ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 22-24 April 2015, i6doc.com publ., ISBN 978-287587014-8. Available from http://www.i6doc.com/en/.

Advances in Learning Analytics and Educational Data Mining

Mehrnoosh Vahdat[1,2,*], Alessandro Ghio[3], Luca Oneto[1], Davide Anguita[3],
Mathias Funk[2] and Matthias Rauterberg[2]

1 - DITEN - University of Genova
Via Opera Pia 11A, I-16145 Genova - Italy

2 - Department of Industrial Design - Technical University Eindhoven
P.O. Box 513, 5600 MB Eindhoven - The Netherlands

3 - DIBRIS - University of Genova
Via Opera Pia 13, I-16145 Genova - Italy

Abstract. 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 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 methods of development and integration of more personalized, adaptive, and interactive educational environments. This has motivated the organization of the *ESANN 2015* Special Session in *Advances in Learning Analytics and Educational Data Mining*. Here, a review of research and practice in LA and EDM is presented accompanied by the most central methods, benefits, and challenges of the field. Additionally, this paper covers a review of novel contributions into the Special Session.

1 Introduction

In recent years, there has been increasing interest in Learning Analytics (LA) and Educational Data Mining (EDM) among both researchers and practitioners of the field. By emerging the computer-assisted learning systems and automatic analysis of educational data, many efforts have been carried out in order to enhance the learning experience [1]. In 2011, the Horizon report claims for a fruitful future of LA [2]: LA is considered as a great help to discover the hidden information and patterns from raw data collected from educational environments [3]. This is one of the reason motivating the raising awareness on LA, and stimulating the vital strengthening of its connections with data-driven research fields like Data Mining and Machine Learning (ML).

[*] This work was supported in part by the Erasmus Mundus Joint Doctorate in Interactive and Cognitive Environments, which is funded by the EACEA Agency of the European Commission under EMJD ICE FPA n^o 2010-0012.

The combination of LA, as a new research discipline with a high potential to impact the existing models of education [3], and EDM, a novice growing research area to apply Data Mining methods on educational data [4], leads to new insights on learners' behavior, interactions, and learning paths, as well as to improve the technology enhanced learning 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 [5]. These opportunities have been strengthened by a huge shift in the availability of the data resources. In this regard, the availability of such resources is an inspiring motivation for growing research in the area: examples are PSLC DataShop and educational enriched data from MOOCs [6]. These available databases are also considered as benchmarks to advance the current methods and algorithms through comparison to other algorithms [7].

The ever growing importance and centrality that LA and EDM in educational processes, also in the industrial domain (e.g. in corporate training [8] and Human-Resources Analytics [9]), motivated the organization of the *ESANN 2015* Special Session in *Advances in Learning Analytics and Educational Data Mining*. In this context, the introductory paper presents the main motivations and concepts of EDM and LA and their applications in practice. Additionally, we present some successful cases of the field as a proof of LA/ EDM application into various educational problems, and the growing need to decrease the gap from research to practice. In this paper, the current state of the art of the field, the importance of LA/ EDM, as well as their benefits and challenges are explored. Additionally, a review of the common applied methods in current research and their applications including the novel contributions from the Special Session are covered in this work.

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 contains the Technology Enhanced Learning (TEL) areas of research such as EDM, recommender systems, and personalized adaptive learning [1]. 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" [10]. 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" [11]. 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 [12].

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 [6, 12, 13].

2.1 Benefits

The benefits of LA and EDM are explained further in many studies. For instance, Unesco policy brief explains the LA benefits in micro, meso, and macro levels covering various stakeholders [14]. These stakeholders are considered in three main groups: educators, learners, and administrators. Educators are responsible to design and plan the educational systems, and they are the most aware of the students' learning process, their needs, and common mistakes. Therefore, the availability of real-time feedback into the performance of learners is a great help for this group to adapt their teaching activities to the students' needs. Learners benefit from recommendation and feedback on their learning activities, resources, and paths. The type of feedback given to the students can be motivating and encouraging. Finally, administrators are dealing with decision-making and budget allowance, and can influence the process of improving the systems and learning resources [15, 16].

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 [13, 17]. In fact, these two research areas are considered complementary and opportunities and challenges are offered by both [18]. Figure 1 shows the LA/ EDM process concerning the data collection, processing, and giving feedback to the learners as a purpose of intervention and optimizing the learning outcome.

2.2 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 are emerged from various fields of analytics, data mining, and statistics, it is challenging for them to obtain the connections with cognition, metacognition, and pedagogy, which are indeed mandatory references

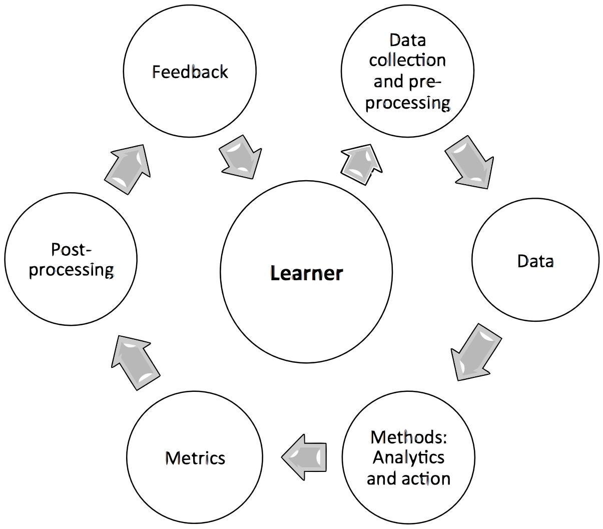
Feedback
Data collection and pre-processing
Post-processing
Learner
Data
Metrics
Methods: Analytics and action

Fig. 1: An LA/ EDM process starts with learner who his data is collected and analyzed, and after post processing, feedback and interventions are made in order to optimize learning (based on [1, 19]).

in understanding the learning processes. Researchers need to pay attention to learning sciences to enforce effective pedagogy and improve learning design [11].

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 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 [11, 13, 20, 21].

3 LA and EDM methods

We survey in this section the most common approaches, adopted by LA and EDM. A propaedeutic step is represented by reviewing data types 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), Per-

ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 22-24 April 2015, i6doc.com publ., ISBN 978-287587014-8. Available from http://www.i6doc.com/en/.

sonal Learning environments (PLE), game-based and simulation-based learning environments (e.g. [22]). 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 datasets are also rich data sources for LA and EDM research [1]. Data gathered from these environments include: 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. Other data categories are concerned with the student profile such as age, background, and interests in many LA and EDM cases [13, 20].

Exploring Data Mining methods adopted in educational context, 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 Data Mining method used in the field is classification. Clustering, regression, and discovery with models are followed respectively in its ranking [18]. 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.

These methods have been explained in further detail 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 [6, 13, 15]. 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 [13]. In the following section, applications of LA and EDM will be explained in further detail to illustrate the common use of these fields.

4 Applications of LA and EDM

The strong points of LA and EDM have been stressed by their applications in various studies. These applications address a wide verity of goals like those explained in introduction. One of the main objectives consists in maintaining and improving courses by giving hints to the educators and administrators. By analyzing the students' learning behavior in the process of problem solving, recommendations are generated to guide students about the content and tasks. Another common application is to predict the students' grades or to classify their behavior based on their learning outcome. Prediction of learner's performance can be used for analyzing the domain structure and its quality measurement. Furthermore, student modeling is applied to detect the students' states such as motivation, 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 [10]. Among these 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 [23]. Another application that is visible in many studies, is the prediction of learner performance which can lead to the increase of student and teacher awareness of learners' situation and enhancement of provided feedback and assessment services [18].

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. 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 school. In such studies, LA is applied to gather and interpret students' data for increasing the reflection and deep learning [24, 25]. 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 selections of pedagogy based on seeing data of students' interactions [26].

Despite applications of LA and EDM in various educational contexts are numerous, it is challenging to find many successful cases which proves that the application of LA and EDM actually leads to the improvement of educational outcome, by measuring the learning behavior and performance of the students. Studies like [27, 28] 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 [29]. 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 a LA dashboard name StepUp! And shows that the evaluation process is a complex task and critical in LA research [30]. eLAT is another example of a LA Toolkit that enables teachers to explore the students' learning behavior, and allow them monitor the students through visualizations [31].

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[1], PELARS[2], LEAs BOX[3], Next-Tell[4], WeSpot[5], and WatchMe[6]. The presence of the success cases of LA and EDM are strong in these initiatives. For instance in LACE FP7 European project which aim to gather and integrate the active communities in LA/EDM research and practice, created a knowledge base of evidence that captures effective evidence of the field and assign them particular evidence types and sectors, highlighting works like [32] as a positive evidence for improving teaching.

Various studies that explored the LA and EDM opportunities [1, 10, 12, 15, 13] have predicted 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 the educators, since currently 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 are needed as it was covered as an important challenge in our paper. 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.

5 Contributions to the ESANN 2015 Special Session on Advances in Learning Analytics and Educational Data Mining

The ESANN 2015 Special Session on *Advances in Learning Analytics and Educational Data Mining* collected research results from ten groups, dealing with issues related to both theoretical and practical application of LA and EDM approaches in different domains. The topics, which the accepted papers cope with, cover a vast range of techniques, ranging from descriptive approaches to predictive methods.

An interesting application domain of descriptive LA and EDM methodologies is related to ITS: when dealing with programming, such learning supporting systems mostly rely on pre-coded feedback provision, such that their applicability is restricted to modeled tasks. This causes researchers to report large authoring times for preparing and classifying the learning material. In [33], the authors investigate the suitability of ML techniques to automate this process by means of a presentation of similar solution strategies from a set of stored examples: for this purpose, the authors propose to apply structure metric learning methods, which can be used to compare Java programs.

[1] http://www.laceproject.eu

[2] http://www.pelars-project.eu

[3] http://www.leas-box.eu

[4] http://next-tell.eu

[5] http://wespot.net

[6] http://www.project-watchme.eu

Paper [34] deals, instead, with the interesting parallelisms, which exist 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 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 a previous study, which introduced Human Rademacher Complexity (HRC): in this work, the authors perform exploratory experiments towards studying Human Algorithmic Stability (HAS), which allows overcoming some drawbacks of HRC.

Another important domain of application for LA and EDM is related to enhancing learning in occupational training programs. By gathering empirical data on multiple factors that can affect learning for work, and by applying computational approaches in order to describe, identify, and understand preconditions of effective learning, LA and EDM can be implemented to effectively get insights towards supporting social capital. The authors in [35] propose to combine theory- and data- driven approaches towards maximizing the usefulness of LA and EDM in this domain.

As analyzed so far, descriptive methodologies are numerous and can be used for different purposes. Some of the most used descriptive analytics approaches belong to the domain of clustering methods, which have been shown their effectiveness in several heterogeneous domains, both in academia and in industry. In [36], 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.

While the papers, introduced so far, deal with descriptive approaches, predictive methods have a remarkable importance as well in the field of supporting education and learning. An interesting application of predictive methodologies is related to pupils, which do not finish their secondary education. There exist studies indicating that ML can be used to predict high-school dropout, which allows for triggering early interventions. Authors present, in [37], the first large-scale study of high-school dropout. It considers pupils who finished at least their first six months of Danish high-school education, and the goal was to predict dropout in the next three months.

Last but not least, we highlighted, in this tutorial paper, the importance of predicting the performance of students based on their behavior and their characteristics. In a blended learning course, in particular, participant's note taking activity reflects learning performance, and 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 authors in [38]. 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.

References

[1] M. A. Chatti, A. L. Dyckhoff, U. Schroeder, and H. Thüs. A reference model for learning analytics. *International Journal of Technology Enhanced Learning*, 4(5):318–331, 2012.

[2] L. Johnson, R. Smith, H. Willis, A. Levine, and K. Haywood. The 2011 horizon report. austin, texas: The new media consortium, 2011, 2011.

[3] G. Siemens. Learning analytics: envisioning a research discipline and a domain of practice. In *International Conference on Learning Analytics and Knowledge*, 2012.

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

[5] W. Greller and H. Drachsler. Translating learning into numbers: A generic framework for learning analytics. *Educational Technology & Society*, 13(3):42–57, 2012.

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

[7] K. Verbert, N. Manouselis, H. Drachsler, and E. Duval. Dataset-driven research to support learning and knowledge analytics. *Educational Technology & Society*, 15(3):133–148, 2012.

[8] L. Servage. Strategizing for workplace e-learning: some critical considerations. *Journal of Workplace Learning*, 17(5/6):304–317, 2005.

[9] Grace M Endres and Lolita Mancheno-Smoak. The human resource craze: Human performance improvement and employee engagement. *Organization Development Journal*, 26(1), 2008.

[10] C. Romero, S. Ventura, M. Pechenizkiy, and Ryan S. Baker. *Handbook of educational data mining*. Data Mining and Knowledge Discovery Series. Boca Raton, FL: Chapman and Hall/CRC Press, 2010.

[11] R. Ferguson. Learning analytics: drivers, developments and challenges. *International Journal of Technology Enhanced Learning*, 4(5):304–317, 2012.

[12] G. Siemens and R. Baker. Learning analytics and educational data mining: towards communication and collaboration. In *international conference on learning analytics and knowledge*, 2012.

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

[14] S. Buckingham Shum. Learning analytics, 2012.

[15] C. Romero and S. Ventura. Educational data mining: A survey from 1995 to 2005. *Expert systems with applications*, 33(1):135–146, 2007.

[16] G. Siemens and P. Long. Penetrating the fog: Analytics in learning and education. *Educause Review*, 46(5):30–32, 2011.

[17] W. He. Examining students' online interaction in a live video streaming environment using data mining and text mining. *Computers in Human Behavior*, 29(1):90–102, 2013.

[18] Z. Papamitsiou and A. Economides. Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence. *Journal of Educational Technology & Society*, 17(4):49–64, 2014.

[19] D. Clow. The learning analytics cycle: Closing the loop effectively. In *International Conference on Learning Analytics and Knowledge*, 2012.

[20] A. Del Blanco, A. Serrano, M. Freire, I. Martínez-Ortiz, and B. Fernández-Manjón. E-learning standards and learning analytics. can data collection be improved by using standard data models? In *Global Engineering Education Conference*, 2013.

[21] K. Gyllstrom. *Enriching personal information management with document interaction histories*. PhD thesis, The University of North Carolina at Chapel Hill, 2009.

[22] M. Vahdat, L. Oneto, A. Ghio, G. Donzellini, D. Anguita, M. Funk, and M. Rauterberg. A learning analytics methodology to profile students behavior and explore interactions with a digital electronics simulator. In *Open Learning and Teaching in Educational Communities*, pages 596–597, 2014.

[23] A. Peña-Ayala. Educational data mining: A survey and a data mining-based analysis of recent works. *Expert systems with applications*, 41(4):1432–1462, 2014.

[24] M. Schön, M. Ebner, and G. Kothmeier. It's just about learning the multiplication table. In *International Conference on Learning Analytics and Knowledge*, 2012.

[25] M. Ebner, M. Schön, and B. Neuhold. Learning analytics in basic math education–first results from the field. *eLearning Papers*, 36:24–27, 2014.

[26] C. Brooks. A data-assisted approach to supporting instructional interventions in technology enhanced learning environments. In *PhD Thesis*, 2013.

[27] J. A. Larusson and B. White. *Learning Analytics: From Research to Practice*. Springer Publishing Company, Incorporated, 2014.

[28] D. Gašević, S. Dawson, and G. Siemens. Let's not forget: Learning analytics are about learning. *TechTrends*, 59(1):64–71, 2015.

[29] K. E. Arnold and M. D. Pistilli. Course signals at purdue: Using learning analytics to increase student success. In *International Conference on Learning Analytics and Knowledge*, 2012.

[30] J. L. Santos, K. Verbert, S. Govaerts, and E. Duval. Addressing learner issues with stepup!: an evaluation. In *International Conference on Learning Analytics and Knowledge*, 2013.

[31] A. Lea Dyckhoff, D. Zielke, M. Bültmann, M. A. Chatti, and U. Schroeder. Design and implementation of a learning analytics toolkit for teachers. *Educational Technology & Society*, 15(3):58–76, 2012.

[32] M. Sao Pedro, R. Baker, and J. Gobert. Incorporating scaffolding and tutor context into bayesian knowledge tracing to predict inquiry skill acquisition. In *International Conference on Educational Data Mining, Memphis, TN*, 2013.

[33] B. Paassen, B. Mokbel, and B. Hammer. Adaptive structure metrics for automated feedback provision in java programming. In *Proc. of the European Symposium on Artificial Neural Networks (ESANN)*, 2015.

[34] M. Vahdat, L. Oneto, A. Ghio, G. Donzellini, D. Anguita, M. Funk, and M. Rauterberg. Human algorithmic stability and human rademacher complexity. In *Proc. of the European Symposium on Artificial Neural Networks (ESANN)*, 2015.

[35] V. Kalakoski, H. Ratilainen, and L. Drupsteen. Enhancing learning at work. how to combine theoretical and data-driven approaches, and individual, behavioural, and organizational levels of data? In *Proc. of the European Symposium on Artificial Neural Networks (ESANN)*, 2015.

[36] M. Saarela and T. Kärkkäinen. Weighted clustering of sparse educational data. In *Proc. of the European Symposium on Artificial Neural Networks (ESANN)*, 2015.

[37] N.-B. Şara, R. Halland, C. Igel, and S. Alstrup. High-school dropout prediction using machine learning: A danish large-scale study. In *Proc. of the European Symposium on Artificial Neural Networks (ESANN)*, 2015.

[38] M. Nakayama, K. Mutsuura, and H. Yamamoto. The prediction of learning performance using features of note taking activities. In *Proc. of the European Symposium on Artificial Neural Networks (ESANN)*, 2015.