Data Mining in Education- An Experimental Study

Dina Abdulaziz AlHammadi
PhD Student
College of Computer and Information Sciences
King Saud University
Riyadh, Saudi Arabia

Mehmet Sabih Aksoy
Professor
College of Computer and Information Sciences
King Saud University
Riyadh, Saudi Arabia

ABSTRACT
Data mining first proved to be beneficial to business related fields such as marketing and consumer related service enhancements. Slowly it has made its way toward other fields such as medicine, science, engineering, and education. The focus of this paper is to review several applications of data mining in education and their benefits, present some classification techniques, test some sample data, and then evaluate them against some selected criteria.

General Terms
Data mining

Keywords
Data mining, education, inductive learning

1. INTRODUCTION
Data mining became an essential asset to many disciplines ranging from military, business intelligence and marketing, science, engineering, and education. These fields make use of data mining by analyzing large data sets and extracting knowledge that is meaningful and beneficial for future use.Data mining can be used in two ways: descriptive, and predictive.Descriptive means extracting patterns and finding trends. While predictive, means is trying to predict a future scenario, or the outcome of certain values in random situations. It has been applied in predicting customer purchases in supermarkets based on associated items. And it would be interesting to experiment with data mining in education. There are many opportunities ready to be investigated such as advising students which courses they should register if they are hoping for a high GPA. It can be used to recommend the best majors for each student based on their courses, and the achievements of past students. Also, identifying student's best skills and suggesting suitable courses, of finding weak skills and advising appropriate extra courses or tutoring. Even gathering national placement exam records and predicting student performance in college. The possibilities for data mining in education and the data to be reaped are endless. In this paper, the research that has been applied to education related environments will be explored such as course management, national placement tests, education improvements, student's success or failure in courses, and so on.

Fig1: DIKW Pyramid

As it can bee seen in Figure1, Data is known facts but hold no meaning until processed into information. Through data mining information is retrieved and knowledge is extracted that can benefit the education field greatly. As knowledge is enhanced wisdom is gained. However,as the researcher moves up the pyramid he will find it difficult to program or write an algorithm. The main text of this study consists of six sections. Section 2 describes related work. Section 3 briefly explains data mining techniques. Section 4 shows an experimental study of data mining in education field. In Section 5 the results obtained by applying six different algorithms to the same data is discussed. Finally, the last section is the conclusion.

2. RELATED WORK
In 2007, Vranice et al.[1] explore several data mining algorithms on students' data on a certain first year course in Croatian university and see if they can successfully predict the success of the students attending the same courses next year. They tried three different implementation algorithms: clustering, association rules, and visual exploratory analysis. Their results were not very different, however they feel that their data set is still small, and perhaps in the future they would try to test a large data set and also include more detailed information of students to raise prediction success.

In India, where Banumathi and Pethalakshmi [2] present a novel approach for upgrading Indian Education through data mining techniques. They based their idea on providing a good foundation on school education which leads to good foundation for higher education. They applied the clustering technique and predict student's behavior which increases the literacy rate of the nation. They developed the UCAM clustering algorithm for numeric data. It is based on manipulating the threshold to reduce the overhead of fixing

cluster size which occurs in K-Means, it improves the scalability and decreases the clustering error.

In a South African university, Ehlers et al. [3], proposed a decision support system for research management in higher education. Although they were not able to use the text of the research, instead only the metadata can be mined, although little, some beneficial knowledge was extracted. They were still able to successfully identify the research focus, intensity, and synergy.

Data mining was also applied for predicting academic trends and patterns as explained by Parack et al. in [4]. They applied two techniques: K-Means (clustering) and Apriori (association rules) for classifying students' profiles and groups. The authors were successful in grouping students and identifying patterns and behaviors.

An interesting paper [5] uses data mining to predict student's GPA and academic dismissal in a Learning Management System (LMS). Nasiri, Vafaei and Minaei tested two algorithms: regression analysis for predicting GPA, and C5.0 for predicting academic dismissal. However, due to the algorithms dependability on data distribution, any small variation in data may lead to a different conclusion. And that is why the authors encouraged their enhancement by including association rules.

In Malaysia, Wook et al., presented a paper which predicts student's academic performance in a computer science department at the National Defense University of Malaysia (NDUM) [6]. The authors use two different techniques: Artificial Neural Networks (ANN) and the hybrid of clustering and decision trees.

Shi et al., proposed data association mining technology in managing university curriculum [7]. They used the Apriori algorithm to extract the item sets, then the association rules to extract the knowledge rules. They concluded that, if a student were successful at certain courses, this would influence their success in other courses, such as the relation between mathematics and physics.

Again in India, Bunkar et al., presented a paper where they applied data mining techniques to predict the performance improvement of graduate students using classification [8]. Three techniques are discussed: ID3, C4.5, and Classification And Regression Trees (CART). The authors were able to identify students that are most likely to fail and give them proper counseling and guidance.

3. DATA MINING TECHNIQUES

3.1. C4.5

Input: an attribute-valued dataset *D*

1. Tree = {}
2. **if***D* is "pure" OR other stopping criteria met then terminate
3. **end if**
4. **for all** attribute a $\mathbf{I}$ D **do**
5. Compute information-theoretic criteria if we can split on *a*
6. **end for**
7. a_{best}= Best attribute according to above computed criteria
8. Tree = Create a decision node that test a_{best}in the root
9. D_v= Induced sub-datasets from *D* based on a_{best}
10. **for all**D_v**do**
11. $Tree_v$ = C4.5 (D_v)
12. Attach $Tree_v$ to the corresponding branch of Tree
13. **end for**
14. **return** Tree

Fig2: C4.5 Algorithm [9]

This algorithm was first introduced by Quilan in 1993 [9][10]. It surpasses ID3 due to the fact that it deals with unknown values, tree pruning, and improving use of continuous attributes.

Similar to ID3 by calculating entropy and information gain for the attribute values,the algorithm is able to decide on the root node and move on to the leafs. C4.5 has evolved into C5.0 the commercial version. Figure 2 explains briefly the C4.5 algorithm.

3.2 K-Means Clustering

Clustering is grouping similar elements together. It is a simple unsupervised learning algorithm. It is based on minimizing the distance between the centroid and the data. The user first inputs the number of clusters and based on the number of attributes and centroid calculation each item is grouped in appropriate cluster[9][11].

3.3. KStar

A lazy classification algorithm first introduced by Cleary and Trigg in 1995 [12]. It is an instance-based algorithm which uses entropy as distance measure. Instance-based Learners are based on comparing the test instance with the training classification. It has the ability to deal with numeric and nominal values as well as missing values.

3.4. RULES-3 EXT

Rule Extraction System was first introduced by Pham and Aksoy in 1995[13] and later improved by many researchers. One the improved versions of RULES family was introduced by Mathkourin 2009 [14]. The algorithm is called RULES3-EXT. It allows the user to control the size of the rule set,it has the ability to deal with large datasets, it can extract general rules, it deals with numerical values, and unknown values. Unlike ID3, it refrains from any complex calculation and focuses on the rule generalization and extraction process. Figure 3 lists the steps followed by RULES-3-EXT classification algorithm.

Step 1.	Define ranges for the attributes which have numerical values and assign labels to those ranges
Step 2.	Set the minimum number of conditions (N_{cmin}) for each rule
Step 3.	Take an unclassified example
Step 4.	$N_c = N_{cmin}$-1
Step 5.	IfN_c<N_a then N_c=N_c+1
Step 6.	Take all values or labels contained in the example
Step 7.	Form objects which are combinations of N_c values or labels taken from the values or labels obtained in Step 6
Step 8.	If at least one of the objects belongs to a unique class then form rules with those objects; ELSE go to Step 5
Step 9.	Select the rule which classifies the highest number of examples
Step 10.	Remove examples classified by the selected rule
Step 11.	If there are no more unclassified examples then STOP; ELSE go to Step 3

Fig 3: RULES-3 EXT Algorithm [14]

3.5. Naïve Bayes

It is a type of supervised learning introduced by Thomas Bayes. It is a classifier that is based on independence assumptions. The training set is provided, and the result is conditional probability tables. However, each attribute is considered separately [9][15].

3.6. Simple Logistics

It is used for building linear logistic regression models. A standard classification tree with logistic regression models for all nodes is built, while pruning some of the subtrees. Also with the use of LogitBoost different perspectives of combination between tree induction and logistic regression can be viewed. LogitBoost uses iterative refinement and adds more variables to build its logistic model [16][17].

4. EXPERIMENTAL STUDY

On the Internetthere is a collection of data repository that serves the purpose of many studies. An interesting data repository was found as well as the application programs for the six techniques mentioned in this paper.The found data is regarding The Irish Educational Transitions [18][19][20][21][22]. The data consisted of six attributes as followed: Gender, Drumcondra Verbal Reasoning Test Score (DVRT), Educational level attained; 11 levels, Leaving Certificates; Taken or not, Prestige score for father's education, and finally the Classification which is the student's school type; secondary, vocational, or primary terminal leaver.

The dataset has 500 instances 250 were chosen for training while other 250 were for testing.

5. DISCUSSION

WEKA [23] was chosen asthe test simulator and tested five different data mining tools, and RULES-3 EXT. Table 1 presents the results of the tests obtained by applying the six selected algorithms to the same training and test examples. FromTable 1, it can easily be seen thatall algorithms except K-means had the worst accuracy (46.8%).Kstar, Simple Logistics and Naïve Bayes functioned almost at the same level of accuracy at 80%, 83.6%, and 82% respectively. However, C4.5 did very well with only three misclassified examples at 98.8% accuracy, and RULES-3 EXT performed best with only one misclassified example and 99.6% accuracy.

C4.5 and Simple Logistics both resulted in a10-leaf tree, and while C4.5 had only 3 misclassified examples, the latter had 41. RULES-3 EXT although extracted twice as many rules, was the most successful in reducing the misclassified examples to only one.

Several education related data were tried, but the results weren't satisfying for all algorithms, the highest being at %50. This has led to more searching for education related data, and successfully landing the Irish Educational Transitions data. It is believed that the variation of the training data has led to the success of the algorithms, excluding K-Means. Inductive learning algorithms rely on the quality of the set, not the quantity. Moreover, it is suggested that the data set should have a lot of possible variations, and no contradicting information such as two similar data may lead to two or more different classifications. This will throw a curve at the learning algorithm where rules will be created to satisfy all possible rules, perhaps even several rules each with different number of conditions.

6. CONCLUSION

After this study the conclusionis that inductive learning algorithms (C4.5 and RULES-3 EXT) work better with this type of data which is a combination of numerical and nominal types. It is interesting to produce future work in education and data mining, and apply RULES-3 EXT as data mining algorithm of choice, however, the difficulty of education mining lies with the data to be harvested and preprocessed for use, as there is a lack in saving the required information for a better mining process.

It is encouragedthat educational institutions should start saving anonymous student data for research purposes, and make it available in data sets. The next step for data mining is grasping the prediction of student success on a national level. Every nation is trying to enhance its education level, and for that purpose, education mining is an essential.

Table 1. Data Mining Technique Evaluation

Algorithm	No.Trainingexamples	No. Testexamples	No.Rules	No.Unclassifiedexamples	No. Misclassifiedexamples	Accuracy (%)
C4.5	250	250	10	0	3	%98.80
K-Means	250	250	NA	0	133	%46.8
K*	250	250	NA	0	50	%80
RULES-3 EXT	250	250	22	0	1	%99.6
Naïve Bayes	250	250	NA	0	45	%82
Simple Logistics	250	250	10	0	41	%83.6

7. ACKNOWLEDGMENTS

The authors would like to thank Research Center in the College of Computer and Information Sciences, King Saud University for its support to complete this study.

8. REFERENCES

[1] Vranic, M., D. Pintar, and Z. Skocir.*The use of data mining in education environment.*In*Telecommunications, 2007.ConTel 2007. 9th International Conference on*. 2007.

[2] Banumathi, A. and A. Pethalakshmi. *A novel approach for upgrading Indian education by using data mining techniques.*In*Technology Enhanced Education (ICTEE), 2012 IEEE International Conference on*. 2012.

[3] Ehlers, K., et al. *A Decision Support System for Institutional Research Management in Higher Education: Data Mining to Determine Research Focus, Intensity and Synergy.* In*Computational Science and Engineering, 2009. CSE '09. International Conference on*. 2009.

[4] Parack, S., Z. Zahid, and F. Merchant. *Application of data mining in educational databases for predicting academic trends and patterns.* In*Technology Enhanced Education (ICTEE), 2012 IEEE International Conference on*. 2012.

[5] Nasiri, M., B. Minaei, and F. Vafaei. *Predicting GPA and academic dismissal in LMS using educational data mining: A case mining.* In*E-Learning and E-Teaching (ICELET), 2012 Third International Conference on*. 2012.

[6] Wook, M., et al. *Predicting NDUM Student's Academic Performance Using Data Mining Techniques.* In*Computer and Electrical Engineering, 2009. ICCEE '09. Second International Conference on*. 2009.

[7] Feng, S., M. Qi, and M. Di. *The application of data association mining technology in university curriculum management.* In*Robotics and Applications (ISRA), 2012 IEEE Symposium on*. 2012.

[8] Bunkar, K., et al. *Data mining: Prediction for performance improvement of graduate students using classification.* In*Wireless and Optical Communications Networks (WOCN), 2012 Ninth International Conference on*. 2012.

[9] Xindong, W. and V. Kumar, *The Top Ten Algorithms in Data Mining*2009: Chapman and Hall/ CRC Press. 232.

[10] Quinlan, J.R., *C4.5: programs for machine learning*1993: Morgan Kaufmann Publishers Inc. 302.

[11] Liu, Y., Z. Bi, and Y. Gao. *Method of assessing operational skill of equipment support based on k-means clustering*. In*Emergency Management and Management Sciences (ICEMMS), 2010 IEEE International Conference on*. 2010.

[12] Cleary, J.G. and L.E. Trigg. *K^*: An Instance-based Learner Using an Entropic Distance Measure*.

[13] Pham, D. and M. Aksoy, *RULES: A simple rule extraction system.*Expert Systems with Applications, 1995. **8**(1): p. 59-65.

[14] Mathkour, H., *RULES3-EXT: Improvements of RULES3 Induction Algorithm.* Math. Comput. Appl, 2009. **15**(3).

[15] John, G.H. and P. Langley. *Estimating continuous distributions in Bayesian classifiers*. In*Proceedings of the eleventh conference on uncertainty in artificial intelligence*. 1995. Morgan Kaufmann Publishers Inc.

[16] Sumner, M., E. Frank, and M. Hall, *Speeding up logistic model tree induction.* Knowledge Discovery in Databases: PKDD 2005, 2005: p. 675-683.

[17] Landwehr, N., M. Hall, and E. Frank, *Logistic Model Trees.* 2004.

[18] Greaney, V. and Kelleghan, T. (1984). *Equality of Opportunity in Irish Schools.* Dublin: Educational Company.

[19] Kass, R.E. and Raftery, A.E. (1993). Bayes factors and model uncertainty. *Technical Report no. 254*, Department of Statistics, University of Washington. Revised version to appear in Journal of the American Statistical Association.

[20] Raftery, A.E. (1988). *Approximate Bayes factors for generalized linear models.* Technical Report no. 121, Department of Statistics, University of Washington.

[21] Raftery, A.E. and Hout, M. (1985). *Does Irish education approach the meritocratic ideal? A logistic analysis.* Economic and Social Review, 16, 115-140.

[22] Raftery, A.E. and Hout, M. (1993). *Maximally maintained inequality: Expansion, reform and opportunity in Irish schools.* Sociology of Education, 66, 41-62.

[23] Hall, M., et al., *The WEKA data mining software: an update.* ACM SIGKDD Explorations Newsletter, 2009. **11**(1): p. 10-18.