ISSN: 2321-7782 (Online)
Volume 4, Issue 1, January 2016
International Journal of Advance Research in Computer Science and Management Studies
Research Article / Survey Paper / Case Study
Available online at: www.ijarcsms.com

A new framework of Kmeans algorithm by combining the dispersions of clusters

Amruta S. Suryavanshi[1]
Computer Department
TSSM's Bhivarabai Sawant College of Engineering and Research, Pune, India

Prof. Anil D. Gujar[2]
Computer Department
TSSM's Bhivarabai Sawant College of Engineering and Research, Pune, India

Abstract: ***Kmeans-type clustering aims at partitioning a data set into clusters such that the objects in a cluster are compact and the objects in different clusters are well separated. However, most kmeans-type clustering algorithms rely on only intracluster compactness while overlooking intercluster separation. A series of new clustering algorithms by extending the existing kmeans-type algorithms is proposed by integrating both intracluster compactness and intercluster separation. First, a set of new objective functions for clustering is developed. Based on these objective functions, the corresponding updating rules for the algorithms are then derived analytically.***

The new algorithm with new objective function to solve the problem of intracluster compactness and intercluster separation has been proposed. Proposed FCS based algorithm works simultaneously on both i.e. intracluster compactness and intercluster separation and it will give a better performance over existing k-means.

Key words: Clustering, Data Mining, Feature weighting Kmeans, Fuzzy Compactness and Separation, Fuzzy C Mean.

I. INTRODUCTION

Clustering is a basic operation in many applications in nature, such as gene analysis, image processing, text organization, and community detection. It is a method of partitioning a data set into clusters such that the objects in the same cluster are similar and the objects in different clusters are dissimilar according to certain predefined criteria. The kmeans-type clustering algorithms are a kind of partitioning method [2], which considers only the similarities among the objects in a cluster by minimizing the dispersions of the cluster. The representative ones of these algorithms include basic kmeans, automated variable weighting kmeans (Wkmeans), attributes-weighting clustering algorithms (AWA).All these methods have the same characteristic: The features must be evaluated with the big weights if the dispersions of the features in a data set are small. In essence, the discriminative capability of a feature not only relates to the dispersion, but also associates with the distances between the centroids, i.e., intercluster separation. The intercluster separation plays an important role in supervised learning methods [6] [7].

A series of new clustering algorithms by extending the existing kmeans-type algorithms is proposed by integrating both intracluster compactness and intercluster separation. These algorithms are E-kmeans, E-Wkmeans, and E-AWA, which extend basic kmeans, Wkmeans, and AWA.

II. LITERATURE SURVEY

In existing system many approaches have been projected. Here survey of some of papers is done. Brief surveys of kmeans-type clustering from two aspects are given: [8] [9]

1) No Wkmeans-type algorithms

2) Vector Wkmeans-type algorithms

A. No-Wkmeans-Type algorithm

No Wkmeans-type algorithms are divided in two parts: Without intercluster separation and with intercluster separation.

1) No Wkmeans-type algorithms Without intercluster separation:

Let X = {X1, X2,... Xn} be a set of n objects. Object Xi = {xi1, xi2,. . . , xim} is characterized by a set of m features (dimensions). The membership matrix U is a n × k binary matrix, where uip= 1 indicates that object i is allocated to cluster p, otherwise, it is not allocated to cluster p. Z = {Z1, Z2,. . . ,Zk} is a set of k vectors representing the centroid of k clusters. The basic kmeans relies on minimizing an objective function [10].

$$P(U, Z) = \sum_{p=1}^{k} \sum_{i=1}^{n} \sum_{j=1}^{m} u_{ip}(x_{ij} - z_{pj})^2$$

Subject to

$$u_{ip} \in \{0, 1\}$$

U and Z can be solved by optimizing the objective function.

Basic kmeans algorithm has been extended in many ways. Steinbach et al. [10] proposed a hierarchical divisive version of kmeans, called bisecting kmeans, which recursively partitions objects into two clusters at each step until the number of clusters is k [10].

2) No Wkmeans-type algorithms With intercluster separation:

To obtain the best k (the number of clusters), some validity indexes [3] which integrate both intracluster compactness and Intercluster separation are used in the clustering process. Yang et al. and Wu et al. [11] proposed a fuzzy compactness and separation (FCS) algorithms which calculates the distances between the centroids of the cluster and the global centroids as the intercluster separation. The promising results are obtained since FCS is more robust to noises and outliers than traditional fuzzy kmeans clustering.

B. Vector Wkmeans-type algorithms

A major problem of No Wkmeans type algorithms lies in treating all features equally in the clustering process. In practice, an interesting clustering structure usually occurs in a subspace defined by a subset of all the features. Therefore, many studies attempt to weight features with various methods [4].

1) Vector Wkmeans Type Algorithm without Intercluster Separation:

Automated variable Wkmeans is a typical vector weighting clustering algorithm, which can be formulated as,

$$P(U, W, Z) = \sum_{p=1}^{k} \sum_{i=1}^{n} u_{ip} \sum_{j=1}^{m} w_j^{\beta}(x_{ij} - z_{pj})^2$$

Subject to

$$u_{ip} \in \{0, 1\}, \sum_{p=1}^{k} u_{ip} = 1, \sum_{j=1}^{m} w_j = 1, 0 \le w_j \le 1$$

Where W is a weighting vector for the features.

De Sarboet al. [4] first introduced a feature selection method; SYNCLUS which partitions features into several groups and uses weights for feature groups in the clustering process.

The algorithm needs a large amount of computational cost. It may not be applicable for large data sets.

2) Vector Wkmeans Type Algorithm with Intercluster Separation:

De Soete [5] proposed an approach to optimize feature weights for ultrametric and additive tree fitting. This approach calculates the distances between all pairs of objects and finds the optimal weight for each feature. However, this approach requires high-computational cost since the hierarchical clustering method used to solve the feature selection problem in this approach needs high-computational cost.

C. Extensions of Kmeans

1) Extension of Basic Kmeans (E-Kmeans)

Basic kmeans is a typical clustering algorithm which has been widely used in various data analysis. However, it considers only the distances between centroids and objects, i.e., intracluster compactness. To utilize intercluster separation, the global centroid of a data set is introduced. Different to the basic Kmeans, E-kmeans, is expected to minimize the distances between objects and the centroid of the cluster that the objects belong to, while maximizing the distances between centroids of clusters and the global centroid. As shown in figure *Z*0 is the global centroid and *Z*1, *Z*2 are the centroids of cluster 1, cluster 2, respectively [1].

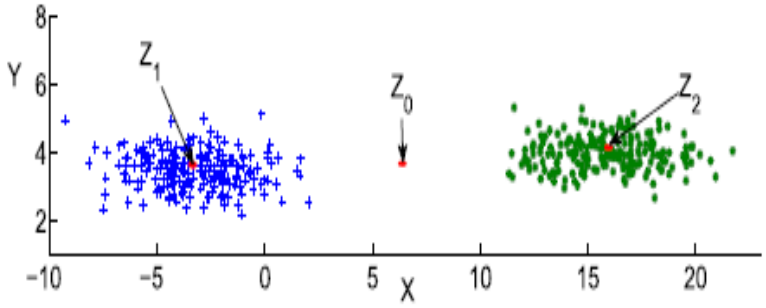

Fig. 1: Effect of Intercluster separation

2) Extension of WKmeans (E-WKmeans)

Basic kmeans and E-kmeans treat all the features equally. However, features may have different discriminative powers in real-world applications. Wkmeans algorithm evaluates the importance of the features according to the dispersions of a data set. E-Wkmeans algorithm considers the dispersions of a data set and the distances between the centroids of the clusters and the global centroid simultaneously while updating the feature weights [1].

3) Extension of AWA (E-AWA)

In Wkmeans and E-Wkmeans, the same feature in different clusters has the same weight. The same feature in different clusters, however, has different weights in most real-world applications [1].E-AWA solves this problem under the condition of utilizing both intracluster compactness and intercluster separation.

III. PROPOSED SYSTEM

A. Contribution and Objective

To implement the existing k-means algorithm for intracluster compactness and intercluster separation with modified objective functions. This will integrate the intracluster and intercluster distance to new objective functions.

B. Proposed Work

The main feature of proposed framework lies in the addition of the information of intercluster separation in a clustering process. At present, most traditional kmeans type algorithms only utilize the intracluster compactness. On the contrary, our proposed framework synthesizes both the intracluster compactness and the intercluster separation.

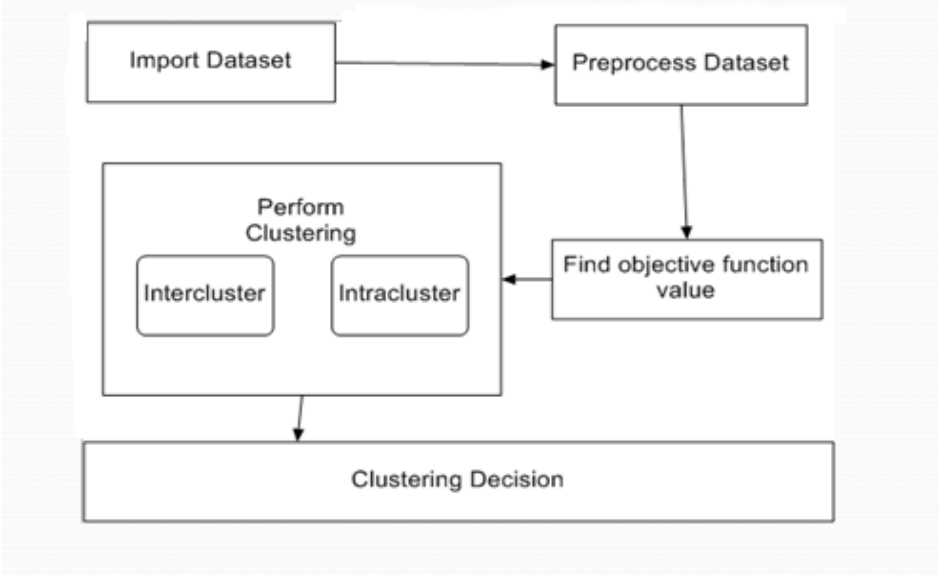

Fig. 2: Architecture of proposed system

Proposed work gives the modified approach based on FCM (Fuzzy C Mean) which consider the intracluster compactness and intercluster separation simultaneously.

Let X={x1,x2,…,xn} be the set of an s dimensional data set. Consider μ $(x1)_1$…… μ $(x)_c$ are fuzzy c-partitions $\mu_{ij} = \mu_i(Xj)$ represent the degree that the data point xj belongs to cluster i. a1,…a_n are the cluster centres. We can define a objective function as below.

$$J_{FCM}(\mu, a) = \sum_{j=1}^{n} \sum_{i=1}^{c} \mu_{ij}^{m} \left\| x_j - a_i \right\|^2$$

Where weighting factor m represents the degree of fuzziness.

Proposed system modifies the fuzzy compactness and separation algorithm based on fuzzy scatter matrix. It adds penalized term to the existing work as below.

$$J_{FCS}(\mu, a) = \sum_{j=1}^{n} \sum_{i=1}^{c} \mu_{ij}^{m} \left\| x_j - a_i \right\|^2 - \sum_{j=1}^{n} \sum_{i=1}^{c} \eta_i \mu_{ij}^{m} \left\| a_i - \bar{x} \right\|^2$$

Where parameter $\eta_i >= 1$, It is known that the $J_{FCS}=J_{FCM}$ When $\eta_i = 0$. System will consider the following equation which minimizes the objective functions.

$$\mu_{ij} = \left(\left\| x_j - a_i \right\|^2 - \eta_i \left\| a_i - \bar{x} \right\|^2 \right)^{\frac{-1}{m-1}} \Bigg/ \sum_{k=1}^{c} \left(\left\| x_j - a_k \right\|^2 - \eta_k \left\| a_k - \bar{x} \right\|^2 \right)^{\frac{-1}{m-1}}$$

And

$$a_i = \frac{\sum_{j=1}^{n} \mu_{ij}^{m} x_j - \eta_i \sum_{j=1}^{n} \mu_{ij}^{m} \bar{x}}{\sum_{j=1}^{n} \mu_{ij}^{m} - \eta_i \sum_{j=1}^{n} \mu_{ij}^{m}},$$

$$\eta_i = \frac{(\beta/4)\min_{i \neq i} \|a_i - a_i\|^2}{\max_k \|a_k - \bar{x}\|^2}, 0 \leq \beta \leq 1.0$$

C. Proposed Fuzzy Compactness and Separation Algorithm

Input: Data set

Output: Clusters

1. Import Dataset.
2. Normalize dataset.
3. For all data points in dataset
4. Calculate value of J_{fcs}
5. Compute the value of μ_{ij} for x_i
6. If μ_{ij} has good degree for cluster a_j
7. Add x_i to a_j cluster.
8. Else
9. Check for other clusters centers
10. End for
11. Return clusters

Where,

$J_{fcs=}$ Objective Function by using FCS

$\mu_{ij=}$ Degree to which object x_i belong to cluster a_j

IV. Conclusion

A new framework for kmeans-type algorithms to include the impacts of the intracluster compactness and the intercluster separation in the clustering process is proposed. Three extensions of kmeans type algorithms by integrating both intracluster compactness and intercluster separation are given and new objective functions based Fuzzy Compactness and Separation are proposed. The extending algorithms are able to produce better clustering results in comparison to other algorithms.

Acknowledgement

I have taken efforts in this survey on a new framework of Kmeans clustering algorithm by combining the dispersion of cluster. However, it would not have been possible without the kind support and help of many individuals. I am highly indebted to Prof. A. D. Gujar for his guidance and constant supervision as well as for providing necessary information regarding this approach.

References

1. Xiaohui Huang, Yunming Ye, and Haijun Zhang "Extensions of Kmeans-Type Algorithms: A New Clustering Framework by Integrating Intracluster Compactness and Intercluster Separation", IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, VOL. 25, NO. 8, AUGUST 2014.
2. J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques. San Mateo, CA, USA: Morgan Kaufmann, 2011.
3. S. Das, A. Abraham, and A. Konar, "Automatic clustering using an improved differential evolution algorithm," IEEE Trans. Syst., Man Cybern., A, Syst. Humans, vol. 38, no. 1,pp. 218–237, Jan. 2008.

4. W. De Sarbo, J. Carroll, L. Clark, and P. Green, "Synthesized clustering: A method for amalgamating alternative clustering bases with differential weighting of variables," Psychometrika, vol. 49, no. 1, pp. 57–78, 1984.
5. G. De Soete, "Optimal variable weighting for ultrametric and additive tree clustering," Qual. Quantity, vol. 20, no. 2, pp. 169–180, 1986.
6. M. Al-Razgan and C. Domeniconi, "Weighted clustering ensembles," in Proc. SIAM Int.Conf. Data Mining, 2006, pp. 258–269.
7. X. Chen, Y. Ye, X. Xu, and J. Zhexue Huang, "A feature group weighting method for subspace clustering of high-dimensional data," Pattern Recognit., vol. 45, no. 1, pp. 434–446, 012.
8. A. Jain, "Data clustering: 50 years beyond k-means," Pattern Recognit. Lett., vol. 31, no. 8, pp. 651–666, 2010.
9. R. Xu and D. Wunsch, "Survey of clustering algorithms," IEEE Trans. Neural Netw., vol.16, no. 3, pp. 645–678, May 2005.
10. M. Steinbach, G. Karypis, and V. Kumar, "A comparison of document clustering techniques," in Proc. KDD Workshop Text Mining, vol. 400. 2000, pp. 525–526
11. K. Wu, J. Yu, and M. Yang, "A novel fuzzy clustering algorithm based on a fuzzy scatter matrix with optimality tests," Pattern Recognit. Lett. vol. 26, no. 5, pp. 639–652, 2005.

AUTHOR(S) PROFILE

Amruta S. Suryavanshi, is currently pursuing M.E (Computer) from Department of Computer Engineering, Bhivarabai Sawant College of Engineering and research, Savitribai Phule Pune University, Pune, Maharashtra, India -411007. She received her B.E (Information Technology) Degree from D. Y. Patil college of Engineering and Technology, Shivaji University, Kolhapur, Maharashtra, India -416113. Her area of interest is Data Mining.

Anil D. Gujar, received the M.Tech. (IT) degree from the Department of Information Technology, Bharati Vidyapeeth University, Pune, Maharashtra, India. He is currently working as Asst. Professor with Department of Computer Engineering, Bhivarabai Sawant College of Engineering and Research, Pune, Maharashtra, India.

