Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 3, Issue. 7, July 2014, pg.579 – 584

RESEARCH ARTICLE

Performance of Students Evaluation in Education Sector Using Clustering K-Means Algorithms

S.Ganga
M.Phil Research Scholar
Dept. of Computer Science
Alagappa University
Karaikudi
TamilNadu
gangas8872@gmail.com

Dr. T.Meyyappan
Professor
Dept. of Computer Science
Alagappa University
Karaikudi
TamilNadu
meyyappantt@alagappauniv.ac.in

Abstract— Data Mining is the crucial step to find out previously unknown information from huge relational databases. Many techniques and algorithms are there in data mining, namely Association rules, clustering, and classification and prediction techniques. Each of the techniques holds its particular characteristics and behavior. In this paper, the authors proposed a new method with prime focus on clustering technique. The database for the specific set of students was collected. The clustering is made on some detailed manner and the results were produced. The clustering algorithm used here is the K-Means clustering algorithm, to find the nearest possible and cluster the similar group. The advantage of the proposed methodology is that, the cluster groups can be controlled, modified and accessed with ease. The dataset experimented contained 180 records with 63 attributes. The cluster group enables the dataset to be visualized in multiple dimensions with ease of access. The experimental analysis showed how the K-Means algorithm can be used especially to improve engineering students performance in higher education. This document gives formatting instructions for authors preparing papers for publication in the Proceedings of an IEEE conference. The authors must follow the instructions given in the document for the papers to be published. You can use this document as both an instruction set and as a template into which you can type your own text.

Keywords— K-Means Clustering, Data Mining, Cluster Groups, Nearest Neighbor, Relational Databases

I. INTRODUCTION

Data analysis is a process in which raw data is prepared and structured so that valuable information can be extracted from it. Data Mining is the task of identifying valid information from high voluminous databases that are unknown in various aspects. The KDD (Knowledge discovery in Database) is the primary process from which various steps includes data preprocessing, removal of irrelevant data, avoiding duplication of the data, interpretation of the data, mining valid information from the data, pictorial representation and processing of the data taken place.

To mine the unknown data, various methods and techniques were used such as the Association rules, pattern mining, classification technique, clustering technique, prediction, Supervised and unsupervised learning etc., Data analysis has multiple facets and approaches, encompassing diverse techniques under an array of names, in different business, science, and social science domains [1]. Data mining processes have integrated techniques from multiple disciplines such as, statistics, machine learning, database technology, pattern recognition, neural networks, information retrieval and spatial data analysis[6].

The K-Means Algorithm focused in this paper is the clustering technique, which is the supervised learning, and partition type data mining algorithm. Clustering is the process of partitioning or grouping a given set of patterns into disjoint clusters. This is done such that patterns in the same cluster are alike and patterns belonging to two different clusters are different.[7]. It is further categorized into Agglomerative and point assignment. The key operation is that formed has well defined cross points over the centric point.

Several algorithms have been proposed in the literature for clustering: ISODATA, CLARA, CLARANS, DBSCAN, PAM etc. The k-means method remains as the better for many applications that involves multiple datasets and attributes. The rest of the paper is organized as follows. Section 2 consists of the literature review related to the proposed work. Section 3 contains the proposed technology with the K- Means clustering algorithm, and the new procedure for the implementation. Section 4 contains the experimental results and their discussions. Section 5 consists of the conclusion.

II. RELATED WORK

Khaled Alsabti, [7] proposed an efficient K – means clustering algorithm is provided. The algorithm given by the author, organizes all the patterns in a k-dtree structure such that one can find all the patterns which are closest to a given prototype efficiently. All the prototypes are potential candidates for the closest prototype at the root level. This approach can be applied recursively until the size of the candidate set is one for each node. The experimental results demonstrate that the scheme can improve the computational speed of the direct k-means algorithm by an order to two orders of magnitude in the total number of distance calculations and the overall time of computation.

Chunfei Zhang and Zhiyi Fang, [3] proposed an improved K-means clustering algorithm, in which it analyses the advantages and disadvantages of the traditional K-means clustering algorithm elaborates the method of improving the K-means clustering algorithm based on improve the initial focal point and determine the K value. Simulation experiments prove that the improved clustering algorithm is not only more stable in clustering process, at the same time, improved clustering algorithm to reduce or even avoid the impact of the noise data in the dataset object to ensure that the final clustering result is more accurate and effective.

Tamir Tassa and Dror J. Cohen, [10] proposed the use of anonymization using the sequential clustering technique, to provide privacy to highly sensitive datasets. The network setting is divided into multiple data holders. Hence the nearest holders were clustered and data transmission taken place. This will reduce the time need considerably.

H.S.Behera, [2] proposed a new Hybridized K-Means Clustering Based Outlier Detection Technique is proposed in which uses k-means clustering algorithm to cluster the data sets and outlier finding technique (OFT) to find out outlier on the basis of density based and distance based outlier finding technique. K-means has sensitivity over outlier data but can be still used with OFT for the detection of outlier data. Partitioning Around Medoids (PAM) clustering technique has been used for clustering the data and the clustered data is used for outlier detection as discussed in Mohammedd Belal AlZoubi.[8]

Ravindra Jain,[11] proposed a hybrid clustering algorithm for data mining is provided. In the work, a hybrid clustering algorithm based on K-mean and K-harmonic mean (KHM) is described. The result is tested on five different dataset. Its performance is compared with the traditional K-means & KHM algorithm. The result shows that the output obtained from proposed hybrid algorithm is much better than the traditional K-mean and KHM algorithm.

III.PROPOSED METHODOLOGY

The clustering algorithm, which is the partitioning model, is used for implementation. The K – means clustering algorithm is taken into use. The provided dataset is the student database consists of 180 records with 65 attributes that hold the year wise subjects, their mark status, report etc. The detailed study of the algorithm and the implementation procedure is as follows:

A The K – Means Clustering Algorithm

Cluster analysis divides the data into groups that are both meaningful and useful to the end user. The clustering is grouping of similar items based on some metrics. The grouping takes place on the nearest neighbor pattern. The procedure for the algorithm is briefly given. Distribute all objects to K number of different cluster at random. Calculate the mean value of each cluster, and use this mean value to represent the cluster, and then Re-distribute the objects to the closest cluster according to its distance to the cluster center. After, distribution is taken place update the mean value of the cluster. That is to say, calculate the mean value of the objects in each cluster and finally, calculate the criterion function, until the criterion function converges.

The K- mean algorithm is considered as one of the classic algorithm to solve the cluster process and it is a simple and faster algorithm to do. Since, the complexity is minimal in the algorithm, particularly for the large Itemsets to form the clusters, this algorithm is preferred. It produces local optimum solution to the given large database. The algorithm is at the same time very simple and quite scalable, as it has linear asymptotic running time with respect to any variable of the problem. K-means clustering is a method of cluster analysis which aims at the partition of n observations (x1, x2, …, xn), where each observation is a d-dimensional real vector that are grouped into k clusters in which each observation belongs to the cluster with the nearest mean [9].

B Procedure for clustering student Database

The student database containing 180 records as considered for mining. The total number of attributes considered here is 65. The working procedure to calculate the mean and clustering of the dataset is given as follows:

Step 1: Read the given student Database.
Step 2: Preprocess the dataset by avoiding duplicate records and remove irrelevant data.
Step 3: Calculate the File size, total number of records and the attributes from the database.
Step 4: Split and cluster the total records into three different clusters namely 1,2,3.
Step 5: Form the clusters based using on the series order three clusters now holds 60 records each.
Step 6: The individual cluster group is again clustered, or partitioned into two ways.
Step 7: Find students those with high CGPA & Arrears.
Step 8: Partition the data by CGPA and Arrear levels based on their low and high mark values.
Step 9: Calculate mean of high CGPA and partition the data based on mean value.
Step 10: Determine the sum of individual subject marks based on their cluster groups.
Step 11 : Clusters are generated for CGPA and Arrear data.

Thus by using the K- means clustering algorithm the biggest dataset of the students report is easily clustered. The main advantage of clustering the dataset is that the time required to process and access each dimension is reduced minimally by choosing the mean. The cluster groups thus formed is highly accessible to end users, since it is partitioned and then clustered. The user from these clusters can now view the mark wise high scores, subject wise, series wise, Year wise scores and the student's percentage cluster group.

C Proposed Architecture:

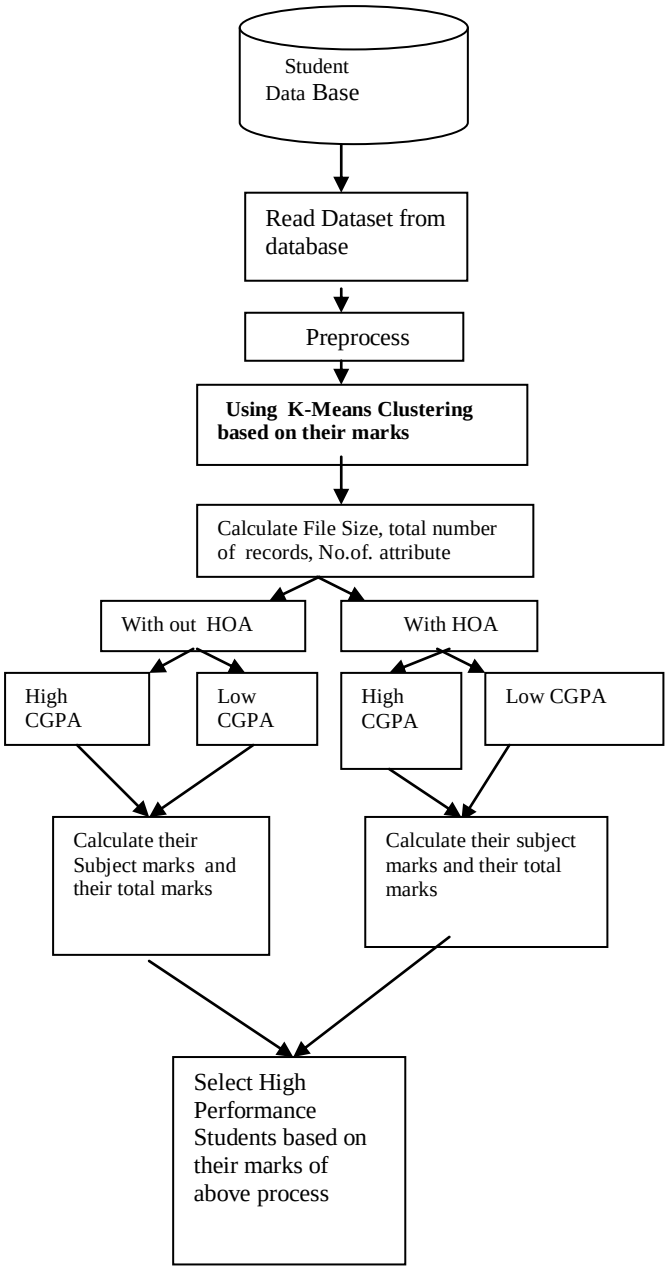

Fig 3.1 Block Diagram of the proposed Architecture

IV. EXPERIMENTAL RESULTS

The results of the proposed method are tabulated and plotted. The graph shows the various clusters based on the CGPA obtained by the students with arrear and without arrear. Another Chart shows the subject cluster of the students.

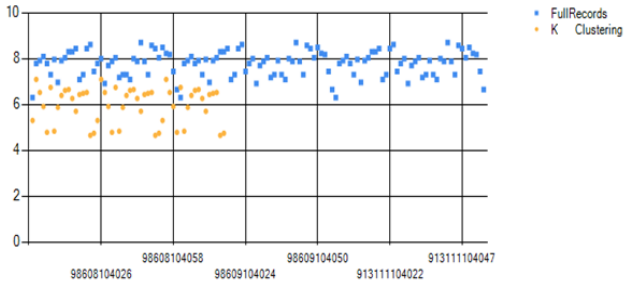

Fig 1 Clustering of CGPA and Arrear Data Chart

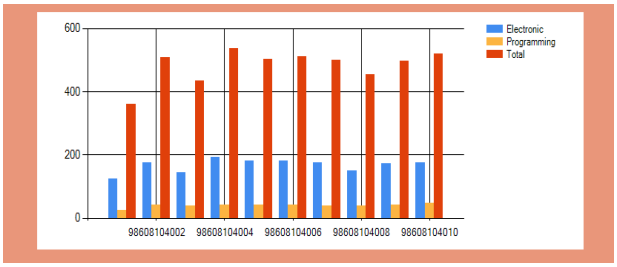

Fig 2 Clustering of student subject Chart

Table 1: Students Related Variables

Variable	Description
ID	Students Identity Number
Name	Students name
Sex	Students Sex
Sub	More than 60 Subjects marks for 4 years.
Tot_mark	Total Marks of all subjects
CGPA	Cumulative Grade Point Access

- Tot-Mark – It is declared as Total mark of all the Subjects of the course.
- CGPA - Cumulative Grade Point Access in total marks of the course and it is declared as response variable. It ia also split into 7 class values:A-90%-100,B-80%-89,C-70%-79%,D-60%-69%,E-50%-59%,AB-Absent,RA-Re appear.

V. CONCLUSION

The proposed work uses the K- means clustering algorithm and forms various cluster groups from the student database. The cluster groups make it flexible for the end user, to search for the particular query with less effort. The partitioned dataset, allow ease of access in multiple dimension. The proposed work, enables the user to determine the performance evaluation of every student based on their mark, subject, year and series wise. The experimental results reveals clusters based on mean Value.

REFERENCES

[1] H.J.Adèr,. (2008). Chapter 14: Phases and initial steps in data analysis. In H.J. Adèr & G.J. Mellenbergh (Eds.) (with contributions by D.J. Hand), Advising on Research Methods: A consultant's companion (pp. 333-356). Huizen, the Netherlands: Johannes van Kessel Publishing.

[2] H.S.Behera, Abhishek Ghosh, Sipak ku. Mishra, "A New Hybridized K-Means Clustering Based Outlier Detection Technique For Effective Data Mining", in International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 4, April, pp:287 – 292.

[3] Chunfei Zhang, , Zhiyi Fang, "An Improved K-means Clustering Algorithm", in Journal of Information & Computational Science, vol: 10, No: 1,2013, pp: 193–199.

[4] Chonghui GUO, Li PENG, "A Hybrid Clustering Algorithm Based on Dimensional Reduction and K-Harmonic Means," IEEE 2008.

[5] J. Han and M. Kamber, "Data Mining Concepts and Techniques, Morgan Kaufmann Publishers," San Francisco, CA, pp. 335–391, 2001.

[6] Irshad Ullah, "Data Mining Algorithms And Medical Sciences", in International Journal of Computer Science & Information Technology (IJCSIT), Vol 2, No 6, December 2010, pp127 – 136.

[7] Khaled Alsabti, Sanjay Ranka, Vineet singh, "An Efficient K-Means Clustering Algorithm", pp: 1-6.

[8] "An Effective Clustering-Based Approach for Outlier Detection" by Mohammed Belal Al- Zoubi European Journal of Scientific Research Vol.28 No.2 (2009).

[9] .Raed T. Aldahdooh and Wesam Ashour, "DIMK-means ―Distance-based Initialization Method for K-means Clustering Algorithm", in International Journal. Intelligent Systems and Applications, 2013, Vol :02, pp: 41-51.

[10] Tamir Tassa and Dror J. Cohen, "Anonymization of Centralized and Distributed Social Networks by Sequential Clustering", In Ieee Transactions On Knowledge And Data Engineering, Vol. 25, No. 2, February 2013 pp:311-324.

[11] Zhaoxia Tang, K-means clustering algorithm based on improved genetic algorithm, in: Journal of Chengdu University (Natural Science Edition), Vol: 30, No :2, 2011,pp: 162-164.

