ISSN:0975-9646

Educational Data mining for Prediction of Student Performance Using Clustering Algorithms

M. Durairaj [#1], C. Vijitha [*2]

[#1] *Assistant Professor, School of Computer Science, Engineering and Applications, Bharathidasan University, Tiruchirappalli, India*
[*2] *Research Scholar, School of Computer Science, Engineering and Applications, Bharathidasan University, Tiruchirappalli, India*

Abstract: In recent years, the biggest challenges that educational institutions are facing the explosive growth of educational data and to use this data to improve the quality of managerial decisions. Educational institutions are playing an important role in our society and playing a vital role for growth and development of nation. Prediction of student's performance in educational environments is also important as well. Student's academic Education details & performance is based upon various factors like personal details, social, psychological etc. Educational database contain the useful information for predicting a students' performance, rank factor & details. The data mining techniques are more helpful in classifying educational database. Educational data mining concerns with developing methods for discovering knowledge from data, that comes from educational institutions. The Data Mining prediction has allowed a decision making tool which can facilitate better resource utilization in terms of students performance. In our college, the student details have been taken for analysis and data mining methods have been employed to get vital information. The work aims to develop a trust model using data mining techniques which mines required information, so that the present education system may adopt this as a strategic management tool.

Keywords- Academic performance, Data mining, Data classification, Clustering, Student's result database.

1. INTRODUCTION

Data mining is data analysis methodology used to identify hidden patterns in a large data set. It has been successfully used in different areas including the educational environment. Educational data mining is an interesting research area which extracts useful, previously unknown patterns from educational database for better understanding, improved educational performance and assessment of the student learning process. Evaluating students' performance is a complex issue, which can't be restricted for the grading. Reasons of good or bad performances belong to the main interests of teachers, because they can plan and customize their teaching program, based on the feedback. Data mining is one of the approaches, which can provide an effective assistance in revealing complex relationships behind the grades. Figure 1 explains different stages of our proposed methodology for mining information on educational data for academic decision making.

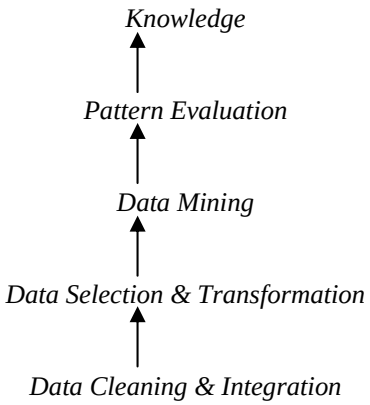

Fig. 1. Different stages of data mining process

2. LITERATURES ON STUDENTS' PERFORMANCE ANALYSIS

Azhar Rauf, et. al. [1] proposed a method known as K-mean clustering, it calculates initial centroids instead of random selection, due to which the number of iterations is reduced and elapsed time is improved. Jaideep Vaidya [2] proposed a privacy preserving K-means clustering method over vertically partitioned data when different web sites contain different attributes for a common set of entities. N.Sivaram [3] surveyed the applicability of clustering and classification algorithms for recruitment data mining techniques that fit the problems which are determined. A study has been made by applying K-means, fuzzy C-means clustering and decision tree classification algorithms to the recruitment data of an industry.

Md.Hedayetul Islam Shovon [4] presented a paper on prediction of student academic performance by applying K-means clustering algorithm. The student's evaluation factor like class quizzes, mid and final exam assignment are studied. It is recommended that all these correlated information should be conveyed to the class advisor before the conduction of final exam. This study will help the teachers to reduce the drop out ratio to a significant level and improve the performance of students. SajadinSembiring [5] discussed the application of Smooth Support Vector Machine (SSVM) classification and kernel k-means clustering techniques. The results of this study reported a model of student academic performance predictors by employing psychometric factors as variables predictors. Oyelade, O. J [6] presented a method of using

K-means clustering algorithm for the prediction of Students' Academic Performance. The ability to monitor the progress of students' academic performance is a critical issue to the academic community of higher learning. This paper is aims to present a systematic review on different clustering techniques applied for educational data mining to predict academic performance of students and its implications. Trilok Chand Sharma [7] presented a data mining techniques to process a dataset and identify the relevance of classification test data. Mining tools to solve large amounts of problems such as classification, clustering, association rule, neural networks, it is an open access tools directly communicates with each tool or called from java code to implement using this. Shilpa Dhanjibhai Serasiya[8] presented a how problems of classification and prediction can be solved using class association rules. In the simulation on WEKA, we have used selected classification techniques to propose the appropriate result from our training dataset. Swasti Singhal[9] presented a the steps of how to use WEKA tool for these technologies. It provides the facility to classify the data through various algorithms.

D. Kabakchieva[10] give a brief discussion about The specific objective of the research work is to find out interesting patterns in the available data that could contribute to predicting student performance at the university based on their personal and pre-university characteristics. This is a first attempt of applying data mining in the Bulgarian educational sector.

3. MATERIALS AND METHODS

3.1 Data Set

We collected our college students' real time data that describing the relationships between learning behavior of students and their academic performance. Sample data is illustrated in Figure 2. The data contain students' details of different subject marks in semester wise have been recorded and subjected to the data mining process.

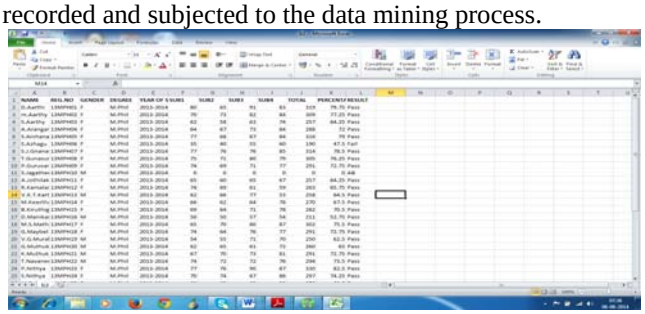

Fig.2 Example data set

3.2 Data processing methods

Data Clustering:

Clustering is a process which divides data into groups of similar objects. From a machine learning perspective, clusters correspond to hidden patterns and search for clusters is categorised as unsupervised learning. From a practical perspective, clustering plays an outstanding role in data mining applications such as scientific data exploration, information retrieval and text mining, spatial database applications, Web analysis, CRM, marketing, medical diagnostics, computational biology, and many others. The K-means algorithm, probably the best one of the clustering algorithms proposed, is based on a very simple idea: Given a set of initial clusters, assign each point to one of them, and then each cluster center is replaced by the mean point on the respective cluster. These two simple steps are repeated until convergence. The objective of this k-means test is to choose the best cluster center which is to be the centroid. The k-means algorithm requires the change of nominal attributes in to numerical. The clustering method produced a model with five clusters.

Data Classification:

Data classification is the categorization of data for its most effective and efficient use. In a basic approach to storing computer data, data can be classified according to its critical value or how often it needs to be accessed, with the most critical or often-used data stored on the fastest media while other data can be stored on slower (and less expensive) media. This kind of classification tends to optimize the use of data storage for multiple purposes - technical, administrative, legal, and economic. Data can be classified according to any criteria, not only relative importance or frequency of use. For example, data can be broken down according to its topical content, file type, operating platform, average file size in megabytes or gigabytes, when it was created, when it was last accessed or modified, which person or department last accessed or modified it, and which personnel or departments use it the most. A well-planned data classification system makes essential data easy to find. This can be of particular importance in risk management, legal discovery, and compliance with government regulations.

3.3. Data mining methodologies

K-Means Clustering Algorithm:

K-mean clustering algorithm, clusters are fully dependent on the selection of the initial cluster centroids. K data elements are selected as initial centers and then the distances of all data elements are calculated by Euclidean distance formula. Data elements having less distance to centroids are moved to the appropriate cluster. The process is continued until no more changes occur in clusters. The following figure 4 shows steps of the basic K-mean clustering algorithm steps.

INPUT: Number of desired clusters *K*
Data objects D= {d1, d2…dn}
OUTPUT: A set of *K* clusters
Steps:

1) Randomly select k data objects from data set D as initial centers.
2) Repeat;
3) Calculate the distance between each data object di (1 <= i<=n) and all *k* clusters C j(1 <= j<=k) and assign data object di to the nearest cluster.
4) For each cluster j (1 <= j<=k), recalculate the cluster center.
5) Until no change in the center of clusters.
6) Time complexity of K-mean Clustering is represented
7) by O(*nkt*)

Note: Where *n* is the number of objects, *k* is the number of clusters and *t* is the number of iterations.

Figure 4. Steps of basic K-mean clustering algorithm

The design of the system requires the complete understanding of the problem domain. The data sets and the input attributes are determined through knowledge engineering in an IT industry. The process involves defining the problem, identifying relevant take holders, and learns about current solutions to the problem. It also involves learning domain-specific terminology, description of the problem and restrictions of it. In this step, interviews were conducted to the domain experts to obtain required information to solve the problem, knowledge extraction was made with the collected information and a knowledge base was built. The knowledge base construction comprises collection of sample data, and deciding which data will be needed in respect to data mining knowledge discovery goals including its format and size. In this work, the students mark analysis performed is shown in Figure 3.

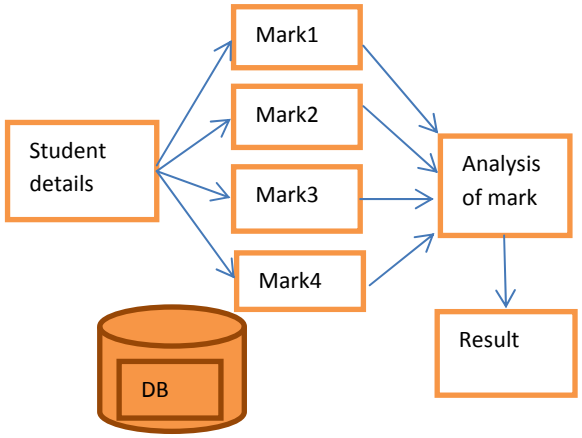

Fig.3. Students' mark analysis

Naviebayes clustering

A Naive Bayes classifier assumes that the value of a particular feature is unrelated to the presence or absence of any other feature, given the class variable. A naive Bayes classifier considers each of these features to contribute independently to the probability that this fruit is an apple, regardless of the presence or absence of the other features. An advantage of naive Bayes is that it only requires a small amount of training data to estimate the parameters necessary for classification. Because independent variables are assumed, only the variances of the variables for each class need to be determined and not the entire confusion matrix. It is an independent feature model.

The algorithm is composed of the following steps:

Algorithm:

1. Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.
2. Assign each object to the group that has the closest centroid.
3. When all objects have been assigned, recalculate the positions of the K centroids.
4. Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized and can be calculated.

Navie bayes probabilistic algorithm

The probability model for a classifier is a conditional model

$$p(C|F_1,\ldots,F_n)$$

Over a dependent class variable C with a small number of outcomes or classes, conditional on several feature variables F_1 through F_n. The problem is that if the number of features n is large or if a feature can take on a large number of values, then basing such a model on probability tables is infeasible. We therefore reformulate the model to make it more tractable. In this paper, the cluster instances are calculated by using the simple K-means clustering algorithm. The clustering instances in a confusion matrix are as shown in Table 1.

	Pos	Neg
pos	0	1
neg	1	37

Table.1. Clustering instances in confusion matrix

4. Results And Discussion

The tool WEKA 3.7.9 is used in this work to implement data mining approach, which is a collection of open source of many data mining and machine learning algorithms, including pre-processing on data, classification, and clustering and association rule extraction. It is a Java based open source tool created by researchers at the University of Waikato in New Zea-land (See Figure 4).

Fig.4. Weka implementation tool

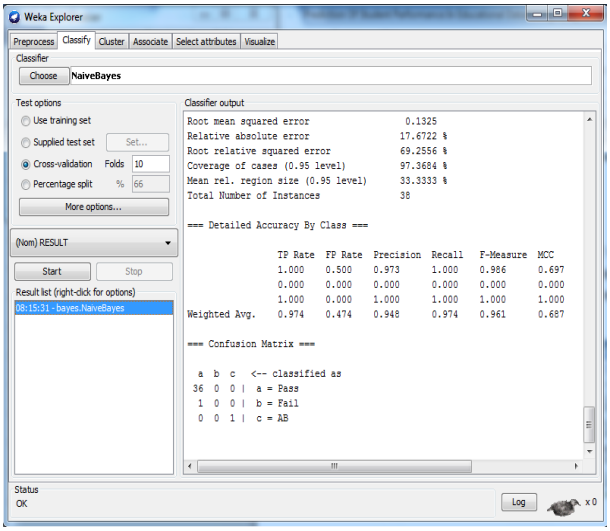

Fig.5. Applying Naviebayes algorithm

We applied Navie bayes alrogithm to analyse data. The figure 5 shows that the correctly and incorrectly classified values and kappa statistic values of students' data. The kappa statistic is a generic term of several similar measures of agreement used with categorical data.

$$K=\frac{P(A)-P(E)}{1-P(E)}$$

By using this formula the kappa statistic value can be calculated.

4.1 Performance Measures

There are some parameters on the basis of which we evaluated the performance of the classifiers such as TP rate, FP rate, Precision, Recall F- Measure and ROC area. The Accuracy of a classifier on a given test set is the percentage of test set tuples that are correctly classified by the classifier. The Error Rate or Misclassification rate of a classifier M, which is 1-Acc (M), where Acc (M) is the accuracy of M. The Confusion Matrix is a useful tool for analysing how well your classifier can recognize tuples of different classes. The sensitivity and specificity measures can be used to calculate accuracy of classifiers. Sensitivity is also referred to as the true positive rate (the proportion of positive tuples that are correctly identified), while Specificity is the true negative rate (that is, the proportion of negative tuples that are correctly identified). These measures are defined as follows

$$\text{Sensitivity}=\frac{t-pos}{pos}$$

$$\text{Specificity}=\frac{t-neg}{neg}$$

$$\text{Precision}=\frac{t-pos}{t-pos+f-pos}$$

where *t-pos* is the number of true positives tuples that were correctly classified, *pos* is the number of positive tuples, *t-neg* is the number of true negatives tuples that were correctly classified, *neg* is the number of negative tuples, and *f-pos* is the number of false positives tuples that were incorrectly labelled. It can be shown that accuracy is a function of sensitivity and specificity:

$$\text{Accuracy}=\text{Sensitivity}\frac{\text{pos}}{\text{pos+neg}}+\text{Specificity}\frac{\text{neg}}{\text{pos+neg}}$$

TP rate: It is the proportion of actual positives which are predicted as positive. The formula is defines as,

$$\text{Tp Rate}=\frac{tp}{tp+fn}$$

Where *tp* stands for true positive and *fn* stands for false negative.

FP rate: It is the rate of negatives tuples that are incorrectly labelled. The formula is defined as,

$$\text{FP rate of class "yes"}=\frac{\text{fn}}{(\text{fn+tn})}$$

$$\text{FP rate of class "no"}=\frac{\text{fp}}{\text{tp+fp}}$$

Data mining application on students' data for performance analysis is carried out using Weka tool. The screen shot of data analysis through Weka tool is demonstrated in figure 6.

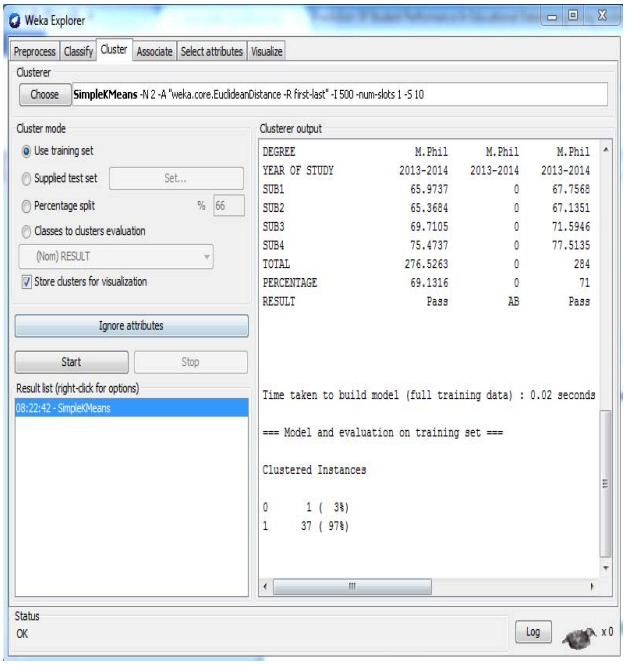

Fig.6. K-means Clustering Algorithm

5. RESULT AND DISCUSSION

The graph (See figure 7) denotes the results of students' performance and status of students' analysis displayed through ROC (Receiver operator Characteristics or Receiver operator Curve). The plot area of true positive and false positive are shown in figure 7. The subsequent confusion matrix is depicted in table 2.

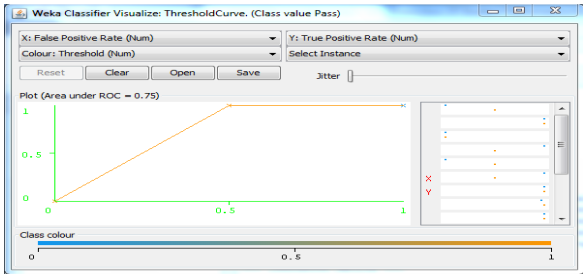

Fig.7. Prediction of ROC Curve

a	B	c
36	0	0
0	1	0
0	0	1

- a-Pass
- b-Fail
- c-Absentees

Table.2. Confusion Matrix

From our total number of 300 students record dataset, we chosen sample 38 students record for our analysis. The confusion matrix demonstrates number of pass, fail and absentees for a particular examination. Number of pass students are 36. Number of Fail student is 1. Number of absentees is 1. The data analysis is performed with the methods of precision, recall and f-measure. These three methods are explained below:

Precision

Prediction is a calculation of positive predicted values precision, which is the fraction of retrieved documents that

are relevant. The precision is calculated using the formula as:

$$\text{Precision}=\frac{|\{\text{relavant documents}\}\Lambda\{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|}$$

$$\text{Precision}=\frac{tp}{tp+fp}$$

Precision takes all retrieved documents into account, but it can also be evaluated at a given cut-off rank, considering only the topmost results returned by the system. This measure is called precision at n.

Recall

Recall in information retrieval is the fraction of the documents that are relevant to the query and that are successfully retrieved. The formula for recall is as given below.

$$\text{Recall}=\frac{|\{\text{relavant documents}\}\Lambda\{\text{retrieved documents}\}|}{|\{\text{relavant documents}\}|}$$

$$\text{Recall}=\frac{tp}{tp+fn}$$

F-Measure

This is a measure that combines precision and recall, a harmonic mean of precision and recall, is known as the traditional F-measure.

$$F=2.\frac{\text{Precision.Recall}}{\text{Precision+Recall}}$$

Technique	TP rate	FP rate	Precision	Recall	F-M	MCC
Decision tree	0.947	0.474	0.922	0.947	0.934	0.660
Navie bayes	1.000	0.000	1.000	1.000	1.000	1.000

Table.3. Comparison of Weighted Average for various techniques

In Table 2, the comparison of two techniques Decision tree and Navie bayes is illustrated. The comparison shows that the Navie bayes algorithm gives more accurate results than decision tree.

6. CONCLUSION

Using K-Means clustering algorithm, we predicted the pass percentage and fail percentage of the Overall students appeared for a particular examination. The results show the students' performance and it is seems to be accurate. The comparison between Naviebayes algorithm and decision stump tree technique shows that the Navie bayes techniques produce accurate result than the other and it is measured using confusion matrix. The results are predicted within 0 seconds. This simple analysis works show that the proper data mining application on student's performance data can be efficiently used for vital hidden knowledge / information retrieval from the vast data, which can be used for the process of decision making by the management of an educational institution. This paper also concludes with that for data mining application for effective and faster results prediction, classification and clustering can be done through Weka implementation tool.

REFERENCES:

[1] Azhar Rauf, Sheeba,"Enhanced K-Mean Clustering Algorithm to Reduce Number of Iterations and Time Complexity", *Middle-East Journal of Scientific Research*, Vol. 12 (7), Pp. 959-963, 2012.

[2] Jaideep Vaidya, "Privacy Preserving K-Means Clustering over Vertically Partitioned Data", *In proceeding of SIGKDD* '03, Washington, DC, USA, August 24-27, 2003.

[3] N. Sivaram, "Applicability of Clustering and Classification Algorithms for Recruitment Data Mining", *International Journal of Computer Applications*, Vol. 4(5), July 2010.

[4] Md. Hedayetul Islam Shovon, "Prediction of Student Academic Performance by an Application of K-Means Clustering Algorithm", *International Journal of Advanced Research in Computer Science and Software Engineering*, Vol. 2(7), July 2012.

[5] Sajadin Sembiring, "Prediction Of Student Academic Performance by an Application of Data Mining Techniques", *International Conference on Management and Artificial Intelligence* IPEDR, IACSIT Press, Vol. 6, 2011.

[6] Oyelade, O. J, "Application of k-Means Clustering algorithm for prediction of Students' Academic Performance", *(IJCSIS) International Journal of Computer Science and Information Security*, Vol.7, 2010.

[7] Trilok Chand Sharma," WEKA Approach for Comparative Study of Classification Algorithm", (IJARCCE*) International Journal of Advanced Research in Computer and Communication Engineering*, Vol. 2(4), April 2013.

[8] Shilpa Dhanjibhai Serasiya," Simulation of Various Classifications Results using WEKA", (*International Journal of Recent Technology and Engineering* (IJRTE), Vol. 1(3), August 2012.

[9] Swasti Singhal," A Study on WEKA Tool for Data Preprocessing, Classification and Clustering", *International Journal of Innovative Technology and Exploring Engineering* (IJITEE) ,Vol. 2(6), May 2013.

[10] D. Kabakchieva, "Analyzing University Data for Determining Student Profiles and Predicting Performance", *Cybernetics and Information Technologies*, Vol.1(3), March 2013.