International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Quality Assessment of Secondary Level Students using K-means Clustering

Baidyanath Sou

Assistant Professor, Department of Computer Science, Jagannath Kishore College, Purulia, West Bengal

Abstract

In this article, the main focuses on the evaluation of secondary level students. For this purpose, primary data was obtained individually from a few secondary level schools in the district of Purulia, West Bengal. In most school education systems in India, every level of student is evaluated in terms of marks gained in any sort of examination, whether written or oral. In this paper, the assessment procedure is proposed in a different direction. The entire procedure is examined utilizing five factors connected with secondary level students including performance of annual examination. Hence these variables(factors) are placed in a Domain space for analysis using K-Means clustering algorithm and deterministic model for achieving the solution with good outcome. The performance of the students as measured by this approach will be beneficial to those students who are hesitant to talk personally about their shortcomings, which will be recognized automatically by this method.

Keywords: Student Assessment, Clustering, K-Means, deterministic model.

1. Introduction

Data analysis is the process of preparing and structuring raw data to ensure that important information may be derived from it. Data mining is the task of identifying reliable information from large databases that are unclear in many ways [1]. Knowledge discovery is the primary procedure in which multiple phases such as data preprocessing, elimination of irrelevant data, preventing data duplication, data interpretation, extraction reliable information from the data, graphic representation, and data evaluating take place [2][3]. Various approaches and techniques, such as association rules, classification technique, clustering technique, prediction, Supervised and unsupervised learning, and so on, were utilized to extract the unknown data. The K-Means approach is a supervised learning and partition type data mining approach that is the topic of this paper. Clustering is the technique of grouping or separating a set of patterns into distinct clusters [4]. This is done so that patterns within the same cluster are similar and patterns between clusters are distinct.[5]. It is further divided into two types: agglomerative and point assignment. The crucial operation is that the cross points created across the centric point are well defined. The Elbow technique is one of the most widely applied approaches for determining the best value of K-clusters in a data set. It used this strategy by combining the classic K-means algorithm with the Elbow method to provide an optimal method of counting optimal number of clusters [6] [7].

The primary goal of this study was to develop a novel clustering model based on the eight parameters including performance of annual examination. This method shows quality of a student depends not only the performance of examination, but also other parameters need to consider. Following the grouping of

students, an improvement plan needs to develop for each group of students, stressing the areas where each student was underperforming.

The remainder of the paper is structured as follows: Section 2 describes the suggested technology, which uses the elbow method, K-Means clustering algorithm, as well as the new implementation procedure. The experimental results and their explanations are presented in Section 3. The conclusion is found in Section 5.

2. Materials and Methods

2.1. Research Data

For the formation of a Database, a considerable number of higher secondary Standard Schools in Purulia, West Bengal has been considered. A total 50 students have been chosen randomly from class 10 standard.

2.2. Proposed Method

The main goal of this research study was to optimize the number of clusters of K-Mean methods in the suggested system while using the Elbow technique to specify the number of clusters during the evaluation procedures.

There are Four phases of processing were applied to the data sets (Figure 1).

Step 1: The first stage was scaling, which involves applying standardization/normalization to dataset features that have a bigger magnitude variance than others.

Step 2: The Elbow approach was then used to the dataset to determine the best number of K-clusters.

Step 3: The third step was to execute the K-means algorithm to clusters based on their performance.

Step 4: The total score is assessed using a deterministic model, in which the collective evaluation in each cluster size is assessed by aggregating the mean of individual marks in each cluster.

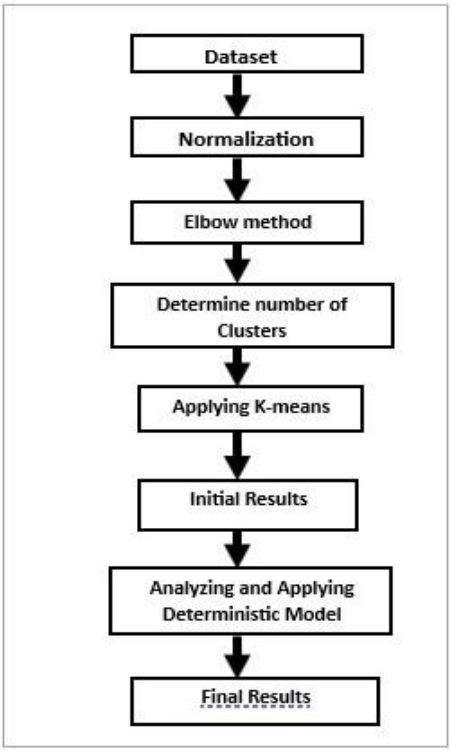

Figure 1: Flowchart of proposed model

2.2.1 Data Modelling

To assess a student, seven parameters has been chosen: C1 to C4 along with the performance of annual examination (C5). The parameters are C1: Attendance in School, C2: Result in different class test, C3: Financial background, C4: Discipline in school. The parameters are rated with 10-point scale. The parameter has been allocated such that if a student has 80% attendance in one academic year, then he/she will score 8 out of 10. If a student scored outstanding in different class test in one academic year, then he/she will score 10 out of 10. Based on the financial condition of a student, he/she will score between 1 to 10. Similarly discipline of a student also score between 1 to 10 .C5 shows the marks obtain in annual examination which is out of 500. To successfully execution this method, C5 parameter must be scaled out of 10. Scaling datasets is a means of normalizing the range of independent variables that is required by many machine learning algorithms. Hence. By subtracting the mean value of each feature and scaling it by dividing non-constant features by their standard deviation, the shape of the distribution and centralization of the data are ignored. The standard deviation is represented in this method as follows [8]:

$$Z = \frac{x-\mu}{\sigma}$$

where, Z= normalized value

X= raw value of the data point

μ = population mean

σ = population standard divisor for the dataset

2.2.2. K-Means Method

In 1967, James MacQueen developed the K-means methodology, which is a form of partitioning/clustering method [9]. It is one of the easiest data mining partitioning/clustering algorithms that uses the Euclidean distance function. The functional objectives of this method are to reduce the cluster performance index, the error sum of squares, and the error threshold, which are the underpinning of discovering the best value of k divisions to achieve a certain criterion. The K-means has several advantages, including simplicity, efficiency, and rapidity [10]. However, this strategy is heavily reliant on starting data points and the diversity in choosing first samples, which are typically aimed at different results [11]. The algorithm is illustrated by the Hartigan-Wong algorithm [12]. The total within-cluster variance is defined as the sum of the squared distances between objects and their corresponding centroid.

$$W(C_k) = \sum_{x_i \in C_k} (x_i - \mu_k)^2$$

where $W(C_k)$= *the total within-cluster variation,* x_i = *a data point in the cluster* C_k *and* μ_k =*the mean value of the points allocated to the cluster* C_k.

Therefore, the total within-cluster variation is defined as follows:

$$\sum_{K=1}^{k} W(C_k) = \sum_{k=1}^{k} \sum_{x_i \in C_k} (x_i - \mu_k)^2$$

2.2.3. Elbow Method

The Elbow method examines the percentages of variation shown as a function of the ideal number of clusters in the K-means algorithm [13]. The key idea is to start with k = 2 and keep increasing it by one point while computing the cluster and the cost of training [14]. The cost will significantly decrease for some k value, and then the cost will climb as the k point is raised higher. The k value to search for is the

point at which the cost drop turns to a cost increase and look like an elbow [15]. The Elbow method is defined as the within-groups sum of squares (WSS) where distance calculated statistically from the group means to the same cluster centroid is the squared average distance of all data points for a cluster.

$$WSS = \sum_{i=1}^{m} (x_i - c_i)^2$$

The K-means and Elbow methods together can locate the value of k at the best cluster to determine k as the number of clusters produced. Within the K-means methodology, the Elbow method is used to select the best number of k clusters for data categorization. The sum of squares error can be used to express the Elbow technique [16][17].

$$SSE = \sum_{k=1}^{k} \sum_{x_i \in sS_k} \|x_i - c_k\|_2^2$$

where k is the number of clusters that comprised c, which is the i th cluster, with x being the data provided at each cluster.

3. Result and Discussion

The experiment is implemented in R programming language which is used to analysis statistical data [18]. This software is appropriate for data modelling and contains predefined machine learning libraries. After successfully execution of elbow method on the dataset, the result shows the optimal number of clusters is 5 shown in figure 2. Choosing the number of cluster, K-means algorithm is executed on the dataset. For k = 5, the cluster size in cluster 1 is 7, cluster 2 is 11, cluster 3 is 8, cluster 4 is 18 and cluster 5 is 6. The graphical representation of overall performance of the constructed cluster is presented in figure 3. The dataset dimensions are consisted of N x M matrices, where N represent the number of rows consisting of the number of students and M represents columns numbers representing the parameters of students. The overall performance of the students is evaluated using the deterministic model [19] shown below, where cluster size is evaluated by summing the average of each student's scores in each cluster.

$$\frac{1}{N}\left(\sum_{j=1}^{N}\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)\right)$$

where, N= Total number of students in a cluster and n= Number of parameters in the data set.

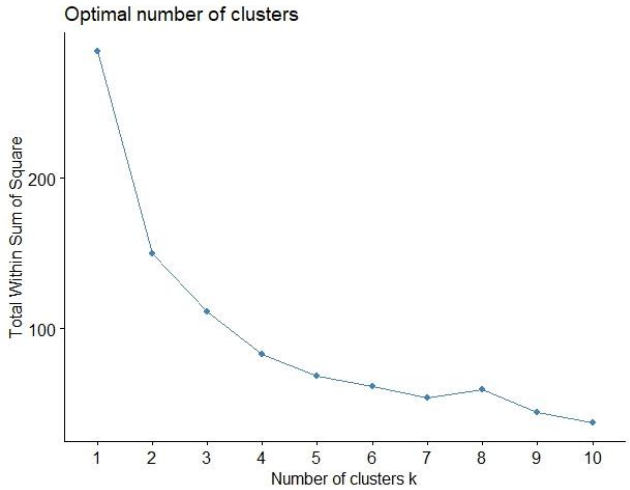

Figure 2: result of the elbow method

After applying deterministic model, the overall performance of cluster 1 is 83.8 %, cluster 2 is 74.2%, cluster 3 is 62.5%, cluster 4 is 58.2%, and cluster 5 is 38.3% shown in table 2. It has been observed that cluster 1 which has 7 students levelled as "excellent", not only preconformed well in annual examination but also scored well in all categories. In cluster 2 comprised of 11 students levelled as "very good" performed well most of the categories. In cluster 3 which has 8 students levelled as "Good", performed well in some categories and in some category's performance not good. In cluster 4 levelled as "fair", performed not good in most of the categories but some students performed well in annual examination. Similarly in cluster 5 which has 6 students levelled as "poor", performed low in all most all the categories although only 1 student performed well in annual examination. It observes that to assess a quality of a student, he/she must perform not only in examination but also other parameters.

Table 1: Performance index

80 and above	Excellent
70 to 79	Very Good
60 to 69	Good
50 to 59	Fair
Below 50	Poor

Table 2: clusterwise Performance Table

Cluster	Cluster size	Overall Performance
1	7	83.8
2	11	74.2
3	8	62.5
4	18	58.2
5	6	38.3

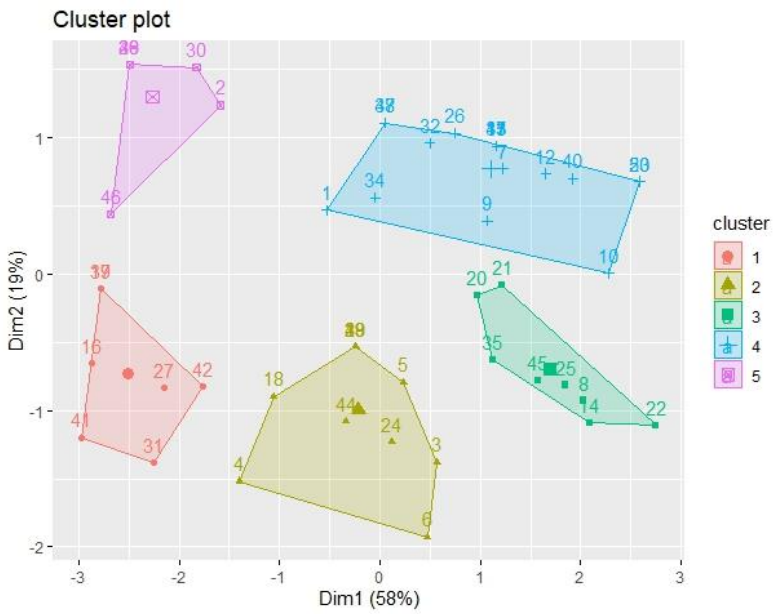

Figure 3: Graphical representation of the result

4. Conclusion

This article presented a simple and qualitative way for comparing the prediction power of a clustering algorithm. It illustrated the method using the k-means clustering algorithm together with the deterministic model in in the dataset collected from secondary school. and offered a numerical analysis of the results for performance or quality assessment. It also enables academic planners to monitor each student's performance and progress level in their higher education and aids in decision support on the future academic session.

References

1. Che, Dunren, Mejdl Safran, and Zhiyong Peng. "From big data to big data mining: challenges, issues, and opportunities." Database Systems for Advanced Applications: 18th International Conference, DASFAA 2013, International Workshops: BDMA, SNSM, SeCoP, Wuhan, China, April 22-25, 2013. Proceedings 18. Springer Berlin Heidelberg, 2013.
2. Idri, Ali, et al. "A systematic map of medical data preprocessing in knowledge discovery." *Computer methods and programs in biomedicine* 162 (2018): 69-85.
3. Ristoski, Petar, and Heiko Paulheim. "Semantic Web in data mining and knowledge discovery: A comprehensive survey." *Journal of Web Semantics* 36 (2016): 1-22.
4. Bansal, Arpit, Mayur Sharma, and Shalini Goel. "Improved k-mean clustering algorithm for prediction analysis using classification technique in data mining." International Journal of Computer Applications 157.6 (2017): 0975-8887.
5. Alsabti, K, et al. "An efficient k-means clustering algorithm." ,1997.
6. Mautz, Dominik, et al. "Discovering non-redundant k-means clusterings in optimal subspaces." Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining. 2018.
7. Syakur, M. A., et al. "Integration k-means clustering method and elbow method for identification of the best customer profile cluster." IOP conference series: materials science and engineering. Vol. 336. IOP Publishing, 2018.
8. Mertens, Willem. *Quantitative data analysis*. Springer, 2017.
9. Tang, JingLei, et al. "Weed identification based on K-means feature learning combined with convolutional neural network." Computers and electronics in agriculture 135 (2017): 63-70.
10. MacQueen, James B. "On the Asymptotic Behavior of k-means." Defense Technical Information Center 10 (1965).
11. Jamadi, Nur Athirah, et al. "Privacy Preserving Data Mining Based on Geometrical Data Transformation Method (GDTM) and K-Means Clustering Algorithm." International Journal of Innovative Computing 8.2 (2018).
12. Yedla, Madhu, Srinivasa Rao Pathakota, and T. M. Srinivasa. "Enhancing K-means clustering algorithm with improved initial center." International Journal of computer science and information technologies 1.2 (2010): 121-125.
13. Amaral, Getulio JA, et al. "K-means algorithm in statistical shape analysis." Communications in Statistics—Simulation and Computation® 39.5 (2010): 1016-1026.
14. Ghayekhloo, Mohadeseh, et al. "A novel clustering approach for short-term solar radiation forecasting." Solar Energy 122 (2015): 1371-1383.

15. Sujatha, S., and A. Shanthi Sona. "New fast k-means clustering algorithm using modified centroid selection method." International Journal of Engineering Research & Technology (IJERT) 2.2 (2013): 1-9.
16. Antunes, Mário, et al. "Knee/elbow point estimation through thresholding." 2018 IEEE 6th International Conference on Future Internet of Things and Cloud (FiCloud). IEEE, 2018.
17. Marutho, Dhendra, Sunarna Hendra Handaka, and Ekaprana Wijaya. "The determination of cluster number at k-mean using elbow method and purity evaluation on headline news." 2018 international seminar on application for technology of information and communication. IEEE, 2018.
18. Reimann, Clemens, et al. Statistical data analysis explained: applied environmental statistics with R. John Wiley & Sons, 2011.
19. Uusitalo, Laura, et al. "An overview of methods to evaluate uncertainty of deterministic models in decision support." Environmental Modelling & Software 63 (2015): 24-31.