Advances in Computer Science and Information Technology (ACSIT)
p-ISSN: 2393-9907; e-ISSN: 2393-9915; Volume 3, Issue 4; April-June, 2016, pp. 275-280
© Krishi Sanskriti Publications
http://www.krishisanskriti.org/Publication.html

Enhancement to Asymmetric Clustering

Amneet Kaur[1] and Sheetal Kalra[2]

[1,2]CSE DEPT. Guru Nanak Dev University Regional Campus Jalandhar, India
E-mail: [1]arora.amneet@gmail.com, [2]sheetal.kalra@gmail.com

Abstract—*Data mining or the knowledge discovery process is a technique for analyzing voluminous amount of data. It is considered important technology in areas like market analysis and management, financial data analysis, Fraud detection, biological data analysis and other various scientific applications. Clustering is a technique in data mining which groups similar objects into clusters. There are various approaches in clustering and performance of clustering depends on ability of algorithms to find hidden useful knowledge. In an asymmetric clustering, data is partitioned in which similar and dissimilar data is separated out. The partitions can be affirmed vigorously and usually run on a single cluster at a time. In the previous model discussed in the paper the time taken to produce clustering results was much lower and decreases the performance of clustering. In this paper we have proposed a model to improve asymmetric clustering results by combining both mean shift and K-means normalization algorithm. Final Clustering results are plotted and performance parameters are compared between previous and proposed method. The proposed model shows performance improvement with increase in accuracy and reduced execution time and noise level.*

Keywords*: Asymmetric; Clustering; CURE; Fig; K-means; K-N means; Mining; N-cut*

1. INTRODUCTION

Data mining or the knowledge discovery process is a technique for analyzing voluminous amount of data and extracting valuable information from it [1]. It is very useful in industries to mine large amount of data. Data mining tools generate results effectively in less time and give accurate results. It is considered important technology in areas like market analysis and management, financial data analysis, Fraud detection, biological data analysis and other various scientific applications. Clustering is a technique in data mining which groups similar objects into clusters. It has significant applications in various fields like information retrieval, outlier detection, image processing etc. There are various approaches in clustering and performance of clustering depends on ability of algorithms to find hidden useful knowledge. Few include CURE (Clustering using representatives), K-means, genetic K-means, Clara, Dbscan, Clarans etc. K-means is also most widely used algorithm which works by selecting initial number of clusters and a centroid [3]. K-means is chosen as it works well with large data sets. And its ability to generate results with simplicity in observable speed. Disadvantages of K-means includes that it does not give good clustering results when clusters are of different size and density. Various data points exist on which K-means takes super polynomial time. Even k-means cannot handle dirty data and missing values. Researchers has developed various methods to improve the accuracy of k-means clustering algorithm. It uses the Euclidean distance to calculate the centroid of the clustered data. This method is not effective when more data is added, as the distance between various objects is not accurately calculated.

A. Advantages of K-means Normalization
   a) Eliminate the redundant data
   b) Valid and reliable data gives good clustering results
   c) It improves accuracy and efficiency.

So it is one of the essential step before performing clustering. There are many k-means normalization methods [4] to normalize the clustering results like Z-score, Min-max and Decimal Scaling. The Usage of best method depends upon the dataset which needs to be normalized. Here we have used Min- Max normalization method as it performs linear transformation on the dataset. The normalized results improve the clustering quality and give more accurate results. Mean Shift is a powerful non parametric iterative algorithm which can be used for various purposes like finding modes, clustering etc. Mean Shift was introduced by Fukunaga and Hostetler which was further applied in various fields like Computer Vision. Mean shift associates with the nearby dataset's probability density function. For each data point, it defines a window around it and calculate the mean of the data point. Then it shifts the center of the window towards the mean and repeat the algorithm till it converges. After each iteration, window shifts towards the denser region of the dataset.

At the high level, we can specify Mean Shift [11] as follows:

- Fix a window around each data point.
- Compute the mean of data within the window.
- Shift the window to the mean and repeat till convergence.

B. Comparison with K-Means

a) The most important difference between k-mean and mean shift is that in K-means the number of clusters are known beforehand and shape of cluster is either spherical or elliptically. Mean shift algorithm does not make any assumption about clusters. Also, it works well with arbitrarily shaped clusters as it based on density estimation.

Initializations in k-means is very sensitive. A wrong initialization can result in wrong cluster and even can

ALGORITHM: Steps of N-K means Algorithm [4]

INPUT: A dataset with d dimensions

OUTPUT: Clusters

1. Initially load data set.

2. Find out the minimum and maximum values of each feature from the dataset.

3. Normalize datasets with maximum and minimum values using equation: v' = v-min(e)/ max(e)- min(e) where, min(e) and max(e) are the minimum and the maximum values for attribute E.

4. Pass the number of clusters and generate initial centroids.

5. Generate clusters.

b) delay convergence. Mean shift is not sensitive and is fairly robust to initialization.

The N-cut is the algorithm is used for segmentation or to divide similar and dissimilar datasets. In this work, N-cut algorithm [9] is applied on dataset which is clustered and to improve the cluster quality, dataset will be further clustered and in each cluster uniqueness will be calculated using the equation

$$\mathrm{ncut}(A,B) = \frac{w(A,B)}{w(A,V)} + \frac{w(A,B)}{w(B,V)}$$

In the equation A and B are two clusters and w is the weight on each cluster. V is the variance of uniqueness between two clusters. In an asymmetric clustering [12], the partitions can be affirmed vigorously and usually run on a single cluster at a time. The task which is specific to an appropriate cluster can be routed to that cluster. The clustering architecture is contrary to the typical stateless server farm where the whole application is simulated across machines. In an asymmetric cluster, business logic is divided into partitions, where every partition can be the singular accessory of a set of underlying data. As a result, each node in the cluster implement its own local cache resulting in high reading and writing performance without maintaining a distributed cache between cluster nodes.

C. Features of asymmetric cluster

a) applications can declare named partitions at any point while it's running

b) partitions are highly available uniquely named singletons and run on a single cluster member

c) incoming work for a partition is routed to the cluster member hosting the partition

This paper is organized as follows:

In Section II, a description of the literature survey is done, which covers the work done by various authors to improve asymmetric clustering algorithm. Then in Section III, our previous model is described followed by experimental results. In Section IV, proposed model is described where a comparison is shown between traditional N-K-means clustering algorithm and Mean shift process. Lastly, the conclusion is stated in section V.

2. LITERATURE REVIEW

Garg S & Sharma A K (2013) introduces in his paper various data mining algorithms [13]. He conducted analysis of various data mining algorithms. He stated that a single algorithm cannot be applied to all the application due to suitability of data types. So, correct data mining algorithm needs to be selected for particular application.

Padhy N Mishra D & Panigrahi R (2012) gave a complete overview of data mining and areas where it can be used [14]. They have stated various data mining techniques: Decision tree and rules, classification methods and nonlinear regression etc. and also areas where data mining can be done to get information which can be used for making decisions. Areas include Healthcare, Education Systems, CRM, Web Education, Sports data mining, E-Commerce etc. The various data mining techniques are used to extract the useful patterns.

Guha S, Rastogi R & Shim K (2001, June). Proposed clustering algorithm called CURE [2] which uses combination of both random sampling and partitioning that efficiently handles large amount of dataset and also effectively filter out outliers. Furthermore, algorithm makes use of multiple representative points for each cluster to assign data points. This enabled to correctly label points when sizes of clusters vary or are non-spherical clusters. The experiment results proved that quality of cluster is much better than existing algorithms.

Shih M Y, Jheng J W & Lai L F (2010) explained about various clustering algorithms [15] have been developed diverse domains in which data is stored in form of groups. The work in these clustering algorithms is either on pure numeric data or on pure categorical data, and on the mixed categorical and numeric data types. However, the existing clustering algorithms has some disadvantages or weakness, the two-step method adds attribute to cluster with integrated hierarchical and partitioning clustering algorithm. This method explains

the relationships between items and improves the weaknesses in single clustering algorithms. Experimental analysis shows that accurate and strong results can be obtained by applying this method to cluster mixed numeric and categorical method.

Carlsson G, Mémoli F, Ribeiro A & Segarra S. (2014) proposed hierarchical quasi clustering method [16], a generalization of hierarchical clustering for asymmetric networks where the asymmetry of the input data is preserved by the output structure. When clustering asymmetric networks, requiring the output to be symmetric as in hierarchical clustering might be undesirable. Hence, we defined quasi-dendrogram, a generalization of dendrogram that admits asymmetric relations, and developed a theory for quasi-clustering methods.

Virmani D, Taneja S & Malhotra G (2015). The K-means is most widely used clustering algorithm for large amount of data. But traditional k means algorithm does not generate good clustering results as final clusters are effected due to automatic initialization of centroids. This paper proposed an efficient algorithm [4] in which we first pre-process our dataset using the normalization technique and generate effective clustering results. Standardisation is achieved by assigning weights to each attribute value. Proposed algorithm is better than traditional K-means algorithm in terms of execution time and speed.

Patel, V. R., & Mehta, R. G. (2011, December) has purposed Mk-means algorithm [5] which automatically initializes the centroid and also calculate the performance of MK-means along with normalization techniques and cleaning method which shows better results for clustering.

Mahmud, M. S., Rahman, M. M., & Akhtar, M. N. (2012) proposed a heuristic method [17] to increase the accuracy of clusters and find initial Centroids. Experimental results show that proposed method increase the cluster quality of k-means clustering algorithm.

Zhang, C., & Xia, S. (2009, January). Proposed a method [6] in which initially centres for clusters are selected and is input to the k-means. User does not require to give number of clusters.

Wu S, Feng X, & Zhou W (2014) has proposed a new technique based on sparse representation vectors [19], using two weight matrices. This construction gives detailed information about coefficient vectors to analyse similarity between two objects. The experiments evaluated shows reliable performance of proposed algorithm and promise wide applicability.

Cheng, Y. (2005). It is shown that mean shift [10] is a process with kernel constructed on a surface. In Gaussian kernels, mean shift works as a gradient mapping. Convergence is studied for mean shift iterations. Cluster analysis if treated as a deterministic problem is used for finding a fixed point of mean shift that characterizes the data. Applications in clustering and Hough transform are demonstrated. Mean shift is also considered as an evolutionary strategy that performs multistate global optimization.

Derpanis K. G. (2005). The main concept for mean shift [19] is to treat all the points in the d-dimensional feature space as a probability density function where center moves towards the dense regions in the space corresponding to the local maxima of the dense region or modes of the underlying distribution. For each data point in the D-dimensional space, a gradient ascent procedure is performed on the estimated local density distribution until it is convergence. The stationary points in this procedure corresponds to modes of the distribution. Furthermore, the data points associated (at least approximately) with the same stationary point belong to the same cluster.

3. PRELEMINARIES

In this section we will describe the previous model which is being used in order to improve the asymmetric clustering results. (Fig1) [19]

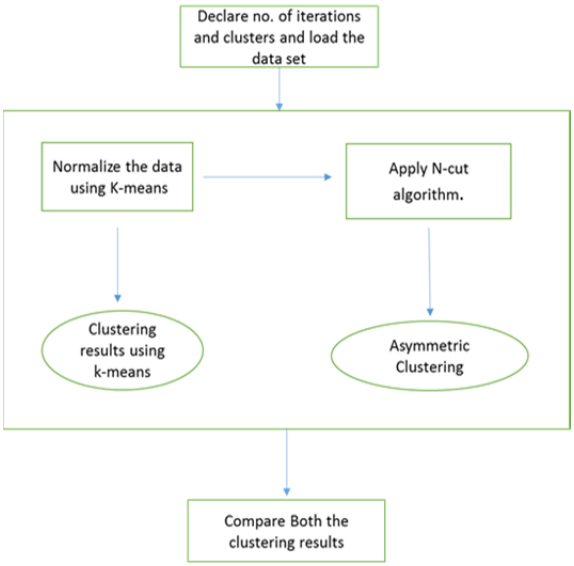

Fig. 1: Previous model using normalized K-means and N-cut algorithm

A. Steps for previous model

a) Load dataset- Firstly, Declare the number of rows, columns and iterations. In this step dataset will be loaded to perform clustering operation

b) Normalize data using K-means- Secondly, apply K-means algorithm and Divide the data by Finding centroid of each dataset segmented and further normalize to improve the clustering results.

c) Clustering results using K-means- Normalized results are plotted on a graph.

d) Apply N-cut algorithm- N-cut algorithm is applied to further group similar and dissimilar data differently.
e) Asymmetric Clustering- Asymmetric clustering results are plotted and compared with Normalized k-means.

B. Experiments and results

The experiments are conducted on IRIS dataset. The dataset is given as input to basic K-means and normalized. The clustering results are calculated in case of normalized k-means(Fig2) and after that N-cut algorithm is applied. Final clustering results are obtained(Fig3). The accuracy of clusters is improved by applying n-cut algorithm and also execution time to generate clusters is reduced.

The results are implemented in MATLAB [8] which is widely used in all areas of research universities, and also in the industry. MATLAB is useful for solving mathematics equations such as numerical integration equations, linear algebra. It is one of the simplest programming languages for writing mathematical programs.

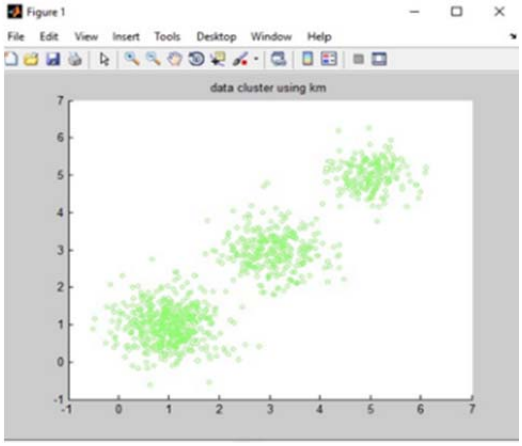

Fig. 2: Data Cluster using K-means

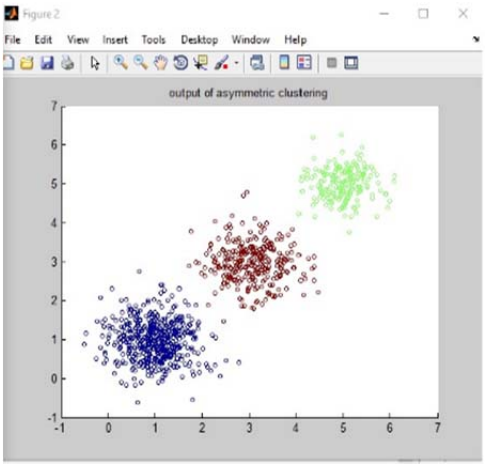

Fig. 3: Clustering results using K-means normalization + N-cut

C. Disadvantages of previous model

1. The previous model does not give good clustering results.
2. The accuracy achieved in this model is very low and needs to be improved.
3. The time taken to generate cluster is much lower.
4. The noise level is quite high and needs to be reduced to improve the cluster quality.

4. PROPOSED MODEL

This model will describe new technique being used to improve accuracy of the asymmetric clustering and also reduce the noise levels. (Fig4)

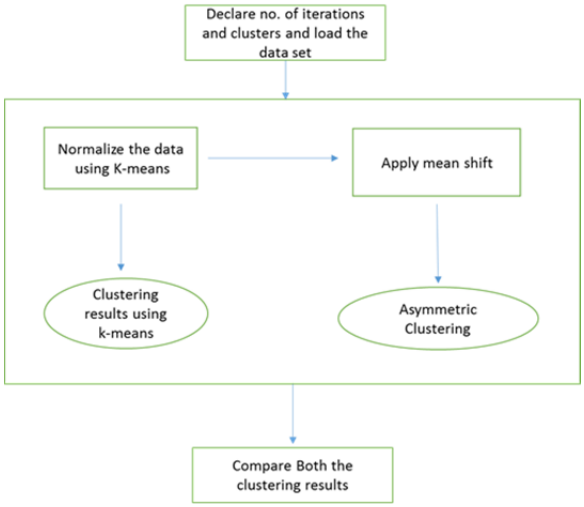

Fig. 4: Proposed model for using mean shift algorithm

A. Explanation of Proposed Flowchart

a) Declare rows, columns, integration and load dataset- This is the first set of algorithm in which the number of rows and columns are defined for the dataset. The second condition defines number of iteration to improve the cluster quality. In the step of the flowchart the dataset will be loaded to perform clustering operation
b) Apply K-mean Normalization and asymmetric Clustering- K-mean normalization method is applied for clustering using the normalization equation. The normalization equation when implemented with k-mean the cluster quality can be improved.
c) Apply mean shift- In this step, two operations are performed. In the first step mean shift algorithm is applied on the loaded dataset. In the mean shift algorithm, the mean value is calculated on the dataset and left shift operation is performed to simplify the operation of clustering.
d) Apply MCL and n-cut- The MCL is the Markova clustering algorithm, which is the unsupervised clustering graph based algorithm. This algorithm is fast

and reliable and has good cluster quality. The main concept behind this algorithm is mathematical theory behind it, its positioning in cluster analysis and graph clustering Issues concerning scalability, implementation, and benchmarking, and performance criteria for graph clustering in general and finally N-cut is applied to further improve the cluster quality.

e) Plot and make clustering and normalize- In this method clustered data will be plotted. When the data is plotted, the method of normalization will be applied on the plotted data to improve the cluster quality.

f) Start of iteration, mean shift insertion and affinity insertion- In these steps of flowchart, the iterations which are defined in start of flowchart. The process of mean shift is calculated which are inserted on every iteration and with each iteration cluster quality had been improved.

B. Experiments and results

The experiments are conducted on IRIS dataset. The dataset is given as input to basic K-means and normalized. The clustering results are calculated in case of normalized k-means and after that mean shift technique is applied. The accuracy of Asymmetric cluster is improved by applying the new technique. The accuracy is increased by 5%.

The First graph shows loaded data points (Fig5) by giving value for number of clusters and iterations. Second graph (Fig6) shows the results for Normalized K-means and further technique is applied for affinity matrix and mean shift and cluster quality is improved (Fig7). The table (Table 1) shows the improved performance results.

The results are implemented in MATLAB [8] which is widely used in all areas of research universities, and also in the industry. MATLAB is useful for solving mathematics equations such as numerical integration equations, linear algebra. It is one of the simplest programming languages for writing mathematical programs.

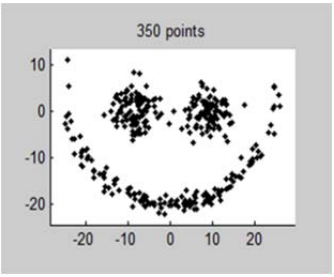

Fig. 5: Load Dataset

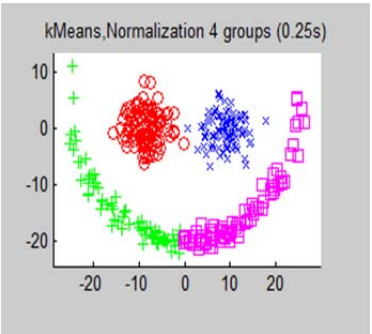

Fig. 6: K-means normalization

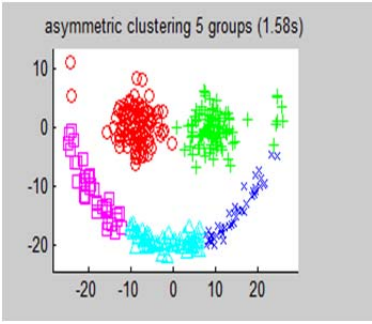

Fig. 7: Final Clustering results using mean shift

Table 1: Performance results of previous and purposed model

Algorithm	Accuracy	Time	noise
Normalized K-means + N-cut	70.22%	1.00318	1.093
Normalized K-means+ mean shift	92.86%	1.00206	0.123

5. CONCLUSION

The combination of two algorithms is used in order to improve clustering accuracy. The model is proposed to improve accuracy and noise level of asymmetric clustering. Existing asymmetric clustering algorithms and methods does not give good clustering results. The new model reduces the execution time than previous developed model and increases the performance. The accuracy achieved by k-means normalization + N-cut is 70.22%, while accuracy in case of K-means normalization + mean shift is 90.15%. The purposed model will select number of clusters and plot the clustering results on 2-D plane. The technique of mean shift is used which is better and give good clustering results in terms of accuracy, execution time and noise level.

REFERENCES

[1] http://www.tutorialspoint.com/data_mining
[2] Guha S, Rastogi R & Shim K (2000). *U.S. Patent No. 6,092,072*. Washington, DC: U.S. Patent and Trademark Office.
[3]http://home.deib.polimi.it/matteucc/Clustering/ tutorial_html/kmeans.html
[4] Virmani D, Taneja S & Malhotra G (2015). Normalization based K-Means Clustering Algorithm.
[5] Patel R V, Mehta G R, "Performance Analysis of MK-means Clustering Algorithm with Normalization Approach", World Congress on Information and Communication Technologies, 2011, pp. 974-979
[6] Zhang Chen, Xia Shixiong, "K-means Clustering Algorithm with improved Initial Center", Second International Workshop on Knowledge Discovery and Data Mining,2009, pp. 790-792
[7] Singh R V, Bhatia M P S, "Data Clustering with Modified K means Algorithm", IEEE-International Conference on Recent Trends in Information Technology, ICRTIT 2011, pp. 717-721
[8]http://in.mathworks.com/products/matlab/? requestedDomain=in.mathworks.com
[9]http://pages.cs.wisc.edu/~dyer/cs766/slides/ ncut/ncut-2up.pdf
[10] Cheng Y (1995). Mean shift, mode seeking, and clustering. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *17*(8), 790-799.
[11]http://homepages.inf.ed.ac.uk/rbf/CVonline/LOCAL_COPIES/TUZEL1/MeanShift.pdf
[12] Gao J, Goodman J T, Cao G & L, H. (2002, July). Exploring asymmetric clustering for statistical language modelling. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics* (pp. 183-190). Association for Computational Linguistics.
[13] Garg S & Sharma A K (2013). Comparative Analysis of Various Data Mining Techniques on Educational Datasets. *International Journal of Computer Applications*, *74*(5).
[14] Padhy N, Mishra D & Panigrahi R (2012). The survey of data mining applications and feature scope.
[15] Shih M Y, Jheng J W & Lai L F (2010). A Two-Step Method for Clustering Mixed Categroical and Numeric Data. Tamkang Journal of Science and Engineering, 13(1), 11-19.
[16] Carlsson G, Mémoli F, Ribeiro A & Segarra S (2014). Hierarchical Quasi-Clustering Methods for Asymmetric Networks.
[17] Mahmud M S, Rahman M M & Akhtar M N (2012, December). Improvement of K-means clustering algorithm with better initial centroids based on weighted average. In *Electrical & Computer Engineering (ICECE), 2012 7th International Conference on* (pp. 647-650). IEEE.
[18] Wu S, Feng X & Zhou W (2014). Spectral clustering of high-dimensional data exploiting sparse representation vectors. *Neurocomputing*, *135*, 229-239.
[19] Derpanis KG (2005) Mean shift clustering, Lecture Notes