Analyzing the Student's Academic Performance by using Clustering Methods in Data Mining

Sreedevi Kadiyala, Chandra Srinivas Potluri

Abstract - Analyzing the engineering student academic performance is not an easy task for the community of high learning. The performance of student during their first year in college is a turning point in their educational path and usually encroaches on their final percentage in a decisive manner. The student's evaluation factors like class quiz, internals, and performance in the lab will be studied. This analyzing will help the teachers to reduce drop out ratio to a significant level and improve their performance in the final exam. Statistics plays an important role in analyzing and evaluating the performance in college in order to make appropriate academic decisions. Academic decisions will result changes in their performance which need to be assessed periodically over span of time. The performance parameters chosen can be viewed at individual student, class, department and college levels. Major clustering methods are used to extract meaningful information and to develop the significant relationship among the variables stored in large data sets. In this paper, we present a procedure based on decision tree of data mining techniques and k-means partitioning of clustering methods which helps to enhance the quality of educational system by analyzing and improving student's performance.

Keywords - Database; Data mining; Data ware house; Data clustering; Decision tree; K-Means clustering algorithm; Partitioning methods; Student performance analysis.

1.INTRODUCTION

Clustering methods in data mining have been applied in many applications such as fraud detection, banking, academic performance and instruction detection. A method may have futures from several categories. Many factors could act as a barrier to student attaining and maintaining a high percentage that reflects the overall academic performance during their tenure in college. These factors could be targeted by the faculty members in developing strategies to improve student learning and academic performance by the way of monitoring and analyzing the progression of their performance. With the help of classical partitioning methods and decision tree of data mining techniques it is possible to discover the key characteristics for future predictions.

The aim of portioning methods in clustering is to partition students into homogenous groups according to their characteristics and abilities. These applications can help both the instructors and student to improve the quality education. Analyze different factors effect a students learning behavior and performance during academic career using K-means clustering algorithm and decision tree in an higher educational institute. Decision tree analysis is a popular data mining technique that can be used to explain different variables like attendance ratio and grade ratio. Clustering is one of the basic techniques often used in analyzing data sets. This study makes use of cluster analysis to segment students in to groups according to their characteristics and use decision tree for making meaningful decision for the students.

- *Sreedevi Kadiyala is currently pursuing Ph.D in Computer Science in JNT University, India. Working as an Asst. Professor in Debre Berhan University, Ethiopia. E-mail: sreedevikadiyala@gmail.com*
- *Chandra Srinivas Potluri is currently pursuing M.Phil in Computer Science, Bharathiar University, India. Working as an Asst. Professor in Debre Berhan University, Ethiopia. E-mail: pcsvas@gmail.com*

2.REALATED WORK

***2.1 Data Base*:** A data base is a collection of data usually associated with some organization or enterprise. Data in a data base are usually viewed to have a particular structure or schema with which it is associated.

2.2 Data Warehouse: Data warehouse is a form of storage system (data base) where large volume of data is stored in such a way that retrieving desirable information from the system is very easy and reliable. Data warehouse is stored in different locations, so that it does not collide with transactional data base systems, which store day-to-day information and provides solutions to sophisticated queries, which involve many computations to be performed at finger level of granularity.

2.3 Data Mining: Data mining is a process of extracting knowledge from massive volume of data. It refers to a way of finding significant and useful information from an organizations data base. The knowledge which is extracted can include pattern types, association rules and different trends. Data mining is confined to a particular organization instead it has a technique to explore the knowledge hidden in any data. Data mining software allow the users to analyze data from different dimensions categorizes it and summarized the relationships, identified during the mining process. The different techniques used from retrieving out data are artificial intelligence, statistical and mathematical techniques and pattern recognition techniques. Data mining techniques can be differentiated by their different model functions and representation, preference criterion

and algorithms. Additionally, data mining systems provide the means to easily perform data summarization and visualization, aiding the security analysis in identifying areas of concern. The models must be represented in some form. Common representations for data mining techniques include rules, decision trees, linear and non-linear functions, instance based examples and probability models.

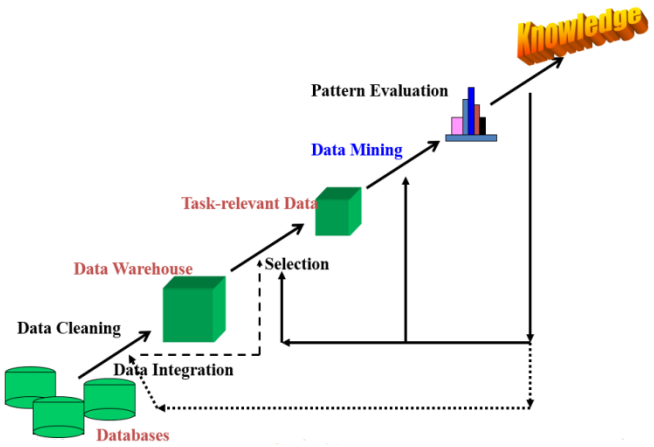


Fig 1: Steps of Knowledge Extraction

3. DATA CLUSTERING

It is the process of grouping a set of physical or abstract objects into classes of similar objects. A cluster is a collection of data objects that are similar to one another with in the same cluster and are dissimilar to the objects in other clusters. Data clustering is alternatively referred to an unsupervised learning and statistical data analysis. Cluster analysis is an important human activity. Cluster analysis has been widely used in numerous applications including pattern recognition, data analysis, image processing and market research. Clustering is a descriptive task that seeks to identify homogenous group objects based on the values of their attributes. Clustering has many requirements like scalability, dealing with different types of attributes, discovery of clusters with arbitrary shape, minimal requirements for domain knowledge to determine input parameters, ability to deal with noisy data, high dimensionality, interpretability and usability. Clustering techniques can be broadly classified into many categories; partitioning, hierarchical, density-based, grid-based, model-based algorithms.

3.1 K-Means clustering algorithm:

K-Means is one of the simplest unsupervised learning algorithms used for clustering. Given D, a data set of n objects, and k, the number of clusters to form, a partitioning algorithm organizes the objects into k partitions ($k \leq n$), where each partition represents a cluster. The clusters are formed to optimize an objective partitioning criterion, such as a dissimilarity function based on distance. The algorithm and flow-chart of K-means clustering is given below…

Algorithm 1: Basic K-means Algorithm

1. Select K points as the initial centroids.
2. repeat
3. Form K clusters by assigning all points to the closest centroid.
4. Recompute the centroid of each cluster.
5. until The centroids do not change

From the algorithm it is implements in 4 steps:

1. Select K centroid (Can be K values randomly, or K data points randomly)

2. Partition objects into k subsets. An object will be clustered into class J if it has the smallest distance with this class mean compared to the distance with the other class mean

3. Compute the new centroids of the clusters of the current partition. The centroid of the jth cluster is the center (mean point) of the data point whose cluster index is found to be the center of class j in the above step.

4. Go back to Step 3, stop when the process converges.

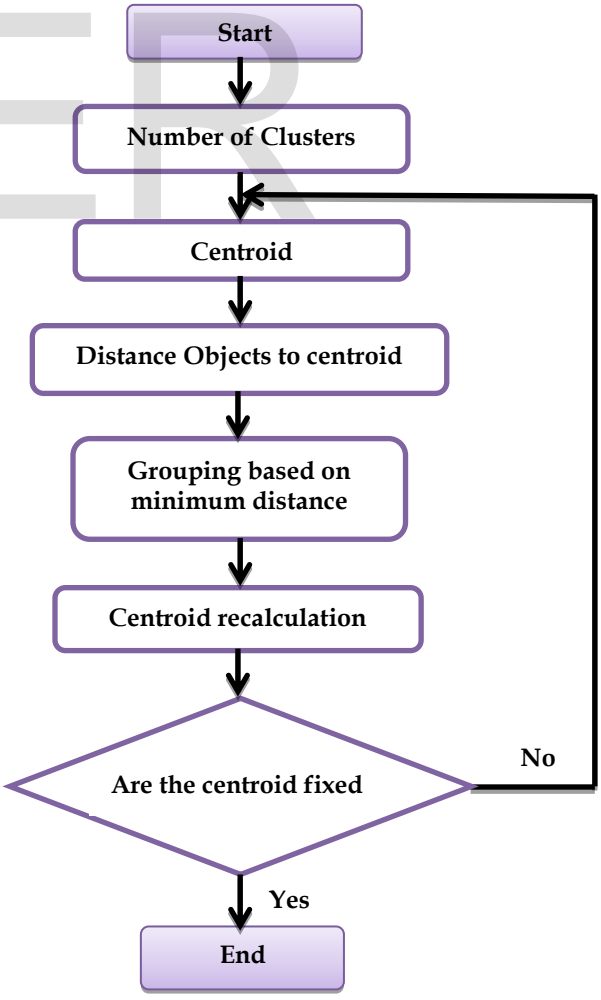


Fig 2: Flow Chart of K-Means Clustering

4. DECISION TREE

Decision tree induction can be integrated with data warehousing techniques for data mining. A decision tree is a predictive node ling technique used in classification, clustering and prediction tasks. A decision tree is a tree where the root and each internal node are labeled with a question. The arcs emanating from each node represents each possible answer to the associated question. Each leaf node represents a prediction of a solution to the problem under consideration. The basic algorithm for decision tree induction is a greedy algorithm that constructs decision trees in a top-down recursive divide-and-conquer manner. Decision tree algorithm generates a decision tree from the given training data.

Algorithm 2: Decision Tree Algorithm

1. Create a node N
2. If samples are all of the same class, C then
3. Return N as a leaf node labeled with the class C;
4. If attribute-list is empty then
5. Return N as a leaf node labeled with the most common class in samples.
6. Select test-attribute, the attribute among attribute-list with the highest information gain;
7. Label node N with test-attribute;
8. For each known value a_i of test-attribute.
9. Grow a branch from node N for the condition test attribute = a_i;
10. Let Si be the set of samples for which test-attribute = a_i;
11. If Si is empty then
12. Attach a leaf labeled with the most common class in samples;
13. Else attach the node returned by generate-decision-tree (S_i, attribute-list-attribute);

Each internal node tests an attribute, each branch corresponds to attribute value, and each leaf node assigns a classification.

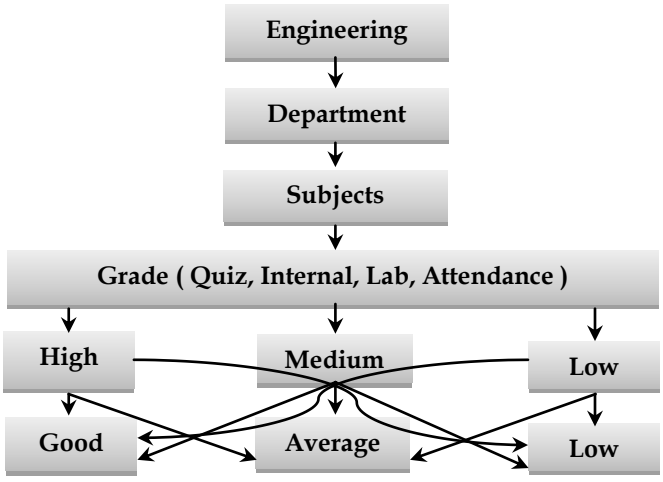


Fig 3: Decision Tree

From 200 students of training data table 1 show 15 samples are shown in the form of training data.

Table 1: Training Sample

No.	Percentage Nominal	internals Nominal	attendence Nominal	quiz Nominal	lab Nominal
1	80	10	10	y	good
2	81	12	9	y	good
3	92	13	10	y	good
4	75	9	10	y	average
5	81	15	10	y	good
6	84	13	9	n	average
7	96	15	10	y	good
8	62	10	9	y	good
9	53	8	5	n	average
10	45	6	6	y	average
11	81	15	10	y	good
12	75	13	10	y	good
13	84	15	9	y	good
14	80	12	10	y	good
15	96	15	10	y	good

If we apply K-Means partitioning method in clustering algorithm on the training data then we can group the students in three classes "High" "Medium" and "Low" according to their new percentage. The table and graph is shown below.

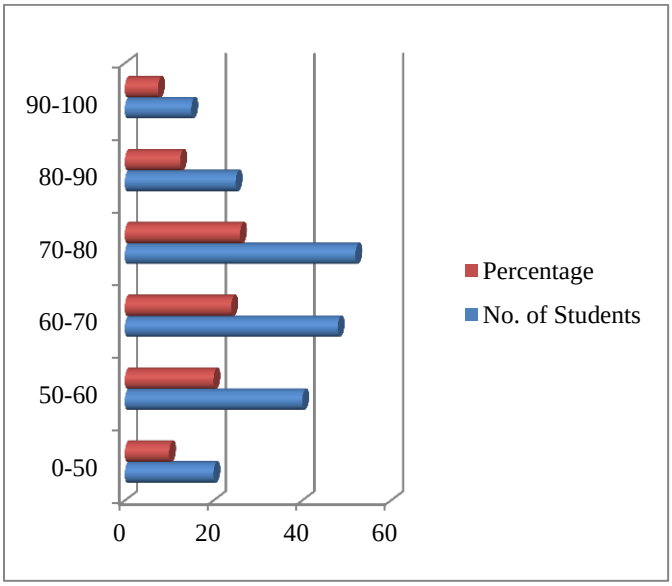
Table 2: Percentage of students according to their exam result

Groups	Exam Percentage	No of Students	Percentage of Students
1	0-50	20	10
2	50-60	40	20
3	60-70	48	24
4	70-80	52	26
5	80-90	25	12.5
6	90-100	15	7.5

In the above table I cluster students among their final result that is percentage 0-50 we have 10% student. From 50-60 we have student percentage is 20% and so on as shown in the table 2.

The graphical representation of final exam result and percentage of students among the students are shown in the graph 1.

Graph 1: Percentage of exams and Number of students



After clustering the students, we group the students into three categories, that is high, medium and low. It is shown in the table 3.

Table 3: Representation of classes

Class	**Percentage in sub**	**No of students**	**Percentage of stud**
High	>=85	38	19
Medium	>=60 and <85	95	47.5
Low	<60	67	33.5

After the pattren is classified from the decision tree we can obtain the specific knowledge discovery to form the knowledge basd system. Similarly the same data mining process can be done to the teacher for classifying their perfromance which help in improving the educational system.

Table 4: Decisions based on the students final results

Groups	Final Exam Percentage	Effort
G1	90-100	He/She is a very good student, there is no need to take any special care. Teacher can give exposure for career growth.
G2	80-90	He/She is a good student. Motivate the student to improve the performance more.
G3	70-80	Is not so good. Need to take care by giving assaignments and conducting quiz.
G4	60-70	Is a medium student. Teacher should take care by giving assaignments, conducting extra classes and more practicals.
G5	50-60	Is a normal student. Need lot of practise not only in reading but also in writing. Need to give more assaignments and make up classes, important topics etc.,
G6	0-50	Is a lower standard student. Need a lot of practise of class work. Need to take more attention, assignments, make up classes, more practical classes, daily tests etc.,

5. CONCLUSION

In this study we make use of data mining process in students data base using K-Means partioning method in clustering algorithms and decision tree technique to analyze and improve the quality of engineering education. The managements can use some techniques to improve the course outcomes to improve the knowledge. We hope that the information generated after the implementation of data mining and data clustering technique may be helpful for the teacher as well as for students. This work may be useful for teacher to analyze and improve the students performance; reducing failing ratio by taking appropriate steps at right time to improve the quality of education. For future work, we hope to refine the technnique in order to get more valuable and accurate outputs, useful for teachers to improve the students learning outcomes.

6. REFERENCES

[1]P.V.Subbareddy and Vuda Sreenivasarao, "The result oriented process for process for students based on distributed data mining", International journal of advanced computer science and applications, Vol. 1, No.5.Nov-2010, pp-22-25.

[2] N.V Anand Kumar, G.V.Uma "Improving Academic Performance of Students by Applying Data MiningTechnique" European Journal of Scientific Research, Volume 34, Issue 4, pp-526-534

[3] Oyelade, Oladipupo & Obagbuwa, "Application of K-means clustering algorithm for prediction of student's academic performance.", IJCSIS2010,vol.7,No.1, pp-292.

[4] Shaeela Ayesha, Tasleem Mustafa, Ahsan Raza sattar and M.Inayat khan, " Data mining model for higher education system", European Journal of Scientific Research,Vol.43 No.1,pp-24-29.

[5] Bindiya M varghese, Jose Tome J, Unnikrishna A and Poulose Jacob K, "Clustering student data to characterize performance patterns", International journal of Advanced computer and applications, Special Issue,pp-138-140.
[6] Kifaya(2009) Mining student evaluation using associative classification and clustering.
[7]Data Mining: Concepts and Techniques, Han JKamber M, Morgan Kaufmann Publishers, 2001.
[8] Introduction to DATA MINING, Tan P., Steinbach M.,Kumar V, Pearson Education, 2006.
[9]Mining student's data to analyze E-learning behavior: A case study-Alaa el-Halees.
[10] Data mining with SQL server 2005 by Zhao Hui.Maclennan.J, Wihely publishing.Inc-2005.