An Exploratory Study on Data Mining in Education: Practiced Algorithms and Methods

Ancelmo Castro, Leandro Garcia, David Prata, Marcelo Lisboa, and Monica Prata

***Abstract*—The aim of this paper is to develop an exploratory research, based on articles available in scientific databases and related to the theme "Data Mining in Education" in order to ascertain which are the algorithms and methods used. The keywords used for the study were: "Data mining and education". After a process for selecting the publications related to the theme, we analyzed the methods and algorithms commonly used for these studies. Results showed the most cited databases, along the methods and algorithms practiced in Educational Data Mining field. We concluded that the decision tree was the main method applied, probably because of the graphic knowledge representation, which could help experts to better interpret the elicited evidences, and to postulate causes and effects.**

***Index Terms*—Algorithms, data mining, education, methods, exploratory research.**

I. Introduction

As the first computer systems have emerged, the idea of storing data has grown considerably. Over the years, due to the need to store data in much larger quantities, this feature has become increasingly evident [1]. This concept of data mining has been widely used in the field of education. Data Mining in Education (EDM) is linked to the development, research and computerized application of some methods to detect patterns in large educational data collection that otherwise would be difficult or impossible to be analyzed, due to the huge volume of data [2].

In the process of mining, or Data Mining, the pursuit of knowledge is accomplished through standards and exploratory study among its instances. To implement this technique, automated methods and algorithms - based on building decision trees - are used in order to improve the analysis process [3]. And these algorithms DM (data mining) usually require parameters and have to offer appropriate values to obtain good results models. Thus, educators must have knowledge to find the correct settings [4].

The solution to this problem is to use decision support systems, facilitating tools, recommendation engines DM parameters and algorithms to automate and facilitate all processes.

After the exploratory study, the paper presents methods and algorithms most widely used in data mining for Education.

Manuscript received 29 August 2015; revised December 17, 2015. This work was supported in part by the Federal University of Tocantins (UFT).

A. Castro is with the Federal Institute of Tocantins (IFTO), CO 77.950-000 Brazil (e-mail: audioespacial@gmail.com).

L. Garcia, D. Prata, M. Lisboa, and M. Prata are with the Federal University of Tocantins (UFT), CO 77.001-009 Brazil (e-mails: lggarcia@mail.uft.edu.br, ddnprata@uft.edu.br, marcelolisboarocha@yahoo.com.br, pratamonica7@gmail.com).

II. Methodology

The theme chosen for this article is: "An Exploratory Study on Data Mining: Practices used in Education". As soon as the choice was made, a question was generated.

This work intends to have the answer based on an exploratory study. According to the Cochrane Handbook [5] publishing model, having a question is especially important on the process of structuring the research. To formulate a question is the first step, which guides the researcher to the remaining steps for the exploratory research.

The question for this study is: What are the algorithms and methods used on Data mining in Education?

The process of the exploratory study followed the steps below:

A. Step 1: Exploratory Research Project

After defining the question, an exploratory research project was set up and all the steps have been established to provide a clear and objective answer to the question.

B. Step 2: In Search of Articles

The review did a search on articles that were compatible with the question, seeking words which have also been defined as the keywords for this study: "Data Mining and Education."

The search of identifying all the studies was carried out in the following databases:

1) IEEE Xplore - http://www.ieee.org/web/publications/xplore/
2) ScienceDirect - www.sciencedirect.com/
3) ACM Digital Library - http://portal.acm.org
4) Google Scholar - http://scholar.google.com.br/
5) Microsoft Academic - http://academic.research.microsoft.com/
6) ISI Web of Science - http://www.isiknowledge.com

C. Step 3: Selection of Articles

The selection of data was done taking into account the reading of the abstract of the articles and their relationship with the theme.

Useful information from of each article was collected to answer the main question of this article. The information obtained in the readings of each article was systematized according to the criteria established for the selection.

The study selection criteria were defined as follows:

1) Inclusion criteria

- Relation to the subject;

- Publications in magazines and periodicals;
- Studies published from 2000 on.

2) Exclusion criteria

- • No relation to the theme;
- • Duplicate existing products in different versions. The most complete versions of these studies were included.

D. Step 4: Data Collection

The information extracted from the selected articles was transferred to a table containing the attributes in separate columns. These attributes are: name of the publication, description, database, institution, year of publication, quotes, results, methods and algorithms used. Among these attributes, the most relevant for this study are the methods used in the research and the algorithm (s) used in data mining. Data collection was organized according to the following table I.

TABLE I: DATA COLLECTION FOR THE EXPLORATORY STUDY

Attribute	Description
Publication name	Article Title
Description	Proposal Article
Data Bases	Data Base where the study was published
Institution	Institution(s) which the authors are linked
Country	Country(ies) where the author of each institutions
Publication Year	Year in which the article was published
Citations	Number of citations
Result	Result(s) that the article generated
Methods	Methods used in the research
Algorithm	Algorithm(s) used in the study

E. Step 5: Study and Presentation of Data

The studies selected for the collection were analyzed in order to obtain data that could answer the question of this study in a very convincing way. The data obtained from these studies were arranged and grouped so as to generate graphics that could present the results of more simplified form.

F. Step 6: Interpretation of Results

The results at the end of the exploratory study helped to identify which are the algorithms and methods used in Data Mining in Education. All data in Articles, institutions, authors of publications references were used in this study according to the information published and contained in each article.

III. RESULTS AND DISCUSSION

After the filtering process of the selected articles, 34 articles remained. These 34 articles were the basis for this study and were used to answer the questions of the research.

Repeated articles were excluded in the process of filtering the items selected - which appeared in different databases. Although some articles could be dealing with the subject of the research, they were irrelevant for this study because its contents were not targeted to meet the objectives thereof.

The relationship between found, selected, removed, and included articles, were found in Table II.

TABLE II: SELECTED ARTICLES INCLUDED FOR THE STUDY

	Key-Words: Data Mining and Education			
	Found Articles	**Selected Articles**	**Excluded Articles**	**Included Articles**
ACM Digital Library	17	10	2	8
ScienceDirect	25	14	3	11
IEEE Xplore	85	19	11	8
Google Scholar	136	35	29	6
Microsoft Academic	107	20	19	1
ISI Web of Science	8	4	4	0
TOTAL	**378**	**102**	**68**	**34**

From the 34 articles selected for obtaining the data, 11 articles (32.35%) were removed from the base Science Direct -http: //www.sciencedirect.com/. We can highlight the ACM Digital Library -http bases: //portal.acm.org and IEEE Xplore ([6]-[13]) - http://www.ieee.org/web with 8 articles (23.52%) each. Google Scholar - http://scholar.google.com.br/ had selected 6 articles (17.64%). Microsoft Academic ([14]) - http://academic.research.microsoft.com/ had selected 1 article (2,94%) (Fig. 1).

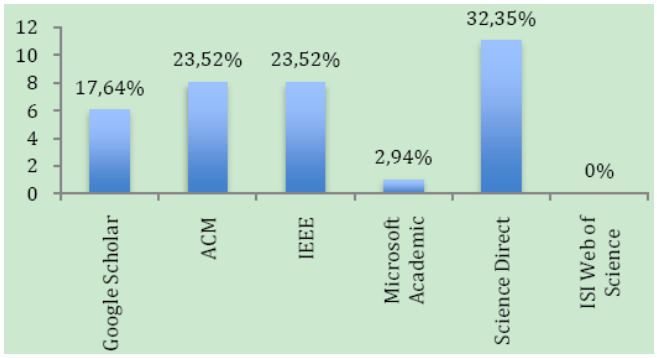

Fig. 1. Selected articles in each database.

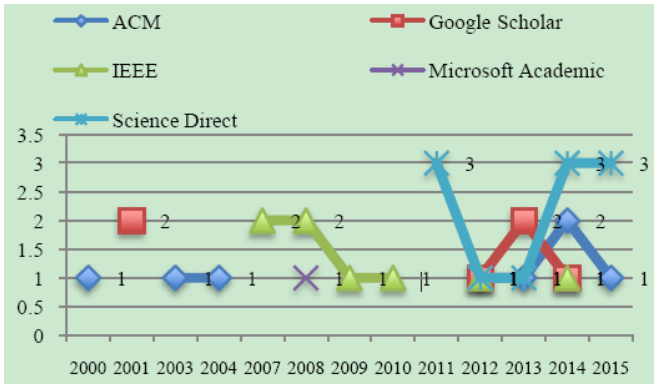

Fig. 2. Number of selected articles in each database over the years.

The extracted articles from the Science Direct base ([15]-[25]), which had the largest number of selected works (11 in total), were published between 2011 and 2015 (Fig. 2).

The Google Scholar bases - with six articles selected for this research ([26]-[31]) - and ACM ([32]-[39]), with 8 items, had their works published between 2000 and 2004 and

between 2012 and 2015.

Another evident point was the fact that the works taken from the ACM and Google Scholar databases were the most cited publications (Fig. 3), while in Science Direct base - which had the largest number of articles selected for this study - had a greater number of articles with fewer citations.

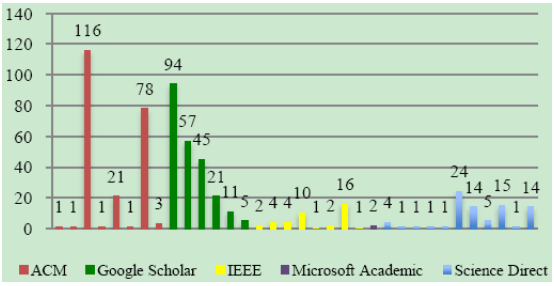

Fig. 3. Number of citations for each article with their databases.

Fig. 4 shows a more detailed result, listing the number of citations of articles published over the years.

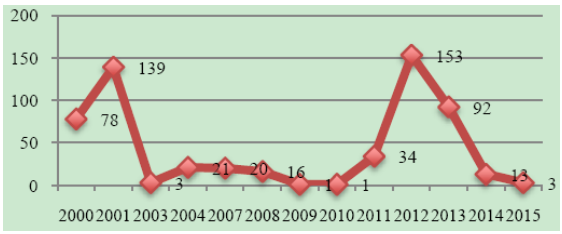

Fig. 4. Number of citations over the years.

The largest number of citations occurred in 2001 and 2012 (139 and 153, respectively). Other items used in this study published in the years 2000 and 2013 were also well cited (Fig. 4).

In this study, several data mining algorithms have been used to assist in modeling the data. Among them, four algorithms were more frequent for this study: C5.0, C4.5, K-Means, and Classification and Regression Trees (CART). The C5.0 is an evolution of C4.5. The C4.5 algorithm is an optimized version of Iterative Dichotomiser (ID3) [40]. The algorithm C5.0, C4.5, J48, and CART, are a decision tree method, and were used by 45% of the selected studies (Fig. 5).

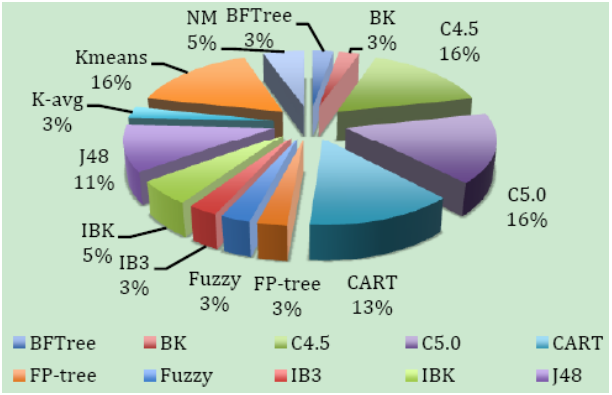

Fig. 5. Algorithms used for data mining in education.

The well known algorithms C4.5, C5.0 and K-means can be highlighted for data mining in education. The methods of Classification and Regression Trees (CART) and J48 (JAVA implementation of decision tree based on C4.5 algorithm) are also well known methods, and succeeded in considerable numbers for this survey: CART was used in 13% of articles, and J48 appeared in 11% [41].

In this study, the C5.0 algorithm peaked in 2012, used in three publications. But the K-means was present in works published between 2007 and 2015.

Recently, J48 C5.0, C4.5, K-means and CART have long been used in Data Mining. The C5.0 algorithm was the most used in 2012 (Fig. 6).

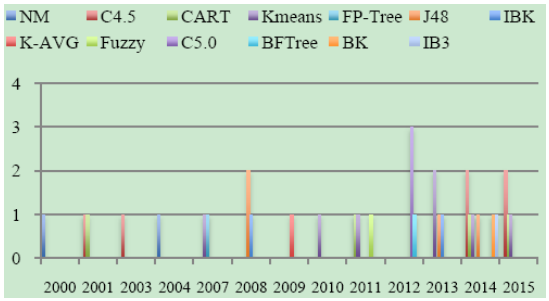

Fig. 6. Algorithms used in this study over the years.

IV. Conclusion

Results showed that different algorithms and methods are used in data mining for education. Many of these algorithms are essential for classification and data control on a large scale.

The algorithms C5.0, C4.5, and K-means, were highlighted methods in this study. For 48% of the articles, these three mining algorithms are most commonly used to analyze data in large scale, especially in education.

Decision tree was the main method applied, with 45% of the algorithms applied, probably because of the graphic knowledge representation, which could help experts to better interpret the elicited evidences, and to postulate causes and effects.

Appendix

Articles included in the exploratory research.

References

[1] C. O. Camilo and J. C. da Silva, "Data mining: Concepts, tasks, methods and tools," Institute of Computer Science Federal University of Goiás, Technical Report, 2009, p. 4.

[2] C. Romero, S. Ventura, M. Pechenizky, and R. Baker, *Handbook of Educational Data Mining*, DFL: Chapman and Hall/CRC Press, 2010.

[3] D. Hand, H. Mannila, and P. Smyth, *Principles of Data Mining*, MIT Press, 2001, p. 53.

[4] D. L. Olson and D. Delen, *Advanced Data Mining Techniques*, Springer, 2008, p. 3-8.

[5] A. A. Castro, H. Saconato, F. Guidugli, and O. A. C. Clark. (2002). Course Online and Systematic Review. Sao Paulo: LED Saconato Meta-Analysis. [Online]. Available: http://www.virtual.epm.br/cursos/metanalise

[6] R. S. Baker and T. College, "Educational data mining: An advance for intelligent systems in education," *IEEE Intelligent Systems*, 2014.

[7] A. Banumathi and A. Pethalakshmi, "A novel approach for upgrading Indian education by using data mining techniques," in *Proc. 2012 IEEE International Conference on Technology Enhanced Education (ICTEE)*, pp. 1-5, 2012.

[8] J. Huang, A. Zhu, and Q. Luo, "Personality mining method in web based education system using data mining," in *Proc. 2007 IEEE International Conference on Grey Systems and Intelligent Services*, pp. 155-158, 2007.

[9] S. L. S. Lihua, Z. Y. Z. Yongsheng, and Z. Z. Z. Zhonglei, "Research on data mining in college education," in *Proc. 2008 International Conference on Computer Science and Software Engineering*, vol. 5, pp. 385-388, 2008.

[10] C. S. C. Song and K. M. K. Ma, "Applications of data mining in the education resource based on XML," in *Proc. 2008 International Conference on Advanced Computer Theory and Engineering*, pp. 943-946, 2008.
[11] M. Vranic, D. Pintar, and Z. Skocir, "The use of data mining in education environment," in *Proc. 2007 9th International Conference on Telecommunications*, pp. 243-250, 2007.
[12] F. Wu, "Discussion on experimental teaching of data warehouse & data mining course for undergraduate education," in *Proc. 2012 7th International Conference on Computer Science and Education*, (Iccse), pp. 1425-1429, 2012.
[13] Q. Z. Q. Zhiming and H. W. H. Wei, "Application of combined grey neural network and data mining in information technology education," in *Proc. 2009 Second International Conference on Education Technology and Training*, pp. 163-166, 2009.
[14] V. P. Bresfelean, "Data mining applications in higher education and academic intelligence management," MPRA (Munich Personal RePEc Archive), No. 21235, pp. 214-226, 2008.
[15] R. Campagni, D. Merlini, R. Sprugnoli, and M. C. Verri, "Data mining models for student careers," *Expert Systems with Applications*, vol. 42, March 2015.
[16] M. Chalaris, S. Gritzalis, M. Maragoudakis, C. Sgouropoulou, and A. Tsolakidis, "Improving quality of educational processes providing new knowledge using data mining techniques," *Procedia — Social and Behavioral Sciences*, vol. 147, pp. 390-397, 2014.
[17] K. Dejaeger, F. Goethals, A. Giangreco, L. Mola, and B. Baesens, "Gaining insight into student satisfaction using comprehensible data mining techniques," *European Journal of Operational Research*, vol. 218, no. 2, pp. 548-562, 2012.
[18] A. P. Alfiani and F. A. Wulandari, "Mapping student's performance based on data mining approach (A case study)," *Agriculture and Agricultural Science Procedia*, vol. 3, pp. 173-177, 2015.
[19] S. Natek and M. Zwilling, "Student data mining solution-knowledge management system related to higher education institutions," *Expert Systems with Applications*, vol. 41, no. 14, pp. 6400-6407, 2014.
[20] A. Peña-Ayala, "Educational data mining: A survey and a data mining-based analysis of recent works," *Expert Systems with Applications*, vol. 41, no. 4-1, pp. 1432-1462, 2014.
[21] B. Sen and E. Ucar, "Evaluating the achievements of computer engineering department of distance education students with data mining methods," *Procedia Technology*, vol. 1, pp. 262-267, 2012.
[22] B. Şen, E. Uçar, and D. Delen, "Predicting and analyzing secondary education placement-test scores: A data mining approach," *Expert Systems with Applications*, vol. 39, no. 10, pp. 9468-9476, 2012.
[23] N. A. Shukor, Z. Tasir, and H. van der Meijden, "An examination of online learning effectiveness using data mining," *Procedia — Social and Behavioral Sciences*, vol. 172, pp. 555-562, 2015.
[24] C. H. Weng, "Mining fuzzy specific rare itemsets for education data," *Knowledge-Based Systems*, vol. 24, no. 5, pp. 697-708, 2011.
[25] W. Xing, R. Guo, E. Petakovic, and S. Goggins, "Participation-based student final performance prediction model through interpretable genetic programming: Integrating learning analytics, educational data mining and theory," *Computers in Human Behavior*, vol. 47, pp. 168-181, 2015.
[26] R. B. Bhise, S. S. Thorat, and A. K. Supekar, "Importance of data mining in higher education system," *Journal of Humanities and Social Science*, vol. 6, no. 6, pp. 18-21, 2013.
[27] W. Hämäläinen *et al*. (2004). Data mining in personalizing distance education courses. *World Conference on Open Learning and Distance Education*. [Online]. *16*. Available: http://www.cs.helsinki.fi/u/whamalai/articles/viscosproj.pdf
[28] L. Jing, "Data mining and knowledge management in higher education," in *Proc. Workshop Associate of Institutional Research International Conference, Toronto*, pp. 1-18, 2002.
[29] C. Romero and S. Ventura, "Data mining in education," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 3, no. 1, pp. 12-27, 2013.
[30] S. K. Yadav, B. Bharadwaj, and S. Pal, "Mining education data to predict student's retention: A comparative study," vol. 10, no. 2, pp. 5, 2012.
[31] E. Yukselturk and C. Education, "Predicting dropout student: An application of data mining methods in an online education program," *European Journal of Open, Distance and e-Learning*, vol. 17, no. 1, pp. 118-133, 2014.
[32] T. Calders and M. Pechenizkiy, "Introduction to the special section on educational data mining," *ACM SIGKDD Explorations Newsletter*, vol. 13, no. 2, pp. 3, 2012.
[33] A. Hajra, D. Birant, and A. Kut, "Improving quality assurance in education with web-based services by data mining and mobile technologies," in *Proc. the 2008 Euro American Conference on Telematics and Information Systems*, pp. 1-7, 2008.
[34] I. Jormanainen, "An open approach for learning educational data mining," *ACM International Conference Proceeding Series*, pp. 203-204, Nov. 2013.
[35] A. London and T. Németh, "Student evaluation by graph based data mining of administrational systems of education," in *Proc. 15th International Conference on Computer Systems and Technologies*, pp. 363-369, 2014.
[36] Y. Ma, B. Liu, C. Wong, P. Yu, and S. Lee, "Targeting the right students using data mining," *Discovery and Data Mining*, pp. 457-464, 2000.
[37] A. H. M. Ragab, A. Y. Noaman, A. S. Al-Ghamdi, and A. I. Madbouly, "A comparative analysis of classification algorithms for students college enrollment approval using data mining," in *Proc. the 2014 Workshop on Interaction Design in Educational Environments (IDEE '14)*.
[38] G. Siemens and R. S. J. Baker, "Learning analytics and educational data mining: Towards communication and collaboration," in *Proc. the 2nd International Conference on Learning Analytics and Knowledge*, pp. 252-254, 2012.
[39] S. E. Sorour, "Correlation of topic model and student grades using comment data mining," pp. 441-446, 2011.
[40] X. Wu, V. Kumar, J. R. Quinlan , J. Ghosh, Q. Yang, and H. Motoda, *Top 10 Algorithms in Data Mining*, Springer-Verlag London Limited, 2007.
[41] S. R. Librelotto and P. M. Mozzaquatro, "Mining algorithms analysis J48 and Apriori applied to detect the quality of life and health indicators," *Interdisciplinary Journal of Teaching, Research and Extension*, vol. 1, 2013, p. 27.

Ancelmo Frank Coêlho Castro was born in São João do Piauí, Brazil on September 30, 1983. Mr. Castro teaches at Institute Federal do Tocantins (IFTO) in Araguatins-TO since 2011 year. He completed his graduation in information systems at Federal Institute of Piauí (IFPI), in 2008 year and his specialization in networking and system security, at INTA-CE College, in 2010 year.

He is currently coordinating the technical course in networks (IFTO) and is a student of the master degree in computational model at Federal University of Tocantins (UFT). His research interests are education, data mining and system development.

Leandro Garcia graduated in biological sciences medical modality (biomedicine), Federal University of São Paulo in 1999 year and a Ph.D. in biological sciences (cell and molecular biology) by cellular and molecular biology department (CEL) at the University of Brasilia in 2004 year.

He has experience in bioinformatics, with emphasis in molecular biophysics, acting on the following topics protein folding, hydrophocicity, energy function, and biochemistry. Also he has experience in bioinformatics with an emphasis on development of virtual learning environments for health care courses. He is currently associate teacher in medical course at Federal University of Tocantins (UFT) and participates as a mastermind teacher in a master in computational modeling at UFT.

David Nadler Prata was born in Goiânia, Brazil on September 18, 1965. Dr. Prata completed his bachelor of computer science in 1992. Then on, he went to complete his specializing in academician. He worked as a system analyst to Tocantins Government, being in charge for the accountability and financial systems. Later, he successfully completed his master degree in computer science from Campina Grande Federal University with application research in education in 2000 year. He coordinated graduate and undergraduate courses in computer science at Alagoas Faculty in Maceio, Brazil. He was allotted to Federal University of Alagoas in 2006. Then, he moved to Federal University of Tocantins. His doctoral was developed in part at Carnegie Mellon University, USA, completed in 2008. He is currently coordinating a master degree in computational model. His research interests are education and ecosystems.

Marcelo Lisboa holds a degree in computer science from the Catholic University of Petrópolis (1994), the masters in Computer from Fluminense Federal University (1997), master science in electrical engineering from the Federal University of Rio de Janeiro (1999) and Ph.D. in electrical engineering from the Federal University of Rio January (2008).

He is currently reviewer of *INFOCOMP Journal of Computer Science*, and associate teacher 4 from Federal University of Tocantins. He has experience in the area of Computer Science, mainly in the following topics metaheuristics, combinatorial optimization, mathematical programming, computer networking and high performance computing.

Monica Nadler Prata concluded her doctoral in literary studies at the University of North Carolina (USA) 2002. Currently, she works at the State Department of Education of the Federal District on the Board of Human Rights, Brazil.