ISSN (Print) : 0974-6846
Indian Journal of Science and Technology, *Vol 9(44), DOI: 10.17485/ijst/2016/v9i44/98858, November 2016* ISSN (Online) : 0974-5645

A Novel Technique to Enhance K-Mean Clustering using Back Propagation Algorithm

Kirti Juneja* **and Rohit Katyal**

Computer Science Engineering Department, Chandigarh University, Gharuan - 140413, Punjab, India; keet.bhasin@gmail.com, katyalrohit77@yahoo.com

Abstract

Objectives: In this work, improvement in the k-mean clustering is proposed in terms of accuracy and execution time. **Methods/Statistical Analysis:** The clustering is the technique which is used to analyze the data in the efficient manner. In the recent times various clustering algorithms has been proposed which are based on different type of clusters. Among the proposed algorithms k-mean performs batter in terms of accuracy and execution time. In the k-mean clustering algorithm, the dataset is loaded and from which number of attributes and members are analyzed. The arithmetic mean is calculated from the dataset which will be the central point. The central point is considered as the referral point and Euclidian distance to calculate to all other points in the dataset. The Euclidian distance is calculated by taking single central point due which accuracy of clustering is reduced. In the work, back propagation technique is applied which will calculate Euclidian distance multiple times and achieve maximum accuracy of clustering. **Findings:** The k-mean clustering is the efficient clustering technique which clusters the similar and dissimilar type of data. The k-mean clustering calculates data similarity by calculating the Euclidian distance from the central point. The central point is calculated by taking arithmetic mean of the dataset. When the dataset is complex and large in size, k-mean clustering is not able to drive exact relation between the data points due which some points are left uncluttered. The back propagation algorithm is used to drive accurate relationship between the data points. In the back propagation algorithm input values and actual clustering is derived and this process continues unless maximum accuracy is achieved of clustering. The proposed algorithm is implemented in MATLAB and it is being analyzed that accuracy is increased upto 15 percent and execution time is reduce 2 seconds as compared to existing k-mean algorithm.

Keywords: Arithmetic Mean,Back Propagation,Clustering, K-mean, Neural Networks

1. Introduction

Data mining is the technology of discovering interesting patterns from large amount of data. It is extraction of implicit previously unknown and potentially useful information from data. Data mining is also called as extraction of hidden patterns. It is also known as knowledge mining from data, knowledge extraction, data/ pattern analysis, data archaeology and data dredging. It may fully automate or semi-automated process to discover knowledge that is useful for user[1].

There are different data clustering algorithms. Arrangement is the most usually connected data mining strategy, which utilizes an arrangement of pre-classified case to build up a model that can group the number of inhabitants in records on the loose. Fraud detection and credit risk applications are especially appropriate to this kind of analysis. Regression strategy can be adjusted for predication. Regression analysis can be utilized to demonstrate the relationship between one or more independent variables and dependent variables[2].In data mining independent variables are traits definitely known and reaction variables are what one needs to predict. Lamentably, some genuine issues are not just predictions. Neural network is an arrangement of associated info/yield units and every association has a weight present with it. Amid the learning stage, network learns by modifying weights in order to have the capacity to predict the right

* *Author for correspondence*

class labels of the input tuples. Neural networks have the remarkable capacity to get meaning from convoluted or imprecise data and can be utilized to concentrate designs and distinguish patterns that are too perplexing to ever be seen by either people or other computer systems[3].

K-means clustering calculation goes under centroid based clustering where every cluster is spoken to by a solitary mean vector which may not as a matter of course is an individual from the data set. In this calculation the quantity of clusters to be shaped is settled to 'k'. At that point discover the k cluster focuses. After that, allot the articles to the nearest cluster focus in a manner that every item will be relegated to one and one and only cluster. It gives an estimate to a NP-hard combinatorial optimization issue. It is an unsupervised calculation. Info given to the calculation is "k" which remains for the quantity of clusters the client planned to have[4]. The k-means clustering calculation comprises of:

1. Finding the initial centroids and
2. Assigning each data point or observation to an appropriate cluster.

In the first k-means calculation the centroids for every cluster is chosen haphazardly. Subsequent to settling the centroids, then dole out every perception to the nearest centroid taking into account the distance between the data points and the initial centroids. This structures the initial partition. The centroids for every cluster are then computed following the centroids may change because of the incorporation of new data points. Same methodology is rehashed for including the datapoints. The process is continued until the centroids will never change. This is the convergence criterion for k-means.

There are a variety of variations present in the k-mean algorithm. Such as[5]:

Variants for centroid initialization: Traditional k-means randomly picks up initial centroids. Because of random selection of centroids, it does not guarantee unique clustering results and may find a local minimum solution. To overcome this drawback of traditional k-means several variants have been proposed.

Variants Based Distance Metrics: k-means method is based on Euclidean distance or Euclidean metric, which is commonly used to evaluate proximity between objects. It works well when a dataset has compact clusters. Because of the usage of this distance metric k-means finds spherical or ball-shaped clusters in data.

Variants for improving k-means Accuracy: In most of the cases the experimenter has some background knowledge about which instances should be and should not be grouped together. This information is expressed in terms of instance-level constraints and is given as input. The modified algorithm assigns an object to a cluster only when none of the constraints are violated. If a legal cluster is not found for an object, then method returns empty cluster[6].

The author[7]proposed non metric distance measure, based on live symmetrywhich is used to measure cluster soundness. For this thrasholding technique is applied first to extracts object from the original image the object pixels are transferred to be the data patterns. Object pixels are labeled by applying the fuzzy clustering algorithm and number of objects are determining by applying proposed validity measure. To define performance of proposed measure simulation results are used. The author[8] described comparison between various clustering techniques like partitioning method hierarchical method, density based method, grid method. Clustering algorithm are mainly used to manage data, categorized data for data compression, model creation and also used for outlier discovery etc. Main motive each clustering technique is to find cluster center that represent each cluster. Partitioning method like k-mean clustering algorithm is used for large datasets, as number clusters is increased its performance is also increased. But its use is limited to numeric values. In the paper they[9] provides new method to improve accuracy and performance k-mean clustering that is a ranking based method. Analysis done on existing k-mean clustering approach which is fit in with some threshold value and ranking method which is weighted page ranking applied on k-mean algorithm. In this in links and out links are used to compare performance in the form of execution time of clustering. Weighted page rank algorithm with k-mean provides better result than existing k-mean algorithm. It takes less computational time then existing k-mean algorithm.The author[10] provides a new method to improve accuracy and performance k-mean clustering that is a ranking based method. Analysis done on existing k-mean clustering approach which is fit in with some threshold value and ranking method which is weighted page ranking applied on k-mean algorithm. In this in links and out links are used to compare performance in the form of execution time of clustering. Weighted page rank algorithm with k-mean provides better result than existing k-mean algorithm. It takes less computational

time then existing k-mean algorithm. The researcher[11] implemented K-means along with Genetic algorithm for dimensionality reduction and support vector machine to classify the data set. K-means algorithm is used to remove outliers and the noisy data. The optimal features are selected by using the genetic algorithm and then Support vector machine classifies the reduced data space using 10 fold cross validation technique. Genetic algorithm selects different features from original set of feature during each run. To obtain consistent results, the experiment was performed 50 times. The result shows that the purposed model achieves the accuracy of 98.82%.

The author[12] proposed a prediction model for medical data with missing value imputation techniques, then analyzing these techniques by using K-means algorithm and choosing the best among them. Thus this model improves the quality of data by using the best imputation technique. Methods such as case deletion, most common method, concept most common, K-means clustering imputation, k-nearest neighbor etc. are applied to fill the missing data values in the data. The efficiency is calculated on three data sets namely Hepatitis, Wisconsin Breast Cancer and Pima Indians Diabetes from the UCI repository. This model achieved accuracy of 99.82% for Diabetes data set, 99.39% for Breast Cancer and 99.08% for Hepatitis data set. For Diabetes and Hepatitis data sets Concept Most Common (CMC) is chosen as the best method, and for Breast Cancer Case deletion is selected as best missing value imputation method. They proposed[13] an optimized version of k-mean that reduces the problems of re-distribution of the data elements those will remain part of the same cluster during the next iteration. After a number of iterations only a few number of data elements change their cluster. While assigning the data element to the cluster there is no need to visit the entire data set, but just a small list of data objects. The implementation showed up to 70% reduction of the running time. The author[14] has examined those algorithms like genetic algorithm, PSO, ANN that can be used in predicting heart disease. Combining these algorithms with the data mining techniques such as clustering, classification etc. or by combining these algorithms with one another will give better performance and accuracy.

2. Proposed Work

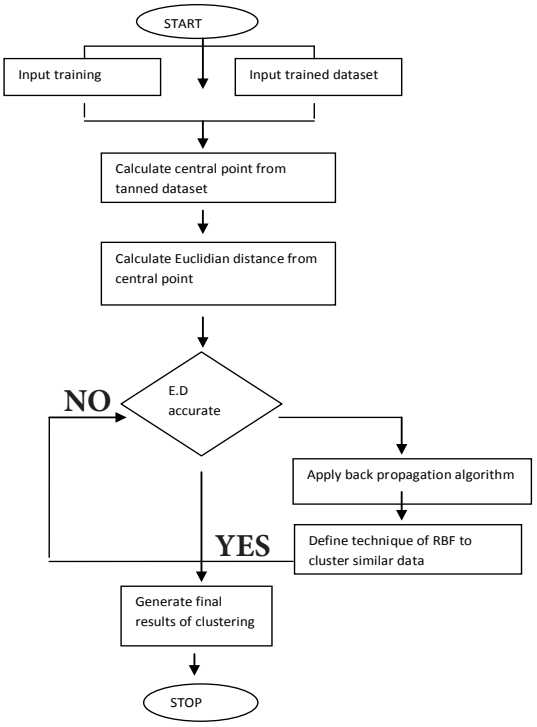
The k-mean clustering algorithm is the algorithm to cluster similar and dissimilar type of data from the input dataset[15]. To cluster similar type of data, two steps are followed in k-mean clustering. In the first step arithmetic mean is calculated and that point is centriod point. In the second step, Euclidian distance is calculated to all other points in the dataset. The points which have similar distance are clustered together and other is not. In the k-mean clustering some points are wrongly clustered due to in accurate calculation of Euclidian distance. The main objective of this work is to calculate Euclidian distance accurately and reduce time of clustering. When the Euclidian distance is calculated accurately, it directly leads to improve in cluster quality[16]. To improve cluster quality technique of back propagation is implemented in which Euclidian distance is calculated multiple times, until maximum accuracy is achieved. To achieve maximum accuracy technique of back propagation is implemented[17]. The back propagation algorithm searches for the minimum of the error function in weight space utilizing the strategy for gradient plunge. The combination of weights which minimizes the error function is thought to be an answer of the learning issue[18,19]. Since this strategy requires calculation of the gradient of the error function at every iteration step, one must ensure the continuity and differentiability of the error function. Clearly one needs to utilize a sort of activation function other than the progression function utilized as a part of perceptrons, in light of the fact that the composite function created by interconnected perceptrons is spasmodic, and in this manner the error function as well[20].

To improve cluster quality following steps are followed:-
Step:-

- Input of tanning dataset for clustering
- Selection of centriod point by calculating arithmetic mean of the dataset
- Calculate Euclidian distance from the centriod point
- It is the Euclidian distance is calculated accurately then show final clustering result
- If the Euclidian distance is not calculated accurately, then iterative process of back propagation is executed until maximum accuracy is achieved
- As the maximum accuracy final result of clustered in displayed

3. Results and Discussions

The proposed idea will be implemented in MATLAB which is widely used in all areas of research universities, and also in the industry.



(a)

Table 1. Table of dataset
(a) Flowchart of methodology

Parameters	Values
Dataset Name	Iris Fisheries
No of attributes	5
No of members	151

As illustrated in Table 1, the dataset which is used to implement k-mean clustering is defined in the table. In the defined table number of attributes and member function are defined

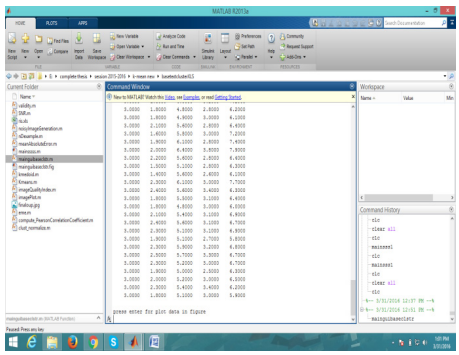


Figure 1. Data clustered.
As Figure 1 illustrates thatthe K-mean is the algorithm in which the data will be clustered according to Euclidian distance. The random center points had been selected from the data. The Euclidian distance will be calculated from the data centers to other points and points will be clustered accordingly. The output of the clustered will be shown in the 2D plane. When the data will be shown in 2D plan, some points which are very close to each other cannot be shown which reduce the cluster quality.

As shown in Figure 1, the loaded dataset is shown on the command window and on the command window attributes and their values are shown which need to be clustered

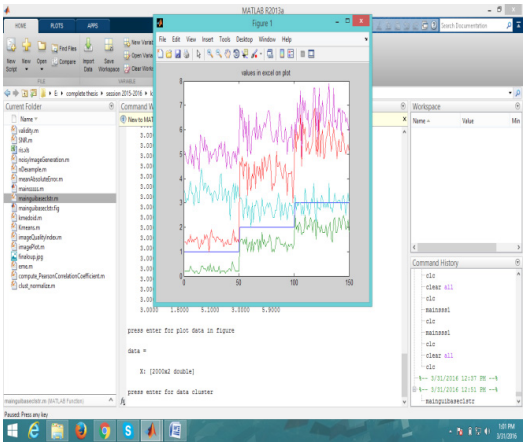


Figure 2. Plotting of data.
As illustrated in Figure 2, the dataset which is read with xlsread commands of MATLAB will be plotted on the 2-D plane

As shown in Figure 2, the dataset which is loaded for clustering and that dataset are plotted on the 2-D plane

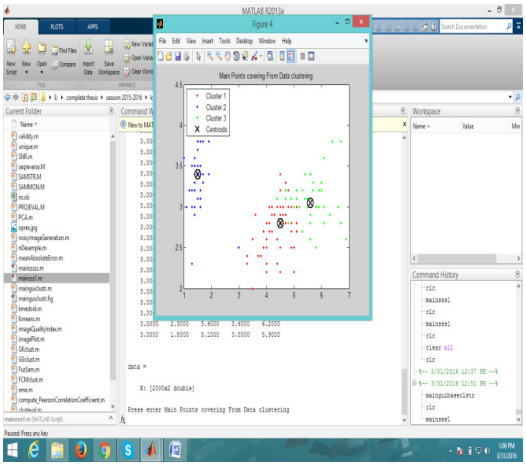


Figure 3. Vornoalierepresentation.
As shown in Figure 3, the dataset which is used in the previous figure will be clustered using the hybrid type of k-mean clustering algorithm. When the dataset will be clustered using hybrid algorithm cluster quality will be improved and each point in the dataset will be shown on voronlie plane for better analysis of dataset

As shown in Figure 3, the clusters are defined using k-mean clustering. The clusters which are made is of veronica type.

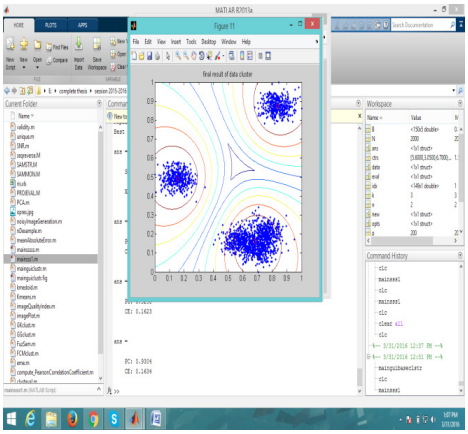


Figure 4. Final output.

As shown in theFigure 4, to improve accuracy of k-mean clustering the points which are remained unclustered are clustered using the technique of normalization. The normalization calculate equalidian distance in the iterative mmner. When high accuracy of clustering is achieved the ouput is shown in the form of clusters

As illustrated in Figure 4, the data which is clustered is shown and final results of clustering is represented on 2-D plane.

Table 2. Compression of proposed and existing technique

Parameter	Existing Algo	Proposed Algo
Accuracy	87.5 percent	93 percent
Time	3.2 second	2.5 second

As shown in figure 2, the comparison between the proposed and existing work is done and it is shown that proposed algorithm has 93 percent accuracy and in the existing algorithm accuracy is 87.5 percent. In the proposed work, execution time is reduced to 2.5 second from 3.2 seconds

As shown in Table 2, the comparison between the proposed and existing technique is done and it is represented that proposed technique performs well in terms of accuracy and execution time.

4. Conclusion and Future Scope

In k-mean clustering, dataset is loaded and from the loaded dataset central points are selected according to defined number of clusters. The central point acts as reference point and from which Euclidian distance is calculated and according to Euclidian distance members are assigned to each cluster. Due to user defined cluster values, some points of the dataset are remained un-clustered which reduced accuracy of clustering.In this paper enhanced K-Mean clustering is defined to improve accuracy. The technique of backpropagation is applied which is used to calculate Euclidian distance in the iterative manner to improve its accuracy.The accuracy is achieved upto 93 percent in the proposed and in the existing k-mean algorithm it is 87.5 percent. In future, technique of mean shift can be applied to improve clustering results.

5. References

1. Bharati M, Ramageri R. Data mining techniques and applications. Indian Journal of Computer Science and Engineering. 2004; 1(4):301–5.
2. Charu C, Aggarwal A, Chandan K, Reddy R. Data clustering: Algorithms and applications. Springer-Verlag Berlin Heidelberg; 2015.
3. Choudhary A. Survey on K-means and its variants International Journal of Innovative Research in Computer and Communication Engineering. 2016.
4. Coates A, Andrew YNG. Learning feature representations with K-means.Springer LNCS.2012; 7700:561–80.
5. Babu GP, Murty MN.A near-optimal initial seed value selection in K-means algorithm using a genetic algorithm. Elsevier Science Publishers. 1993; 14(10):763–9.
6. Philip D, Khazenie HN. Classification of multispectral remote sensing data using a back-propagation neural network.IEEE Transactions on Geoscience and Remote Sensing. 1992; 30(1):1–15.
7. Chou C, Hsieh YZ,Su MC, Chu YL.Extracting and labeling the objects from an image by using the fuzzy clustering algorithm and a new cluster validity. International Journal of Computer and Communication Engineering. 2013; 2(3):1–3.
8. Joshi A,Kaur R. A review: Comparative study of various clustering techniques in data mining.International Journal of Advanced Research in Computer Science and Software Engineering. 2013; 3(3):1–3.
9. Singh A,KaurN. To improve the convergence rate of K-means clustering over K-means with weighted page rank algorithm. International Journal of Advanced Research in Computer Science and Software Engineering. 2013; 3(8):1–5.
10. Jain A, Rajavat A, Bhartiya R. Design, analysis and implementation of modified K-mean algorithm for large data-set to increase scalability and efficiency. Fourth International Conference on Computational Intelligence and Communication Networks.2012.
11. Santhanam T, Padmavathi MS. Application of K-Means and Genetic Algorithms for Dimension Reduction by Integrating SVM for Diabetes. Diagnosis Procedia Computer Science. 2012;47:76–83.
12. Purwar A, Singh SK. Hybrid prediction model with missing value imputation for medical data. Expert Systems with Applications. 2015;42(13):5621–31.
13. Poteras CM, Mihaescu MC, Mocanu M.An optimized ver-

sion of the K-Means clustering algorithm.Federated Conference onComputer Science and Information Systems (FedCSIS), IEEERomania; 2014. p. 695–9.
14. Kumari VA, Chitra R. Classification of diabetes disease using support vector machine.International Journal of Engineering Research and Applications. 2014;3(2):1797–1801.
15. Nawi NMRS, Ransing R, Salleh MNM, Ghazali R, Hamid NA. An improved back propagation neural network algorithm on classification problems. Database Theory and Application, Bio-Science and Bio-Technology. 2010; 118:177–88.
16. Verma V, Bhardwaj S, Singh H. A hybrid K-mean clustering algorithm for prediction analysis. Indian Journal of Science and Technology. 2016 Jul; 9(28):1–5.
17. Kaur S, Kaur A. Detection of malware of code clone using string pattern back propagation neural network algorithm. Indian Journal of Science and Technology. 2016 Aug; 9(33):1–12.
18. Jayadurga R, Gunasundari R. A novel approach in vehicle object classification system with hybrid of central and hu moment features using back propagation algorithm. Indian Journal of Science and Technology. 2016 Jul; 9(26):1–7.
19. Ashok V, Singh SR, Nirmalkumar A. Determination of blood glucose concentration by back propagation neural network. Indian Journal of Science and Technology. 2010 Aug; 3(8):1–3.
20. Fathima KAR, Raghavendiran TA. A novel intelligent unified controller for the management of the Unified Power Flow Controller (UPFC) using a single back propagation feed forward artificial neural network. Indian Journal of Science and Technology. 2014 Jan; 7(8):1155–69.