



A Study on the Academic Performance of the Students by Applying K-Means Algorithm

G. Narasinga Rao, M.Ramachandra

Computer Science Engineering Department

GMR Institute of Technology

GMR Nagar, Rajam, Srikakulam District, Andhra Pradesh
India

Abstract: Many Educational Institutions want their students to be placed in reputed companies and hence they focus on the academic performance of their students. If they could able to know about the student's academic performance, then they can focus more on those students whose academic performance is poor. They can take special care on those students whose academic performance is poor and make them prepare for the placement. In this paper, we present a study on the academic performance of the students of a reputed college by applying k-means algorithm of Data Mining which helps them to know about the performance of their students.

Keywords: Data mining, Data Set, Data objects, Clustering, K-means.

I. Introduction

Data Mining is the process of analyzing data available in large data sets and obtains knowledge out of it. The knowledge so obtained is known as the Knowledge Discovery in Databases (KDD). This knowledge helps us in different ways. A data set is a collection of data objects. Each data object represents a record. This data object consists of any number of attributes which specifies the characteristics of the data object. Any Educational Institution consists of students with different categories. We can categorize students based on several properties like attendance, percentage of marks secured, attitude, intelligence etc., to judge their academic performance. In this paper, we present a study on the academic performance of the students based on their attendance, percentage of marks secured. In order to divide the students into different groups based on their attendance, percentage of marks secured, we use k-means algorithm as a simple and efficient tool to cluster them into groups.

II. Methodology

Given a set of n data objects, we will choose k initial centroids from the n data objects where k is a user-specified parameter, namely, the number of clusters desired. Each data object is then assigned to the closest centroid, and each collection of data objects assigned to a centroid is a cluster. The centroid of each cluster is then updated based on the data objects assigned to the cluster. We repeat the assignment and update steps until no data object changes clusters.

Basic k-means algorithm

1. Select k data objects as initial centroids
2. repeat
3. Form k clusters by assigning each data object to its closest centroid
4. Recompute the centroid of each cluster
5. Until centroids do not change.

III. Results

We applied the k-means algorithm on the data set of a reputed Institution consisting of 49 students by considering the percentage of marks secured and attendance as attributes of each of the student data object. Here we consider each student data object as an ordered pair (a, b) where 'a' represents the percentage of marks secured and 'b' represents the attendance of each of the student. Since we consider the percentage of marks and attendance as the attributes of each of the student object, we can categorize the percentage of marks secured as highest, medium and least. Also we can categorize the attendance as highest, medium and least. Therefore the performance of each of the student by means of performance index can be categorized as shown in the table below.

Table 1

Serial No	Percentage of Marks	Attendance%
1	Highest	Highest
2	Highest	Medium
3	Highest	Least
4	Medium	Highest

5	Medium	Medium
6	Medium	Least
7	Least	Highest
8	Least	Medium
9	Least	Least

Out of the 9 categories of the performance index, we can exclude the combination of (Highest percentage of marks, least attendance) and (Less percentage of marks, Highest attendance) which is generally not possible and which are not there in the data set that is considered. Hence we can have 7 possible cases as a performance index and hence we choose 7 initial centroids in the first step for assigning each student data object to one of the 7 clusters.

Out of 49 student's data that we have, we have chosen 7 data objects as initial centroids from the data set that is observed. The following table gives the list of initial centroids chosen from the given data set.

Table 2

Serial No	Percentage of Marks	Attendance
1	70	92
2	72	81
3	56	90
4	55	79
5	55	65
6	45	66
7	46	74

By choosing the above 7 data objects as initial centroids and applying the k-means algorithm, we obtain the following results from the student data set consisting of 49 students. 10% of the students fall in the category of (Highest percentage of marks, Highest attendance%), 27% of the students fall in the category of (Highest percentage of marks, Medium attendance%), 16% of the students fall in the category of (Medium percentage of marks, Highest attendance%), 37% of the students fall in the category of (Medium percentage of marks, Medium attendance%), 2% of the students fall in the category of (Medium percentage of marks, Least attendance%), 4% of the students fall in the category of (Least percentage of marks, Least attendance%) and 4% of the students fall in the category of (Least percentage of marks, Least attendance%). From the above statistics, we can conclude that 90% of the students secure highest and medium percentage of marks if their attendance is medium and highest i.e., if the students attend classes regularly, then their performance can be improved.

IV. Conclusion

In this paper, we present a study on the academic performance of students by considering the percentage of marks secured by the students and attendance percentage of the students as attributes of each of the student data object to monitor the performance of the students. We demonstrated this by applying k-means algorithm on a student data set of a reputed institution for total number of 49 students. We came to a conclusion that 90% of the students secure good marks if their attendance is good. Hence the institution can focus on the student's attendance as an important factor so that the results of their students could be improved.

References

- [1] Larose, D. T., "Discovering Knowledge in Data : An Introduction to Data Mining", Hoboken, NJ, USA: Wiley, 2005.
- [2] S.Sujit Sangsiri, M.Bhosle and K. Sail, "Factors that affect academic performance among pharmacy students," American Journal of Pharmaceutical Education, 2006
- [3] N. V. Anand Kumar and G. V. Uma, "Improving Academic Performance of Students by Applying Data Mining Technique,"
- [4] R. Baker, "Data mining for education," in *International Encyclopedia of Education*, B.McGaw, P. Peterson, and E. Baker, Eds., 3rd ed. Oxford, U.K.: Elsevier, 2010.
- [5] Garcia, E. P. I., & Mora, P. M., "Model Prediction of Academic Performance for First Year Students", *Paper presented at the Artificial Intelligence (MICAI)*, 10th Mexican International Conference on, Nov. 26 2011-Dec. 4 2011.
- [6] Sachin, R. B., & Vijay, M. S, "A Survey and Future Vision of Data Mining in Educational Field", Paper presented at the Advanced Computing & Communication Technologies (ACCT), Second International Conference on 7-8 Jan. 2012.