ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 6, June 2013

Data Mining Techniques in EDM for Predicting the Pupil's Outcome

S. Saranya[1], N.Tamilselvi[2] , P.Usha[3], M.Yasodha[4], V.Padmapriya[5]

Assistant Professor, Department of Computer Science, Dr.N.G.P Arts and Science College

Coimbatore, Tamil Nadu[1]

Department of Computer Science, Dr.N.G.P Arts and Science College, Coimbatore, Tamil Nadu[2,3,4,5]

Abstract— In recent era, growth of higher education has increased massively. Many new institutions and graduation schools are being established by both the private and government sectors for the growth of education and welfare of the students. Each institution aims at producing higher and exemplary graduation rates by employing various teaching and grooming methods. But still there are certain cases of unemployment that exists among the medium and low risk students. This paper aims to describe the use of data mining techniques to improve the efficiency of academic performance in the educational institutions. Various data mining techniques such as clustering, decision tree, association and rule induction, nearest neighbors, neural networks, genetic algorithms, exploratory factor analysis and stepwise regression can be applied to the higher education process, which in turn helps to improve pupils' performance. These approaches fit to provide a model to the problem domain that takes place in the educational systems.

Keywords— Data Mining, KDD, EDM, ANN, Decision Trees, Association, Clustering.

I. INTRODUCTION

Data mining is simply coined as "mining the hidden and useful information from the database". Data are generally stored in various formats like text, images, audio, video, animated scripts etc in the repository. The resulting amount of data in database represents an untapped resource.

The information can be extracted from these data, which could then be converted to valuable knowledge with data mining techniques. A process for converting large amounts of data to valuable information is the Knowledge Discovery in Databases(KDD). The input provided is theses techniques will be the data and the output obtained will be the useful information as desired by the user.

Most authors have various definitions for data mining and KDD process. Goebel and Gruenwald define KDD as the non-trivial process of identifying potentially useful, ultimately understandable patterns in Data and data mining as the extraction of patterns or models from the observed data.

Han and Kamber define data mining as the process of discovering 'hidden images', patterns and knowledge within large amount of data and making predictions for outcomes or behaviors.

Berzal et al define KDD as the non-trivial extraction of potentially useful information from a large volume of data

Gartner Group defines data mining as the process of discovering meaningful new correlation, patterns and trends by shifting through large amount of data stored in repositories and by using pattern recognition technologies as well as statistical and mathematical techniques.

Adriaans and Zantinge define KDD as non-trivial extraction of implicit, previously unknown and potentially useful knowledge from data

Data mining, also popularly referred to as knowledge discovery from data (KDD), is the automated or convenient extraction of patterns representing knowledge implicitly stored or catchable in large databases, data warehouses, the Web, other massive information repositories, or data stream where the information is implicit(although previously known) and data mining as a generic term that uncovers research results ,techniques and tools used to extract useful information from large

Rubenking explains Data mining is the process of automatically extracting useful information and relationships from immense quantities of data. In its purest form, data mining doesn't involve looking for specific information. Rather than starting from a question or hypothesis, data mining simply finds patterns that are already present in the data.

In institutional research, data mining uses large numbers of variables and aims at identifying data patterns that can shed light on student behavior. The basic steps that are included in KDD process are depicted in the diagram as follows:

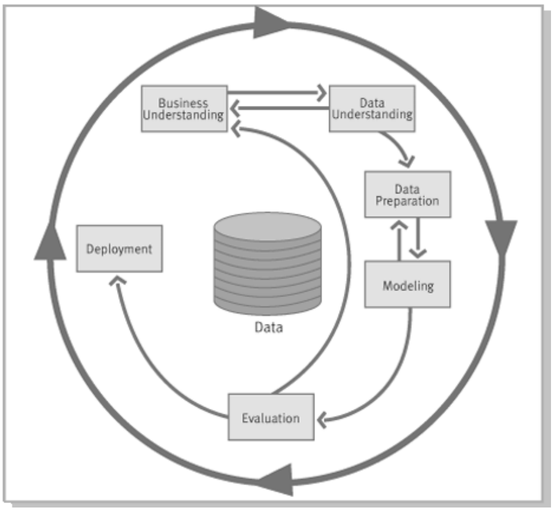

Fig.1. KDD Process

II.KNOWLEDGE DISCOVERY IN DATABASE

Data mining is being used in several applications like Health, Insurance, Banking, Security, Education, Research, Artificial intelligence and computation etc., but it has extended its wings in the field of education tremendously from 1990 to present.

Educational Data Mining (EDM) is an emerging discipline, concerned with developing methods for exploring the unique types of data that come from educational settings, and using those methods to better understand students, and the settings which they learn in. A key area of EDM is mining student performance. Another key area is mining enrollment data. The areas of EDM application are: Providing feedback for supporting instructors, Recommendations for students, Predicting student performance, Student modeling and detecting undesirable student behaviors, Grouping students, constructing courseware, Planning and scheduling.

EDM uses various techniques like classification, Decision trees, Rule induction, Nearest neighbors, Neural networks, Clustering, Genetic algorithms, Exploratory factor analysis, Stepwise regression. This paper describes the usage of data mining techniques and algorithms to mine the hidden and implicit knowledge from the database and develops a probable model for the educational domain.

III. LITERATURE REVIEW

Data mining has evolved its research very well in the field of education in a massive amount. This tremendous growth is mainly because it contributes much to the educational systems to analyze and improve the growth of students as well as the pattern of education. Various works had been done by a large number of scientists to explore the best mining technique for performance monitoring and assessment. Some of the related works are listed down to have a better understanding of what should be carried on in the future for further growth.

Han and Kamber [1] describes data mining software that allow the users to analyze data from different dimensions, categorize it and summarize the relationships which are identified during the mining process.

Galit [2] gave a case study that use students data to analyze their learning behavior to predict the results and to warn students at risk before their final exams.

Al-Radaideh, et al [3] applied a decision tree model to predict the final grade of students who studied the C++ course in Yarmouk University, Jordan in the year 2005. Three different classification methods namely ID3, C4.5, and the Naïve Bayes were used. The outcome of their results indicated that Decision Tree model had better prediction than other models.

Khan [4] conducted a performance study on 400 students comprising 200 boys and 200 girls selected from the senior secondary school of Aligarh Muslim University, Aligarh, India with a main objective to establish the prognostic value of different measures of cognition, personality and demographic variables for success at higher secondary level in science stream. The selection was based on cluster sampling technique in which the entire population of interest was divided into groups, or clusters, and a random sample of these clusters was selected for further analyses. It was found that girls with high socio-economic status had relatively higher academic achievement in science stream and boys with low socio-economic status had relatively higher academic achievement in general.

Hijazi and Naqvi [5] conducted as study on the student performance by selecting a sample of 300 students (225 males, 75 females) from a group of colleges affiliated to Punjab university of Pakistan. The hypothesis that was stated as "Student's attitude towards attendance in class, hours spent in study on daily basis after college, students' family income, students' mother's age and mother's education are significantly related with student performance" was framed. By means of simple linear regression analysis, it was found that the factors like mother’s education and student’s family income were highly correlated with the student academic performance.

Cortez and Silva [6] attempted to predict failure in the two core classes (Mathematics and Portuguese) of two secondary school students from the Alentejo region of Portugal by utilizing 29 predictive variables. Four data mining algorithms such as Decision Tree (DT), Random Forest (RF), Neural Network (NN) and Support Vector Machine (SVM) were applied on a data set of 788 students, who appeared in 2006 examination. It was reported that DT and NN algorithms had the predictive accuracy of 93% and 91% for two-class dataset (pass/fail) respectively. It was also reported that both DT and NN algorithms had the predictive accuracy of 72% for a four-class dataset.

Pandey and Pal [7] conducted study on the student performance based by selecting 600 students from different colleges of Dr. R. M. L. Awadh University, Faizabad, India. By means of Bayes Classification on category, language and background qualification, it was found that whether new comer students will performer or not.

Bray [8], in his study on private tutoring and its implications, observed that the percentage of students receiving private tutoring in India was relatively higher than in Malaysia, Singapore, Japan, China and Srilanka. It was also observed that there was an enhancement of academic performance with the intensity of private tutoring and this variation of intensity of private tutoring depends on the collective factor namely socio-economic conditions.

IV. EDM METHODOLOGIES

In data mining literature, numerous frameworks have been devised to gather and arrange the data for building model. Generally methodologies are designed on the basis of the nature of the problem, in which an In-depth and complete study of the problem proves to be effective for deciding the type of technique and algorithm to be applied on the dataset pertaining to provide a prompt solution for that problem. The methodology should be chosen in a manner such that it should be capable of yielding an accurate prediction of the performance which should be satisfactory for both the institutions and the pupil's. Various reviews, survey and journals helps to establish a proper methodology for identifying the best methodological pattern for the problem. One such model is the CRISP-DM (Cross-Industry Standard Process for data mining) which was proposed in the 1990s by European consortium of companies. This methodology consists mainly of six steps: understanding the higher education objective, collecting the educational data, preparing the data, building the models, evaluating the model using one of the evaluation methods, and finally deployment which using the model for future prediction of the student performance, similar to that of the KDD process. The problem domain should be carefully understood and data can be gathered depending upon the problem. The collected data can be cleaned and prepared for extracting information by eliminating the redundant and inconsistent data, noisy and missing values and then they are used to build the model. The prepared data are divided into two sets namely testing and training set, the former one is used to model the data and the latter one to validate the developed model. The models are designed and built by using the data mining techniques and finally they are ensured for well formalness and then deployed in the respective platform. As mentioned earlier data mining is the extraction of hidden information from the repositories that are efficiently mined by means of a number of techniques and algorithms. The most important among them are Classification, Prediction, Association, Clustering and Artificial Neural Networks.These techniques are used generally in almost all applications with the basic concepts and significance described below:

CLASSIFICATION

Classification is the most commonly applied data mining technique, which employs a set of pre-classified examples to develop a model that can classify the population of records at large. This approach frequently employs decision tree or neural network-based classification algorithms. The data classification process involves learning and classification. In Learning the training data are analyzed by classification algorithm. In classification test data are used to estimate the accuracy of the classification rules. If the accuracy is acceptable the rules can be applied to the new data. The classifier-training algorithm uses these pre-classified examples to determine the set of parameters required for proper discrimination. The algorithm then encodes these parameters into a model called a classifier. Decision Tree based Methods, Rule-based Methods; Memory based reasoning, Neural Networks, Naïve Bayes and Bayesian Belief Networks, Support Vector Machines

ASSOCIATION

Association and correlation is usually to find frequent item set findings among large data sets. This type of finding helps businesses to make certain decisions, such as catalogue design, cross marketing and customer shopping behavior analysis. Association Rule algorithms need to be able to generate rules with confidence values less than one. However the number of possible Association Rules for a given dataset is generally very large and a high proportion of the rules are usually of little (if any) value. The most commonly used algorithm in association is the Apriori algorithm.

PREDICTION

Regression technique can be adapted for prediction. Regression analysis can be used to model the relationship between one or more independent variables and dependent variables. In data mining independent variables are attributes already known and response variables are what we want to predict. Unfortunately, many real-world problems are not simply prediction. Therefore, more complex techniques (e.g., logistic regression, decision trees, or neural nets) may be necessary to forecast future values. The same model types can often be used for both regression and classification. For example, the CART (Classification and Regression Trees) decision tree algorithm can be used to build both classification trees (to classify categorical response variables) and regression trees (to forecast continuous response variables). Neural networks too can create both classification and regression models.2 types of regression analysis are available: Simple, Multiple.

CLUSTERING

Clustering can be said as identification of similar classes of objects. By using clustering techniques we can further identify dense and sparse regions in object space and can discover overall distribution pattern and correlations among data attributes. Classification approach can also be used for effective means of distinguishing groups or classes of object but it becomes costly so clustering can be used as preprocessing approach for attribute subset selection and classification. Clustering includes 2 types namely Partitional and hierarchical clustering which are used for the discovery of clusters with arbitrary shape.

NEURAL NETWORKS

ANN's are non-linear predictive models that learn through training and resemble biological neural networks in structure. Neural network is a set of connected input/output units and each connection has a weight present with it.

During the learning phase, network learns by adjusting weights so as it will be able to predict the correct class labels of the input data. Neural networks have the remarkable ability to derive meaning from complicated or imprecise data and can be used to extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques.. Neural networks are best at identifying patterns or trends in data and well suited for prediction or forecasting needs. Some of the algorithms that are used in the above techniques are listed below:

Genetic algorithms: Optimization techniques that use process such as genetic combination, mutation, and natural selection in a design based on the concepts of natural evolution.

Decision trees: Tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. The main purpose of the decision tree is to expose the structural information contained in the data. Specific decision tree methods include Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID). CART and CHAID are decision tree techniques used for classification of a dataset. They provide a set of rules that can be applied to a new (unclassified) dataset to predict which records will have a given outcome. CART segments a dataset by creating 2-way splits while CHAID segments using chi square tests to create multi-way splits. CART typically requires less data preparation than CHAID. ID3 (Induction Decision Tree),a recursive procedure that is used to construct a decision tree from data .A series of improvements to ID3 culminated to the decision tree approach called C4.5 which deals with numeric attributes, missing values and noisy data.

Nearest neighbor method: A technique that classifies each record in a dataset based on a combination of the classes of the k record(s) most similar to it in a historical dataset (where k 1). Sometimes called the k-nearest neighbor technique.

Rule induction: The extraction of useful if-then rules from data based on statistical significance.

For better understanding the usage of data mining techniques in EDM, refer below:

teaching method for high-risk students so that there won't be any drop-outs in the consecutive semesters

- Prediction- Predict which group of students are likely to score/fail, Predict the attitude/behavior for both the groups, Predict the continual factors/tasks that affects the progression of high-risk group, predict the factors that will influence the talents of the low-risk.
- Association-Associate the training period of the low-risk and high-risk students, Association of the scores obtained by both the groups when taught individually and as a crew, Association of semester wise scores in performance evaluation, Association of students endurance towards their motto, Association of the individual and teamwork done by the students .
- Clustering-Groups low-risk and high-risk students in separate groups, groups students with similar and dissimilar behavior, groups students with similar learning methods
- Other techniques- Identifying and enhancing the teaching methods, presenting seminars, reviews at eh end of each week for assessing the level of students, building models to find similar patterns of student.

Therefore said patterns can be used for deciding the performance by the educational institutes and the teachers for the growth of education.

V. CONCLUSION

As the growth of education is going beyond the expectations, it is must to enrich the career of the students by providing them a valuable education which would best meet their educational and career motto. This paper hopes to provide a survey of the techniques available in Data mining, how the data are mined, how it can be used in the higher education, identifying the best pattern for the defined problem and what type of methodology should be used to resolve it.

REFERENCES

[1] J. Han and M. Kamber, "Data Mining: Concepts and Techniques," Morgan Kaufmann, 2000.

[2] Galit.et.al, "Examining online learning processes based on log files analysis: a case study". Research, Reflection and Innovations in Integrating ICT in Education 2007.

[3] Q. A. AI-Radaideh, E. W. AI-Shawakfa, and M. I. AI-Najjar, "Mining student data using decision trees", International Arab Conference on Information Technology (ACIT'2006), Yarmouk University, Jordan, 2006.

[4] Z. N. Khan, "Scholastic Achievement of Higher Secondary Students in Science Stream", Journal of Social Sciences, Vol. 1, No. 2, 2005, pp. 84-87.

[5] S. T. Hijazi, and R. S. M. M. Naqvi, "Factors Affecting Student's Performance: A Case of Private Colleges", Bangladesh e-Journal of Sociology, Vol. 3, No. 1, 2006.

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

[7] 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.

[8] M. Bray, The Shadow Education System: Private Tutoring And Its Implications For Planners, (2nd ed.), UNESCO, PARIS, France, 2007.