ISSN: 2250-3676
Volume-3, Issue-4, 165-169

ROBUST NEW DISTANCE KERNELIZED APPROACH TO DISTRIBUTED CLUSTERING

Deepika Singh[1], Anjana Gosain[2]

[1] Research Scholar, USICT GGSIP University sector- 16C New Delhi -110078 India, deep.16feb84@gmail.com
[2]Professor, USICT GGSIP University sector- 16C New Delhi -110078 India, anjana_gosain@hotmail.com

Abstract

Clustering has become one of the most widely used tasks in analyzing the vast amount of data. In clustering the given datasets are grouped in similar sets where the data points of one group are dissimilar to the data points belonging to other groups. Fuzzy clustering is based on the clustering process which forms the soft clusters. The clustering algorithm DKFCM- new identifies outliers by using density of points in the data-set before creating clusters. It is a density oriented kernelized technique to fuzzy c-means algorithm based on new distance measure. But all these algorithms are traditional clustering algorithms. Traditional clustering algorithms require all the data sets to reside at the single location In today's business scenario data sets usually resides at different data locations, for this distributed clustering algorithms are implemented. This paper presents a distributed version of DKFCM- new algorithm named as robust new distance kernelized approach to distributed clustering.

Index Terms: *clustering, fuzzy clustering, kernel function, distributed clustering, centroid, membership function.*

--- *** ---

1 INTRODUCTION

Clustering is an unsupervised classification process of the data elements. Fuzzy clustering result in overlapping data clusters where an object may belong to more than one cluster at a time. With each object is associated the membership degree set which defines the association of an object to a particular cluster. Higher value of this membership degree indicates the close belongingness of the object in a particular cluster. The most widely used fuzzy clustering algorithm is Fuzzy c-means algorithm [1]. There are number of variants of FCM algorithm. FCM can only detect hyper spherical cluster because for the distance measure it uses Euclidean distance. Many other distance measures have been proposed by the researchers like Mahalanobis distance measure, kernel based distance measure to identify non-hyper spherical clusters.

Density Oriented Fuzzy C-Means (DOFCM) [2], [3] identifies outliers using density of points in the data-set before creating clusters. DOFCM produces 'n+1' clusters with 'n' good clusters and one outliers cluster. Further DKFCM – new algorithm was proposed in [4]. It incorporates the kernel function and the distance measure proposed by Tsai and Lin [5]. All these above clustering algorithm work only for the single centralized data. Due to the increase in number of autonomous data sites there is a need for efficient distributed clustering techniques. In distributed clustering, the data resides at different geographical sites.

Distributed clustering implies applying the same or different clustering algorithms at each local subset and then combining this local clustering knowledge to form the global clustering. Distributed clustering offers better scalability, increases data security and offer better response time as compared to the centralized clustering process.

In this work we have proposed a distributed version of DKFCM – new algorithm which efficiently clusters the data distributed at different locations and also removes the noise from the clusters formed at local sites. The paper is organized as follows. In Section 2, we have given the various versions of Fuzzy c – means algorithm and in Section 3, we have focus on the main algorithm i.e. the Distributed Density oriented Kernelized approach to Fuzzy C – means with new Distance metric. Finally, in section 4 we concluded the paper.

2. RELATED STUDIES

2.1 Fuzzy c- means algorithm (FCM)

Fuzzy c-means [1] is the first and most popular fuzzy clustering algorithm. Consider a set of unlabeled data set $X = \{x_1, x_2, \ldots.., x_n\}$, $x_i \in R^p$, where 'n' is the number of data sets and '*p*' is the dimension of data set vectors (features). FCM assumes that the number of clusters '*c*' is known in priori and it focuses on minimizing the objective function (J_{FCM}) as

$$J_{FCM} = \sum_{i=1}^{c} \sum_{k=1}^{n} u_{ik}^{m} d_{ik}^{2} \tag{1}$$

where
n: the number of patterns in X
c: the number of clusters
U: the membership function matrix; the elements of U are u_{ik}
u_{ik}: the value of the membership function of the ith data set belonging to the k^{th} cluster
d_{ik}: the distance from x_i to v_k viz. $d_{ik} = \|x_k - v_i\|$
V: the cluster centor vector
m: the exponent on u_{ik} to control fuzziness or amount of clusters overlap, m=2 is used in this paper
The FCM algorithm minimizes the objective function with the constrained on U

$$\sum_{i=1}^{c} u_{ik} = 1; \;\; i = 1,2, \ldots ..n \tag{2}$$

Minimization of J_{FCM} is performed by a fixed point iteration scheme known as the alternating optimization technique. The conditions for local extreme for (1) and (2) are derived using Lagrangian multipliers:

$$u_{ik} = \frac{1}{\sum_{j=1}^{c} \left(\frac{d_{ik}}{d_{jk}}\right)^{\frac{2}{m-1}}} \quad \forall\ k, i \tag{3}$$

where 1≤ i ≤ c; 1≤ k ≤ n and

$$v_i = \frac{\sum_{k=1}^{n}(u_{ik}^{m} x_k)}{\sum_{k=1}^{n}(u_{ik}^{m})} \quad \forall\ i \tag{4}$$

The FCM algorithm iteratively optimizes $J_{FCM}(U,V)$ with the continuous update of U and V, until $|U^{(l+1)} - U^{(l)}| <= \varepsilon$, where '$l$' is the number of iterations.

2.2 Kernelized Fuzzy c- means (KFCM)

In KFCM the kernel function is used for calculating the distance of the data points from the centroid. Kernel function transforms the linear algorithms into their equivalent non-linear algorithms. It uses a mapping function $\Phi(x)$, which defines a non-linear transformation: $x \rightarrow \Phi(x)$. Given an unlabeled dataset $X = \{x_1, x_2, \ldots, x_n\}$in the p- dimensional space R^p , let Φ be a non – linear mapping function from this input space to a high dimensional feature space H:

$$\Phi : R^p \rightarrow H, x \rightarrow \Phi(x)$$

The dot product in the high dimensional feature space can be calculated through the kernel function $K(x_i,x_j)$ in the input space R^p.

$$K(x_i,x_j) = \Phi(x_i)\ \Phi(x_j) \tag{5}$$

In KFCM, the objective function is modified as:

$$J_{KFCM} = \sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik}^{m} \parallel \Phi(x_k) - \Phi(v_i)\|^2 \tag{6}$$

where $\|\ \Phi(x_k) - \Phi(v_i)\|^2$ is the square distance between $\Phi(x_k)$ and $\Phi(v_i)$. The distance in the feature space is calculated through the kernel in the input space as follows:

$$\begin{aligned} \Phi(d_{ki}^{2}) = \|\ \Phi(x_k) - \Phi(v_i)\|^2 &= (\Phi(x_k) - \Phi(v_i))(\ \Phi(x_k) - \Phi(v_i)) \\ &= \Phi(x_k)\ \Phi(x_k) - 2\ \Phi(x_k)\ \Phi(v_i) + \Phi(v_i)\ \Phi(v_i) \\ &= K(x_k,x_k) - 2K(x_k,v_i) + K(v_i,v_i) \end{aligned} \tag{7}$$

For positive kernel width $K(x,x)$ = 1. Thus (6) can be written as

$$J_{KFCM} = \sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik}^{m} \parallel 1 - K(x_k,v_i) \parallel^2 \tag{8}$$

Minimizing (8) under the constraint of U, we get

$$u_{ik} = \frac{1}{\sum_{i=1}^{c}\left(\Phi_{d_{ik}^{2}} / \Phi_{d_{ki}^{2}}\right)^{1/(m-1)}} \tag{9}$$

$$v_i = \frac{\sum_{k=1}^{n} u_{ik}^{m} x_k}{\sum_{k=1}^{n} u_{ik}^{m}} \tag{10}$$

2.3 Density Oriented Fuzzy c- means (DOFCM)

DOFCM results into 'n+1' clusters with 'n' good clusters and one noise cluster. It identifies outliers on the basis of density of points in the data. Neighborhood membership of point 'i' in the data set X is defined as

$$M_{neighborhood}^{i} = \frac{\eta_{neighborhood}^{i}}{\eta_{max}} \tag{11}$$

where
$\eta^{i}_{neighborhood}$: is the number of points in the neighborhood of point i
η_{max} : is the maximum number of points in the neighborhood of any point in the data set

α : is the threshold value, it should be close to zero and selected from the range of $M_{neighborhood}$ values

Consider point '*i*' in the data set '*X*', then if

$$M^{i}_{neighborhood} = \left\{ \begin{array}{ll} < \alpha & \text{outlier} \\ \geq \alpha & \text{non-outlier} \end{array} \right\} \tag{12}$$

DOFCM modifies the objective function as

$$J_{DOFCM(U,V)} = \sum_{i=1}^{c+1} \sum_{k=1}^{n} u_{ki}^{m} d_{ki}^{2} \tag{13}$$

Here the number of clusters is '*c*+1' since one is the noise cluster. Membership function u_{ki} is modified as

$$u_{ki} = \begin{cases} \frac{1}{\sum_{j=1}^{c} (d_{ki}/d_{ji})^{2/(m-1)}} \forall k,i & \text{if } M^{i}_{neighborhood} \geq \alpha \\ 0(zero) & \text{if } M^{i}_{neighborhood} < \alpha \end{cases}$$

(14)

Updating of centroid is same as in FCM as (4). The constraint on fuzzy membership is extended to

$$0 \leq \sum_{i=1}^{c} u_{ki} \leq 1, \quad i = 1,2,\ldots,n \tag{15}$$

2.4 Fuzzy c-means with new distance metric (FCM - σ)

Tsai and Lin [5] proposed FCM – σ, using a new distance measure which is defined as

$$\hat{d}_{ki}^{2} = \frac{\|x_k - v_i\|^2}{\sigma_i} \tag{16}$$

Here σ_i is the weighted mean distance of cluster '*i*' and is calculated as

$$\sigma_i = \left\{ \frac{\sum_{k=1}^{n} u_{ki}^{m} \|x_k - v_i\|^2}{\sum_{k=1}^{n} u_{ki}^{m}} \right\}^{1/2} \tag{17}$$

The objective function for *FCM*- σ is minimizes as

$$J_{FCM-\sigma} = \sum_{i=1}^{c} \sum_{k=1}^{n} u_{ik}^{m} \frac{\|x_k - v_i\|^2}{\sigma_i} \tag{18}$$

The membership function and cluster centers are calculated as

$$u_{ik} = \frac{1}{\sum_{j=1}^{c} \left(\hat{d}_{ik}/\hat{d}_{jk}\right)^{2/(m-1)}} \forall k,i \tag{19}$$

where $1 \leq i \leq c;\ 1 \leq k \leq n$ and

$$v_i = \frac{\sum_{k=1}^{n} (u_{ik}^{m} x_k)}{\sum_{k=1}^{n} (u_{ik}^{m})} \forall i \tag{20}$$

2.5 Kernel Fuzzy c-means with new distance metric (KFCM – σ)

Traditional FCM and FCM – σ works well for linearly separable data sets. The observed data sets can be transformed to higher dimensional feature space through a non-linear mapping function. KFCM – σ uses the following new distance measure

$$\hat{\Phi}_{d_{kc}^{2}} = \frac{\|\Phi(x_k) - \Phi(v_i)\|^2}{\Phi_{\sigma_i}} = \frac{\Phi_{d_{kc}^{2}}}{\Phi_{\sigma_i}} \tag{21}$$

$$\Phi_{\sigma_i} = \left\{ \frac{\sum_{k=1}^{n} u_{ki}^{m} \Phi_{d_{kc}^{2}}}{\sum_{k=1}^{n} u_{ki}^{m}} \right\}^{1/2} \tag{22}$$

here $\Phi_{\sigma i}$ is the weighted mean distance of cluster i in the mapped feature space.

KFCM – σ algorithm minimizes the objective function as

$$J_{KFCM-\sigma} = \sum_{i=1}^{c} \sum_{k=1}^{n} u_{ik}^{m} \|\Phi(x_k) - \Phi(v_i)\|^2 \tag{23}$$

The KFCM – σ membership function satisfies the following relation

$$\sum_{i=1}^{c} u_{ik} = 1, \quad i = 1,2,\ldots,n \tag{24}$$

Where

u_{ik} : is the membership of the data set 'x_k' in cluster '*i*'

$\|\Phi(x_k) - \Phi(v_i)\|^2$: is the square distance between x_k and v_i

Minimizing (23) w.r.t U, as per (7) we get

$$u_{ik} = \frac{1}{\sum_{i=1}^{c} \left(\Phi_{d_{kc}^{2}} / \Phi_{d_{ki}^{2}}\right)^{1/(m-1)}} \tag{25}$$

$$v_i = \frac{\sum_{k=1}^{n} u_{ik}^{m} x_k}{\sum_{k=1}^{n} u_{ik}^{m}} \tag{26}$$

2.5 Density oriented Kernelized approach to fuzzy C-means with new Distance metric (DKFCM - new)

DKFCM – new algorithm [4] constructs noiseless clusters. It produces ‘n+1’ clusters with ‘n’ good clusters and one noise cluster. It first identifies outliers and then apply clustering to produce noiseless clusters.

DKFCM – new algorithm defines density factor, called neighborhood membership. This density factor measures density of an object in relation to its neighborhood. The non – outlier clusters has to contain at least a minimum number of other data points in the neighborhood of a given radius of each point in the data set.

$$M^{i}_{neighborhood} = \frac{\eta^{i}_{neighborhood}}{\eta_{max}} \qquad (27)$$

where,
$\eta^{i}_{neighborhood}$: is the number of points in the neighborhood of data point ‘i’
η_{max} : is the maximum number of points in the neighborhood of any point in the data set
Let the point ‘s’ is in the neighborhood of point ‘i’, so ‘s’ will satisfy
$\{s \; \varepsilon \; X \mid dist(i,s) \le r_{neighborhood}\}$ (28)

Outlier in DKFCM – new is defined as a point whose neighborhood membership is less than the threshold value ‘α’.

$$M^{i}_{neighborhood} = \left\{ \begin{matrix} < \alpha & \text{outlier} \\ \ge \alpha & \text{non-outlier} \end{matrix} \right\} \qquad (29)$$

‘α’ can be selected from the range of $M_{neighborhood}$ values and should be close to zero.
DKFCM – new minimizes the objective function as

$$J_{DKFCM\text{-}new(U,V)} = \sum_{i=1}^{c+1} \sum_{k=1}^{n} u^{m}_{ik} \frac{\left\| 1 - \exp\left(- \sum \left| x^{a}_{i} - x^{a}_{j} \right|^{b} / h^{2} \right) \right\|^{2}}{\Phi_{\sigma_i}} \qquad (30)$$

here u_{ki} us calculated as

$$u_{ki} = \begin{cases} \dfrac{1}{\sum_{j=1}^{c} \left(\dot{\Phi}_{d^2_{kc}} / \dot{\Phi}_{d^2_{ki}} \right)^{1/(m-1)}} \forall k,i & \text{if } M^{i}_{neighborhood} \ge \alpha \\ 0(zero) & \text{if } M^{i}_{neighborhood} < \alpha \end{cases} \qquad (31)$$

The constraint on fuzzy membership is extended to

$0 \le \sum^{c}_{i=1} u_{ki} \le 1, \; i = 1,2,\ldots\ldots,n$ (32)

DKFCM – new calculates the cluster centers as

$$v_i = \frac{\sum_{k=1}^{n} u^{m}_{ki} K(x_k, v_i) x_k}{\sum_{k=1}^{n} u^{m}_{ki} K(x_k, v_i)} \qquad (33)$$

Algorithm:

Input parameters: Data-set(X), Number of clusters(i=c+1), Number of Iterations, Stopping criteria(C), fuzziness index(m).

Output: Cluster centroids matrix, Outlier vector, and Membership matrix.

Identification of outliers:

Step 1: for i=1,2,3,…,n; do:
a) Calculate the number of points in the neighborhood of each point i.e. $\eta^{i}_{neighborhood}$
b) Select η_{max}
c) Compute neighborhood membership, $M^{i}_{neighborhood}$, for each point using (27)

Step 2: Select Thershold value ‘α’ based upon density of points in the data-set from the whole range of neighborhood membership values.
Step 3: With the given value of ‘α’, identify outliers using (29)
Clustering process
Step 4: Determine initial centroids v_k.
Step 5: Initialize the membership u_{ki} and update the memberships of all the outliers to zero
Step 6: for n=1,2,3,….,max_iter; do:
a) Update all centroids v^{n}_{i} using (33)
b) Update all membership values u_{ki} using (31)
c) Compute objective function (O^{n}) using (30)
d) Compute $E^{n} = \max | O^{n} - O^{n-1} |$, if $E^{n} \approx C$, stop;
Else n=n+1

3. THE PROPOSED TECHNIQUE

3.1 Robust Kernelized approach to Distributed Clustering by incorporating new distance measure (DDKFCM - new)

After explaining FCM, KFCM, DOFCM, FCM – σ, KFCM – σ, DKFCM – new, we are now in position to construct the distributed version of DKFCM – new algorithm. In distributed environment data does not reside at a single site instead it is distributed at multiple sites. It involves heavy cost and time if the data of multiple sites is transmitted at a single site for the clustering process. Also, usually the data is so large that practically it becomes infeasible to transfer all the data at the single site. Therefore, distributed clustering algorithms are used to solve such clustering problems.

DDKFCM – new is designed for the distributed environment. It works as follows: First, at each local site DKFCM – new algorithm is implemented to form the clusters locally. Second, this local centroid information is transmitted to the global site. Third, the global site retransmits this local centroid information to all the other local sites so that local sites can update their respective cluster centroids accordingly. This procedure is repeated until there is no change in the centroid positions. DDKFCM – new algorithm removes the affect of noise on the cluster centroid locations.

Proposed Algorithm (DDKFM - new)

Input parameters: Data – set(X), Number of Clusters (i = c+1), number of Iterations, Stopping criteria (€), fuzziness index(m), number of data sites.

Output: Cluster centroids matrix, Outlier vector, and Membership matrix of each local data site.

```
Step1:  communicate cluster prototypes from each data site to all others;
Step2:  For each data site D[$ii$], $ii$=1,….,P do
Step3:  [Identification of outliers at each data site]
        for $i$=1,2,3,…,$n$; do
            a)  calculate the no of points in the neighborhood of each point i.e. $\eta^{i}_{neighborhood}$
            b)  select $\eta_{max}$
            c)  compute neighborhood membership, $M^{i}_{neighborhood}$ for each point using (27)
Step4:  Select threshold value '$\alpha$' based upon density of points in the data-set from the
        whole range of neighborhood membership values.
Step5:  with the given value of '$\alpha$', identify outliers using (29)
Step6:  end //for loop of Step3
Step7:  For each data site D[$ii$], $ii$=1,….,P do
Step8:  [Local Clustering process]
        Determine initial centroids $v_{k.}$
Step9:  Initialize the membership $u_{ki}$ and update the membership of all the outliers to
        zero.
Step10: for $n$ = 1,2,3,…,max_iter ; do:
            a)  Update all centroids $v_i^n$ using (33)
            b)  Update all membership values $u_{ki}$ using (31)
            c)  Compute objective function ($O^n$) using (30)
            d)  Compute $E^n$ = max | $O^n$ - $O^{n-1}$|, if $E^n \approx C$, stop;
                Else $n$=$n$+1
Step11: end // for loop of Step7
Step12: communicate the local centroids $v_i^n$ to the global site.
Step13: repeat Step10 to Step 12 until cluster prototypes do
        not significantly change between two consecutive
        iterations.
```

4. CONCLUSION

In this paper we first discussed the various versions of Fuzzy c – means algorithm which includes the basic FCM algorithm. Then we discussed the Kernelized FCM approach and Density based FCM algorithm. Thereafter, we explained FCM algorithm with new distance measure, Kernel FCM with new distance measure and then Density oriented kernelized approach to FCM using new distance measure. Due to the emerging trend of distributed data sets residing at different sites, we proposed a distributed version of robust kernel approach proposed by prabhjot [4]. DDKFCM – new algorithm uses the new distance measure which is based on kernel function. In future work we will give the implementation of DDKFCM – new algorithm.

REFERENCES

[1]. J.C. Bezdek, Pattern Recognition With Fuzzy Objective Function Algorithms. Norwell, MA: Kluwer, 1981.

[2]. Prabhjot Kaur, Anjana Gosain, 2010. Density oriented approach to identify outliers and get noiseless clusters in fuzzy c-means. In: 2010 IEEE International Conference on Fuzzy Systems, July 18-23, Barcelona, Spain.

[3]. Kaur, P.,Gosain,A., 2011.A density oriented fuzzy c-means clustering algorithm for recognizing original cluster shapes from noisy data. Int. J. Innovative Computer Appl.3(2),77–87.

[4]. Prabhjot Kaur, A.K. Soni, Anjana Gosain, 2013. Robust kernelized approach to clustering by incorporating new distance measure. International Journal Elsevier, Engineering Applications of Artificial Intelligence, 883- 847.

[5]. D.M. Tsai, C.C.Lin, 2011. Fuzzy c-menas based clustering for linearly and non-linearly separable data. Pattern Recognition 44(2011), 1750-1760.

BIOGRAPHIES

Deepika Singh, completed her M.Tech (CSE) from Guru Gobind Indraprasth University, New Delhi. She has done MCA from IGNOU and BCA from Dr. B.R. Ambedkar University, Agra. She has qualified Net 2012, Gate 2011, Gate 2010. Her technical and research interests include data mining, fuzzy logic and databases.

Dr. (Mrs.) Anjana Gosain is working as Associate Professor in University school of information and Communicaion Technology. She obtained her Ph.D. from GGS Indraprastha University & M.Tech in Information Systems from Netaji Subhas Institute of Technology (NSIT) Delhi. Prior to joining the school, she has worked with computer science department of Y.M.C.A institute of Engineering, Faridabad (1994 - 2002). She has also worked with REC kurukshetra. Her technical and research interests include data warehouse, requirements engineering, databases, software engineering, object orientation and conceptual modeling. She has published around 50 research papers in International / National journals and conferences.