ISSN: 2321-7782 (Online)
Volume 2, Issue 7, July 2014
International Journal of Advance Research in Computer Science and Management Studies
Research Article / Survey Paper / Case Study
Available online at: www.ijarcsms.com

Data Mining Algorithms and their applications in Education Data Mining

Shafiq Aslam[1]
IT Department
University of Gujrat
Gujrat – Pakistan

Imran Ashraf[2]
Lecturer, IT Department
University of The Punjab
Gujranwala – Pakistan

Abstract: ***Data mining is efficiently used to extract potential patterns and associations for discovering the hidden knowledge from data that is collected from different sources. In this review emphasis is put on data mining algorithms used in field of Education mining, to highlight the need and consequently the application of data mining in this field. An importance of extracted knowledge in education domain has been discussed. Algorithms; C4.5, SVM, EM, PageRanker, Naïve Bayes, Apriori and CART are important. This paper presents the significance of use of these algorithms in Education field. These algorithms have been ranked high by IEEE International Conference on Data Mining in 2006.***

Keywords: Data mining; Apriori; CART; PageRanker; Naive Bayes; SVM.

I. INTRODUCTION

A lot of research is going on in the field of data mining algorithms. These algorithms are used to extract knowledge out of data in order to devise new and innovative strategies. This applies to machine-learning process as well. Research institutions use mining to improve their managerial standard and make decision making effective. For these reason algorithms in data mining has such an important role to play [1, 25]. Education data mining is a major research field also known as EDM. It aims at devising and using algorithms to improve educational results and explain educational strategies for further decision making.

This paper discusses some of the data mining algorithms to education related areas. These algorithms are applied to discover knowledge from educational data and study those attributes which can contribute to higher the performance [1, 26]

II. DATA MINING

Data Mining [1] is the field of computer science that intents to find out different potential factors and patterns to help make decisions. Data mining focuses on different fields as, databases machine learning [1, 2] and machine intelligence (known as artificial intelligence) which are very fast developing fields.

A. Dtata Mining Tasks

There are a few tasks perform by data mining [5]:

1- Data Analysis: These techniques are interactive and visual for exploring data without any clear idea.

2- Descriptive Modeling: gives picture of data and includes models these models tell the relationship between different objects.

3- Predictive modeling: It is prediction of unknown values of different variables from different known variables.

4- Discovering Patterns and Rules: It means to spot behaviors like fraud detection, instant change in behavior by transaction [2, 28].

5- Content retrieval: It is finding pattern- pattern of interest, commonly used for text mining.

B. *Dtata Mining Methods*

Data mining methods [27] are broadly categorized as:

1- Classification

2- Clustering,

3- Association Rule mining

4- Temporal Data Mining

5- Time Series Analysis

6- Web Mining [4]

These methods are used differently to categorize data.

C. *Dtata Mining Models*

Data mining Models [7] are:

1- predictive

2- descriptive

Predictive model [6, 7] are used to predict about unknown values from known values of different variables, like Classification, Prediction, Time series Analysis [4] etc. Descriptive model [6,7] highlights pattern or relationship in examined data, like Clustering, Summarization, Sequence discovery [4] etc.

III. USE OF ALGORITHMS AND EDUCATIONAL MINING

" Educational Data Mining " (E D M) is a new and big discipline, focusing on the research techniques to search and make data that is retrieved from different places and sources; like data warehouse and flat files. Following algorithms have been used in educational data to make various calculations extensively.

TABLE I
Algorithms and Their Percentage Usage in Data Mining

Algorithms	**Usage**
Clustering	62%
Decision Trees, Ruler 180	59%
Regression 163	57%
Statistics 150	47%
Visualization 135	40%
Support Vector (SVM) 92	38%
Text Mining	27%

IV. PREVIOUS WORK

Even being a recent research field, Education data mining has enough work done. Education data mining has become a potential strength of the educational institutions. Romero and Ventura [14], have had a survey on educational data mining between 1995 and 2005 concluding that educational data mining is a promising area of research and it has a specific requirements not presented in other domains. Thus, work should be oriented towards educational domain of data mining.

There are different domains of DM;

a- Statistical

b- Web mining

1- Clustering, Classification[15]

2- Association rule mining

3- Sequential pattern mining [14,15]

A second view point on education data mining is given by Baker [20]

1- Classification

2- Regression

3- Density estimation

a- Relationship mining

1- Association rule

2- Correlation

3- Sequential patteren

These types are famous among the researchers in Data mining rest two good and simple categories from a data mining perspective [38-42].

V. IMPORTANT ALGORITHMS IN DATA MINING

1. **Classification algorithms** – for predicting one or more discrete variables, based on the other attributes in the dataset
2. **Regression algorithms** – to predict continuous variables, such a profit to loss, based on other attributes in the dataset
3. **Segmentation algorithms** – to divide data to groups, or clusters properties.
4. **Association algorithms** – finding correlations between different attributes in a dataset.
5. **Sequence analysis** – to summarize sequences or episodes in data, as a Web path

A. Prominent Papers for Algorithms

To investigate the development and history of EDM, one way is to look at the articles and related papers. Citations for the selected papers are taken from Google Scholar. Table1.Top 5 most cited papers.

TABLE III
Most Sited Papers for Algorithms

Article	Citation	Rating
Zaiane, O Web usage mining for a better web learning	110	1st
Zaiane, O Building a recommended agent for e-learning	89	2nd

VI. APPLICATIONS OF ALGORITHMS IN EDUCATION MINING

Number of universities and students is increasing day by day; we think that data mining technology can help improving the education standard and consequently causing high ratio of successful candidate, low ratio of students' drop-out and maximizing education system efficiency [37]. Following is a detail of the algorithms used in education mining.

A. C 4.5

A classifier system takes input from the cases described by values and attributes and output a classifier that can accurately predict classes of new cases. C 4.5 is a descendant of CLS [39] and IDE [39] ,creates classifier and generated decision tree. It can also make classifier in most comprehensive rule-set forms.

1. *Use in Web Education*

To improve courses, the relationships are discovered and association patterns are searched out among the used data picked up during students' learning sessions [32]. This data can prove to be useful for authors and teachers of the respective courses who are to decide what modifications will be helpful and appropriate to improve the effectiveness of the course [10].

TABLE IIIII
Data mining questions in the business sector and their higher researched equivalents using C4.5 algorithm

Business Questions	**Higher Researched Equivalents**
Who are the valued buyers?	Which object is taken mostly?
Who are repetitive buyers?	Which objects are returned?
Who are deviated buyers?	Which customers are "persiters"?
Which customers are likely to defect?	Which courses are like to attract the customers

TABLE IVV
Classification of Data

Name	**Gender**	**Height (m)**	**Output A**	**Output B**
James	M	3	tall	Medium
John	M	2	Tall	Medium
Magy	F	1.7	Short	Medium
Marathon	F	1.8	Medium	Tall
Stephine	F	1.85	Medium	Tall
Wentworth	M	2.2	Tall	Tall
whythe	F	1.75	Medium	medium

Above table defines classification of data into different height groups, such as similar data occurs in one class or category.

2. *Creating a Decision Tree*

TABLE V
Data for making decision tree, deciding best attribute

Appearance	**Hot**	**Air pressure**	**Suny**	**Action**
Light	Hot	Low	Right	False
Light	Cold	High	Right	Yes
Cloudy	Cold	Low	Not right	No
Airy	Medium	High	Not right	No
Cloudy	Cold	High	Right	Yes
Sunny	Hot	Not normal	Right	Yes
Rain	Cold	Normal	Right	yes

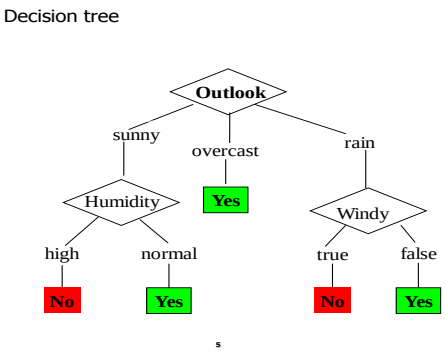

Fig. 1 a tree structure representing set of decisions.

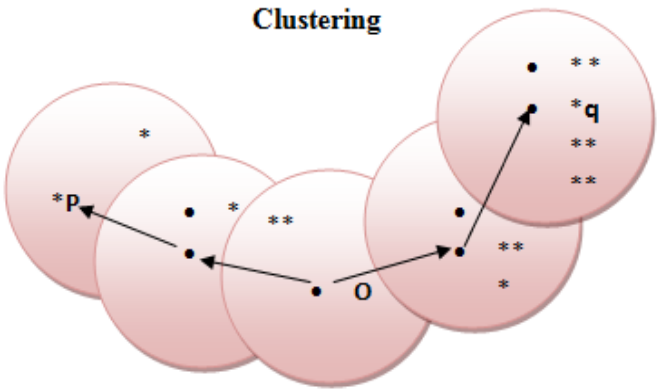

Fig. 2 Density based clustering methods

B. *Support Vector Machine (SVM)*

Support Vector Machine (S.V.M) [36] is considered an efficient tool to train data. If offers accurate methods among algorithms. SVM is the most worked upon algorithm for training purposes and a lot of research is still going on. SVM can find classification function in two- class learning tasks. The metric for"best" can be realized geometrically. It is good because of its generalization ability.

a- To perform linear regression

b- Rank elements

C. *Apriori*

A popular way to find frequent data item sets [30] is by comparing explosion. After data items are obtained then we can easily generate association rules.

There are different steps for that.

1- Generate Ck+1 for item sets of k+1

2- Calculate support

3- Put items for minimum support for fk+1

D. *The Expectation Maximization Algorithm*

It provides a flexible mathematical approach to modelling and clustering of data on randomly observed basis. This can be used to cluster continuous data. Expectation Maximization algorithm is used to model distribution of random phenomenal data.

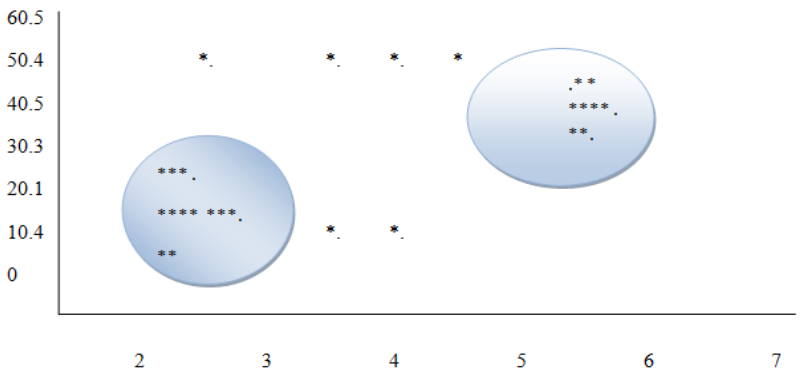

Fig. 3 EM clustering of Old Faithful Data

E and M parts EM algorithms give (ML) estimation of normal components. On (k+1) turn of EM, log of the data is taken.

E. Page Ranker

Page Ranker [10] was given forth by Brin Karruy Page in 1998. On this algorithm's basis they built Google, which has an excellent success ratio. It produces a static ranking of different web pages in sense that pager value is determined offline and does not depend on the online queries.

Pager Ranker formula:

1- A hyperlink point the value of the page is an implicit conveyance for authority. Thus more links a page receives more prestige it has.

2- Pages that go to I is also considered good

Page Rank algorithm given by Lawrence Page and Sergey Brin is in a lot of publications. It is as under…

PR (A) = (1-d) + d (pr (ti) / C (ti) + …..... + PR (Tn) /C (tn))

Where

- PR(A) is PageRank of page A
- PR (Ti) is PageRank
- C (Ti) is number of outbound
- d a damping

VII. CONCLUSION

Data mining is a tremendously vast area that includes employing different techniques and algorithms for pattern finding. The algorithms discussed in this paper are the ones used in education mining. These algorithms have shown a remarkable improvement in strategies like course outline formation, teacher student understanding and high output and turn out ratio. ICDM conference encourages employment and development of algorithms helpful in data mining. An appreciable research is still being done on various algorithms. I hope this review paper appreciates the current algorithm researchers and inspires the new ones to explore further.

References

1. IBM Research .Knowledge Discovery and Data Mining.IBM Corp. [online].Available: http://domino.research.ibm.com/comm/research.nsf/pages/r.kdd.html
2. Botia, J.A., Garijo, M.y Velasco ,J.R., Skarmeta, A.F., "A Generic Data mining System basic design and implementation Guidelines" A technical Project Report of CYCYT project of Spanish Government .1998.website: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.53.1935
3. Last, M.,and Kandel , A., 2001 "Data mining for Process and Quality Control" Kluwer Academic ,page 179-205
4. Chapman , P.,Clinton, J., Kerber, R., Khabaza, T., Reinzrtz, T., Shearer, C. and Wirth , R., "CRISP-DM 1.0 : Step by Step data mining guide,NCR Systems Engineering Copenhagen (USA and Denmark), Daimler Chrysler AG (Germany) , SPSS Inc. (USA) and UHRA Verzekeringenen Bank Group B.V (The Netherlands),2000".
5. Lasore,D.T., "Discovering Knowledge in Data: An introduction to Data Mining", ISBN 0-471-66657-2, John Wiley & Sons, Inc,2005.
6. Fayyad, U., Piatetsky-Shapiro, G., and Smyth P., "From Data Mining to Knowledge Discovery in Databases," AI Magazine ,American Association for Artificial Intelligence,1996.
7. [7]Bernstein , A and Provost, F., "An intelligent Assistant for Knowledge Discovery Process". Working paper of the Center for Digital Economy Research ,New York University and also Presented at the IJCAI 2001 Workshop on Wrappers for Performance Enhancement in Knowledge Discovery in Databases.
8. Halteren. H. V., Oostdijk N., "Linguistic profile of texts for the purpose of language verification, The ILK research group, Tilburg centre for Creative Computing and the department of Communication and Information Sciences of faculty of Humanities, Tilburg University ,The Netherlands.
9. Antonie, M. L., Zaiane, O. R., Coman, A., "Application of Data Mining Techniques for Medical image Classification", Proceedings of The Second International Workshop on Multimedia Data Mining (MDM/KDD2001) in conjuction with ACM SIGKDD conference ,San Francisco, August 26,2001.
10. Romero, C., Ventura, S. and De-Bra, P. "Knowledge Discovery with Genetic Programming for Providing Feedback to Courseware Authors, Kluwer Academic Publishers, Printed in the Netherlands, 30/08/2004."
11. Anjewierden, A., Koll offel, B., and Hulshof C., "Towards educational data mining :using data mining methods for automated chats analysis to understand and support inquiery learning process". International Workshop on Applying Data Mining in e-learning. ADML'07. Vol-=305, page no23-32issi, Lassithi Crete Greece, 18 September, 2007.
12. Kumar, V. (2011). "An Empirical Study of the Applications of Data Mining Techniques in Higher Education." IJACSA - International Journal of Advanced Computer Science and Applications, 2(3), 80-84. Retrieved from http://ijacsa.thesai.org.

13. Jahdhav. S. R., and Kumbargoudar, P., "Multimedia Data Mining in Digital Libraries: Standards and Features READIT 2007, pp 54-59
14. Romero, C. and Ventura, S. (2007) 'Educational data Mining: A Survey from 1995 to 2005', Expert Systems with Applications (33), pp. 135-146
15. Mansur, M. O. , Sap, M. and Noor , M. (2005) 'Outlier Detection Technique in Data Mining: A Research Perspective', In Postgraduate Annual Research Seminar.
16. Heikki, Mannila, "Data mining: machine learning, statistics, and databases", IEEE, 1996.
17. Alaa el-Halees, "Mining students data to analyze e-Learning behavior: A Case Study", 2009.
18. Han, J. and Kamber, M., "Data Mining: Concepts and Techniques, 2nd edition." The Morgan Kaufmann Series in Data Management Systems, Jim Gray, Series Editor,2006.
19. El-Halees, A. "Mining Students Data to Analyze Learning Behavior: A Case Study", The 2008 international Arab Conference of Information Technology (ACIT2008) –Conference Proceedings, University of Sfax, Tunisia, Dec 15- 18.
20. Mansur, M. O. , Sap, M. and Noor , M. "Outlier Detection Technique in Data Mining: A Research Perspective", In Postgraduate Annual Research Seminar 2005.
21. J. R. Quinlan, "Introduction of decision tree: Machine learn", 1: pp. 86-106, 1986.
22. B.K. Bharadwaj and S. Pal. "Data Mining: A prediction for performance improvement using classification", International Journal of Computer Science and Information Security (IJCSIS), Vol. 9, No. 4, pp. 136-140, 2011.
23. Shaeela Ayesha, Tasleem Mustafa, Ahsan Raza Sattar, M. Inayat Khan, "Data mining model for higher education system", Europen Journal of Scientific Research, Vol.43, No.1, pp.24-29, 2010.
24. Galit.et.al, "Examining online learning processes based on log files analysis: a case study". Research, Reflection and Innovations in Integrating ICT in Education.
25. Romero, C., Ventura, S. and Garcia, E. (2008) 'Data mining in course management systems: Moodle case study and tutorial', Computers & Education, vol. 51, no. 1, pp. 368-384.
26. Kumar, V. and Chadha, A. (2011) "An Empirical Study of the Applications of Data Mining Techniques in Higher Education", International Journal of Advanced Computer Science and Applications, vol. 2, no. 3, pp. 80-84.
27. J. Han and M. Kamber, "Data Mining: Concepts and Techniques," Morgan Kaufmann, 2000.
28. Sheikh, L., Tanveer, B. and Hamdani ,S. (2004) "Interesting Measures for Mining Association Rules", IEEE-INMIC –Conference Proceedings.
29. Vashishta, S. (2011). Efficient Retrieval of Text for Biomedical Domain using Data Mining Algorithm. IJACSA - International Journal of Advanced Computer Science and Applications, 2(4), 77-80
30. U. K. Pandey, and S. Pal, "A Data mining view on class room teaching language", (IJCSI) International Journal of Computer Science Issue, Vol. 8, Issue 2, pp. 277-282, ISSN:1694-0814, 2011

