



## A Survey Paper on Modified Approach for Kmeans Algorithm

Authors

**Amruta S. Suryavanshi<sup>1</sup>, Prof. Anil D. Gujar<sup>2</sup>**

<sup>1</sup>TSSM's Bhivarabai Sawant College of Engineering and Research, Pune, India  
 Email: [amrutasuryavanshi@gmail.com](mailto:amrutasuryavanshi@gmail.com)

<sup>2</sup>TSSM's Bhivarabai Sawant College of Engineering and Research, Pune, India  
 Email: [anil.ranje@yahoo.com](mailto:anil.ranje@yahoo.com)

### 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 Kmeans.*

**Keywords-** Clustering, Data Mining, E-kmeans, E-Wkmeans, E-AWA, Feature weighting Kmeans, Fuzzy Compactness and Separation, Fuzzy C Mean

### 1. 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 <sup>[2]</sup>, 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 <sup>[6][7]</sup>.

A series of new clustering algorithms by extending the existing kmeans-type algorithms is given 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 <sup>[1]</sup>.

### 2. EXISTING WORK

In existing system authors investigate the potential framework of the kmeans-type algorithms by integrating both intracluster compactness and intercluster separation <sup>[1]</sup>. System proposes three algorithms: E-kmeans, E-Wkmeans, and E-AWA,

which extend basic kmeans, Wkmeans, and AWA, respectively. The desirable features of algorithms can be concluded as follows.

1) The generality of the extending algorithms encompasses a unified framework that considers both the intracluster compactness and the intercluster separation [1]. System develops a new framework for kmeans-type algorithms to include the impacts of the intracluster compactness and the intercluster separation in the clustering process.

2) The proposed framework is robust because it does not introduce new parameters to balance the intracluster compactness and the intercluster separation. The proposed algorithms are also more robust for the input parameter which is used to tune the weights of the features in comparison to the original kmeans-type algorithms.

3) The extending algorithms are able to produce better clustering results in comparison to the state-of-the-art algorithms in most of cases since they can utilize more information than traditional kmeans-type algorithms such that our approaches have capability to deliver discriminative powers of different features.

### 3. 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.

#### 3.1 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  $X = \{X_1, X_2, \dots, X_n\}$  be a set of  $n$  objects. Object  $X_i = \{x_{i1}, x_{i2}, \dots, x_{im}\}$  is characterized by a set of  $m$  features (dimensions). The membership matrix  $U$

is a  $n \times k$  binary matrix, where  $u_{ip} = 1$  indicates that object  $i$  is allocated to cluster  $p$ , otherwise, it is not allocated to cluster  $p$ .  $Z = \{Z_1, Z_2, \dots, Z_k\}$  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.

#### 3.2 Vector WKmeans Type Algorithm

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

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

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.

### 3.3 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.

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.

## 4 MOTIVATION

Most of existing clustering methods consider only intracluster compactness. So, the weights of features are updated according to the dispersions of the cluster but this does not work well in certain circumstances. It may be ineffective to evaluate the weights of the features using only the dispersions of a data set. Under this condition, the intercluster separation can play an important role in distinguishing the importance of different features.

## 5 PROPOSED WORK

We propose a new framework of kmeans-type algorithm by combining the dispersions of the clusters which reflect the compactness of the intracluster and the distances between the centroids of the clusters indicating the separation between clusters. We will give the complete proof of convergence of the extending algorithms based on the new framework.

### 5.1 Problem Statement

In our proposed work we implement the new algorithm with new objective function to solve the problem of intracluster compactness and intercluster separation. Our proposed FCM based FCS algorithm works simultaneously on both i.e. intracluster compactness and intercluster separations. Proposed algorithm will give us a better performance over existing k-means.

## 6 CONCLUSION

The paper presents a new approach for the intercluster and intracluster clustering. This paper presents the different approaches and extension of k-means algorithm used for the intercluster and intracluster clustering. The paper proposed a new approach based on FCM called as FCS which increases the intercluster distance and minimizes the intracluster compactness.

## 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.