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.

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

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

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.

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.](image)

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.](image)

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.](image)

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.](image)

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

S. K. Yadav, B. Bharadwaj, and S. Pal, "Mining education data to...

C. Romero and S. Ventura, "Data mining in education,...


T. Calders and M. Pechenizkiy, "Introduction to the special section on educational data mining."


S. R. Librelotto and P. M. Mozaquarro, "Mining algorithms analysis 14th and Apriori applied to detect the quality of life and health indicators," Interdisciplinary Journal of Teaching, Research and Extension, vol. 1, 2013, p. 27.

Auncelo Frank Coelho 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 (IFI), in 2008 year and his specialization in networking and system security, at INTA-CE College, in 10 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 1990 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, hydropathy, 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 specialization 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.