Available online at www.sciencedirect.com

ScienceDirect

Procedia Technology 10 (2013) 596 – 603

Procedia Technology

International Conference on Computational Intelligence: Modeling Techniques and Applications (CIMTA) 2013

A Novel Selective Scale Space Based Fuzzy C-Means Model for Spatial Clustering

Parthajit Roy[a], J. K. Mandal[b,*]

[a]*Department of Computer Science, The University of Burdwan, Burdwan, India-713104*
[b]*Department of Comp. Sc. & Engg., The University of Kalyani, Nadia, India-741235*

Abstract

This paper proposed a novel Scale Space Filter based Fuzzy C-Means algorithm for clustering spatial data. The number of clusters, C , in present case is known in advance. The Scale Space filter is used for better separability of the data which are not linearly separable and in the present paper the same is used to selective parameters for betterment to meet the complexity-accuracy tradeoff. The Xie-Beni validity index is used as Objective Function of the model to check the quality of the clusters produced. The Results are tested on Standard iris data. The analysis and comparative study with existing algorithms has also been drawn.

Selection and peer-review under responsibility of the University of Kalyani, Department of Computer Science & Engineering

Keywords: Fuzzy C-Means, Clustering, Space Scaling, Xie-Beni, Benchmark Data Set.

1. Introduction

Pattern Recognition or Pattern Classification an emerging and promising field in the modern day computing and in the field of Pattern Recognition, clustering is one of the interesting and difficult task. Clustering means unsupervised grouping of objects where labeled data for training is not available [1]. This restriction has made the clustering more challenging. Clustering mainly deals with the topology of the data set, proximity of the data set or some kind of density or distribution measures. Learning by only these shot of tools is a major focus area of pattern clustering.

There are five major steps of data clustering proposed by Jain et al[1]. These are

- Feature Extraction
- Introduction of the idea of proximity
- Clustering of grouping
- Data abstraction
- Validation of cluster

*Corresponding author
E-mail address: jkm.cse@gmail.com (J. K. Mandal).

Selection and peer-review under responsibility of the University of Kalyani, Department of Computer Science & Engineering
doi:10.1016/j.protcy.2013.12.400

Given n objects $\{o_1, o_2, \cdots, o_n\}$ to be clustered the features of the of the objects are to be extracted first. If there are m different attributes that represents an object, the feature vector for every object will be m dimensional vectors. Let the features of each objects are represented by $S = p_1, p_2, p_3, \cdots, p_n$ where every $p_i \in R^m$. i.e. the feature of any object is m dimensional vector. From now onward it will be assumed that a feature vector p_i is the representative of the actual object o_i. After the feature is extracted, the concept of proximity is introduced through distance vectors. There are different distance measures like Mahalanobis distance, Minkowski distance available but the most frequently used distance measure is Euclidean distance. Having the feature vectors and the distance in hand, the objective of the clustering is to find out a partition matrix $F(S)$ of order $C \times n$ representing the clustering into C number of clusters where each cell of the matrix $u_{ij}, \forall i = 1, \cdots, n \ \ and \ j = 1, \cdots, C$ has a value representing the degree of membership of i^{th} object to the j^{th} cluster[2].

Duda et al [3] discussed a solid foundation of classification whereas Everitt et al [4] explored the clustering from theoretical aspects. An excellent survey on clustering up to 1999 is done by Jain et al[1]. An outstanding survey on clustering algorithms is done by Xu and Wunsch II[5].

The most popular clustering algorithm is Fuzzy C-Means algorithm[5] where the partition matrix will take fuzzy membership values. i.e. in case of Fuzzy C-Means algorithm, one object may belong to more than one clusters with variable degree of membership values. Fuzzy C-Means takes an approach which is called known cluster number approach. Here, it is assumed that the total number of clusters is known in advanced and the algorithm proposes an initial random cluster center for the set of points and then subsequently updates the clusters till the desired accuracy. Robust survey on Fuzzy clustering techniques is done by Baraldi and Blonda [6][7] in 1999. An effort for generalization of Fuzzy C-Means is done by Zhu et al [8].

Scale Space Filter[9] is a technique of data refinement where the data sets are transformed to higher dimension. In this paper we have developed a Space-Scale Filter based Fuzzy C-Means algorithm. The idea is, given a set of objects with several parameters, the objective is to separate them according to their proximity and separability measures. As there is no scope for labeled data in clustering, the proximity-separability is the major thing to be exploited. The Scale Space Filter does exactly the same thing. It transforms the data set into higher dimension for better proximity-separability. i.e. near points, after transformation, will come nearer and far points will scatter further.

The model endeavors to proceed further. Given the data set, even if we apply the Scale-Space Filters to all of the parameters separately then also we may not get a fruitful result. There are reasons behind this is, even if there are m number of parameters in a data set, not all the parameters are equally strong for clustering. Some features are very strong for clustering where as some may be such that all values are very close to one another and hence, direct application of Scale Space Filters to all of the parameters will not be wise. If a particular parameter or a set of parameters are not well distributed along the line of clustering then the Scale Space Filtering can only be applied to those parameters. On the other hand, if a particular parameter is well distributed for clustering, then the application of scale space filtering may scatter the parameters and the proximity of clustering, which is the key factor in any clustering, may be lost. This novel approach is called selective approach.

Secondly, application of scale space filtering to some parameters instead of that to all parameters will increase the computational efficiency of the mode also.

In the proposed technique, a novel, Standard Deviation based, selective Scale Space Filtering approach for Fuzzy C-Means algorithm has been proposed to improve the quality of the result as well as the performance of the overall model. This approach identifies the parameters which are not well distributed by calculating the Standard Deviations. The key concept behind this is that, if the Standard Deviation is high, then the parameter values are well scattered. So, Scale Space will not be applied to those. It is only those parameters where the S.D. values are lower, the application of Scale Space Filters will be more fruitful. Some benchmark value for S.D. can be set depending upon the application area.

Rest of the paper is organized as follows. Section 2 explains the mathematical foundations. Section 3 proposes the model along with its working principle. The result and discussion are given in section 4. Future scope and the possible up gradation of the proposed model is given in section 5 and reference are drawn at the end.

2. Mathematical Background

Suppose are n number of objects to be clustered whose representative feature vectors are $S = \{p_1, p_2, p_3, \cdots, p_n\}$. Let us also suppose that each point has m attributes. So, We can assume that each point is m dimensional vector in R^m. In case of Fuzzy C-Means, it is assumed that there are C number of clusters (A *priori* which is essential for FCM). Given such information, the generic Fuzzy C-Means technique is stated using algorithm 1.

Algorithm 1 Fuzzy C-Means Algorithm

Input: $S = \{p_1, p_2, \cdots, p_n\}$ ▷ Points to be clustered.
Output: $C = \{k_1, k_2, \cdots, k_C\}$ ▷ The final cluster centers.

```
declare F_{n×C}S as Matrix
declare ObjectiveValue as Real
INITIALIZECLUSTERCENTERS(C)
while ObjectiveValue ≤ Benchmark do
    POPULATEFUZZYPARTITIONMATRIX(F_{n×C}, C, S)
    UPDATECLUSTERCENTERS(C, F_{n×C})
    ObjectiveValue ← OBJECTIVEFUNCTION(C, F_{n×C})
end while
```

The fuzzy membership should be such that

$$\sum_{k=1}^{n}\sum_{i=1}^{C}\mu_{ik} = n$$

with $0 \le \mu_{ik} \le 1 \forall i = 1, 2, \cdots, C$ *and* $\forall j = 1, 2, \cdots n$, $0 < \sum_{k=1}^{n}\mu_{ik} < 1$ and $\sum_{i=1}^{C}\mu_{ik} = 1$

The validity of the cluster in present paper is measured by Xie-Beni index[10]. This index is a function of Variation v and Minimum Separation of the clusters centers. Variation v can be stated using equation 1[11].

$$v(P,C,X) = \sum_{i=1}^{C}\sum_{j=1}^{n}\mu_{ij}^2 E^2(c_i, x_j) \tag{1}$$

where c_i are cluster centers and x_j are feature vectors. The separation measure can be stated using equation 2.

$$d(C) = min_{i\neq j} E^2(c_i - c_j) \tag{2}$$

where the $E(.,.)$ ia the Euclidean norm and finally the Xie-Beni index (XB) can be given using equation 3.

$$XB(P,C)_X = \frac{v_X(P,C)}{n\, d(C)} = \frac{\sum_{i=1}^{C}\sum_{j=1}^{n}\mu_{ij}^2 E^2(c_i, x_j)}{n\, min_{i\neq j} E^2(c_i - c_j)} \tag{3}$$

The Xie-Beni index always shows its inclination towards compactness of a single cluster as well as separability of different clusters. The value of membership functions for every cluster points to every cluster centers is given in equation 4[11].

$$\mu_{i,k} = \frac{1}{\sum_{j=1}^{C}\left(\frac{E(c_i - x_k)}{E(c_j - x_k)}\right)^{\frac{2}{m-1}}}, 1 \le i \le C, 1 \le k \le n \tag{4}$$

The Center update rule[11] is done through the equation 5

$$c_i = \frac{\sum_{k=1}^{n}(\mu_{ik})^m x_k}{\sum_{k=1}^{n}(u_{ik})^m} \tag{5}$$

3. Proposed Model

In the present paper a novel variation of the traditional Fuzzy C-Means algorithm has been proposed. The model has proposed a preprocessing which is responsible for better separability of the feature vectors. This is popularly known as Space Scaling or Scale Space Filtering.

3.1. Scale Space Based Fuzzy C-Means

Given a set of vectors, it may so happen that the vectors are not very good for clustering. We used to say that the clusters are not linearly separable. The idea of scale space filter is that some transformation function will be given to the parameters so that the clusters will become linearly separable from the transformed data set. Some previous work on clustering using scale space filtering can be found in [9]. The proposed models have adapted this Scale Space filtering in the Fuzzy C-Means model for better performance. The Scale Space based Fuzzy C-Means model passes the input parameters through Scale Space Filters and gets the data in higher dimension. The purpose of this function is to scale the space.

If for a particular parameter the values are very close to each other then it is not an effective parameter for clustering. mathematically if for some parameter i, the corresponding values of the feature vector $x_j[i]$ for $j = 1,2,\cdots,n$ are such that $\theta - \varepsilon \leq x_j[i] \leq \theta + \varepsilon \ \ \forall j = 1,2,\cdots,n$ for some θ and some very small ε, then we used to say that the parameter is ill parameter. In present paper, this parameter is considered for further tuning, so that the ill parameters, after scale space filtering, will become strong to contribute to the clustering process significantly.

There are several functions that can be used as scale space filters. Gaussian function and polynomial functions are some of them which are very popular. In present model we have applied a Gaussian transformation function

$$f(x,\sigma) = \frac{1}{(\sigma\sqrt{2\Pi})^2} e^{-\frac{E^2(x-x_i)}{2\sigma^2}}$$

The function is applied to all of the parameters individually. The resultant values are then processed and normalized using equation 6. The resultant values are then used in the Fuzzy C-Means algorithm for clustering. As the functions are such that the distance between far values will increase, there is possibility that the clusters which are not linearly separable, will become linearly separable. So, the chance of correct clustering will be better. The proposed Scale Space based Fuzzy C-Means algorithm is stated in algorithm 2.

Algorithm 2 Scale Space Filtering based FCM

Input: $S = \{p_1, p_2, \cdots, p_n\}$ ▷ Points to be clustered.
Output: $C = \{k_1, k_2, \cdots, k_C\}$ ▷ The final cluster centers.

```
declare F_{n×C}S as Matrix
declare ObjectiveValue as Real
for All p_i ∈ S do
    p_i ← GaussianScaleSpace(p_i)
end for
initializeClusterCenters(C)
while ObjectiveValue ≤ Benchmark do
    populateFuzzyPartitionMatrix(F_{n×C}, C, S)
    updateClusterCenters(C, F_{n×C})
    ObjectiveValue ← objectiveFunction(C, F_{n×C})
end while
```

3.2. Standard Deviation based Selective parameter Scale Space FCM

The above mentioned model may be modified further for better efficiency. The idea is as follows,

As the Scale Space has a tendency that the far points will scatter further, it may be the case that the values scatters to such a large extent, that they losses the proximity. the result is two originally close points will end up to be there in separate clusters. That is why not all the parameters are suitable for scale space transformation. For this, in the present paper, a novel Standard Deviation(S.D.) based selective model has been proposed. The idea is, first the S.D. of the values for a particular parameter is considered and then if the standard deviation is low, that means scattering is not good for that parameter, then only the Scale Space transformation can be applied to these parameters. If the S.D. value is large, i.e. the original data set is well scattered, the Scale Space filters will not be applied to these parameters.

As the range of the values of a parameter may vary, from computational point of view, we have normalized the data before passing to the actual model.

Given a particular parameter, the formula for normalization can be stated as follows.

$$x[j] = \frac{x_i[j] - min(X[j])}{max(x[j]) - min(x[j])} - \forall x_i[j] \in X[j] \tag{6}$$

where $x[j]$ is the set of j^{th} parameter of the data set x and $x_i[j]$ is the j^{th} parameter of i^{th} data.

The final model after all kind of preprocessing and internal reorientation has been proposed in the algorithm 3.

Algorithm 3 S.D. based Selective Scale Spaced FCM

Input: $S = \{p_1, p_2, \cdots, p_n\}, sdThreshold$
Output: $C = \{k_1, k_2, \cdots, k_C\}$ ▷ The final cluster centers.

declare $F_{n\times C}S$ **as** *Matrix*
declare $ObjValue$ **as** *Real*
NORMALIZEDATASET(S)
for $All\ \ parameter[i]\ \ of\ \ dataset\ \ S$ **do**
    $sd \leftarrow$ GETSTANDARDDEVIATION($param_i$)
    if $sd < sdThreshold$ **then**
        $p_i \leftarrow$ GAUSSIANSCALESPACE(p_i)
    end if
end for
INITIALIZECLUSTERCENTERS(C)
while $ObjValue \leq Benchmark$ **do**
    POPULATEFUZZYPARTITIONMATRIX($F_{n\times C}, C, S$)
    UPDATECLUSTERCENTERS($C, F_{n\times C}$)
    $ObjValue \leftarrow$ XIEBENIOBJFUNCTION($C, F_{n\times C}$)
end while

4. Result and Discussion

The model has been tested on iris data[12]. Iris data set, made online by UCI machine learning unit, is a benchmark data set for Pattern Classification/Clustering having 150 instances with four parameters Sepal length, Sepal width, Petal length and Petal width.

The data set has three classes namely *'Iris Setosa'*, *'Iris Versicolor'* and 'Iris Virginica' each having 50 instances. There is no missing data and no outliers. Out of these, *Iris Setosa* is linearly separable from the other two and the *Iris Versicolor* and *Iris Virginica* have instance overlapping i.e. they are not linearly separable.

In the present paper two types of errors have been considered for the comparative study of the proposed models with the standard existing model. These are,

- Same class non identified element (Type I).
- Different class identified element (Type II).

For example, in table 1, for Versicolor, the number of actual point is 50, but the associated model has predicted 60. Which does not mean that the error is 10 extra points. By further examination, it reveals that out of 60, 47 points are form the Versicolor class itself, and 13 are from other class or classes. i.e. from the same class it couldn't identify 3 points and has identified 13 points wrongly from other classes. i.e. the total error is 16 points. So, the error is 3 out of 50 for same class not identified and 13 out of rest 100 points for different classes identified. So, the error is $(6+13)\% = 19\%$.

The data set has been normalized first and then Fuzzy C-Means algorithm has been teated on it. Models of the present paper, namely Scale Space based FCM and S.D. based selective Scale Space FCM, has also been tasted on it and has been compared with the Fuzzy C-Means algorithm.

Table 1 shows the result of pure Fuzzy C-Means algorithm on iris data. It is clear that the identification of Setosa is correct. This is very much expected from the distribution of the sample as it is known that the class Setosa is linearly separable, but, for Versicolor and Verginica, the result is not at all satisfactory. From the class distribution also it is known that these two classes are not linearly separable. So, Scale Space theory has been adapted for better separability of the input vectors.

Table 2 shows the result of the Scale Space based Fuzzy C-Means algorithm. The result shows that the clustering is much better than the original pure FCM algorithm. The reason is that we have applied scaling function on the parameters for better separability.

Table 3 shows results of S.D. based selective Scale Space FCM which is applied for selective parameters, namely parameter 1 and parameter 2. Parameters are scaled down by normalizing and fabricated within $[0,1]$ followed by Standard Deviation. The reflection is shown in figure 1.

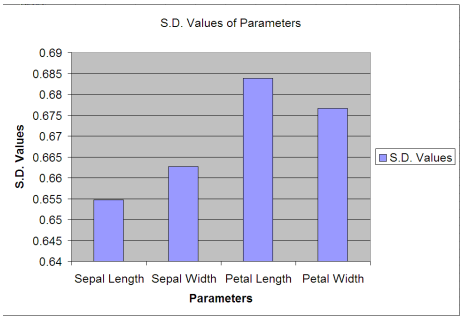

Fig. 1. The Parameterwise Standard Deviation

Table 1. The Result of Pure FCM

Class Name	Actual Distribution	Prediction by Pure FCM			Percentage of Error
		Total Pred	Correct Pred	Wrong Pred	
Setosa	50	50	50	0	0%
Versicolor	50	60	47	13	19%
Verginica	50	40	37	3	29%

Table 2. The Result of Scale Space FCM

Class Name	Actual Distribution	Scale Space FCM			Percentage of Error
		Total Pred	Correct Pred	Wrong Pred	
Setosa	50	54	50	4	4%
Versicolor	50	49	49	0	2%
Verginica	50	47	47	0	6%

Table 3. The Result of Standard Deviation based Scale Spaced FCM

Class Name	Actual Distribution	S.D. based Scale Space FCM			Percentage of Error
		Total Pred	Correct Pred	Wrong Pred	
Setosa	50	50	50	0	0%
Versicolor	50	53	50	3	3%
Verginica	50	47	47	0	6%

Table 4. The Comparative Study of Models on Iris Data

Model Name	Correct Prediction	Wrong Prediction
Pure FCM	89.3%	10.7%
Scale Space FCM	97.3%	2.7
SD Scale Space FCM	**98%**	**2%**

It is clear from the figure 1 that the Sepal length and Sepal width i.e. parameter 1 and 2 has relatively smaller Standard Deviations, which means, their original from are not suitable for clustering. So the scale space based transformation has been applied on these two parameters only. A benchmark S.D. value is considered, below which, the scale spacing can be applied. The results are shown in table 3. The result of S.D. based selective Scale Space based FCM is even better than Space Scaling applied to all of the parameters. The reason is, if Scale Space filter is applied to a parameter having larger S.D., the values will be too far to merge to a single cluster. i.e. there proximity will be lost. The comparative study of the models considered in the present paper is shown in table 4.

5. Conclusion and Future Scope

This paper surveyed the Fuzzy C-Means and analyzed its performance along with the variation of the FCM, i.e. Scale Space filtering based fuzzy C-means algorithm and S.D. based Selective Scale Space FCM algorithm. The results of Fuzzy C-Means with Selective Scale Space filtering obtained better performance. The even better result is achieved by selective Scale Space based FCM. Nevertheless, there are scope of betterment of the proposed model. First of all, its computation time is little bit higher than pure Fuzzy C-Means as the Scale Space filtering is incorporated. Secondly, number of clusters are to be known in advance. Thirdly, if the shape of the clusters are not like hyper-sphere. The problem can further be addressed by introducing the Principal Component Analysis(PCA) where the extra bit of computation can be nullified by the reduction of the dimensionality. Some better searching techniques like GA or Tabu search can also be introduced for even better results.

Acknowledgments

The authors express the deep gratitude to the Department of Computer Science, the University of Burdwan, West Bengal, India and Department of Computer Science & Engineering, the University of Kalyani, West Bengal, India, for providing necessary infrastructure and support for the present work.

References

[1] A. K. Jain, M. N. Murty, P. J. Flynn, Data clustering: A review, ACM Computing Surveys 31 (3) (1999) 264–323.
[2] U. Maulik, S. Bandyopadhyay, Performance evaluation of some clustering algorithms and validity indices, IEEE Transaction on Pattern Analysis and Machine Intelligence 24 (12) (2002) 1650–1654.
[3] R. O. Duda, P. E. Hart, D. G. Stork, Pattern Classification, 2nd Edition, John Wiley, 2008.
[4] B. S. Everitt, D. Stahl, M. Leese, S. Landau, Cluster Analysis, 5th Edition, John Wiley & Sons, 2011.
[5] R. Xu, D. W. II, Survey of clustering algorithms, IEEE Transactions on Neural Network 16 (3) (2005) 645–678.
[6] A. Baraldi, P. Blonda, A survey of fuzzy clustering algorithms for pattern recognitionpart i, IEEE Transaction on Systems, Man and Cybernetics-Part B:Cybernetics 29 (6) (1999) 778–785.
[7] A. Baraldi, P. Blonda, A survey of fuzzy clustering algorithms for pattern recognitionpart ii, IEEE Transaction on Systems, Man and Cybernetics-Part B:Cybernetics 29 (6) (1999) 786–801.
[8] L. Zhu, F.-L. Chung, S. Wang, Generalized fuzzy c-means clustering algorithm with improved fuzzy partitions, IEEE Transaction on Systems, Man , and CyberneticsPart B: Cybernetics 39 (3) (2009) 578–591.
[9] Y. Leung, J.-S. Zhang, Z.-B. Xu, Clustering by scale-space filtering, IEEE Transactions on Pattern Analisys and Machine Intelligence 22 (12) (2000) 1396–1410.
[10] X. L. Xie, G. Beni, A validity measure for fuzzy clustering, IEEE Transactions on Pattern Analisys and Machine Intelligence 13 (8) (1991) 841–847.
[11] U. Maulik, S. Bandyopadhyay, Fuzzy partitioning using a real-coded variable-length genetic algorithm for pixel classification, IEEE Transactions on Geoscience and Remote Sensing 41 (5) (2003) 1075–1081.
[12] R. A. Fisher, UCI machine learning repository.
URL http://archive.ics.uci.edu/ml