On the Noise Distance in Robust Fuzzy C-Means

M. G. C. A. Cimino, G. Frosini, B. Lazzerini, F. Marcelloni

***Abstract*—**In the last decades, a number of robust fuzzy clustering algorithms have been proposed to partition data sets affected by noise and outliers. Robust fuzzy C-means (robust-FCM) is certainly one of the most known among these algorithms. In robust-FCM, noise is modeled as a separate cluster and is characterized by a prototype that has a constant distance δ from all data points. Distance δ determines the boundary of the noise cluster and therefore is a critical parameter of the algorithm. Though some approaches have been proposed to automatically determine the most suitable δ for the specific application, up to today an efficient and fully satisfactory solution does not exist. The aim of this paper is to propose a novel method to compute the optimal δ based on the analysis of the distribution of the percentage of objects assigned to the noise cluster in repeated executions of the robust-FCM with decreasing values of δ. The extremely encouraging results obtained on some data sets found in the literature are shown and discussed.

***Keywords*—** noise prototype, robust fuzzy clustering, robust fuzzy C-means.

I. Introduction

CLUSTERING algorithms are widely used in different engineering and scientific fields such as pattern recognition, data mining, knowledge discovery [1]. A clustering algorithm partitions a data set into homogeneous groups (called *clusters*) in such a way that objects within a cluster are more similar to each other than objects belonging to different clusters. A large amount of clustering algorithms have been proposed in the literature. Most of them assume that data sets are not affected by noise and outliers. In real applications, however, this assumption is often false due to different types of problems which may influence the process of data collection. Thus, clustering algorithms can generate misleading partitions. To overcome this problem, in the last years some clustering algorithms *robust* against outliers and noise have been introduced [2]. Among these, very few consider fuzzy rather than crisp partitions. One of the most known robust fuzzy clustering algorithms is certainly the robust version of fuzzy C-means (robust-FCM) proposed by Davé [3]. In robust-FCM, noise is described as a further cluster, denoted *noise cluster*. The noise cluster is represented by a fictitious prototype that has a constant distance δ from all the data objects. Thus, an object belongs to a real cluster only if there exists a prototype such that its distance from the object is less than δ; otherwise, the object belongs to the noise cluster. We consider an object as belonging to a cluster if its membership value to that cluster is higher than to others. Noise distance δ is obviously a critical parameter of the algorithm and should be determined very carefully. As suggested in [4] and [5], the value of δ should be based on data set statistics: in particular, it should be related to the concept of "scale" in robust statistics [2]. Unfortunately, the proper estimation of this scale is not a trivial task [6] and requires some knowledge of the data, which cannot always be supposed in real clustering applications. To overcome these problems, this paper proposes a completely different approach to compute the most suitable value of δ.

The method is based on executing the robust-FCM with decreasing values of δ and analyzing the distribution of the percentage of objects assigned to the noise cluster. This distribution has an abrupt change of slope when the value of δ is so small that objects naturally belonging to real clusters are classified into the noise cluster. The abrupt change determines the optimal δ.

We describe in detail results obtained on three different noisy data sets. The 100% classification performance obtained on all the data sets proves the effectiveness of the method. Further, we also show that the method can achieve 100% classification performance in absence of noise by determining a distance δ which does not include objects in the noise cluster.

II. The Robust-FCM Algorithm

Fuzzy C-means (FCM) is one of the most used and popular fuzzy clustering algorithms. FCM partitions a data set minimizing the Euclidean distance between each point (strongly) belonging to a cluster and the prototype of the cluster. Though several examples of application of FCM to real clustering problems have proved the good characteristics of this algorithm with respect to stability and partition quality, it is well-known in the literature that FCM is not robust against noise and outliers.

Robust-FCM resolves this problem by describing noise as a further cluster, the noise cluster. The presence of the noise cluster modifies the objective function of FCM as follows:

Mario G.C.A. Cimino, Graziano Frosini, Beatrice Lazzerini and Francesco Marcelloni are with the *Dipartimento di Ingegneria dell'Informazione: Elettronica, Informatica, Telecomunicazioni*, University of Pisa, Via Diotisalvi 2, 56122 Pisa, ITALY (corresponding author phone: +39 050 2217678; fax: +39 050 2217600; e-mail: f.marcelloni@iet.unipi.it).

$$J(V,U,X)=\sum_{i=1}^{C}\sum_{j=1}^{M}u_{ij}^{m}\cdot d^2(\underline{v}_i,\underline{x}_j)+\sum_{j=1}^{M}u_{*j}^{m}\cdot\delta^2 \qquad (1)$$

where M is the number of objects, C is the number of classes, m is the fuzzification coefficient, $d(\underline{v}_i,\underline{x}_j)$ is the distance between the prototype $\underline{v}_i$ of cluster i and the object $\underline{x}_j$, u_{ij} is the membership degree of $\underline{x}_j$ to the cluster represented by $\underline{v}_i$, and $u_{*j}=1-\sum_{i=1}^{C}u_{ij}$ is the membership of $\underline{x}_j$ to the noise cluster. The minimization of (1) is achieved by updating iteratively the membership degrees u_{ij} and the cluster prototypes $\underline{v}_i$ in accordance with the formulas proposed in [3] until the difference between two consecutive partitions is lower than a pre-fixed real number ε. The formulas are derived under the constraint $\sum_{i=1}^{C}u_{ij}+u_{*j}=1$.

The success of robust-FCM depends on the appropriate choice of the noise distance δ. If δ is too large, robust-FCM degenerates to classical FCM and outliers are forced to belong to real clusters; on the other hand, if δ is too small, a lot of objects can be considered as noise and misplaced into the noise cluster. Though some solutions to automatically determine the optimal value of δ have been proposed in the literature, the estimation of this value is still an open-problem. In the following, we propose a novel approach which has proven to be effective without requiring a preliminary knowledge of the data.

III. Automatic Detection of the Noise Distance

Intuitively, distance δ fixes the boundary of the noise cluster. This boundary can be visualized as a hyper-spherical surface of radius δ around each real cluster. Fig. 1 shows a synthetic data set composed of three different-size clusters and randomly added noise. Here, the boundary is represented by dashed circles. Let us express distance δ as $\delta^2=\lambda\delta_M^2$, where $\lambda\in[0,1]$ and δ_M is a coarse overestimation of the optimal δ. Thus, if $\lambda\to 1$ no object naturally belonging to a cluster is placed into the noise cluster; on the contrary, if $\lambda\to 0$, all objects are members of the noise cluster.

The starting point of our approach is the typical assumption of each robust clustering method: the density of objects within clusters is considerably higher than outside; furthermore, the density within regions of noise is lower than the density in any of the clusters. We apply robust-FCM with $\lambda=1$. Then, we decrease λ and apply robust-FCM for each value of λ: the number of objects belonging to the noise cluster will remain quite low until δ will be so small as to force objects naturally belonging to a cluster to be members of the noise cluster. Since the density within real clusters is considerably higher than within the noise cluster, when this occurs, the number of noise objects increases rapidly. The sudden change of the number of noise objects determines the optimal value λ_{opt} of λ. In Fig. 1, when the radius of the circles is getting so small as to "cut out" some objects of the largest cluster, then the number of noise objects increases suddenly.

Fig. 2 plots the percentage p (white circles in the figure) of objects classified in the noise cluster against λ. To speed up computation, instead of using a pre-fixed step, we update λ by the following rule: $\lambda^{(t)}=\lambda^{(t-1)}/2$, where $\lambda^{(t)}$ and $\lambda^{(t-1)}$ are the values of λ at the current and previous execution of robust-FCM. We can observe that p increases very slowly for the first five values of λ (recall that λ starts from 1 and then decreases); then, at the sixth value, the curve shows a sudden increase. It is interesting to observe that we have the maximum classification rate (white squares in Fig. 2) in proximity to the slope change. We decrease λ and execute robust-FCM until the percentage p becomes higher than 0.5. The computation of further values is useless since, in our hypothesis, the number of noise objects is certainly lower than the number of objects belonging to real clusters. To automate the determination of λ_{opt}, we approximate the distribution of the percentages p with a Pareto curve $p=q\,\lambda^{-s}$ (continuous line in Fig. 2). In the approximation, we do not consider the points with $p = 0$. Indeed, these points are not expressive of the trend of p. The estimation of q and s can be easily obtained by an anamorphosis procedure, which transforms λ and p in logarithmic scale, and therefore by applying a linear least-squared error minimization in the new scale. We choose λ_{opt} as the value of λ where the prime derivative of the Pareto curve is equal to –1, that is, the tangent to the curve forms an angle of $\pi/4$ with the λ axis. This leads to compute λ_{opt} as $\lambda_{opt}=\sqrt[s+1]{sq}$. In Fig. 2, the percentage p and the classification rate in correspondence of λ_{opt} are represented by a black circle and a black square, respectively.

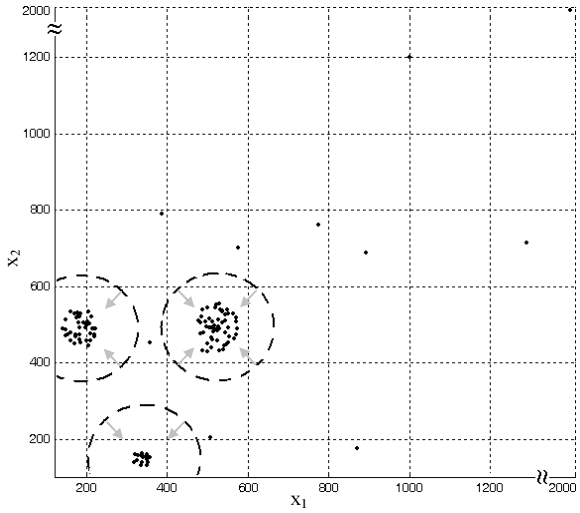

Fig. 1. A synthetic data set with three clusters and ten outliers.

The method can be summarized as follows:

1. Compute $\delta_M^2 = \dfrac{\sum_{j=1}^{M} d^2(\bar{x}, \underline{x}_j)}{M}$, with $\bar{x} = \dfrac{1}{M}\sum_{j=1}^{M} \underline{x}_j$. This value of δ guarantees that no object naturally belonging to a cluster is a member of the noise cluster.
2. Initialize an unsigned integer t to 0 and $\lambda^{(t)}$ to 1. Fix the number C of clusters, the fuzzification coefficient m, and the termination error ε (in the experiments, we used $m = 2$ and $\varepsilon = 0.001$). Execute robust-FCM with an initial random partition. Compute the percentage $p^{(0)}$ of objects belonging to the noise cluster.
3. Increase t and compute $\lambda^{(t)} = \lambda^{(t-1)}/2$.
4. Execute robust-FCM using as initial partition the final partition at iteration t-1.
5. Compute the percentage $p^{(t)}$ of objects belonging to the noise cluster. If $p^{(t)} \le 0.5$ go to item 3.
6. Approximate the values $p^{(t)}$ by a Pareto curve and determine λ_{opt} as described above.

We applied the method to the data set shown in Fig. 1. We achieved 100% classification rate. We executed the robust-FCM 10 times, but thanks to the use of the final partition of a trial as initial partition of the subsequent, robust-FCM converged in very few steps (about 4-5 against 20-30 of the random initialization).

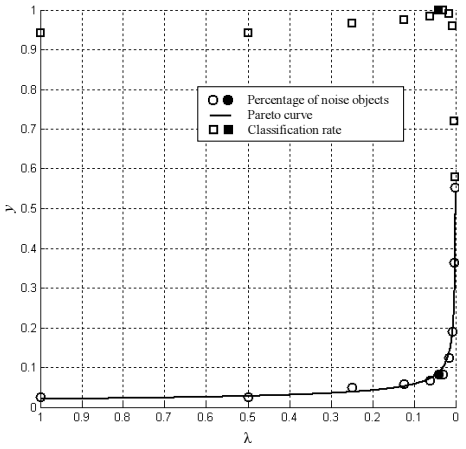

Fig. 2. Percentage of noise objects and classification rate for the data set in Fig. 1.

To verify the reliability of our approach, we eliminated the outliers in the data set in Fig. 1. We wanted to check whether the method was able to determine a value of λ, which did not include objects in the noise cluster. Fig. 3 plots the percentage p of objects classified in the noise cluster and the percentage of correct classifications against λ. Making the prime derivative of the Pareto curve equal to –1, we obtain a λ_{opt} which corresponds to 100% classification rate. No object, therefore, belongs to the noise cluster, thus confirming that our method can be also used in absence of noise and outliers.

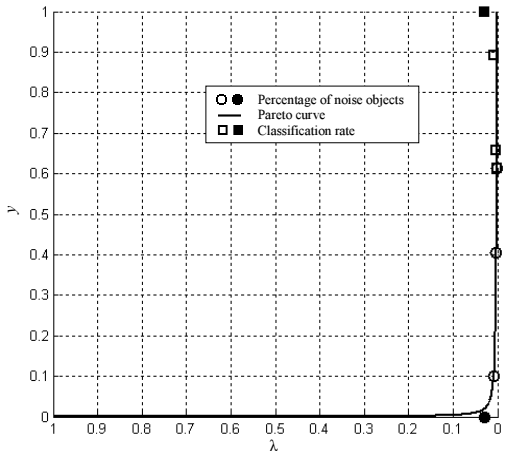

Fig. 3. Percentage of noise objects and classification rate for the data set in Fig. 1 (after eliminating the outliers).

IV. Experimental Results

A. Synthetic dataset

Fig. 4 shows the second data set used in the experiments. The data set is extracted from [7] and consists of two distinct clusters and six outliers (black stars) scattered over the 2-dimensional space. The data set is interesting because the clusters are not strongly compact. Fig. 5 plots the percentage p of objects classified in the noise cluster and the percentage of correct classifications against λ. Since the number of points belonging to each cluster is comparable to the number of outliers, we can observe that p shows two changes of slope: the first change is quite gradual and occurs when the noise cluster starts to include the outliers; the second is abrupt and occurs when points belonging naturally to one of two real clusters are included into the noise cluster. This second variation of the slope identifies the optimal λ, which allows establishing an optimal boundary between the noise cluster and the real clusters. The continuous line in Fig. 5 shows the Pareto curve that approximates the distribution of p. In correspondence of λ_{opt}, we have again 100% classification rate (black square in the figure).

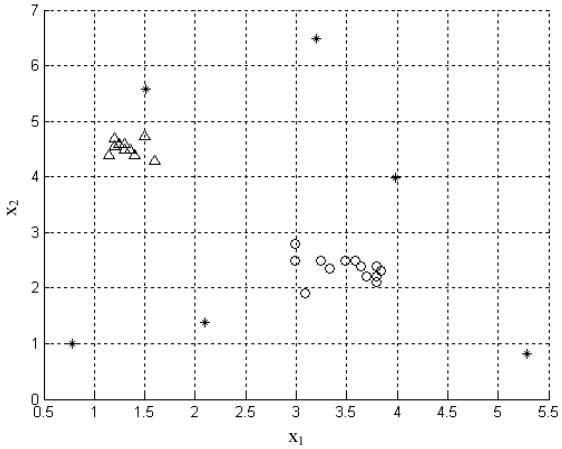

Fig. 4. A synthetic data set with two clusters and six outliers.

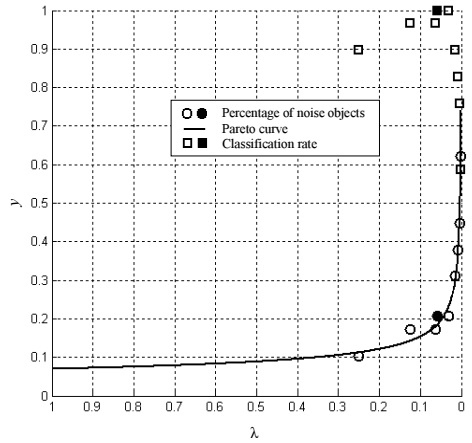

Fig. 5. Percentage of noise objects and classification rate for the data set in Fig. 4.

B. Iris dataset

The last data set is extracted from [7] and is obtained by adding six randomly generated patterns to the well-known Iris data set. Iris consists of 150 patterns characterised by 4 numeric features which describe, respectively, sepal length, sepal width, petal length and petal width. Patterns are equally distributed in three classes of Iris flowers, namely Iris Setosa, Iris Versicolor and Iris Virginica. Fig. 6 shows the Iris data along with the noise in a 3-dimensional feature space. The outliers are represented as black stars. Fig. 7 plots the percentage p of samples classified in the noise cluster and the percentage of correct classifications against λ. In correspondence of λ_{opt} all the outliers are classified in the noise cluster, thus producing 100% correct classification of the outliers and complete separation between noise and data. The slight overlap between classes Versicolor and Virginica does not allow, however, the robust-FCM to achieve 100% total classification rate. This well-known problem depends on the clustering algorithm and also affects FCM when applied to Iris without outliers).

V. Conclusions

In the framework of robust fuzzy clustering algorithms, robust fuzzy C-means proposed by Davé holds a significant position. Though robust-FCM has proved to be effective in identifying noise and outliers, its success strongly depends on the appropriate choice of the noise distance. In this paper, we have proposed a method to automatically perform this choice. The method exploits the typical assumption of all robust clustering algorithms, that is, the density of outliers is lower than the density of the objects in real clusters. We have discussed in detail the application of our method to three data sets. We have shown that the method can achieve optimal performance with a limited computational effort. To further assess the validity of our approach, we performed other experiments on some of the noisy datasets shown in [4] and [5]. Our approach always selected a λ_{opt} able to allow robust-FCM to achieve almost 100% classification rate.

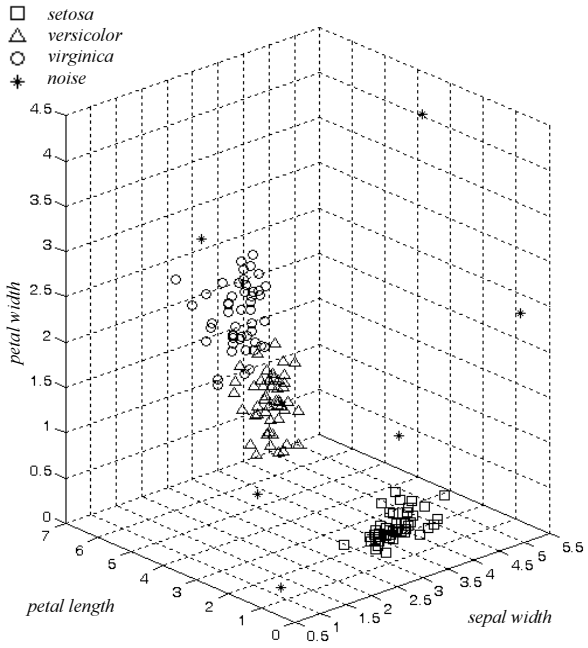

Fig. 6. Iris data set with outliers represented by three-dimensional features.

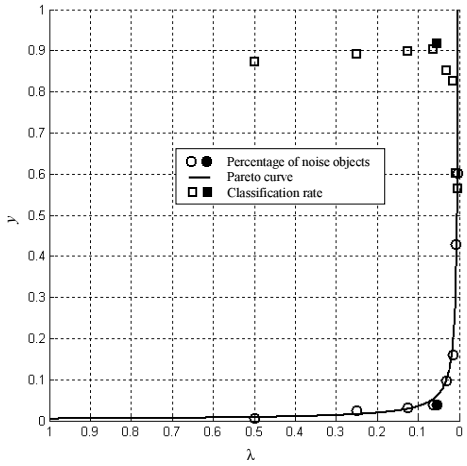

Fig. 7. Percentage of noise objects and classification rate for the data set in Fig. 6.

References

[1] A. K. Jain, M. N. Murty, P. J. Flynn. "Data clustering: A review," *ACM Computing Surveys*, vol. 31, no. 3, 1999, pp. 265-323.

[2] R. N. Davé, R. Krishnapuram, "Robust Clustering Methods: A Unified View", *IEEE Transactions on Fuzzy Systems*, vol. 5, no. 2, 1997, pp. 270-293.

[3] R. N. Davé, "Characterization and detection of noise in clustering", *Pattern Recognition Letters*, vol. 12, no. 11, 1991, pp. 657-664.

[4] R. N. Davé, "Robust Fuzzy Clustering Algorithms", *Second IEEE International Conference on Fuzzy Systems*, 28 March-1 April 1993, vol. 2, pp. 1281-1286.

[5] R. N. Davé, S. Sen, "Robust fuzzy clustering of relational data", *IEEE Transactions on Fuzzy Systems*, Vol. 10, No. 6, pp. 713-727, 2002.

[6] R. N. Davé, S. Sen, "Noise Clustering Algorithm Revisited", *NAFIPS'97*, 21-24 September 1997, pp. 199-204.

[7] S. Pemmaraju, S. Mitra, "Identification of noise outliers in clustering by a fuzzy neural network", *Second IEEE International Conference on Fuzzy Systems*, vol.2, pp. 1269 – 1274, April 1993.