Implementation of an Improved K-Means Clustering Algorithm for Balanced Clusters

T.Thilagaraj[1], Dr.N.Sengottaiyan[2]

[1] Part-Time Ph.D. (Category – B), R&D Centre, Bharathiar University, Coimbatore & Assistant Professor in Computer Applications, Kongu Arts and Science College, Erode, Tamil Nadu, India

[2] Director, Professor in Sri Shanmugha College of Engineering and Technology, Sankari, Tamil Nadu, India

Abstract

Discovering hidden information from a large data set and also it involves in analyzing patterns to make decisions on the correct time. The clustering techniques strictly allow an object to be the member of the cluster and partitioning of the data is also taken place in clustering. The cluster's quality will be assigned according to the techniques implemented on it. The K-Means clustering uses the Euclidean distance formula to find the distance between two points and the clustering will happen by holding minimum distance objects. But the clusters are changed if centroid value is different and the centroid will be chosen randomly is the main drawback. So to overcome this issue the centroid is found by choosing a minimum distance from an initial center point. By using the best centroid the clusters are found so here the centroid is fixed and also the mean is used to make balance clusters. Here the four cluster categories are found to find the performance level of students.

Keywords: *K-Means, Clustering, Data Mining, Balanced cluster*

1. Introduction

Discovering various patterns to make predictions on future activities, the extraction of previously unknown and useful information's are known as data mining. Nowadays it is difficult to make data mining techniques effectively due to enormous database systems and data stores. The main objective of the clustering analysis is to divide the objects according to its categories or level [1]. The supervised and unsupervised are the two categories of clustering analysis are available in data mining. More Human Interactions are needed for supervised clustering and the minimal level of knowledge about the dataset is needed in unsupervised techniques. In classification, the predefined classes are available to furnish the tasks but in clustering it not hold any predefined classes and the final clusters will be found only after completion of the execution. At the end of execution, each object will become a member of any one cluster [2]. The New compact representation is found by summarizes of the given data set with its centroid [3].

One of the most common techniques to extract useful information for practical applications is the K-Means Clustering Algorithm. The simplicity and speed are the two main merits of the K-Means clustering algorithm while dealing with large datasets [4].

The major issue of the K-Means clustering algorithm is that it forms different clusters for different initial centroid values [3]. And also is unable to find the global optimal solution and also the final clusters are not balanced on a few times [4]. This method will calculate the distance between each object from the centroid which is found in the current iteration. Likewise, for every iteration the distance calculation will occur is one of the shortcomings of this method [5]. If the unrelated property exists means the error will occur on finding the characteristics of data cluster [6]. Here one proposed method is introduced to improve the performance of the K-Means Clustering Algorithm.

Every Institution has the responsibility to provide proper training for the students to face their real-life challenges. Matching the applicant with their core related job is the biggest challenge for every institution [7]. The proper training is needed to give to the students according to their levels. To identify their levels the improved clustering technique is implemented on the student performance and it is measured by averaging of Academic, Aptitude, Technical and Interpersonal.

The R-Tool is the statistical programming language which holds more than 70 versions and 14,000 packages in CRAN, the global repository for various processes and it is open source. The Integrated Development Environment, R-Studio helps the researchers to facilitate all requirements and it is easy to create complex data models [8].

2. Related Work

According to Malinen and Fränti [9], the main objective of balanced clustering is to make equal size in all clusters and this is faster than the original K-Means algorithm. The proposed balanced k-means algorithm holds high potential to deal with larger datasets but the mean square error value is slightly higher.

Chang, et al. [10] Here the exclusive lasso model is used to deal with the variables in the same group. The multi-task feature selection is using to make better performance while clustering. The exclusive lasso in k-means and min-cut clustering algorithms are used to find the best data points to forms balanced clusters. The new iterative design is also used to make the function much easier.

Levin [11] Here the balanced clustering technique initially finds the balance indices. The difference between cluster parameters, number of elements to form a cluster, cluster weight, total edges or arcs of a cluster and structure of cluster are found to form a balanced cluster. The balance indices calculated based on the structure of the cluster and its elements, the clustering solutions for sample network and balanced clusters are found for students among several teams are the three examples discussed in this paper.

The author Gupta [12] discussed the development in balanced clustering using various techniques. The balanced clustering is achieved by obtaining its size and density. The balanced k-means clustering algorithm is used here.

3. Proposed Methodology

Input:

sid = {sid1, sid2, sid3,…., sidn} //student id
spv = {x_1, x_2, x_3,…., x_n} //student performance values

Output:

Using mean values and best centroid, the four ranges of clusters found.

Algorithm:

Step 1: Where ‘sid' and ‘spv' are the vectors variables passed to the function placeclust and it will be received by vector variables' and ‘v' correspondingly.
Step 2: The length of ‘v' must be found and store it in ‘n1'. Now find the mean value of ‘v' and if it holds decimal places then use a round function to change as an integer, store the result in the vector ‘cm'.

$$cm = round\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) \quad (1)$$

Step 3: Initialize m1, m2, v1, v2, cfv as zero and assign the mean value ‘cm’ to the vector variables ‘cm1’ and ‘cm2’ for further process.
Step 4: For each i=1 to n.

4.1 For each j=1 to n.
4.2 If student performance value ‘v[j]’ is equal to mean value ‘cm1’ then performance value ‘v[j]’ is assigned to ‘v1’ and mean value ‘cm1’ will be stored in ‘m1’. End if.
4.3 If ‘m1’ not equals to zero, then break, End if.
4.4 End for.
4.5 If ‘m1’ not equals to zero, then break, End if.
4.6 The initial centroid value ‘cm1’ will be decrement by 1.
End for.

Step 5: For each i=1 to n.

5.1 For each j=1 to n.
5.2 If student performance value ‘v[j]’ is equal to mean value ‘cm2’ then performance value ‘v[j]’ is assigned to ‘v2’ and mean value ‘cm2’ will be stored in ‘m2’. End if.
5.3 If ‘m2’ not equals to zero, then break, End if.
5.4 End for.
5.5 If ‘m2’ not equals to zero, then break, End if.
5.6 The initial centroid value ‘cm2’ will be incremented by 1.
End for.

Step 6: To find the best centroid, bm1= cm-m1, bm2=m2-cm, Now compare bm1 and bm2.

Step 7: If bm1 less than or equal to bm2, then assign ‘v1’ as best centroid and store it in ‘cfv’. If ‘bm1’ greater than ‘bm2’ then assign ‘v2’ as best centroid and store it in ‘cfv’.

Step 8: First cluster begins, assign j=1, ‘a1’ and ‘a2’ as zero.

8.1. For each i=1 to n.

8.2 If performance value v[i] less than or equal to best centroid value ‘cfv’ then vector ‘a1[j]’ assigns the student-id value ‘s[i]’ and ‘a2[j]’ assigns the performance value ‘v[i]’. Now Increment the j value by 1. End if.

8.3 End for.

Step 9: Second cluster begins, assign j=1, ‘b1’ and ‘b2’ as zero

9.1. For each i=1 to n.

9.2 If distance value v[i] greater than best centroid value ‘cfv’ then vector ‘b1[j]’ assigns the student-id value ‘s[i]’ and ‘b2[j]’ assigns the performance value ‘v[i]’. Now Increment the j value by 1. End if.

9.3 End for.

Step 10: Now return a1, a2, b1, b2 as list and then convert as a vector for further clustering.

Step 11: Assign sid=a1 and spv=a2, Now call the function placeclust. Go to step1.

Step 12: Now two clusters values will be returned from the function placement. The low range and medium range clusters were found.

Step 13: Assign sid=b1 and spv=b2, Now call the function placeclust. Go to step1.

Step 14: Now another two clusters values will be returned from the function placeclust. That is a high range and very high range clusters. Now the four ranges of clusters were found from the given values using its mean value.

4. Result and Analysis

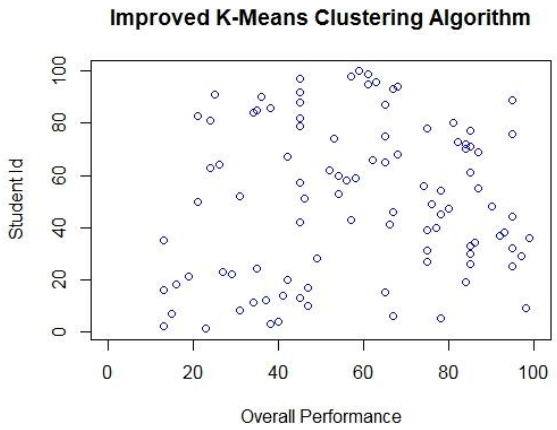


Figure 1: Performance of Students is Plotted according to the Student Id.

In the above figure, the student's performance value is calculated using the average of Aptitude, Technical, Interpersonal and Academic marks. The student performance is

implied as to the x-axis and student Id is implied as to the y-axis. This improved K-Means clustering Algorithm is implemented using R-Tool.

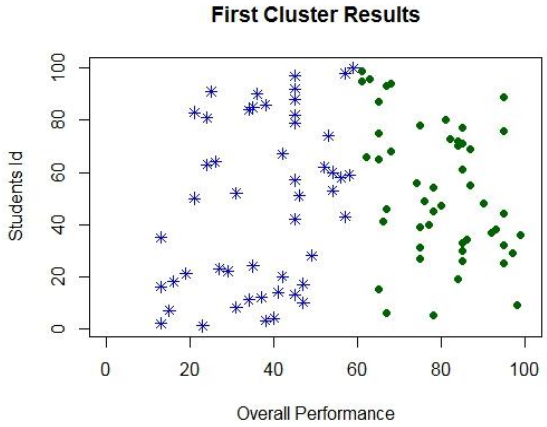


Figure 2: Two clusters are formed using the best centroid value

In the above figure, the two groups of clusters were formed by implementing the above-said algorithm. Here the best centroid value was found to form these clusters.

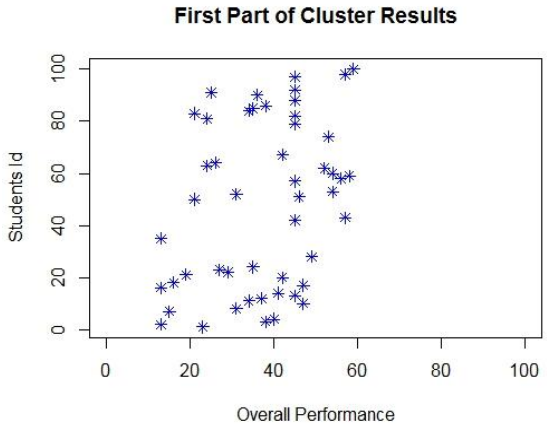


Figure 3: First part of the cluster is separated

In the above figure, the founded cluster is separated for the next level of clustering by using the same methodology.

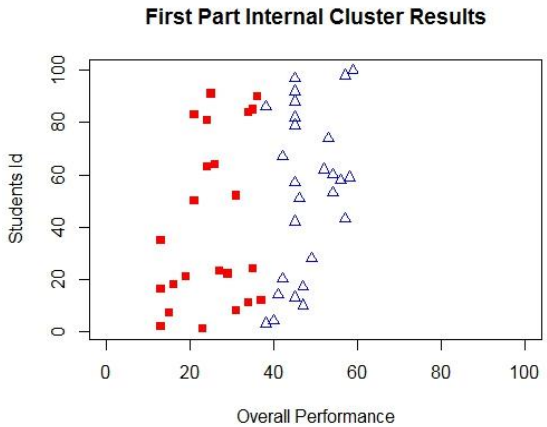


Figure 4: Again two clusters are formed from the first part of a cluster

In the above figure, the first cluster is displayed in a filled square (red) which implies the low range of cluster and second cluster is viewed in triangle point up (blue) which implies a medium range of clusters.

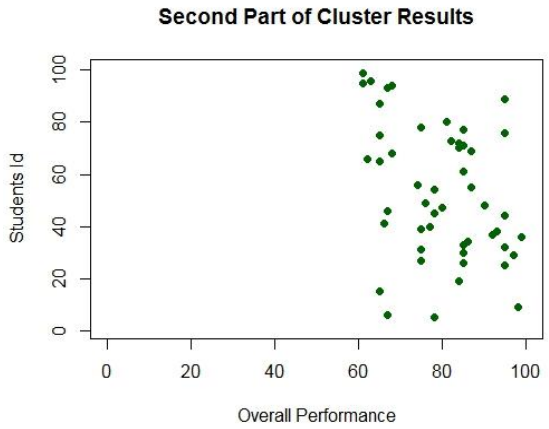


Figure 5: Second part of the cluster is separated

In the above figure, the second part of the cluster is separated for the next level of cluster.

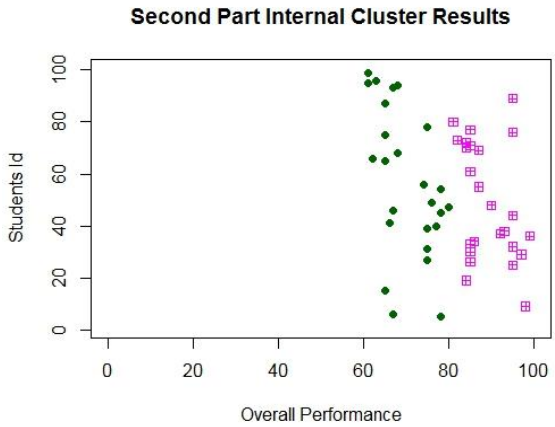


Figure 6: The two clusters are formed from the second part of the cluster

In the above figure, the first cluster is displayed in a solid circle (dark green) which implies the high range of clusters and the square plus (Magenta) implies the very high range of clusters.

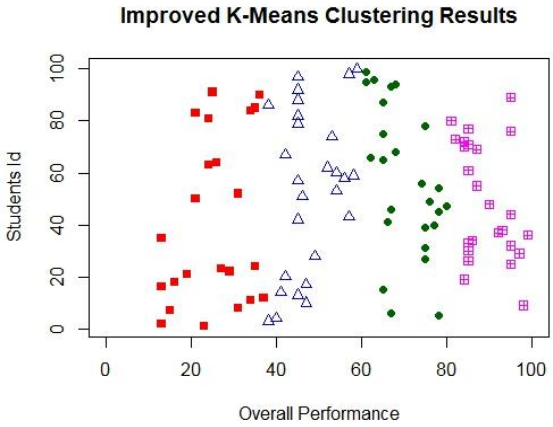


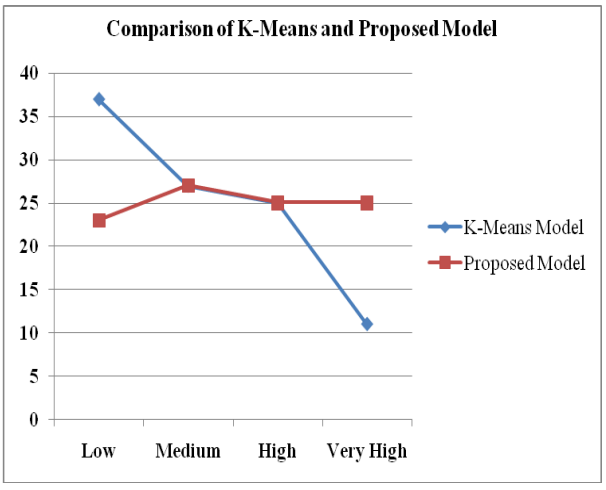
Figure 7: The complete view of all four clusters

In the above figure, the four clusters are filled square represents the low range of cluster, triangle point up represents the medium range of clusters, the solid circle represents the high range of clusters and square plus represents a very high range of clusters.

S.No.	Category	K-Means Model	Proposed Model
1	Low	37	23
2	Medium	27	27
3	High	25	25
4	Very High	11	25

Table 1: Comparison between K-Means Clustering Model and Proposed model

In the above table, the comparison of cluster sizes between two models K-Means and Proposed K-Means clustering. The cluster sizes are balanced in the Proposed K-Means clustering model while comparing with the K-Means clustering model.



Graph 1: Comparison of K-Means with Proposed K-Means Model

In the above graph, the clustering sizes are balanced in the proposed model. The K-Means clustering will form the few clusters with a high range of distance.

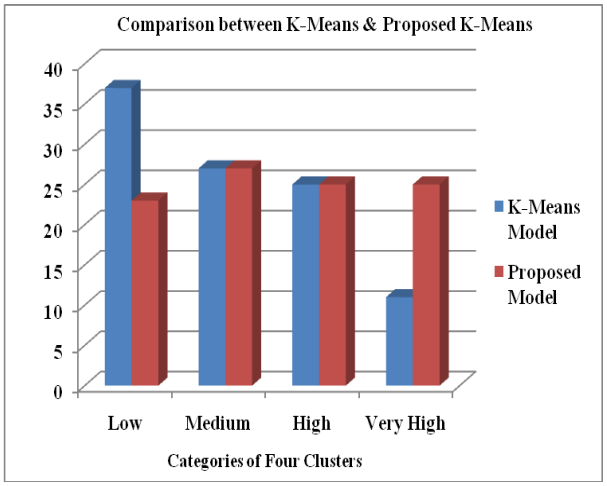


Chart 1: The Cluster sizes and their categories

In the above chart, we can easily find the difference between both models. The proposed model will form the cluster which is almost balanced while comparing with the original K-Means Clustering Algorithm.

5. Conclusion

The Proposed work represents four different ranges of clusters are low, medium, high and very high. This helps the placement officer to categorize the students for placement training in a balanced range. Here the proper allocation of students upon their level and also minimize the cost by maintaining balanced strength in all batches. It is very easy to identify the low-level students and provide the appropriate training to them. While comparing with the original K-Means clustering algorithm the proposed balanced clustering model finds the fixed centroid value and makes unchangeable clusters.

REFERENCES

[1] *G. Usman, U. Ahmad, and M. Ahmad, "Improved k-means clustering algorithm by getting initial centroids," World Applied Sciences Journal, vol. 27, no. 4, pp. 543-551, 2013.*

[2] *N. Kaur, J. K. Sahiwal, and N. Kaur, "Efficient k-means clustering algorithm using the ranking method in data mining," International Journal of Advanced Research in Computer Engineering & Technology, vol. 1, no. 3, pp. 85-91, 2012.*

[3] A. Thammano and P. Kesisung, "Enhancing K-means algorithm for solving classification problems," in 2013 IEEE International Conference on Mechatronics and Automation, 2013: IEEE, pp. 1652-1656.

[4] J. Gu, J. Zhou, and X. Chen, "An enhancement of k-means clustering algorithm," in 2009 International Conference on Business Intelligence and Financial Engineering, 2009: IEEE, pp. 237-240.

[5] S. Na, L. Xumin, and G. Yong, "Research on k-means clustering algorithm: An improved k-means clustering algorithm," in 2010 Third International Symposium on intelligent information technology and security informatics, 2010: IEEE, pp. 63-67.

[6] J. Zhu and H. Wang, "An improved K-means clustering algorithm," in 2010 2nd IEEE International Conference on Information Management and Engineering, 2010: IEEE, pp. 190-192.

[7] S. S. Angadi and G. M. Ravanavar, "A study and analysis of training and placement cells in engineering colleges," International Journal of Management, IT and Engineering, vol. 4, no. 10, p. 162, 2014.

[8] A. Malviya, A. Udhani, and S. Soni, "R-tool: Data analytic framework for big data," in 2016 Symposium on Colossal Data Analysis and Networking (CDAN), 2016: IEEE, pp. 1-5.

[9] M. I. Malinen and P. Fränti, "Balanced k-means for clustering," in Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR), 2014: Springer, pp. 32-41.

[10] X. Chang, F. Nie, Z. Ma, and Y. Yang, "Balanced k-means and min-cut clustering," arXiv preprint arXiv:1411.6235, 2014.

[11] M. S. Levin, "On balanced clustering (indices, models, examples)," Journal of Communications Technology and Electronics, vol. 62, no. 12, pp. 1506-1515, 2017.

[12] S. Gupta, "A survey on balanced data clustering algorithms," International Journal for Women Researchers in Engineering, Science and Management, vol. 2, no. 9, pp. 2611-2614, 2017.