Data Mining in Education: Data Classification and Decision Tree Approach

Sonali Agarwal, G. N. Pandey, and M. D. Tiwari

Abstract—**Educational organizations are one of the important parts of our society and playing a vital role for growth and development of any nation. Data Mining is an emerging technique with the help of this one can efficiently learn with historical data and use that knowledge for predicting future behavior of concern areas. Growth of current education system is surely enhanced if data mining has been adopted as a futuristic strategic management tool. The Data Mining tool is able to facilitate better resource utilization in terms of student performance, course development and finally the development of nation's education related standards. In this paper a student data from a community college database has been taken and various classification approaches have been performed and a comparative analysis has been done. In this research work Support Vector Machines (SVM) are established as a best classifier with maximum accuracy and minimum root mean square error (RMSE). The study also includes a comparative analysis of all Support Vector Machine Kernel types and in this the Radial Basis Kernel is identified as a best choice for Support Vector Machine. A Decision tree approach is proposed which may be taken as an important basis of selection of student during any course program. The paper is aimed to develop a faith on Data Mining techniques so that present education and business system may adopt this as a strategic management tool.**

Index Terms—**Data mining, education data mining, data classification, support vector machine, decision tree.**

I. Introduction

As we are growing in terms of population, technology advancements, literary rate and globalization, education systems are also taking a new shape as business systems. Now Data Mining has become universal tool for strategic management in business organizations as well as social system and organizations. Today out of all infrastructural support for the development of society, education is considered as one of the key inputs for social development. This paper analyzes the data available on student's academic record and student likelihood in terms of placement may be predicted on the basis of entrance examination marks, quantitative ability marks and verbal ability marks by using decision tree approach and decision rule approach. Considering the global opportunities coupled with global competition even in case of education it is essential to admit the best students as far as possible. So their academic performance and subsequent placements are best in the world.

Data mining is useful whenever a system is dealing with large data sets. In any education system, student records i.e. enrollment details, course eligibility criteria, course interest and academic performance may be an important consideration to analyze various trends since all the systems are now computer based information system so data availability, modification and updation are a common process now. Data warehousing may be taken as good choice for maintaining the records of past history. The data warehouse can be easily developed in any education institute with the adaptation of common data standard. Common data standards may eliminate the need of data clarity and modification before loading this for a data warehouse.

An institute with efficient Data Warehousing and Data Mining approach can find out novel way of improving student's behavior, success rate and course popularity. All these effort may finally improve the quality of education, better student intake, better career counseling and overall practices of education system. In Data Mining classification clustering and regression are the three key approaches. Classification is a supervised learning approach in which students are grouped into identified classes [1]. Classification rules may be identified from a part of data known as training data and further it may be tested for rest of the data [2].

The effectiveness of classification approach may be evaluated in terms of reliability of the rule with test data set. Clustering approach is based on unsupervised learning because there are no predefined classes. In this approach data may be grouped together as cluster [2], [3]. The usability of clusters in terms of relevant area may be interpreted by data mining expert. Regression is a data mining approach in which it uses the explanatory variable to predict an outcome variable. For example, performance appraisal of faculty members may be done by regression analysis. Here, faculty qualification, feedback rating, amount of content covered may be taken as explanatory variable and faculty salary, increment, bonus and perks may be estimated as outcome variable so regression may be the best way to setting few important parameters based on existing variables [2].

II. Related Work

As a part of this research work more than 25 papers has been explored and thoroughly studied. Few papers related to

Manuscript received January 9, 2012; revised March 6, 2012.

S. Agarwal is working as Assistance Professor in Indian Institute of Information Technology, Allahabad, A Center of Excellence in IT established by the Human Resource Development Ministry, Government of India, New Delhi.

G. N. Pandey is working as Adjunct Professor in Indian Institute of Information Technology, Allahabad, A Center of Excellence in IT established by the Human Resource Development Ministry, Government of India, New Delhi.

M. D. Tiwari is the Director of the Indian Institute of Information Technology, Allahabad, A Center of Excellence in IT established by the Human Resource Development Ministry, Government of India, New Delhi.

education data mining are highlighted here. Salazar, et al. suggested a clustering and decision rule based Data Mining approach to identify group of clusters, which have been qualitatively described [4]. The papers used 20,000 students' records and WEKA as Data Mining Tool.

M. Ramaswami, *et al.* suggested a CHi-squared Automatic Interaction Detector (CHAID) Based approach to analyze the performance of higher secondary students [5]. The study reveals that few parameters like students school type, location, family background, educational organization, medium of teaching like Hindi or English are prime key factors to predict their performance in higher education.

Cortez, *et al.* proposed a Bayesian Networks based approach for student Data Classification [6]. The research work implement binary data classification, 5 levels Data classification and a regression based analysis for predictive performance calculation of classification algorithms and descriptive knowledge analysis for better quality management in educational organization. Qasem, et al. discussed a decision tree approach based on C4.5, id3 and Naïve Bayes. The research work proposed a Cross Industry Standard Process for Data Mining (CRISP-DM) based classification model for student performance evaluation Model [7].

A critical literature review suggests that majority of the work is based on data classification by using only a specific approach or classifiers. The aim should be not only to find out a solution for the problem specified but there should be some work for identifying the best approach. Here in the present research work few classifiers are taken together and applied to the dataset in order to select best classifier for the indentified problem. There must be some comparative analysis is essential to establish the optimum approach and this is the ultimate goal of present investigation.

III. Proposed Model

A typical Data Warehousing and Data Mining application includes data collection from heterogeneous data sources, cleaning and transformation of data and periodic updation of data field [8]. A Data Warehouse for Education Data Mining may include student personal details, academic details, examination details and accounting details. A student data cube is shown in figure where dimensions are data attributes and their values are stored in the cell. Figure 1 shows a student data cube with name, verbal ability and MAT score as attributes. Data viewing operation i.e. rolls up and drill down may be performed with respect to multi dimensional data cube [9]. For example, a student data cube is proposed for three different attributes i.e. name, MAT score and verbal ability score. Aggregate details of the student are stored in the individual cell of the data cube. . Figure 1 shows a student data cube with name, verbal ability and MAT score as attributes.

Any specific view of data cube is considered as data mart. Department wise, course wise, or subject wise data mart represent a close view of data cube. Here the paper is only limited to a theoretical model of student data cube because it is an essential part of any Data Warehouse [10]. For the implementation of data classification a student dataset of community college database has been taken. The dataset has 4 attributes and 2000 records of student performance details. The attributes are MAT score, verbal ability score, quantitative ability score and likelihood of placement.

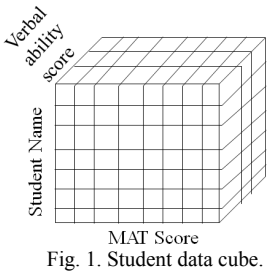

Fig. 1. Student data cube.

Data processing is required to make dataset appropriate for various classification algorithms. Here numerical data sets are converted as nominal datasets. Verbal ability score, quantitative score and MAT score has been converted as category data like 'Low', 'Medium' and 'High'. The likelihood of placement may also be defined as two categories like 'Yes' and 'No'.It is essential to have a suitable Data Mining tool for the purpose of carrying out Data Mining analysis of the data available. The software WEKA is suitable because it is open source. WEKA can efficiently work with limited data [11]. WEKA also provides convenient data preprocessing, cleaning and handling missing values. It takes data from excel file in Comma Separated Values (CSV) format, which is a very common application software to be used in each school / college for initial collection of data [12]. This software contains tools for a whole range of data mining tasks like Data pre-processing, Classification, Clustering, Association and Visualization.

IV. Classification Methods

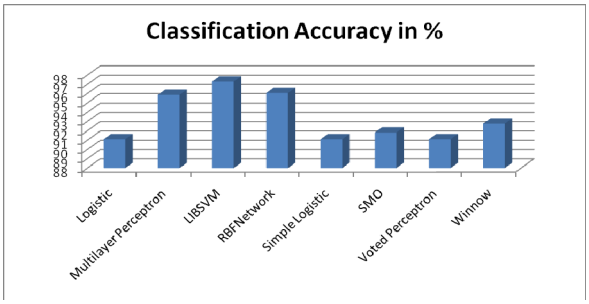

Fig. 2. Classification accuracy for WEKA classifiers.

WEKA has a approximately 40 classifiers divided into 4 groups. With the help of its explorer tool any classifiers may be applied to available data set. The research work has chosen 8 different classifiers for comparative analysis of performance of classifiers. As given in Figure 2 LIBSVM classification accuracy is 97.3 %, RBFNetwork accuracy is 96.05% and Multilayer Perceptron accuracy is 95.85% accuracy. Minimum accuracy is given by Logistic, Simple Logistic and Voted Perceptron classifiers. Another observation is made in terms of root relative square error as shown in Figure 3. It shows the accuracy of the predicted values [13]. Here again LIBSVM has minimum root relative

square error. For LIBSVM the study is further enhanced to observe the performance of various kernel types. It is observed that radial basis kernel has maximum accuracy and Sigmod type kernel as minimum accuracy [14].

TABLE I: THE WEKA CLASSIFIER AND THEIR WORKING PRINCIPLE

Classifiers	Details
Logistic	A multi category logistic regression model is used for Data Classification [15]
Multilayer Perceptron	Here back propagation approach is used to classify the datasets[15]
LIBSVM	A wrapper class for the libsvm tools for implementing Support Vector machine classifier [15], [16]
RBFNetwork	It uses normalized Gaussian radial basis function for data classification [15]
Simple Logistic	It is a simple linear regression model based classifier [15]
SMO	It is an implementation of sequential minimal optimization algorithm for training a support vector classifier[17]
Voted Perceptron	It is based on voted perceptron algorithm by Freund and Schapire.[15]
Winnow	Implements Winnow and Balanced Winnow algorithms by Littlestone [15]

The above study suggests that data with categorical attributes is very well classified if LIBSVM with Radial Basis Kernel has been taken as a best choice for Data classification [18]. A comparative analysis of LIBSVM with various kernel type is shown in Fig. 4.

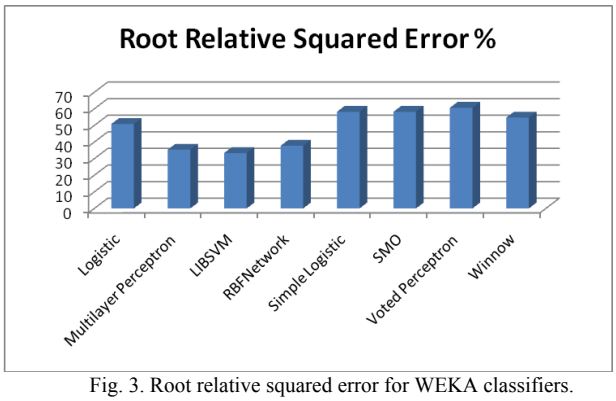

Fig. 3. Root relative squared error for WEKA classifiers.

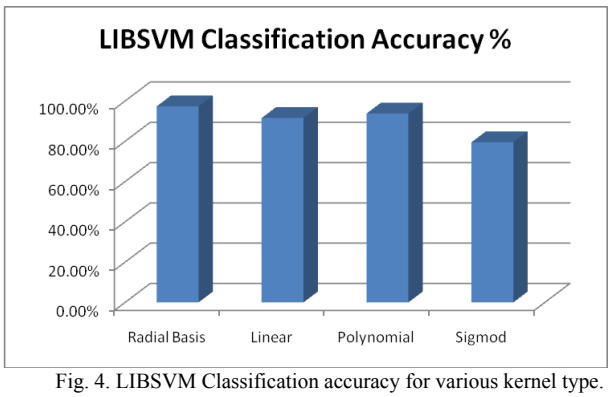

Fig. 4. LIBSVM Classification accuracy for various kernel type.

V. MINING DATA WITH DECISION TREE

A typical decision tree approach is an effective way of generating interesting classification rule. In decision tree approach an attribute which is required for analysis is taken as a starting node. The attribute is first classified in terms of groups and then next important attribute is again taken and classified under certain consideration. Here for the student dataset verbal ability, qantitative ability and MAT score all these parameters are important to predict the student placement.

As per two chosen criteria suppose a graph is plotted with X axis as MAT Score and Y axis as verbal ability score. Figure 5 shows a Scatter Plot of Verbal Ability and MAT Score. Here as result of scatter plot two different classes i.e. successful and failure student in terms of placement represented by + and - sing. The X axis is showing MAT score ranging from 90 to 100 percentile and y axis showing Verbal Ability score ranging from 90 to 100 percentile.

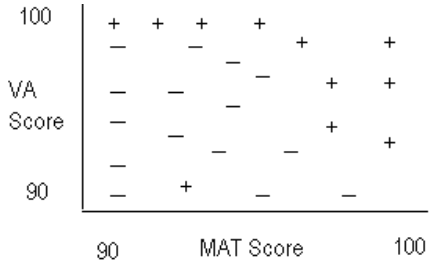

Fig. 5. Scatter plot of verbal ability and MAT score.

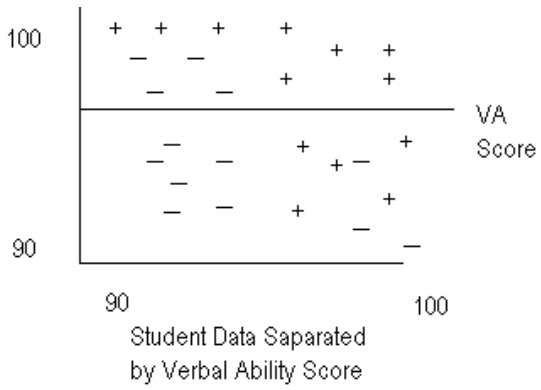

Fig. 6. Student data separated by verbal ability and MAT score.

As per observed range it is clear that cut off marks for verbal ability is 90 percentile for getting admission in the particular Business School. The minimum MAT score taken for admission is 90 percentile and maximum score observed is 100. As per given details of figure 6 a graph is potted for the two dimension. One can easily observe that successful and failure students cannot be analyzed by taking any parameter alone. So MAT score or high verbal ability score may not predict anything about the success rate but when we take the combination of these, may predict rule in terms of success rates. Decision tree approach first takes a parameter which may consider as important decision making parameter. Following are the rules.

1. class YES IF:
   MAT_SCORE in {HIGH} ^ Quant_SCORE in {LOW} ^ VA_SCORE in {LOW} (159)

2. class YES IF:
   MAT_SCORE in {LOW,MEDIUM,HIGH} ^ Quant_SCORE in {HIGH} ^ VA_SCORE in {HIGH} (691)
3. class YES IF:
   MAT_SCORE in {MEDIUM,HIGH} ^ Quant_SCORE in {LOW} ^ VA_SCORE in {HIGH} (268)
4. class NO IF:
   MAT_SCORE in {HIGH} ^ Quant_SCORE in {HIGH} ^ VA_SCORE in {LOW} (59)
5. class NO IF:
   GMAT_SCORE in {LOW,MEDIUM} ^ Quant_SCORE in {LOW,HIGH} ^ VA_SCORE in {LOW} (644)
6. class NO IF:
   GMAT_SCORE in {LOW} ^ Quant_SCORE in {LOW} ^ VA_SCORE in {HIGH} (65)

So if we take rule no. (1) to (4) the likelihood of placement is always high. The rule suggests that a student with High MAT score and high verbal ability score may have batter chances of placement. For any decision making process if more than one parameters are important then a combination of individual parameter can also taken for example, a proper scaling can be taken to scale both parameter as equal range and sum is finally taken as new attribute. Here combination of MAT Score and verbal ability score together may be taken as important consideration at the time of generating rules. The same is clearly verified in figure 7. Figure 8 shows the formation of decision tree based on the rules explored.

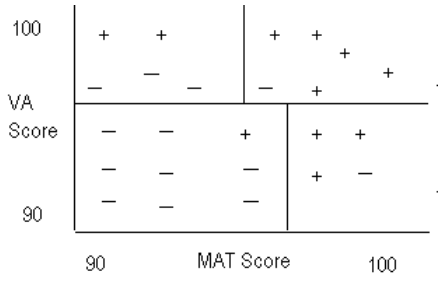

Fig. 7. Student data separated by verbal ability and MAT score.

In some cases the classification line is inclined under certain angle. The inclination indicates that here two parameters are taken and one may have more predictive value than other. So it depends upon availability of data, for any institution. The above example has been taken for calculating the possible placements; other education related processes can also be evaluated in the light of Data mining approaches. The admission policies, retention rate, examination performance, career interest, may also be estimated on the basis of students past practices.

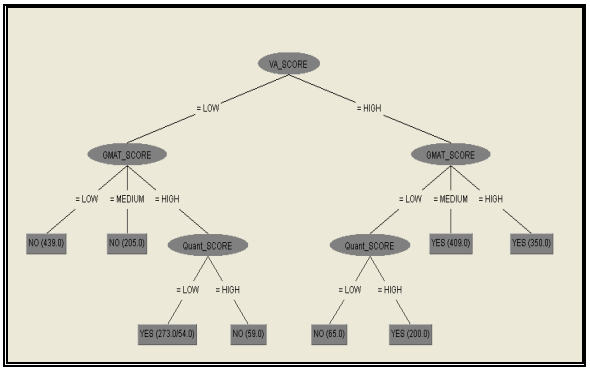

Fig. 8. Decision tree of student data.

VI. Conclusion

Data Mining could be used to improve business intelligence process including education system to enhance the efficacy and overall efficiency by optimally utilizing the resources available. The performance, success of students in the examination as well as their overall personality development could be exponentially accelerated by thoroughly utilizing Data Mining technique to evaluate their admission academic performance and finally the placement. It has been established beyond doubt that placement which is one of the critical issue to the education process could be based on their performance is qualifying examination as well as their performance in the test. This is the obvious conclusion of the present investigation. The research work also suggests that for the given data set LIBSVM with Radial Basis Kernel has been taken as a best choice for data classification.

References

[1] A. Merceron and K. Yacef, "Educational Data Mining: a Case Study," In *C. Looi; G. McCalla; B. Bredeweg; J. Breuker, editor, Proceedings of the 12th international Conference on Artificial Intelligence in Education AIED,* pp. 467–474. Amsterdam, IOS Press, 2005

[2] I. H. Witten and E. Frank, "Data mining: Practical Machine Learning Tools and Techniques," 2nd ed, Morgan-Kaufman Series of Data Management Systems San Francisco Elsevier, 2005.

[3] J. Han and M. Kamber, "Data mining: Concepts and Techniques," (Morgan-Kaufman Series of Data Management Systems). San Diego: Academic Press, 2001.

[4] A. Salazar, J. Gosalbez, I. Bosch, Miralles, and R. Vergara, "A case study of knowledge discovery on academic achievement, student desertion and student retention," Information Technology: Research and Education, 2004. ITRE 2004. *2nd International Conference,* Issue 28, pp. 150 – 154, 2004.

[5] M. Ramaswami *et al.*, "A CHAID Based Performance Prediction Model in Educational Data Mining," *IJCSI International Journal of Computer Science Issues,* vol. 7, Issue 1, no. 1, January 2010, ISSN (Online): 1694-0784.

[6] P. Cortez and A. Silva, "Using Data Mining To Predict Secondary School Student Performance," In *EUROSIS*, A. Brito and J. Teixeira (Eds.), pp. 5-12, 2008.

[7] Q. A. AI-Radaideh, E. M. AI-Shawakfa, and M. I. AI-Najjar, "Mining Student Data using Decision Trees,," International Arab Conference on Information Technology(ACIT'2006), Yarmouk University, Jordan

[8] T. Ohmori, M. Naruse, and M. Hoshi, "A New Data Cube for Integrating Data Mining and OLAP," *Data Engineering Workshop, 2007 IEEE 23rd International Conference*, pp. 896-903, 2007.

[9] C. Surjit and D. Umeshwar, "An Overview of Data Warehousing and OLAP Technology," Retrieved February 2010, from www.cs.brown.edu/courses/cs227/Papers/General/SurajitUmeshSIGMODrecord97.pdf.

[10] L. Jing, "Data Mining as Driven by Knowledge Management in Higher Education" Keynote for SPSS Public Conference, UCSF, 2001.

[11] "WEKA Data Mining Book" (n.d.) http://www.cs.waikato.ac.nz/~ml/weka/book.html.

[12] "WEKA 3: Data Mining Software in Java" (n.d.) Retrieved March 2010 from http://www.cs.waikato.ac.nz/ml/weka/.

[13] A. Chaudhuri, K. De, and D. Chatterjee, "A Comparative Study of Kernels for the Multi-class Support Vector Machine," *icnc, Fourth International Conference on Natural Computation,* vol. 2, pp. 3-7, 2008.

[14] J. Platt, "Fast Training of Support Vector Machines Using Sequential Minimal Optimization," *in Advances in Kernel Methods – Support Vector Learning,* B. Schoelkopf, C. Burges, and A. Smola Eds., pp. 185- 208, 1999.

[15] C.-C. Chang and C.-J. Lin, "LIBSVM -- A Library for Support Vector Machines," 2003.

[16] T.-N. Do and J.-D. Fekete, "Large Scale Classification with Support Vector Machine Algorithms," *in proc. of ICMLA'07, 6th International Conference on Machine Learning and Applications*, IEEE Press, Ohio, USA, pp. 7-12, 2007.

[17] I. H. Witten and E. Frank, and D. Mining, "Practical Machine Learning Tools and Techniques with Java Implementations," Academic Press, 2000.
[18] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, pp. 273–297, 1995.

Dr. Sonali Agarwal is Assistance Professor , Indian Institute of Information Technology, Allahabad, India

Prof. G.N.Pandey is Adjunct Professor , Indian Institute of Information Technology, Allahabad, India

Dr. M.D.Tiwari is Director, Indian Institute of Information Technology, Allahabad, India.