Sabine Rathmayer, Hans Pongratz (Hrsg.): Proceedings of DeLFI Workshops 2015
co-located with 13th e-Learning Conference of the German Computer Society (DeLFI 2015)
München, Germany, September 1, 2015 101

Educational Data Mining / Learning Analytics: Methods, Tasks and Current Trends

Agathe Merceron[1]

Abstract: The 1st international conference on "Educational Data Mining" (EDM) took place in Montreal in 2008 while the 1st international conference on "Learning Analytics and Knowledge" (LAK) took place in Banff in 2011. Since then the fields have grown and established themselves with an annual international conference, a journal and an association each, and gradually increase their overlapping. This paper begins with some considerations on big data in education. Then the principal analysis methods used with educational data are reviewed and are illustrated with some of the tasks they solve. Current emerging trends are presented. Analysis of educational data on a routine basis to understand learning and teaching better and to improve them is not a reality yet. The paper concludes with challenges on the way.

Keywords: Educational data mining, learning analytics, prediction, clustering, relationship mining, distillation of data for human judgment, discovery with models, multi modal analysis, multi-level analysis, natural language processing, privacy, data scientist.

1 Introduction

"Big Data in Education" was the name of a MOOC offered on Coursera in 2013 by Ryan Baker. What means big data in education? To answer this question I consider different sources of educational data following the categorization of [RV 10]. Schools and universities use information systems to manage their students. Take the case of a small- medium European university with 12 000 students and let us focus on the marks. Assuming that each student is enrolled in 6 courses, each semester the administration records 60 000 new marks (including the null value when students are absent).

Many universities and schools use a Learning Management System (LMS) to run their courses. Let us take an example of a small course, not a MOOC, taught for 60 students on 12 weeks with one single forum, and a set of slides and one quiz per week. LMSs record students' interactions, in particular when students click on a resource, write or read in the forum. Assume that each student clicks on average twice each week on the set of slides and the quiz, and 3 times on the forum in the semester. This gives 3060 interactions that are stored for one course during one semester. Let us suppose that the small-medium university from above has 40 degree-programs with 15 courses each. This gives 1 836 000 interactions stored by the LMS each semester.

[1] Beuth Hochschue für Technik, Fachbereich Medieninformatik, Luxemburgerstrasse 19, 13353 Berlin, merceron@beuth-hochschule.de

Another main source of data in education is dedicated software like Intelligent Tutoring Systems that students use to train specific skills in one discipline. A well-known repository for such data is Datashop [KBS10] that contains millions of interactions. On top of those main sources of data, there are various other sources like social media, questionnaires or online forums. These simple considerations show that data in education are big and cannot be analyzed by hand.

Research published the "international educational data mining society" or the "society of learning analytics and research" show that analyzing these data allows us "to better understand students, and the settings which they learn in." [BY 09]. These two societies started around different persons with different research backgrounds [BS 14] but they have similar aims. While research with an emphasis in machine learning appears more in EDM and research with an emphasis on humans appear more in LAK, research that could be published in both conferences grows each year as the citations in this paper make clear. The next section reviews the computational methods used to analyze educational data as listed in [BY 09] and illustrates some of the tasks they solve. Current trends are presented in section three. The conclusion presents challenges of the field.

2 Computational Methods

Methods come mainly from machine learning, data mining and, emerging in the last years, from natural language processing; methods classical in artificial intelligence such as the hill-climbing algorithm are occasionally used [Ba 05].

2.1 Prediction

A major task tackled by prediction methods is to predict performance of students. Predicting performance has several levels of granularity: it can be predicting that there will be no performance at all when students drop-off [WZN13], predicting pass/fail or the mark in a degree [ZBP11, AMP14], pass / fail or the mark in a course [LRV12] or predicting whether a student masters a given skill in a tutoring system [PHA07].

The numerous studies published in that area show that it is indeed possible to predict drop-off or performance in a degree or in a course (MOOCs excluded) with a reasonable accuracy, mostly over 70%. However these studies show also that there is neither one classifier nor a set of features that work well in all contexts though a number of studies indicate that having only socio-economic features and no marks at all lead to a poorer accuracy [GD 06, ZBP11]. Therefore one has to investigate which methods and which features work the best with the data at hand. I take [AMP14] to illustrate such a work. The aim of this work is to predict the class or interval A, B, C, D or E, in which the mark

of the degree lies. The degree is a 4 years Bachelor of Science in Computing and Information Technology in a technical university in Pakistan. Enrollment in this degree is competitive: students are selected based on their marks at High School Certificate (short HSC), average and Math-Physics-Chemistry, and on their performance on a entrance test. Because of this context, drop-off is almost non-existent. Using in particular the conclusions of [GD 06, ZBP11], [AMP14] conjectured that marks only, no socio- economic features, might be enough to predict performance at the end of the degree with an acceptable accuracy. The features that have been used to build several classifiers are HSC marks as well as first year and second year marks in each course. Table 1 shows the classifiers that have achieved an accuracy better than the baseline of 51.92%, the accuracy that is achieved when the majority class C is always predicted.

Classifier	Accuracy / Kappa
Decision Tree with Gini Index	68.27% / 0.49
Decision Tree with Information Gain	69.23% / 0.498
Decision Tree with Accuracy	60.58% / 0.325
Rule Induction with Information Gain	55.77% / 0.352
1- Nearest Neighbors	74.04% / 0.583
Naives Bayes	83.65% / 0.727
Neural Networks	62.50% / 0.447
Random Forest with Gini Index	71.15% / 0.543
Random Forest with Information Gain	69.23% / 0.426
Random Forest with Accuracy	62.50% / 0.269

Tab. 1: Comparison of Classifiers

A unique feature of this work is to take one cohort to train a classifier and the next cohort to test it, as opposed to most of the works reported in the literature which use cross-validation, which means that only one cohort is used to train and test the classifier. The aim of using two successive cohorts is to check how well results generalize over time so as to use the experience of one cohort to put in place some policy to detect weak or strong students for the following cohort. One notices that 1- nearest neighbor and Naives Bayes perform particularly well although they have the drawback of not giving a human interpretable explanation of the results: it is not possible to know whether some courses could act as detectors of particularly poor or particularly good performance.

2.2 Clustering

Clustering techniques are used to group objects so that similar objects are in the same cluster and dissimilar objects in different clusters. There are various clustering techniques and there are many tasks that use clustering. [CGS12] for instance clusters students and find typical participation's behaviors in forums.

[EGL15] clusters utterances and is concerned with classifying automatically dialog acts

also called speech acts within tutorial dialogs. A dialog act is the action that a person performs while uttering a sentence like asking a question ("What is an anonymous class?"), exposing a problem or issue, giving an answer, giving a hint ("Here an interesting link about Apache Ant"), making a statement ("The explanations of the lectures notes are a bit succinct"), giving a positive acknowledgment ("Thanks, I have understood"), etc.. A common way of classifying sentences or posts into dialog acts is to use prediction or supervised methods as done in [KLK10]. First a labeled corpus is built: several annotators label the sentences of a corpus and identify cues or features to choose the dialog act. Support vector machines are reported to do rather well for this kind of task: [KLK10] reports F-score varying from 0.54 for positive acknowledgment (9.20% of the sentences of the corpus) to 0.95 for questions (55.31% of the sentences of the corpus). A major drawback of this approach is getting the labeled corpus, a major work done by hand. Therefore several works such as [EGL15] investigate approaches to classify sentences without the manual labeling. The corpus of [EGL15] comes from a computer-mediated environment to tutor students in introductory programming; moreover, in this case study, students have been recorded by Kinect cameras. Sentences are described by different kinds of features: lexical features (e.g. unigrams, word ordering, punctuation), dialog-context features (e.g. utterance position, utterance length, author of the previous message (student, tutor)), task features (e.g. writing code), posture features (e.g. distance between camera and head, mid torso, lower torso) and gesture features (e.g. one-hand-to-face, two-hands-to-face).

Utterances are clustered using the K-Medoids algorithm and Bayesian Information Criterion (BIC) to infer the optimal number of clusters. For lexical features the distance between two utterances is calculated using their longest common subsequence and for other features using cosine similarity. Several clusterings are performed according to the dialog act of the previous tutor utterance. The majority vote of the utterances in each cluster gives the dialog act or label of that cluster. To classify a new student's utterance, the proper clustering is chosen according to the preceding dialog act of the tutor and the distance between the new utterance and the center of each cluster is calculated. The nearest cluster gives its dialog act to the utterance. Using a manually labeled corpus for evaluation, and a leave-one-student-out cross-validation, an average accuracy of 67% is reported (61.7% without posture and gesture features). Even if these results stay below what is currently achieved with supervised methods, this approach is very promising and continues to improve over earlier similar work such as reported in [VMN12].

2.3 Relationship Mining

[BY 09] divides this category into four sub-categories. Two of them, association rule mining and correlation mining, are illustrated here.

[MY 05] uses the apriori algorithm for association rules to find mistakes that students often make together while solving exercises with a logic tutor. Results include associa-

tions such as "if a student chooses the wrong set of premises to apply a rule, s/he is likely to also make a wrong deduction when applying a rule." Such findings have been used to enhance the tutor with proactive feedback whenever students make a mistake belonging to the found associations. One challenge in using association rules is the big number of rules that algorithms can return and the choice of an appropriate interestingness measure to filter them [MY 10].

[BCK04] conducted observations of students while using a cognitive tutor for middle school mathematics. Observers recorded whether students were on-task or off-task and, when off-task, whether students were in conversation, doing something else, inactive or gaming the system. Gaming the system means that a student uses quickly the hints offered by the tutor and so finishes quickly an exercise as the solution is basically given by the tutor. The calculation of correlations between post-tests and different off-task behaviors revealed that the biggest correlation in absolute value was obtained with gaming the system: -0.38. This is high enough to indicate that gaming the system has a negative impact on learning.

2.4 Distillation of Data for Human Judgment

This category includes statistics and visualizations that help humans make sense of their findings and analyses. Proper diagrams on the proper data help to grasp what happens at a glance and they form the essence of many dashboards and analytics tools such as LeMo [FEM13].

All the works presented so far show that data preparation is crucial. This remains true for visualization. Students and their marks can be visualized by a heat map. Clustering the students according to their marks [AMP15] and using the order given by the clustering to build the heat map helps visualize courses that can act as indicators of good or poor performance.

2.5 Discovery with Models

As noted in [BY 09] this category is usually absent from conventional books about data mining or machine learning. This category encompasses approaches in which the model obtained in a previous study is included in the data to discover more patterns. An interesting illustration is given by the work of [BCR06] and [SPB15]. Building on [BCK04] the work in [BCR06] proposes a detector for gaming the system. This detector uses only data stored in the log files recorded by the cognitive tutor, no other source of data from sensors or cameras. Features include the number of times a specific step is wrong across all problems, the probability that the student knows a skill as calculated by the tutor, various times such as the time taken by the last 3 or 5 actions. Latent Response Models have been used to build the detector. This detector has been shown to generalize to new students and to new les-

sons of the cognitive tutor, and thus can be used to infer whether students who used the cognitive tutor gamed the system without having to actually observe them. The work in [SPB15] investigates the relation between different affects and behaviors and the majors chosen in college. Students who game the system enroll less in Science, Math. & Technology.

3 Current Trends

Over the three last editions of the EDM conferences and the last edition of the LAK conference I observe an increase in the number of papers using following techniques: Natural Language Processing, Multilevel Analysis and Multimodal Analysis.

Textual data produced by learners receive more attention. Methods of natural language processing are mainly used to analyze tutorial dialogs, to model students' reading and writing skills and to understand discussion forums. Multimodal analysis means that data from different sources are aggregated together as illustrated earlier with [EGL15].

Multilevel analysis means that different kinds of data stored by one system are aggregated and analyzed as in the framework “Traces” [Su 15] that is used to analyze interactions of a large community of users in a kind of learning platform offering chats, forums, file uploads and a calendar. “Traces” extracts events from the database and constructs contingency graphs which show the likelyhood that events are related. For example two events like uploading a file and writing a message in a chat might be related by a proximal contingency if they occur close enough in time, or two events like two messages having an overlap in their vocabulary might be related by a lexical contingency. These graphs can be abstracted and folded at several levels, the most general level being a sociogram which represents how actors are related through their contributions. “Traces" can detect session of activities and in these sessions identify the main actors and those who might be disengaged. On a much smaller scale [Me 14] relates the forum level (dialog acts) to the performance level in an online-course taught with a LMS.

4 Conclusion

Big data in education is a reality. There are numerous approaches to analyze educational data, numerous tasks that are tackled and interesting findings that are discovered.

What is not a reality yet is the analysis of educational data on a routine basis to understand learning and teaching better and to improve them. I see at least two challenges on the way. One is privacy. Users of educational software have to trust what happens with their data that systems store and analyze. A reasonable answer is opt-in: interactions are stored only when users opt for it. This can limit the available data, hence the findings that can be made. Another challenge is generalizability: is a classifier for predicting

performance still valid 2 years later or for another degree? My answer is probably not. Validation needs to be checked regularly, which can slow down the adoption of educational data mining or learning analytics in everyday life. Models that are demanding computationally like classifiers for performance or detectors of behaviors have to be continuously re-established by data scientists. Nonetheless I believe that this field will continue to grow and finds its place in everyday education.

Literaturverzeichnis

[AMP14] Asif, R.; Merceron, A.; Pathan, M.K.: Predicting Student Academic Performance at Degree Level: A Case Study, In International Journal of Intelligent Systems and Applications (IJISA), Volume 7, Number 1, S. 49-61, 2015.

[AMP15] Asif, R., Merceron, A. and Pathan, M.K.: Investigating Performance of Students: a Longitudinal Study. In (Blinkstein, A.; Merceron, A.; Siemens, G.; Baron, J.; Marziaz, N.; Lynch, G. Hrsg.) Proceedings of LAK 15, ACM, S. 108-112, 2015.

[Ba 05] Barnes, T.: Q-matrix Method: Mining Student Response Data for Knowledge. In the technical Report (WS-05-02) of the AAAI-05 Workshop on Educational Data Mining. 8 p., 2005.

[BCK04] Baker, R.; Corbett, A.T.; Koedinger, K.; Wagner, A.Z.: Off-task behavior in the cognitive tutor classroom: when students "game the system". In Proc. Of SIGCHI conference of Human Factors in Computing Systems, Vienna, Austria, 383-390, 2004.

[BCR06] Baker, R.; Corbett, A.T.; Roll, I.; Koedinger, K.: Developing a generalizable detector of when students game the system. User Modeling and User-Adapted Interaction, 18(3), 287-314, 2006.

[BS 14] Baker, R.; Siemens G.: Educational data mining and learning analytics. In Sawyer, K. (Ed.) Cambridge Handbook of the Learning Sciences: 2nd Edition, pp. 253-274, 2014.

[BY 09] Baker, R.; Yacef, K.: The State of Educational Data Mining in 2009: A Review and Future Visions", In Journal of Educational Data Mining, Vol. 1(1), 2009.

[CGS12] Cobo, G., Garcia, D., Santamaria, E., Moran, J.A., Melenchon, J., Monzo, C. Using agglomerative hierarchical clustering to model learner participation profiles in online discussion forums. In (Dawson, S., Haythornthwaite, C. Eds.): Proceedings of the 2nd International Conference on Learning Analytics and Knowledge., ACM, S. 248-251, 2012.

[EGL15] Ezen-Can, A.; Grafsgaard, J.F.; Lester, J.C., Boyer, K.E.: Classifying Dtudent Dialogue Avts with Multimodal Learning Analytics. In (Blinkstein, A.; Merceron, A.; Siemens, G.; Baron, J.; Marziaz, N.; Lynch, G. Hrsg.) Proceedings of LAK 15, ACM, S. 280-289, 2015.

[FEM13] Fortenbacher, A.; Elkina, M.; Merceron, A.: The Learning Analytics Application LeMo – Rationals and First Results. In International Journal of Computing, Volume 12, Issue 3, S. 226-234, 2013.

[GD 06] P. Golding, O. Donaldson: “Predicting Academic Performance”, Proceedings of 36th ASEE /IEEE Frontiers in Education Conference, 2006.

[KBS10] Koedinger, K.; Baker, R.., Cunningham, K., Skogsholm, A., Leber, B.; Stamper, J.: “A Data Repository for the EDM Community: The PSLC DataShop,” Handbook of Educational Data Mining, CRC Press, S. 43–56, 2010.

[KLK10] Kim, J.; Li, J.; Kim. T. Towards Identifying Unresolved Discussions in Student Online Forums. In (Tetreault, J., Burstein, J., Leacock, C. Hrsg.): Proceedings of the 5th Wokshop on Innovative Use of NLP for Building Educational Applications. (Los Angeles, CA, USA, June 2010). Association for Computational Linguistics, 84 - 91.

[LRV12] Lopez, M. I., Romero, R., Ventura, V., Luna, J.M. Classification via clustering for predicting final marks starting from the student participation in Forums. In (Yacef, K., Zaïane, O., Hershkovitz, H., Yudelson, M., and Stamper, J. Hrsg.): Proceedings of the 5th International Conference on Educational Data Mining. S. 148-151, 2012.

[MY 05] Merceron, A.; Yacef, K.: Educational Data Mining: a case study. In (C.-K. Looi, G. McCalla, B. Bredeweg and J. Breuker Eds) Proceedings of Artificial Intelligence in Education (AIED2005), S. 467-474, 2005.

[Me 14] Merceron, A.: Connecting Analysis of Speech Acts nd Performance Analysis: a Initial Study. In Proceedings of the Workshop 3: Computational Approaches to Connecting Levels of Analysis in Networked Learning Communities, LAK 2014, Vol-1137, 2014.

[MY 10] Merceron, A.; Yacef, K.: Measuring correlation of Strong Association Rules in Educational Data. In (C. Romero, S. Ventura, M. Pechenizkiy & R.S.J.d. Baker Eds.) the Handbook of Educational Data Mining. CRC Press, S. 245 -256, 2010.

[PHA07] Pardos, Z.; Hefferman, H.; Anderson, B.; Hefferman, C.: “The effect of Model Granularity on Student Performance Prediction Using Bayesian Networks,” In Proceedings of the international Conference on User Modelling, Springer, Berlin, pS 435-439, 2007.

[RV 10] Romero, C.; Ventura, S.: Educational Data Mining: A Review of the State of the Art. IEEE transactions on Systems, Man and Cybernetics, vol. 40(6), S.601-618, 2010.

[SPB15] San Pedro, M.O., R. Baker, N. Heffernan, J. Ocumpaugh: What Happens to Students Who Game the System?. In (Blinkstein, A.; Merceron, A.; Siemens, G.; Baron, J.; Marziaz, N.; Lynch, G. Hrsg.) Proceedings of LAK 15, ACM, S. 36-40, 2015.

[Su 15] Suthers, D.: From Contingencies to Network-level Phenomena: Multilevel Analysis of Activity and Actors in Heterogeneous Networked Learning Environments. In (Blinkstein, A.; Merceron, A.; Siemens, G.; Baron, J.; Marziaz, N.; Lynch, G. Hrsg.) Proceedings of LAK 15, ACM, S. 368-377, 2015.

[VMN12] Rus, V.; Moldovan, C.; Niraula, N.; Graesser, C.C.: Automated discovery of speech act categories in educational games. In Proceedings of the International Conference on Educational Data Mining, S. 25-32, 2012.

[WZN13] Wolff, A.; Zdrahal, Z.; Nikolov, A.; Pantucek, M.: Improving retention: predicting at-risk students by analysing clicking behaviour in a virtual learning environment. In Proceedings of the Third International Conference on Learning Analytics and Knowledge, S. 145-149, 2013.

[ZBP11] Zimmermann, J.; Brodersen, K.; Pellet, J.P.; August, E.; Buhmann, J.M.: Predicting graduate-level performance from undergraduate achievements. In Proceedings of the 4th International Conference on Educational Data Mining, S.357-358, 2011.