Using homogeneous fuzzy cluster ensembles to address fuzzy c-means initialization drawbacks

Erind Bedalli, Ilia Ninka

Abstract— The goal of clustering algorithms is to reveal patterns by partitioning the data into clusters, based on the similarity of the data, without any prior knowledge. The fuzzy approach to the clustering problem, where the fuzzy c-means clustering algorithm (FCM) is one of the eminent representatives, provides more flexibility as it allows data to have partial membership into several clusters simultaneously. Despite the plethora of significant applications of FCM, there are several drawbacks which must be taken into consideration, as they may affect the accuracy and reliability of the obtained clusters. Sensitivity to initialization is one of the drawbacks, which is particularly relevant in the case of data characterized by a large value of variance. In this paper we are providing a cluster ensemble approach to address the issue of the sensitivity to initialization. Our methodology consists of the application of the FCM algorithm with several random initializations and several choices of the fuzziness parameter to obtain multiple partitions, which will be fused into the final consensus matrix by using a fuzzy t-norm. Finally we have experimentally evaluated and compared the accuracy of this methodology.

Index Terms— Fuzzy c-means clustering, fuzzy cluster ensembles, consensus partitioning, fuzzy t-norm, unsupervised learning

1 Introduction

Clustering, as an unsupervised learning form, is an important domain of machine learning with wide applications in various fields of research like business, medicine, pattern recognition, cognitive sciences etc. The key idea about clustering is the distribution of the instances (data points) into several clusters based on their similarity, thus revealing structures inside the data. While in the classification problem some class label information about data is provided, the clustering problem is vastly a data-driven procedure; what makes clustering a very important tool as an initial step in data exploratory analysis [1]. The fuzzy approach to the clustering problem provides more flexibility compared to the crisp (hard) approach, as it allows data points to have partial memberships (a real value in the interval]0,1[ ). Thus a data point does not have to be assigned to exactly one of the clusters as in the hard partitioning, but it may belong simultaneously to several clusters. This approach is not only more flexible, but it also more realistic. Also it manages the problems of uncertainty, imprecision and noise that frequently associate the data from real world scenarios [1], [5].

The fuzzy c-means algorithm is one of the most widely-used fuzzy clustering algorithms. It is expressed as a nonlinear optimization problem of an objective function, solved by a Picard iteration scheme. The algorithm needs several parameters to be specified before starting the operations, like the number of clusters, the fuzzy exponent (fuzziness parameter), the similarity metrics and the scale of precision (tolerance) [4], [5], [6]. Also the partition matrix must be initialized, i.e. some initial values must be chosen as initial centers of the clusters. As the algorithm operates in an unsupervised way it is very sensitive to the choice of the parameters and the initial values of the cluster centers. In certain situations the obtained clusters may be unrealistic and the whole study may be affected. Generally the application of a single clustering procedure involves the risk of inconsistent results due to the biases and assumptions that characterize each clustering algorithm [2], [3].

Instead of applying a single clustering procedure on the data, the application of several clustering procedures and later combining the results into a single final partition is the main idea of the cluster ensemble approach. This technique aims to improve the accuracy and to provide stability to the treatment of the fuzzy clustering problem. The fuzzy cluster ensemble approach typically avoids the disadvantages of a poor initialization and also it is more robust to the presence of noise and outliers in the data [2], [7], [8].

Besides the usage of cluster ensemble approach to improve accuracy and robustness by avoiding the drawbacks that would associate a single clustering algorithm, there are several other successful applications of this approach known in literature. So an important application is in knowledge reuse. A typical example would be the exploit of knowledge in legacy clustering while re-clustering the data. Another important application is in the clustering of distributed data. Communicating the entire set of data over the network would be a very complex task, or it may be even prohibited for privacy or security reasons, while communicating results of clustering can be completed without facing any problem [1], [2].

2 The fuzzy c-means clustering algorithm

FCM is a clustering algorithm which allows data points to belong simultaneously to several of the resulting clusters. This method was developed by Dunn and improved by Bezdek [4] and is one the most widely-used unsupervised learning algorithms. The algorithm is essentially a data-driven procedure, as no prior information (like labels of some data) are provided. The data points are categorized into several clusters (classes) based on some distance metric which estimates the dissimilarity between the data points. Thus the algorithm tends to partition the data elements into several clusters where the elements in the same cluster are closer (smaller distance) to each other than to the elements in the other clusters [5], [6]. The elements

may have partial membership in several clusters, while the sum of the membership values into distinct clusters must be equal to 1. The algorithm tries to minimize the objective function:

$$J = \sum_{i=1}^{n} \sum_{j=1}^{c} \mu_{ij}^{\varphi} \, d^2(x_i, c_j) \quad (1)$$

where n represents the number of elements in the data set, c represents the number of the clusters, c_j the center (prototype) of the j-th cluster, x_i the i-th element, μ_{ij} the membership of the x_i element in the c_j cluster, $d^2(x_i, c_j)$ the square of the distance from x_i to c_j according to some distance metrics (dissimilarity metrics), and φ the fuzzy exponent which varies in $[1, \infty)$. There are several possible choices for the distance metrics like the Euclidean distance, the Manhattan distance, the Minkowski distance, the maximum distance, the Pearson correlation distance etc.

The algorithm consists of three important phases: the initialization of the cluster centers, evaluation of the distances of each data point from each cluster center and the update of the partition matrix. The algorithm takes as input the number of the clusters (c), the fuzzy exponent φ (such that φ > 1) and the tolerance scale ε. The algorithm is described by the given pseudo-code [4], [5]:

1. Pick c (random) points as the initial centers of the clusters.
2. Assign k=1
3. Compute the distance from each point and each center according to chosen distance metrics.
4. Update the partition matrix $U_k = [\mu_{ij}]$, with entries:

$$\mu_{ij} = \frac{d_{ij}^{-\frac{2}{\varphi-1}}}{\sum_{k=1}^{c} d_{ik}^{-\frac{2}{\varphi-1}}} \quad (2)$$

5. Compute the new centers

$$c_i = \frac{\sum_{j=1}^{n} \mu_{ij}^{\varphi} x_j}{\sum_{j=1}^{n} \mu_{ij}^{\varphi}} \quad (3)$$

6. If $||U_k - U_{k-1}|| < \varepsilon$ then TERMINATE, otherwise increment k and jump to step 3.

The final result of the FCM algorithm will be the partition matrix U with dimensions cxN whose entries μ_{ij} show the membership of the x_i element in the c_j cluster. Each of these entries will be a value in the interval [0,1] and the sum of the entries of each column will be equal to 1 (the memberships of each data point in the resulting clusters will sum up to 1).

3 Fuzzy t-norms

The fuzzy t-norms (triangular norms) are considered as generalizations of the classical intersection operator. From the mathematical point of view, a t-norm is a function from $[0,1]^2$ to [0,1] for which the following properties hold [13], [14]:

- $T(0,0) = T(0,1) = T(1,0) = 0$ (boundary conditions)
- $T(x,1) = x, \ \forall x \in [0,1]$ (identity property)
- $T(x,y) = T(y,x), \ \forall x \forall y \in [0,1]$ (symmetry property)
- If $0 \le u \le x \le 1, \ 0 \le v \le y \le 1$ then $T(u,v) \le T(x,y)$ (the monotonic property)
- $T(T(x,y),z) = T(x,T(y,z)) \ \forall x \forall y \forall z \in [0,1]$ (the associative property)

As some particular examples of fuzzy t-norms which are carefully investigated and frequently employed in practice we would mention [9], [13], [14]:

- The minimum t-norm (originating from Gödel–Dummett logic) defined as $T(x,y) = min(x,y)$
- The Lukasiewicz t-norm (originating from Lukasiewicz logic) defined as $T(x,y) = max(x+y-1,0)$
- The product t-norm defined as $T(x,y) = xy$
- Hamacher t-norm defined as $T(0,0) = 0$ and $T(x,y) = xy/(x+y-xy)$ when at least one of x and y is different from 0. Also it can be generalized as: $T(x,y) = xy/(r+(1-r)(x+y-xy))$ for any $r > 0$

In our application we utilize the product t-norm in the second stage of the ensemble in order to combine the multiple partitions generated by the different runs of the clustering algorithm, into the coincidence matrix.

4 Design principles of cluster ensembles

There are many ways of designing a cluster ensemble. Firstly there are several choices that can be made about the selection of the clustering algorithm(s) [2], [7], [8], [10], [11], [12]:

- applying multiple runs of a single clustering algorithm with different initializations
- applying multiple runs of a single algorithm by varying one or some of the parameters of the algorithm (i.e. the fuzzy parameter, the distance metrics etc)
- applying multiple runs of a single clustering algorithm with different sets of features (also known as feature-distributed clustering)
- applying different clustering algorithms (known as heterogeneous or hybrid clustering)
- applying the cluster algorithm(s) on different subsets on the entire data set, for example by re-sampling (known as distributed clustering)

Also there exist many other forms of ensemble design by combining the general approaches mentioned above [7]. In this paper we have used a combination of the first and second techniques, so we apply several runs of a single clustering algorithm (the fuzzy c-means algorithm) with different initializations and with different values of the fuzzy parameter.

At the end of the first stage of the ensemble approach, as a result of several runs of the algorithm(s), several soft partitions of the data are obtained. In the second stage the obtained partition will be combined together to generate a single final partition. Also there are several possible choices that can be made for the second stage, like [1], [7], [10], [15], [16]:

- The re-labeling approach (also known as the direct approach) tries to find relationships among the generated clusters in the different partitions in order to

combine the data points placed in the corresponding clusters.

- The graph-theoretic approaches use concepts of graph theory to determine the final resulting partition. In literature some of the most well-known graph theoretic algorithms for the combination of the partitions are: Cluster-based Similarity Partitioning Algorithm (CSPA), Hyper-graph Partitioning Algorithm (HGPA), Meta-clustering Algorithm (MCLA) and Hybrid Bipartite Graph Formulation (HBGF).
- The feature-based approach handles the results of the applied clustering algorithm(s) as a categorical feature. The group of L features is considered as an ''intermediate feature space'' and another clustering algorithm can be run on it.
- The pair-wise approach (also known as co-association approach) utilizes a coincidence matrix among all pairs of data points. In the case of the hard clustering this matrices will have entries 0 (if the two data points do not belong to the same cluster) or 1 (if they belong to the same cluster). In the case of fuzzy clustering the entries of the matrices are evaluated using t-norms. These matrices are then combined to yield the final clustering.

In this paper we have used the pair-wise approach to create the coincidence matrices for each partition. In the final stage we have to generate the consensus clustering based on the entries of the coincidence matrix. Again there is more than one choice that can be made to complete this stage [7], [9]:

- A threshold value may be used to modify the values of the incidence matrix. So the values above the threshold are set to 1 and the values below the threshold are set to 0. Finally the consensus partition is generated directly from the new values of the incidence matrix.
- The values in the incidence matrix may be interpreted as similarities and a (fuzzy) clustering algorithm may be run on them to generate the consensus partition.

We have employed the second approach in our ensemble model. Although the choice of the clustering algorithm in the final stage is independent of the choice of the clustering algorithm during the first stage, we have decided the use the FCM algorithm in both cases.

5 The Implemented Ensemble Model

There are two main approaches in the design of a cluster ensemble, which are the heterogeneous ensembles and the homogeneous ensembles. In the case of heterogeneous cluster ensembles several distinct clustering algorithms are employed to run on the data for generating the partitions, while in the case of homogeneous cluster ensembles a single clustering algorithm is run several times varying one or more of its parameters. In this study we have used a homogeneous cluster ensemble model based on the fuzzy c-means (FCM) algorithm. In the first stage of the ensemble, multiple partitions are generated by using different random initializations and by running the FCM several times for each initialization with different values of the fuzzy parameter (the values that we have used are 1.5, 2, 2.5 and 3). Later using the pair-wise approach we combine the multiple partitions generated in the first stage, into a coincidence matrix. To evaluate the entries of this matrix we use a t-norm, more specifically the product t-norm. Finally we run the fuzzy c-means again, now on the values of the incidence matrix to find the consensus clustering.

Some of the relevant inputs to be specified before our ensemble model starts operations are: the data set $X = \{x_1, x_2, \dots, x_n\}$, the number of clusters to be generated c, the number of runs of the clustering algorithm 4k (we denote it as a multiple of 4 as for each initialization we will apply the FCM with four distinct values of the fuzzy parameter: 1.5, 2, 2.5 and 3) and the fuzzy t-norm $T(x, y)$. The entire ensemble procedure may be described by the given pseudo-code:

1. For each value of the variable $count$ from 0 to $k-1$ do the steps 2 and 3.
2. Randomly initialize the centers of the clusters.
3. Apply the FCM algorithm on the data using this initialization and respectively the values 1.5, 2, 2.5 and 3 for the fuzzy parameter to generate the corresponding partitions $U_{4count+1}$,$U_{4count+2}$,$U_{4count+3}$ and $U_{4count+4}$.
4. For each value of the variable i from 1 to n do the steps 5 - 6.
5. For each value of the variable j from 1 to n do the step 6.
6. Evaluate the current entry of the coincidence matrix M:

$$m_{ij} = \frac{1}{k}\sum_{r=1}^{k} T(U_{ri}, U_{rj}) \quad (4)$$

7. Apply the FCM algorithm on the rows of matrix M to generate the final consensus partition U^*.

6 Experimental Results

Several experiments were conducted on various data sets to compare the accuracy of our fuzzy clustering ensemble method. For each data set we have firstly applied the FCM algorithm alone and later the ensemble method, with respectively 4, 8 and 12 values of the parameter k (remind that 4k represents the total number of runs of clustering algorithms during the first stage, so in these cases the total number of runs is 16, 32, 48). Seven data sets were used among which five datasets of the UCI machine learning repository (http://archive.ics.uci.edu/ml/), namely Wine, Glass, Yeast, Statlog (Vehicle Silhouettes), Dermatology and two synthetic data sets.

The Wine data set contains the results of a chemical analysis of wines grown in the same region in Italy but derived from three different cultivars. The analysis determined the quantities of 13 constituents found in each of the three types of wines.

The Glass data set contains values of samples of glass, with

attributes representing chemical and optical properties of the instances.

The Yeast data-set contains information about proteins within yeast cells with the class attribute denoting the localization within the cell.

The Statlog (Vehicle Silhouettes) dataset contains features of silhouettes of vehicle seen from many different angles.

The Dermatology dataset contains features of skin conditions for some dermatologic diseases.

The details of the datasets are provided in the following table.

TABLE 1
Details of the Used Datasets.

Dataset	Number of attributes	Number of instances
Wine	13	178
Glass	9	214
Yeast	8	1484
Statlog (V.S.)	18	846
Dermatology	33	366
Synth1	9	250
Synth2	12	320

For each data set we do not use any labeling information while applying the clustering procedures. After the final results are generated we utilize the provided labels in the evaluation of the accuracy of the clustering model. In the next table are summarized the results about the accuracy of the respective algorithms.

TABLE 2
Accuracy of the Clustering Models

Dataset	FCM	Cluster ensembles		
		k=4	k=8	k=12
Wine	0.56	0.59	0.60	0.62
Glass	0.63	0.66	0.64	0.65
Yeast	0.66	0.68	0.69	0.68
Statlog (V.S.)	0.57	0.58	0.61	0.61
Dermatology	0.55	0.55	0.57	0.58
Synth1	0.71	0.72	0.74	0.74
Synth2	0.68	0.70	0.72	0.73

As it seen from the table 2, the cluster ensemble approach is almost always more efficient than a single run of the classical fuzzy c-means algorithm. Another important observation is that when the number of multiple runs of the cluster ensemble data points increases, the accuracy is generally improved, but also rarely the increase in the number of multiple runs of the cluster ensemble may not change or may even slightly deteriorate the accuracy.

6 Conclusions

In this paper we have presented a homogenous cluster ensemble approach for enhancing the accuracy and robustness of the treatment to the fuzzy clustering problem. The clustering algorithms are characterized by drawbacks like biases, sensitivity to the initialization, sensitivity to outliers, sensitivity to noise etc. There is no clustering algorithm to operate superiorly to all datasets.

Our cluster ensemble model consists of three main stages. In the first stage many partitions of the data are obtained by multiple runs of the fuzzy c-means algorithm with distinct random initializations and with various values of the fuzzy parameter. In the second stage the partitions are fused to create the coincidence matrix. Here a fuzzy t-norm was used for the evaluation of the entries of the coincidence matrix. In the last stage the entries of the coincidence matrix are used as data on which the fuzzy c-means operates again to yield the final consensus clustering.

Experimental results on five benchmark data sets and two synthetic data sets provided evidence of a better accuracy of the fuzzy cluster ensemble approach compared to the classical fuzzy c-means clustering.

References

[1] De Oliveira, J. Valente, and W. Pedrycz, eds. *Advances in fuzzy clustering and its applications*. New York: Wiley, pp 69-83, 2007.

[2] Strehl, Alexander, and Joydeep Ghosh. "Cluster ensembles---a knowledge reuse framework for combining multiple partitions." *The Journal of Machine Learning Research* 3 (2003): 583-617.

[3] Strehl, Alexander, and Joydeep Ghosh. "Cluster ensembles-a knowledge reuse framework for combining partitionings." *AAAI/IAAI*. 2002.

[4] Bezdek, James C., Robert Ehrlich, and William Full. "FCM: The fuzzy c-means clustering algorithm." Computers & Geosciences 10.2 (1984): 191-203.

[5] Klawonn, Frank, Rudolf Kruse, and Thomas Runkler. Fuzzy cluster analysis: methods for classification, data analysis and image recognition. New York: John Wiley, pp 35-43, 1999.

[6] Bedalli, Erind, and Ilia Ninka. "Implementation of some cluster validity methods for fuzzy cluster analysis." ISCIM 2013.

[7] Hadjitodorov, Stefan T., Ludmila I. Kuncheva, and Ludmila P. Todorova. "Moderate diversity for better cluster ensembles." Information Fusion 7.3 (2006): 264-275.

[8] Kuncheva, L. I., S. T. Hadjitodorov, and L. P. Todorova. "Experimental comparison of cluster ensemble methods." *Information Fusion, 2006 9th International Conference on*. IEEE, 2006.

[9] Avogadri, Roberto, and Giorgio Valentini. "Ensemble clustering with a fuzzy approach." Supervised and Unsupervised Ensemble Methods and their Applications. Springer Berlin Heidelberg, 2008. 49-69.

[10] Avogadri, Roberto, and Giorgio Valentini. "Fuzzy ensemble clustering based on random projections for DNA microarray data analysis." Artificial Intelligence in Medicine 45.2 (2009): 173-183.

[11] Fern, Xiaoli Zhang, and Carla E. Brodley. "Random projection for high dimensional data clustering: A cluster ensemble approach." *ICML*. Vol. 3. 2003.

[12] Kuncheva, Ludmila I., and Dmitry P. Vetrov. "Evaluation of stability of k-means cluster ensembles with respect to random initialization." Pattern Analysis and Machine Intelligence, IEEE Transactions on 28.11 (2006): 1798-1808.

[13] Klement, Erich Peter, Radko Mesiar, and Endre Pap. "Triangular

norms. Position paper I: basic analytical and algebraic properties." Fuzzy Sets and Systems 143.1 (2004): 5-26.

[14] Butnariu, Dan, Erich Peter Klement, and Samy Zafrany. "On triangular norm-based propositional fuzzy logics." Fuzzy Sets and Systems 69.3 (1995): 241-255.

[15] Kuncheva, Ludmila I., and Stefan Todorov Hadjitodorov. "Using diversity in cluster ensembles." *Systems, man and cybernetics, 2004 IEEE international conference on*. Vol. 2. IEEE, 2004.

[16] Hu, Xiaohua, and Illhoi Yoo. "Cluster ensemble and its applications in gene expression analysis." Proceedings of the second conference on Asia-Pacific bioinformatics-Volume 29. Australian Computer Society, Inc., 2004.

IJSER