A Hybrid K-Mean Clustering Algorithm for Prediction Analysis

Vikas Verma*, Shaweta Bhardwaj and Harjit Singh

School of Computer Science and Engineering, Lovely Professional University, Jalandhar-Delhi G.T. Road, National Highway 1, Phagwara - 144411, Punjab India; vikas.verma@lpu.co.in, shaweta2bhardwaj@gmail.com, harjit.14952@lpu.co.in

Abstract

Background/Objective: The objective of this research is to make improvement in defining the clusters automatically and to assign required clusters to un-clustered points. **Methods/Statistical Analysis:** The main disadvantage of k-mean is of accuracy, as in k-mean clustering user needs to define number of clusters during the start of process. This restriction of predefined number of clusters leads to some points of the dataset remained un-clustered. So by enhancing the cluster technique, the predictions can be improved. We use Iris dataset for the current study and to generate the results using normalization in the methodology which will lead to improvement in accuracy and will reduce clustering time by the member assigned to the cluster. **Findings:** The normalization is used to get better results in the form of finding distance to have exact centroid and to remove noise data which is not needed. We are applying backtracking method to find the exact number of clusters that should be defined to analyze the data in better way. The results shows that there is an improvement in clustering when compared to the existing methodologies.

Keywords: K-mean Clustering, Prediction Analysis, Data Mining, Classification, Clustering, Hybrid Clustering,

1. Introduction

The significant application areas of cluster analysis incorporate data analysis, statistical surveying, pattern recognition, outlier detection, image processing applications such as recognition of credit card fraud¹. Data collecting (or clustering), is an denounce classification method whose purpose is to create groups of objects, or clusters, in such a way that objects in the same cluster are having same properties and those objects in different clusters are quite distinguishable².

There are various types of clustering in data mining³. Partitioning *Clustering* (Figure 1) is a blend of high similarity of the specimens within clusters with high difference between particular clusters. Most partitioning techniques are distance-based.

In Density Based Clustering, the number of partitioning method clusters the objects in the light of the distance among various objects as shown in Figure 2. Spherical shaped clusters are uncovered by these methods and experience problem in finding clusters of arbitrary structures. So for arbitrary shapes new methods are used known as density-based methods which use the notion of density. It helps to discover arbitrary shape clusters. It also handles noise in the data.

Grid Based Clustering quantize the object space into set of cells that help to shape a grid structure. A quick strategy which is autonomous of the data objects however dependent upon cells in every dimension in the quantized space⁴. To form clusters Grid algorithm uses subspace and hierarchical clustering techniques. STING, CLIQUE, Wave cluster, BANG, OptiGrid, MAFIA, ENCLUS, PROCLUS, ORCLUS, FC and STIRR⁵.

In Hierarchical Methods, the hierarchical deterioration of the given group of data objects is completed. It is depicted in Figure 3. It can be appointed by any one of method; agglomerative or divisive in perspective of how this hierarchical decomposition is encircled by that given dataset. In this sort of clustering, it is possible to view partitions at various levels of granularities. One of the examples is flat clustering.

*Author for correspondence



Figure 1. Partitioning Clustering.



Figure 2. Density based Cluster.



Figure 3. Hierarchal Clustering.

In⁶ explained that clustering is the powerful tool which is used in various forecasting tools. The generic methodology of incremental K-mean clustering was proposed for weather forecasting and applied on air pollution of west Bengal dataset.

In⁷ explained that huge data is available in medical field to extract information from large data sets using analytic tool and the data set has been taken from SGPGI.

In⁸ proposed a system named Student Performance Analysis System (SPAS). The proposed system offers

student performance prediction through the rules generated via data mining technique. The data mining technique used in this project is classification, which classifies the students based on students' grade.

Another work in the same field was done to define the ability of the student performance of high learning⁴. An approach was designed to analyze student result based on cluster analysis and use standard statistical algorithm to arrange their score according to the level of their performance. In this paper K-mean clustering is implemented to analyze student result. The model was added up with deterministic model to check student's performance of the system.

In⁹, the web data processing was done on the log files to obtain information of user sessions. Then, clustered were defined based on user sessions using enhanced K-means algorithm to group the users into similar groups or categories. The designed approach works on replacement temporal clustering algorithmic rule known as enhanced K-means clustering algorithmic rule for effective and dynamic grouping of Web users.

In¹⁰ presented an analysis of the prediction of survivability rate of breast cancer patients using data mining techniques. SEER public datasets has been used in this project.

In¹¹ explained that forecasting stock return is one the important subject to be learn for prediction for data analysis. They defined decision tree classifier which is one of the best data mining techniques.

2. Research Methodology

In k-mean clustering algorithm, probability of the most relevant function is calculated and using Euclidian distance formula the functions are clustered¹². Despite being used in a wide array of applications, the k-means algorithm has drawbacks: As many clustering methods, this algorithm says that the clusters k in the database is called as beforehand which are not completely right in real-world application¹³. Moreover, the k-means algorithm is computationally very expensive also¹⁴.

In this work, we will enhance the Euclidian distance formula to increase the cluster quality. Here the proposed work has been published in "International Control and Theory" by explaining the idea of increasing accuracy by using hybrid technique as shown in Figure 4. The enhancement will be based on normalization. In the enhancement two new features will be added shown in Figure 5. The



Figure 4. Flowchart of Selecting Centroid under k-mean Clustering (Proposed Methodology).



Figure 5. Flowchart of Research Methodology.

first point is to calculate normal distance metrics on the basis of normalization. In second point the functions will be clustered on the basis of majority voting. The proposed technique is implemented in MATLAB. In the enhancement two new features will be added. The first point is to calculate normal distance metrics on the basis of normalization. In second point the functions will be clustered on the basis of majority voting. The proposed technique will be implemented in MATLAB.

- First of all, we have started process in which at initial stage we generated data from user end in which we give number of data inputs, which are generated by sigma and random functions.
- When all data has been generated then Simple k-means applied and got result in subplot.
- After applying normalization on that data, we gave scatter data in second subplot.
- Now we applied normalization, in precede in which we read text file data of that generate data after that applied hierarchy k-means before normalization in which we got result in different form rather than first subplot.
- After this process normalization on that process is done in which iterations process started.
- This process is continued until we don't get a nearest point to accurate position with generating data.
- At last calculated their total time in which we got results which shows betterment in accuracy of cluster.

3. Experimental Results

The Eris dataset is used for the research process and to find the results. The original data were highly dimensional, but only 5 attributes has been finally considered on the basis of requirements are shown in the Table 1.

The dataset is loaded using Matlab and taken like as given in Figure 6.

S.No.	Attribute Name	Attribute Description
1	Species_No	Number is given to species
2	Petal_width	Petal is defined by its width
3	Petal_length	Length of petal is considered
4	Sepal Width	Sepal width is given under this attribute
5	Sepal_length	Length of sepals are given under this attribute

The dataset is clustered using the algorithm of k-mean clustering in Figure 7. Clusters are made with closest values and using k-means it made three clusters which have similar properties or can say nearer values.

As shown in Figure 8, the k-mean clustering is improvement to improve cluster quality using the technique of normalization. The dataset is loaded is and it is shown on the command window. The plotted data will be clustered using the algorithm of k-mean clustering. The central points are marked in each cluster.

The plotted data will be clustered using the algorithm of k-mean clustering. The central points are marked in each cluster. The normalization technique is applied to calculate best distance to make high quality clusters as shown in Figure 9. The final output of the clusters is shown on the 2-D plane.

As shown in Figure 10 the k-mean clustering improvement is done in the case of cluster quality using the technique of normalization.

Command	d Windo	w					G
							,
в =							
1.	0000	0.2000	1.4000	3.5000	5.1000		
1.	0000	0.2000	1.4000	3.0000	4.9000		
1.	0000	0.2000	1.3000	3.2000	4.7000		
1.	0000	0.2000	1.5000	3.1000	4.6000		
1.	0000	0.2000	1.4000	3.6000	5.0000		
1.	0000	0.4000	1.7000	3.9000	5.4000		
1.	0000	0.3000	1.4000	3.4000	4.6000		
1.	0000	0.2000	1.5000	3.4000	5.0000		
1.	0000	0.2000	1.4000	2.9000	4.4000		
1.	0000	0.1000	1.5000	3.1000	4.9000		
1.	0000	0.2000	1.5000	3.7000	5.4000		
1.	0000	0.2000	1.6000	3.4000	4.8000		
1.	0000	0.1000	1.4000	3.0000	4.8000		
1.	0000	0.1000	1.1000	3.0000	4.3000		
1.	0000	0.2000	1.2000	4.0000	5.8000		
1.	0000	0.4000	1.5000	4.4000	5.7000		
1.	0000	0.4000	1.3000	3.9000	5.4000		
1.	0000	0.3000	1.4000	3.5000	5.1000		
1.	0000	0.3000	1.7000	3.8000	5.7000		
1.	0000	0.3000	1.5000	3.8000	5.1000		
1.	0000	0.2000	1.7000	3.4000	5.4000		
1.	0000	0.4000	1.5000	3.7000	5.1000		
1.	0000	0.2000	1.0000	3.6000	4.6000		
1.	0000	0.5000	1.7000	3.3000	5.1000		

Figure 6. K-Mean Clustering.



Figure 7. K-Mean Clustering.



Figure 8. K-mean clustering.

Command Window	\odot
Replicate 175, 8 iterations, total sum of distances = 132.5.	^
Replicate 176, 5 iterations, total sum of distances = 132.5.	
Replicate 177, 5 iterations, total sum of distances = 132.5.	
Replicate 178, 9 iterations, total sum of distances = 132.5.	
Replicate 179, 7 iterations, total sum of distances = 132.5.	
Replicate 180, 4 iterations, total sum of distances = 132.5.	
Replicate 181, 5 iterations, total sum of distances = 132.5.	
Replicate 182, 5 iterations, total sum of distances = 132.5.	
Replicate 183, 7 iterations, total sum of distances = 132.5.	
Replicate 184, 6 iterations, total sum of distances = 136.4.	
Replicate 185, 5 iterations, total sum of distances = 132.5.	
Replicate 186, 4 iterations, total sum of distances = 132.5.	
Replicate 187, 5 iterations, total sum of distances = 136.4.	
Replicate 188, 3 iterations, total sum of distances = 166.6.	
Replicate 189, 8 iterations, total sum of distances = 132.5.	
Replicate 190, 4 iterations, total sum of distances = 132.5.	
Replicate 191, 5 iterations, total sum of distances = 132.5.	
Replicate 192, 3 iterations, total sum of distances = 166.6.	
Replicate 193, 5 iterations, total sum of distances = 132.5.	
Replicate 194, 8 iterations, total sum of distances = 132.5.	
Replicate 195, 4 iterations, total sum of distances = 132.5.	
Replicate 196, 3 iterations, total sum of distances = 137.3.	
Replicate 197, 3 iterations, total sum of distances = 166.3.	
Replicate 198, 8 iterations, total sum of distances = 132.5.	
Replicate 199, 5 iterations, total sum of distances = 132.5.	
Replicate 200, 6 iterations, total sum of distances = 136.4.	
Best total sum of distances = 132.5	
fe >>	~
<	>

Figure 9. K-mean clustering.



Figure 10. K-mean clustering (Improvised).

4. Conclusion

We have proposed technique for the enhancement in K-Mean algorithm. According to the member assigned to the cluster for cluster selection. The proposed improvement will lead to better accuracy and reduce clustering time by the member assigned to the cluster to predict disease.

5. References

- Rauf A, Mahfooz M, Khusro S, Javed H. Enhanced K-Mean Clustering Algorithm To Reduce Number of Iterations and Time Complexity. Middle-East Journal of Scientific Research. 2012; 12(7):959–63.
- Osamor VC, Adebiyi A, Oyelade O, Doumbia D. Reducing the Time Requirement of K-Means Algorithm. PLoS ONE. 2012; 7(12):56–62.
- Rajalakshmi K, Dhenakaran SS, Roobin N. Comparative Analysis of K-Means Algorithm in Disease Prediction. International Journal of Science, Engineering and Technology Research (IJSETR). 2015 July; 4(7):1–3.
- Oyelade O, Oladipupo O, Obagbuwa IC. Application of K-Means Clustering Algorithm for Prediction of Students' Academic Performance. International Journal of Computer Science and Information Security. 2010; 7(1):1–4.
- 5. Rani CMS, Narayana T, Sajana K. A Survey on Clustering Techniques for Big Data Mining. Indian Journal of Science and Technology. 2016; 9(3):1–12.
- 6. Weather Forecasting using Incremental K-means Clustering. Date accessed: 2014: Available from: https://arxiv.org/ftp/ arxiv/papers/1406/1406.4756.pdf.

- Yadav AK, Tomar D, Agarwal S. Clustering of Lung Cancer Data Using Foggy K-Means. International Conference on Recent Trends in Information Technology (ICRTIT). 2013; p. 13–18.
- Sa CL, Ibrahim BA, Hossain D. Student performance analysis system (SPAS). Information and Communication Technology for The Muslim World (ICT4M). 2014 Nov; p. 1–6.
- Selvakumar K, Sai Ramesh L and Kannan A. Enhanced K-Means Clustering Algorithm for Evolving User Groups. Indian Journal of Science and Technology. 2015; 8(24):1–8.
- Bellaachia A, Guven E. Predicting Breast Cancer Survivability Using Data Mining Techniques. Software Technology and Engineering (ICSTE). 2010 Oct; 2(1):227–31.
- Qasem A. Al-Radaideh A, Assaf AA, Alnagi A. Predicting Stock Prices Using Data Mining Techniques. The International Arab Conference on Information Technology. 2013; p. 1–8.
- Sundar B, Devi VT, Saravan V. Development of A Data Clustering Algorithm for Predicting Heart. International Journal of Computer Application. 2012 June; 48(7): 888–975.
- Ray S, Rose H, Turi T. Determination of Number of Clusters in K-Means Clustering and Application in Colour Image Segmentation. Proceedings of the 4th International Conference on Advances in Pattern Recognition and Digital Techniques, India. 1999; p. 137–43.
- Yedla M, Srinivasa TM. Enhancing K-means Clustering Algorithm with Improved Initial Center. International Journal of Computer Science and Information Technologies. 2010; 1(2):1–5