press.

Madani, K. and Sabourin, C. (2011). Multi-level cognitive machine-learning based concept for human-like artificial walking: application to autonomous stroll of humanoid robots. *Neurocomputing*, 74(8):1213–1228.

Mar, T., Tikhanoff, V., Metta, G., and Natale, L. (2015). Self-supervised learning of grasp dependent tool affordances on the icub humanoid robot. In *Robotics and Automation*.

Matsuka, T., Sakamoto, Y., Chouchourelou, A., and Nickerson, J. V. (2008). Toward a descriptive cognitive model of human learning. *Neurocomputing*, 71(13):2446–2455.

McAllester, D. A. (1998). Some pac-bayesian theorems. In *Computational Learning Theory*.

McCabe, T. J. (1976). A complexity measure. *IEEE Transactions on Software Engineering*, 2(4):308–320.

McDiarmid, C. (1989). On the method of bounded differences. *Surveys in Combinatorics*, 141(1):148–188.

Medin, D. L. and Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, 85(3):207.

Medler, D. A., Arnoldussen, A., Binder, J. R., and Seidenberg, M. S. (2005). The wisconsin perceptual attribute ratings database. http://www.neuro.mcw.edu/ratings/.

Mohamed, N., Sulaiman, R. F. R., and Endut, W. R. W. (2013). The use of cyclomatic complexity metrics in programming performance's assessment. In *Procedia-Social and Behavioral Sciences*.

Montuori, A. (2012). Reproductive learning. *Encyclopedia of the Sciences of Learning*, 2:2838–2840.

Mukherjee, S., Niyogi, P., Poggio, T., and Rifkin, R. (2006). Learning theory: stability is sufficient for generalization and necessary and sufficient for consistency of empirical risk minimization. *Advances in Computational Mathematics*, 25(1):161–193.

Mukherjee, S., Tamayo, P., Rogers, S., Rifkin, R., Engle, A., Campbell, C., Golub, T. R., and Mesirov, J. P. (2003). Estimating dataset size requirements for classifying dna microarray data. *Journal of Computational Biology*, 10(2):119–142.

Murphy, G. L. (2002). *The big book of concepts*. MIT press.

Murphy, P. and Aha, D. W. (1995). Uci repository of machine learning databases–a machine-readable repository.

Naarmann, S. (2011). Lix. `https://github.com/SimonN/Lix`.

Nakayama, M., Mutsuura, K., and Yamamoto, H. (2015). The prediction of learning performance using features of note taking activities. In *Artificial Neural Networks, Computational Intelligence and Machine Learning*.

Neisser, U. (1976). *Cognition and reality*. San Francisco: Freeman.

Njoo, M. and De Jong, T. (1993). Exploratory learning with a computer simulation for control theory: Learning processes and instructional support. *Journal of Research in Science Teaching*, 30(8):821–844.

Nosofsky, R. M. and Palmeri, T. J. (1998). A rule-plus-exception model for classifying objects in continuous-dimension spaces. *Psychonomic Bulletin & Review*, 5(3):345–369.

Oei, A. C. and Patterson, M. D. (2014). Playing a puzzle video game with changing requirements improves executive functions. *Computers in Human Behavior*, 37:216–228.

Oneto, L., Ghio, A., Ridella, S., and Anguita, D. (2015a). Fully empirical and data-dependent stability-based bounds. *IEEE Transactions on Cybernetics*, 45(9):1913–1926.

Oneto, L., Ghio, A., Ridella, S., and Anguita, D. (2015b). Global rademacher complexity bounds: From slow to fast convergence rates. *Neural Processing Letters*, 43(2):567–602.

Oneto, L., Ghio, A., Ridella, S., and Anguita, D. (2015c). Learning resource-aware classifiers for mobile devices: From regularization to energy efficiency. *Neurocomputing*, 169:225–235.

Oneto, L., Ghio, A., Ridella, S., and Anguita, D. (2015d). Local rademacher complexity: Sharper risk bounds with and without unlabeled samples. *Neural Networks*, 65:115–125.

Opper, M. (1995). Statistical mechanics of learning: Generalization. In *The Handbook of Brain Theory and Neural Networks*.

Opper, M., Kinzel, W., Kleinz, J., and Nehl, R. (1990). On the ability of the optimal perceptron to generalise. *Journal of Physics A: Mathematical and General*, 23(11):L581.

Orduna, P., Almeida, A., López-De-Ipiña, D., and Garcia-Zubia, J. (2014). Learning analytics on federated remote laboratories: tips and techniques. In *IEEE EDUCON*.

Papamitsiou, Z. and Economides, A. (2014). Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence. *Educational Technology & Society*, 17(4):49–64.

Pashler, H. and Mozer, M. C. (2013). When does fading enhance perceptual category learning? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 39(4):1162.

Pechenizkiy, M., Trcka, N., Vasilyeva, E., Van Der Aalst, W. M. P., and De Bra, P. (2009). Process mining online assessment data. In *International Working Group on Educational Data Mining*.

Pedaste, M., Mäeots, M., Leijen, A., and Sarapuu, T. (2012). Improving students' inquiry skills through reflection and self-regulation scaffolds. *Technology, Instruction, Cognition & Learning*, 9(1/2):81–95.

Pedaste, M., Mäeots, M., Siiman, L. A., De Jong, T., Van Riesen, S. A. N., Kamp, E. T., Manoli, C. C., Zacharia, Z. C., and Tsourlidaki, E. (2015). Phases of inquiry-based learning: Definitions and the inquiry cycle. *Educational Research Review*, 14:47–61.

Peña-Ayala, A. (2014). Educational data mining: A survey and a data mining-based analysis of recent works. *Expert Systems with Applications*, 41(4):1432–1462.

Piech, C., Huang, J., Chen, Z., Do, C., Ng, A., and Koller, D. (2013). Tuned models of peer assessment in moocs. In *Educational Data Mining*.

Piech, C., Sahami, M., Koller, D., Cooper, S., and Blikstein, P. (2012). Modeling how students learn to program. In *Computer Science Education*.

Poggio, T., Rifkin, R., Mukherjee, S., and Niyogi, P. (2004). General conditions for predictivity in learning theory. *Nature*, 428(6981):419–422.

Polk, T. A. and Seifert, C. M. (2002). *Cognitive modeling*. MIT Press.

Ponta, D., Anguita, D., Da Bormida, G., and Donzellini, G. (1998). Ten years of activity on computer-aided learning for electronics: Needs, experiences, field evaluation. In *Congreso Sobre Tecnologías Aplicadas a la Enseñanza de la Electrónica*.

Pratheesh, N. and Devi, T. (2013). Influence of learning analytics in software engineering education. In *Emerging Trends in Computing, Communication and Nanotechnology*.

Prince, M. J. and Felder, R. M. (2006). Inductive teaching and learning methods: Definitions, comparisons, and research bases. *Journal of Engineering Education*, 95(2):123–138.

Ramirez-Cano, D., Colton, S., and Baumgarten, R. (2010). Player classification using a meta-clustering approach. In *Computer Games, Multimedia & Allied Technology*.

Rasmussen, E. M. (1992). Clustering algorithms. *Information Retrieval: Data Structures & Algorithms*, 419:442.

Rauterberg, M. (1992). A method of a quantitative measurement of cognitive complexity. In *Human-computer Interaction: Tasks and Organisation*.

Rauterberg, M. (1995). About a framework for information and information processing of learning systems. In *ISCO*.

Rauterberg, M. (1996). How to measure cognitive complexity in human-computer interaction. In *Cybernetics and Systems Research*.

Rauterberg, M., Schluep, S., and Fjeld, M. (1997). How to model behavioural and cognitive complexity in human-computer interaction with petri nets. In *Robot and Human Communication*.

Rauterberg, M. and Ulich, E. (1996). Information processing for learning systems: an action theoretical approach. In *Systems, Man, and Cybernetics*.

Reid, D. J., Zhang, J., and Chen, Q. (2003). Supporting scientific discovery learning in a simulation environment. *Journal of Computer Assisted Learning*, 19(1):9–20.

Romero, C. and Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert Systems with Applications*, 33(1):135–146.

Romero, C. and Ventura, S. (2015). J. a. larusson, b. white (eds): Learning analytics: from research to practice. *Technology, Knowledge and Learning*, 20(3):357–360.

Romero, C., Ventura, S., Espejo, P. G., and Hervás, C. (2008). Data mining algorithms to classify students. In *Educational Data Mining*.

Romero, C., Ventura, S., Pechenizkiy, M., and Baker, R. S. (2010). *Handbook of educational data mining*. CRC Press.

Saarela, M. and Kärkkäinen, T. (2015). Weighted clustering of sparse educational data. In *Artificial Neural Networks, Computational Intelligence and Machine Learning*.

Santos, J. L., Verbert, K., Govaerts, S., and Duval, E. (2013). Addressing learner issues with stepup!: an evaluation. In *Learning Analytics and Knowledge*.

Sao Pedro, M., Baker, R., and Gobert, J. (2012). Improving construct validity yields better models of systematic inquiry, even with less information. In *User Modeling, Adaptation, and Personalization*.

Sao Pedro, M., Baker, R., and Gobert, J. (2013). Incorporating scaffolding and tutor context into bayesian knowledge tracing to predict inquiry skill acquisition. In *Educational Data Mining*.

Sapounidis, T., Demetriadis, S., and Stamelos, I. (2015). Evaluating children performance with graphical and tangible robot programming tools. *Personal and Ubiquitous Computing*, 19(1):225–237.

Şara, N. B., Halland, R., Igel, C., and Alstrup, S. (2015). High-school dropout prediction using machine learning: A danish large-scale study. In *Artificial Neural Networks, Computational Intelligence and Machine Learning*.

Scanlon, E., Anastopoulou, S., Kerawalla, L., and Mulholland, P. (2011). How technology resources can be used to represent personal inquiry and support students' understanding of it across contexts. *Journal of Computer Assisted Learning*, 27(6):516–529.

Schacter, D. L., Gilbert, D. T., and Wegner, D. M. (2010). *Psychology*. Worth Publishers.

Schön, M., Ebner, M., and Kothmeier, G. (2012). It's just about learning the multiplication table. In *Learning Analytics and Knowledge*.

Schwab, J. J. and Brandwein, P. F. (1966). *The teaching of science*. Harvard University Press.

Serrano-Laguna, A. and Fernández-Manjón, B. (2014). Applying learning analytics to simplify serious games deployment in the classroom. In *EDUCON*.

Serrano-Laguna, A., Torrente, J., Moreno-Ger, P., and Fernández-Manjón, B. (2014). Application of learning analytics in educational videogames. *Entertainment Computing*, 5(4):313–322.

Shawe-Taylor, J. and Cristianini, N. (2004). *Kernel methods for pattern analysis*. Cambridge university press.

Shute, V. J., Ventura, M., and Ke, F. (2015). The power of play: The effects of portal 2 and lumosity on cognitive and noncognitive skills. *Computers & Education*, 80:58–67.

Siemens, G. (2012). Learning analytics: envisioning a research discipline and a domain of practice. In *Learning Analytics and Knowledge*.

Siemens, G. and Baker, R. (2012). Learning analytics and educational data mining: towards communication and collaboration. In *Learning Analytics and Knowledge*.

Siemens, G. and Long, P. (2011). Penetrating the fog: Analytics in learning and education. *EDUCAUSE Review*, 46(5):30.

Slotta, J. D., Tissenbaum, M., and Lui, M. (2013). Orchestrating of complex inquiry: three roles for learning analytics in a smart classroom infrastructure. In *Learning Analytics and Knowledge*.

Steinwart, I. (2005). Consistency of support vector machines and other regularized kernel classifiers. *IEEE Transactions on Information Theory*, 51(1):128–142.

Štuikys, V. and Damaševičius, R. (2013). Complexity evaluation of feature models and meta-programs. In *Meta-Programming and Model-Driven Meta-Program Development*.

Subrahmanyam, K. and Greenfield, P. M. (1994). Effect of video game practice on spatial skills in girls and boys. *Journal of Applied Developmental Psychology*, 15(1):13–32.

Taibi, D. and Dietze, S. (2013). Fostering analytics on learning analytics research: the lak dataset. In *The LAK Data Challenge*.

Tang, T. Y. and McCalla, G. (2002). Student modeling for a web-based learning environment: a data mining approach. In *AAAI/IAAI*.

Thomas, J. C. and Richards, J. T. (2009). Achieving psychological simplicity: Measures and methods to reduce cognitive complexity. *Human-Computer Interaction: Design Issues, Solutions, and Applications*, 161:489–508.

Thurau, C., Bauckhage, C., and Sagerer, G. (2003). Combining self organizing maps and multilayer perceptrons to learn bot-behaviour for a commercial game. In *GAME-ON*.

Tian, Y., Ruan, Q., An, G., and Xu, W. (2015). Context and locality constrained linear coding for human action recognition. *Neurocomputing*, 167:359–370.

Trcka, N. and Pechenizkiy, M. (2009). From local patterns to global models: Towards domain driven educational process mining. In *Intelligent Systems Design and Applications*.

Trcka, N., Pechenizkiy, M., and Van Der Aalst, W. M. P. (2010). *Process mining from educational data*. CRC Press.

Triantafillou, E., Pomportsis, A., and Demetriadis, S. (2003). The design and the formative evaluation of an adaptive educational system based on cognitive styles. *Computers and Education*, 41(1):87–103.

Vahdat, M., , Ghio, A., , Oneto, L., Anguita, D., Funk, M., and Rauterberg, M. (2015a). Advances in learning analytics and educational data mining. In *Artificial Neural Networks, Computational Intelligence and Machine Learning*.

Vahdat, M., Carvalho, M. B., Funk, M., Rauterberg, M., Hu, J., and Anguita, D. (2016a). Learning analytics for a puzzle game to discover the puzzle-solving tactics of players. In *Technology Enhanced Learning: Adaptive and Adaptable Learning*.

Vahdat, M., Oneto, L., Anguita, D., Funk, M., and Rauterberg, M. (2015b). Educational process mining (epm): A learning analytics data set. `https://archive.ics.uci.edu/ml/datasets/Educational+Process+Mining+(EPM)%3A+A+Learning+Analytics+Data+Set`.

Vahdat, M., Oneto, L., Anguita, D., Funk, M., and Rauterberg, M. (2015c). A learning analytics approach to correlate the academic achievements of students with interaction data from an educational simulator. In *Design for Teaching and Learning in a Networked World*.

Vahdat, M., Oneto, L., Anguita, D., Funk, M., and Rauterberg, M. (2016b). Can machine learning explain human learning? *Neurocomputing*, 192:14–28.

Vahdat, M., Oneto, L., Ghio, A., Anguita, D., Funk, M., and Rauterberg, M. (2015d). Human algorithmic stability and human rademacher complexity. In *Artificial Neural Networks, Computational Intelligence and Machine Learning*.

Vahdat, M., Oneto, L., Ghio, A., Donzellini, G., Anguita, D., Funk, M., and Rauterberg, M. (2014). A learning analytics methodology to profile students behavior and explore interactions with a digital electronics simulator. In *Open Learning and Teaching in Educational Communities*.

Van Der Aalst, W. M. P. (2011). *Process mining: discovery, conformance and enhancement of business processes*. Springer Science & Business Media.

van der Aalst, W. M. P. and Günth, C. W. (2007). Finding structure in unstructured processes: The case for process mining. In *Application of Concurrency to System Design*.

van der Aalst, W. M. P., Guo, S., and Gorissen, P. (2013). Comparative process mining in education: An approach based on process cubes. In *Data-Driven Process Discovery and Analysis*.

Van der Aalst, W. M. P., Weijters, T., and Maruster, L. (2004). Workflow mining: Discovering process models from event logs. *IEEE Transactions on Knowledge and Data Engineering*, 16(9):1128–1142.

van Dongen, B. F., Busi, N., Pinna, G., and van der Aalst, W. M. P. (2007). *An iterative algorithm for applying the theory of regions in process mining*. Beta, Research School for Operations Management and Logistics.

Van Joolingen, W. R. and De Jong, T. (2003). Simquest. In *Authoring Tools for Advanced Technology Learning Environments*.

van Joolingen, W. R., De Jong, T., Lazonder, A. W., Savelsbergh, E. R., and Manlove, S. (2005). Co-lab: research and development of an online learning environment for collaborative scientific discovery learning. *Computers in Human Behavior*, 21(4):671–688.

Vapnik, V. N. (1998). *Statistical learning theory*. Wiley–Interscience.

Vats, D., Studer, C., Lan, A. S., Carin, L., and Baraniuk, R. (2013). Test-size reduction for concept estimation. In *Educational Data Mining*.

Ventura, M., Shute, V., and Kim, Y. J. (2012). Video gameplay, personality and academic performance. *Computers & Education*, 58(4):1260–1266.

Verbert, K., Manouselis, N., Drachsler, H., and Duval, E. (2012). Dataset-driven research to support learning and knowledge analytics. *Educational Technology & Society*, 15(3):133–148.

Vozniuk, A., Rodrıguez-Triana, M. J., Holzer, A., Govaerts, S., Sandoz, D., and Gillet, D. (2015). Contextual learning analytics apps to create awareness in blended inquiry learning. In *Information Technology Based Higher Education and Training*.

Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244.

Watanabe, S. (1985). *Pattern recognition: human and mechanical*. John Wiley & Sons, Inc.

White, B. Y. and Frederiksen, J. R. (1998). Inquiry, modeling, and metacognition: Making science accessible to all students. *Cognition and Instruction*, 16(1):3–118.

White, B. Y., Shimoda, T. A., and Frederiksen, J. R. (1999). Enabling students to construct theories of collaborative inquiry and reflective learning: Computer support for metacognitive development. *International Journal of Artificial Intelligence in Education*, 10:151–182.

Yip, F. W. M. and Kwan, A. C. M. (2006). Online vocabulary games as a tool for teaching and learning english vocabulary. *Educational Media International*, 43(3):233–249.

Yuan, Y., Guo, Q., and Lu, X. (2015). Image quality assessment: A sparse learning way. *Neurocomputing*, 159:227–241.

Zahn, C. T. (1971). Graph-theoretical methods for detecting and describing gestalt clusters. *IEEE Transactions on Computers*, 100(1):68–86.

Zhang, X. and Li, Y. (2015). Adaptive energy detection for bird sound detection in complex environments. *Neurocomputing*, 155:108–116.

Zhu, X. (2015). Machine teaching: An inverse problem to machine learning and an approach toward optimal education. In *Artificial Intelligence*.

Zhu, X., Gibson, B. R., Jun, K., Rogers, T. T., Harrison, J., and Kalish, C. (2010). Cognitive models of test-item effects in human category learning. In *Machine Learning*.

Zhu, X., Gibson, B. R., and Rogers, T. T. (2009). Human rademacher complexity. In *Neural Information Processing Systems*.

Zuse, H. (1991). *Software complexity*. Walter de Cruyter.

Glossary

API Application Programming Interface
AS Algorithmic Stability
CL Concept Learning
CM Cyclomatic complexity of McCabe
COGS Cognitive Science
DM Data Mining
EDM Educational Data Mining
EPM Educational Process Mining
HAS Human Algorithmic Stability
HCI Human Computer Interaction
HL Human Learning
HRC Human Rademacher Complexity
IBL Inquiry Based Learning
LA Learning Analytics
ML Machine Learning
MOOC Massive Open Online Course
PM Process Mining
RC Rademacher Complexity
TEL Technology Enhanced Learning

List of Publications

Journal Paper

- Vahdat, M., Oneto, L., Anguita, D., Funk, M., and Rauterberg, M. (2016b). Can machine learning explain human learning? *Neurocomputing*, 192:14-28.

Conference Contribution

- Vahdat, M., Carvalho, M. B., Funk, M., Rauterberg, M., Hu, J., and Anguita, D. (2016a). Learning analytics for a puzzle game to discover the puzzle-solving tactics of players. In *European Conference on Technology Enhanced Learning: Adaptive and Adaptable Learning*.

- Vahdat, M., Oneto, L., Ghio, A., Anguita, D., Funk, M., and Rauterberg, M. (2015d). Human algorithmic stability and human rademacher complexity. In *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*.

- Vahdat, M., Oneto, L., Anguita, D., Funk, M., and Rauterberg, M. (2015c). A learning analytics approach to correlate the academic achievements of students with interaction data from an educational simulator. In *European Conference on Technology Enhanced Learning: Design for Teaching and Learning in a Networked World*.

- Vahdat, M., Ghio, A., Oneto, L., Anguita, D., Funk, M., and Rauterberg, M. (2015a). Advances in learning analytics and educational data mining. In *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*.

- Vahdat, M., Oneto, L., Ghio, A., Donzellini, G., Anguita, D., Funk, M., and Rauterberg, M. (2014). A learning analytics methodology to profile students behavior and explore interactions with a digital electronics simulator. In *European Conference on Technology Enhanced Learning: Open Learning and Teaching in Educational Communities*.

- Cardinali, F., Vahdat, M., Anguita, D., Lupo, L. (2014). Promoting Training & Performance Support Analytics at the Manufacturing Workplace. The MAN.TR.A. Model for LACE Project. In *European Distance and E-Learning Network Annual Conference*.

- Hoel, T., Cooper, A. R., Vahdat, M., Cardinali, F. (2014). LACE - Creating a Community on Learner Analytics and Educational Data Mining - Structuring the Discourse. In *European Distance and E-Learning Network Annual Conference*.

Data Set

- Vahdat, M., Oneto, L., Anguita, D., Funk, M., and Rauterberg, M. (2015b). Educational Process Mining (EPM): A Learning Analytics Data Set. https://archive.ics.uci.edu/ml/datasets/Educational+Process+Mining+(EPM)%3A+A+Learning+Analytics+Data+Set.

Curriculum Vitae

Mehrnoosh Vahdat started her PhD research in 2013 in the Interactive and Cognitive Environments programme (ICE, an Erasmus Mundus double-degree). Her PhD research was conducted in collaboration between the University of Genoa, Italy, and the Eindhoven University of Technology, Netherlands. She carried out research in several projects in the fields of Learning Analytics and Educational Data Mining over large amount of data collected from students of engineering and design. She also collaborated with LACE (Learning Analytics Community Exchange) FP7 European project on workplace learning. Her background comes from two international master programs in 'Interactive Media and Knowledge Environments' and 'Digital Library Learning'. She conducted research for her Master theses at the LIRIS laboratory in INSA Lyon, France, on Technology Enhanced Learning, and in the Media Integration and Communication Center, University of Florence, Italy. Her major fields of interest are Learning Analytics, Educational Data Mining, Human-Computer Interactions, and Technology Enhanced Learning.

UNIVERSITÀ DEGLI STUDI
DI GENOVA

TU/e Technische Universiteit
Eindhoven
University of Technology