A new Fuzzy Noise-rejection Data Partitioning Algorithm with Revised Mahalanobis Distance

M.H. Fazel Zarandi, Milad Avazbeigi[1] I.B. Turksen[2]

1 Department of Industrial Engineering, Amirkabir University of Technology
Tehran, Iran
2 Department of Industrial Engineering, TOBB Economy and Technology University
Ankara, Turkey
Email: zarandi@cic.aut.ac.ir, miladavazbeigi@cic.aut.ac.ir, bturksen@etu.edu.tr

Abstract—*Fuzzy C-Means (FCM) and hard clustering are the most common tools for data partitioning. However, the presence of noisy observations in the data may cause generation of completely unreliable partitions from these clustering algorithms. Also, application of the Euclidean distance in FCM only produces spherical clusters. In this paper, a new noise-rejection clustering algorithm based on Mahalanobis distance is presented which is able to detect the noise and outlier data and also ellipsoidal clusters. Unlike the traditional FCM, the proposed clustering tool provides much efficient data partitioning capabilities in the presence of noise and outliers. For validation of the proposed model, the model is applied to different noisy data sets.*

Keywords— Cluster Validity Index (CVI), Fuzzy C-Means (FCM), Possibilistic C-means (PCM), Revised Gustafson-Kessel (GK), Revised Mahalanobis Distance.

1 Introduction

Clustering methods have been extensively used in computer vision and pattern recognition. Fuzzy clustering methods have shown spectacular ability to detect not only volume clusters, but also clusters which are actually thin shells, i.e. curves and surfaces.

Most analytic fuzzy clustering approaches are derived from the fuzzy C-means (FCM) algorithm. FCM uses the probabilistic constraint that the membership of a data point across classes sums to 1. The constraint is used to generate the memberships update equations for an iterative algorithm. The memberships resulting from FCM and its derivative however, do not always correspond to the intuitive concept of belonging or compatibility. Moreover, the algorithms have considerable trouble in noisy environments. Reference [1] summarizes the main problems of classic FCM as follows:

- In order to get the optimal partition, initial locations of the cluster centers should be assigned. The FCM algorithm always converges to a local extreme. Different choices of initial cluster centers may lead to a different extrema.
- The scientific basis for the choice of m, the weighting exponent, is still not clear.
- The optimum number of clusters in the data is assigned a priori. There should be a criterion to assign the optimal number of clusters.

To overcome these problems and drawbacks, first, [2] introduces a new method for fuzzy clustering called possibilistic fuzzy clustering. Their approach differs from the previous clustering methods in that the resulting partition of the data can be interpreted as a possibilistic partition, and the membership values may be interpreted as degrees of possibility of the points belonging to the classes, i.e., the compatibilities of the points with the class prototypes. They construct an appropriate objective function whose minimum will characterize a good possibilistic partition of the data, and derive the membership and prototype update equations from necessary conditions for minimization of the related criterion function. Next nominated work is [1]. Melek et al. in [1] show how their PCM addresses the mentioned problems of the FCM for some example cases. However, their method has some limitations which are the concentration of this paper.

The Rest of the paper is organized as follows: Section 2 discusses the drawbacks of the previous PCM's. Then, in section 3 the proposed method is presented in order to overcome these limitations. Section 4 applies the method for different data sets to verify and validate the proposed method.

2 Limitations of the traditional PCM's

Reference [1] uses Euclidean distance in all data partitioning steps. However, clearly, Euclidean distance usually fails to recognize the appropriate shape of the clusters in complex data sets. Especially, when data include ellipsoidal shapes, Euclidean distance loses sight to those data which should be considered in a cluster from data which should not be included in that cluster.

Moreover, the cluster validity index (*CVI*) applied in [1], suffers from linear behaviour of the Euclidean distance. There are many samples of data sets which the applied *CVI* is not able to recognize the number of clusters correctly.

Another limitation is about noise identification procedure. Application of Euclidean distance can mislead the noise rejection procedure. This problem is discussed in the related section of the proposed algorithm.

While Euclidean distance implies the above mentioned limitations, Mahalanobis distance can mitigate or in some cases can overcome the cited limitations completely. Another advantage of the Mahalanobis distance is that the

ISBN: 978-989-95079-6-8

Mahalanobis distance can identify both spherical and ellipsoidal clusters correctly.

3 Improvements to the traditional PCM's

As quoted in introduction, the proposed method applies the Mahalanobis distance for noise rejection problem. However, when applying mahalanobis distance, some difficulties may occur with covariance matrix.

In this section, first, the problem with calculation of the Covariance Matrix in Mahalanobis distance is discussed. An estimation of Covariance Matrix is introduced to mitigate the problem. Then, new Mahalanobis distance and revised Gustafson-Kessel Clustering are presented based on the estimation. Next, the revised Gustafson-Kessel Clustering method with the new Mahalanobis distance is integrated into the PCM which is presented in [1].

Moreover, a new *CVI* is applied which is based on proximity of two fuzzy sets. The new CVI is independent from the distance type and therefore can be integrated to the method with Mahalanobis distance. This *CVI* is used in [3].

3.1 Revised GK with revised Mahalanobis

In statistics, Mahalanobis distance is a distance measure introduced by P. C. Mahalanobis in 1936. It is based on correlations between variables by which different patterns can be identified and analyzed. It is a useful way of determining similarity of an unknown sample set to a known one. It differs from Euclidean distance in that it takes into account the correlations of the data set and is scale-invariant, i.e. not dependent on the scale of measurements.

Mahalanobis distance can be defined as dissimilarity measure between two random vectors $\vec{x}, \vec{y}$ of the same distribution with the covariance matrix S:

$$D_M(\vec{x},\vec{\mu}) = \sqrt{(\vec{x}-\vec{\mu})^T S^{-1}(\vec{x}-\vec{\mu})}$$

The eigenvalues and eigenvectors of the covariance matrix describe the shape and orientation of the clusters in GK method of clustering which applies mahalanobis distance instead of Euclidean one. When an eigenvalues is zero or when the ration between the maximal and minimal eigenvalues, i.e., the condition number of F, is very large the matrix is nearly singular. In such a case, the inverse of covariance matrix cannot be calculated. Also the normalization to a fixed volume fails, as the determinant (the volume of the covariance matrix) becomes zero and the following formula thus cannot be applied in GK method [4]:

$$\det(F_i)^{1/n} F_i^{-1} \quad (1)$$

A straightforward way to avoid numerical problems is to constrain the ratio between the maximal and minimal eigenvalues such that it is smaller than some predefined threshold. When this threshold exceeds, the minimal eigenvalues is increased such that the ratio equals to the threshold and the covariance is reconstructed by [4]:

$$F = \Phi \Lambda \Phi^{-1} \quad (2)$$

where, Λ is the diagonal matrix containing the limited eigenvalues and Φ is a matrix whose columns are the corresponding eigenvectors.

Fig. 1 shows an example of a data set which cannot be clustered with standard GK. Using new estimation for covariance matrix, the numerical problems would be resolved and GK can appropriately find the correct number of clusters (example from [4]).

The above modification prevents the GK algorithm from running into numerical problems. However, as a result one can get clusters that are extremely long in the direction of the largest eigenvalues and have little relationship with real distribution of data. This can cause over fitting of the data and consequently one obtains a poor model [4].

This problem occurs mainly when the number of data points in a cluster becomes too low. In such a case, the computed covariance matrix is not a reliable estimate of the underlying data distribution [7]. One way to tackle this problem is to

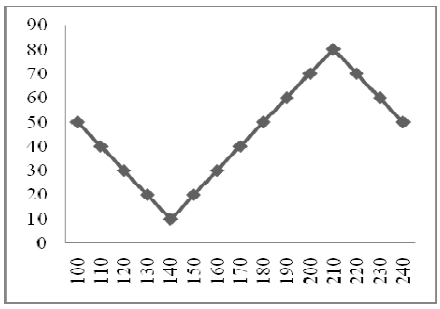

Figure 1: Linear Clusters

limit the ratio between maximal and minimal eigenvalues even further than described in the previous section. This will prevent the extreme elongation of the clusters. Another way is to add a scaled identity matrix to the covariance matrix. Reference [7] and [8] describe several different methods to improve the covariance estimation. Inspired by these methods, [4] proposes the following estimate for the GK algorithm and calculation of Mahalanobis distance:

$$F_i^{new} = (1-\gamma)F_i + \gamma \det(F_0)^{1/n} I \quad (3)$$

where $\gamma \epsilon [0,1]$ is the tuning parameter and F_0 is the covariance matrix of the whole data set. Depending on γ, the clusters are forced to have a more or less equal shape. When γ is 1, all covariance matrices are equal and have the same size, which of course limits the possibility of the algorithm to properly identify clusters.

For the complete description of the revised GK, readers can refer to [4]. The main steps of the algorithm are as follows [4]:

Repeat for *l = 1, 2, …, Ite#*

Step 1: Compute cluster prototypes (means)

$$v_i^{(l)} = \frac{\sum_{k=1}^{N} (\mu_{ik}^{(l-1)})^m Z_k}{\sum_{k=1}^{N} (\mu_{ik}^{(l-1)})^m}, 1 <= i <= K \quad (4)$$

Step 2: Compute the cluster covariance matrices

$$F_i = \frac{\sum_{k=1}^{N} (\mu_{ik}^{(l-1)})^m (Z_k - v_i^{(l)})(Z_k - v_i^{(l)})^T}{\sum_{k=1}^{N} (\mu_{ik}^{(l-1)})^m}, 1 \le i \le K \quad (5)$$

Add a scaled identity matrix:

$$F_i = (1-\gamma)F_i + \gamma \det(F_0)^{1/n} I, 1 \le i \le K \quad (6)$$

Extract eigenvalues λ_{ij} and eigenvectors ϕ_{ij} from F_i. Find $\lambda_{imax}=max_j\lambda_{ij}$ and set:

$$\lambda_{ij} = \frac{\lambda_{i\max}}{\beta} \; \forall j \; for \; which \; \frac{\lambda_{ij}}{\lambda_{i\max}} > \beta \quad (7)$$

Reconstruct F_i by

$$F_i = [\phi_{i1} \dots \phi_{in}] diag(\lambda_{i1}, \dots \lambda_{in}) [\phi_{i1} \dots \phi_{in}]^{-1}$$

Step 3: Compute the distances

$$D_{ikA_i}^2 = (Z_k - v_i^{(l)})^T [\rho_i \det(F_i)^{1/n} F_i^{-1}](Z_k - v_i^{(l)}) \quad (8)$$

where *1<=i<=k* and *1<=k<=N*.

Step 4: Update the partition matrix

$$for \; 1 \le k \le N$$

$$if \; D_{ikA_i} > 0 \; for \; 1 \le i \le K,$$

$$\mu_{ik}^{(l)} = \frac{1}{\sum_{j=1}^{K} (D_{ikA_j} / D_{ikA_i})^{\frac{2}{m-1}}} \quad (9)$$

Otherwise

$$\mu_{ik}^{(l)} = 0 \; if \; D_{ikA_i} > 0, \; and \; \mu_{ik}^{(l)} \in [0,1] \quad (10)$$

with $\sum_{i=1}^{K} \mu_{ik}^{(l)} = 1 \; otherwise.$

Until $\left\| U^{(l)} - U^{(l-1)} \right\| < \varepsilon$.

Improved GK clustering does not have the problems that standard GK may face in some cases. In the next section, applied cluster validity index is described.

3.2 Cluster Validity Index (CVI)

This paper uses a cluster validity index proposed by [9].

Definition1: The relative similarity $S_{rel}(x_j{:}A_p,A_q)$ between two fuzzy sets A_p and A_q at x_j is defined as:

$$\frac{f(x_j{:}A_p \cap A_q)}{f(x_j{:}A_p \cap A_q) + f(x_j{:}A_p - A_q) + f(x_j{:}A_q - A_p)} \quad (11)$$

while $f(x_j{:}A_p \cap A_q) = u_{A_p}(x_j) \wedge u_{A_q}(x_j)$ and $\wedge$ is the minimum operator. Moreover for function of deference:

$$f(x_j{:}A_p - A_q) = Max(0, u_{A_P}(x_j) - u_{A_q}(x_j)) \quad (12)$$

Definition2: The relative similarity between two fuzzy sets A_p and A_q is defined as:

$$S_{rel}(A_p,A_q) = \sum_{j=1}^{n} S_{rel}(x_j : A_p,A_q) h(x_j) \quad (13)$$

where

$$h(x_j) = -\sum_{p=1}^{c} u_{A_p}(x_j) \log(u_{A_p}(x_j)) \quad (14)$$

Here, h is the entropy of datum x_j and $u_{A_p}(x_j)$ is the membership value of x_j to cluster A_p.

Definition3: The cluster validity index is as follows:

$$V(U,V;X) = \frac{2}{c(c-1)} \sum_{p \ne q}^{c} S_{rel}(A_p,A_q) \quad (15)$$

The optimal number of clusters is obtained by minimizing V over the range of c values (Number of clusters).

Moreover, some other popular CVI's such as "Xie and Beni" CVI [5] and "Known" CVI [6] are evaluated. Results show that the new CVI has better results and is able to find the number of clusters correctly in many complex situations.

The inputs to the CVI used in the proposed model are obtained by the revised GK.

4 The proposed method

In this section, we present our method base on the modifications presented in section 3.

For the selection of the weight exponent (m), it is suggested to be chosen far from its both extremes so as to ensure that the cluster validity index shows the optimum number of fuzzy clusters. A fuzzy total scatter matrix is defined in [10] as:

$$S_T = \sum_{k=1}^{N} (\sum_{i=1}^{K} (u_{ik})^m)(x_k - \overline{v})(x_k - \overline{v})^T \quad (16)$$

The trace of the fuzzy total scatter matrix decreases monotonically from a constant value z to zero as m varies from one to infinity. For data partitioning, a suitable value for m is that which gives a value for trace (s_T) equal to z/2 [11]. The constant value z is defined as:

$$z = trace(\sum_{k=1}^{N} [(x_k - \frac{1}{N}\sum_{k=1}^{N} x_k)(x_k - \frac{1}{N}\sum_{k=1}^{N} x_k)^T]) \quad (17)$$

Using the trace value, we determine the value of (m) which is the degree of fuzziness of the system. Next, using the introduced *CVI* and achieved (m), the appropriate number of clusters would be obtained.

After determination of m and number of clusters, we repeat the following procedure, iteratively:

We first implement the revised GK on the data set. To choose the initial cluster centers of the revised GK, we apply simple FCM for initialization of our clustering method. Other methods like [1] have applied AHC. The GK is sensitive to the initial cluster centers and initial membership function values. Our experiments show that unlike FCM, AHC is not a good method for the initialization of the revised GK.

Next, in order to find the data points that are ''too far'' from all cluster centers, [13] proposes the following index for each data point x_j:

$$W_j = \sum_{i=1}^{C} \left\| x_j - v_{hi} \right\|_A \quad (18)$$

where, *j = 1, 2, ... , N*, c is the number of clusters, and N is the number of data. The index W_j is the summation of the distance of the data point x_j to all cluster centers. This gives a measure of how far each data point is from the different cluster centers assigned in the first step of the algorithm. The noise is identified through the data points that have large

ISBN: 978-989-95079-6-8

values of W_j and therefore, a threshold X is assigned to trim these outliers from the data set. The value of the threshold depends on the range of the input data to the algorithm. While [1] applies (18) to find the W_j, we proposed the following equation in order to find the W_j:

$$W_j = \underset{i=1}{\overset{C}{Mean}}(WGHT_i \times \left\| x_j - \upsilon_{hi} \right\|_A) \quad (19)$$

where, "$WGHT_i$" is a weight assigned to the *i-th* distance after that we sorted the distances of x_j from all cluster centers (c). Reference [1] does not consider such weight. It is clear that simply the summation of distances can not provide a good measure for detection of the outliers. Instead the proposed model can assign the largest weight to the smallest distance and the smallest weight to the largest distance vice versa using (19). Weights are the input to the model. After choosing the threshold, [1] computes:

$$z = \frac{\eta_n}{N} \quad (20)$$

where, η_n is the number of noise points and N is the total number of data. The percentage of ''good'' data points, i.e., inliers can then be calculated as:

$$\hat{z} = 1 - z \quad (21)$$

After identifying the percentage of inliers in the data, we compute the corresponding chi-square data distribution value [13]. Then, we calculate the cut-off distance:

$$u_{FC}cut^2 = \upsilon_i \chi^2 \quad (22)$$

where, υ_i is a resolution parameter that depends on the number of clusters, and χ^2 is the chi-square value computed by (21). By knowing the new cut-off distance, the optimum number of clusters, the degree of fuzziness, and the initial location of the clusters centers, we calculate the membership matrix through (23):

$$u_{ij} = \frac{1}{1 + \{\frac{d^2(x_j, \upsilon_i)}{\upsilon_i}\}^{\frac{1}{m-1}}} \quad (23)$$

It should be noted that (20), (21), (22) and (23) are presented in [1]. In the proposed PCM, we use these equations with revised Mahalanobis distance.

5 Results of Experiments

In this section, the proposed model is applied for handling four different cases. The parameters of the proposed model for these cases are summarized in table 1:

- *NOC*: Number of Clusters
- *m* is the degree of fuzziness of system obtained through 16 and 17.
- *Ω* is the cut-off distance
- *Beta and Gamma* are the parameters of (3) and (7).
- *#Itr*: The number of iteration which PCM goes on.
- *W: WGHT* in (19).
- *Error:* If the changes in the value of membership functions were smaller than "*Error*", the algorithm stops.

In all cases, *Beta* is 1.00E+16, *#Itr* is 7 and *Error* is 1.00E-07. It should be noted that the data of the cases and their related configuration are very similar to [9] and [3]. However, we regenerated the data by ourselves.

Table 1: Parameters of the algorithm for different cases

Parameter	***NOC***	***m***	***Ω***	***Gamma***	***W***
Case 1	4	3	210	1.00E-07	[6 2 2 1]
Case 2	4	2	375	1.00E-09	[8 3 2 2]
Case 3	5	2	315	1.00E-01	[8 3 2 2 1]
Case 4	5	3	4590	1.00E-07	[8 4 3 2 1]
Case 5	5	3	805	1.00E-01	[8 4 3 2 1]

It should be mentioned that the experiments can't be applied for the previous noise rejection methods in the literature. This is because of the ellipsoidal forms of data which can't be handled using Euclidean distance. To make this clear, in Fig. 2, the clustering methods based on Euclidean distance apparently fail to recognize the shapes of clusters and noises correctly. That's why, only experiments are applied for the proposed method.

5.1 Case Study I

Every data cloud includes 500 data and 50 data are randomly generated as noise. Main data and noises both have Gaussian distribution.

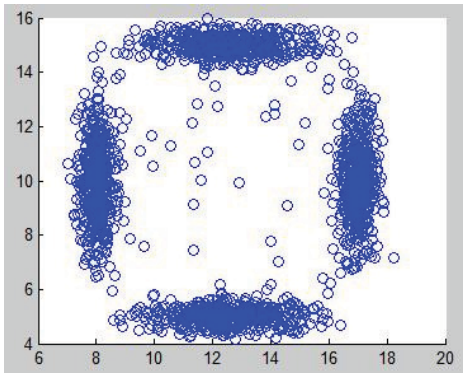

Figure 2: Data before clustering

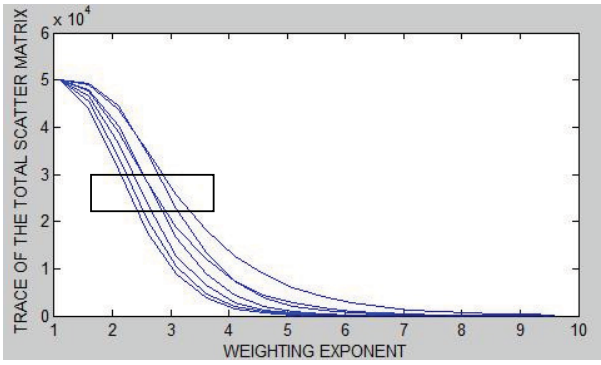

Figure 3: Finding the appropriate *m* in case study I

To determine the exponent of fuzziness (*m*), the total scatter matrix is calculated for different values of m and different

ISBN: 978-989-95079-6-8

number of clusters as is shown in (16). Then using z in (17), the appropriate m is chosen.
The trace value calculated as 5.8315e+004. We use half of trace value (2.9157e+004) to determine the optimum m (Fig. 3).

To choose the correct number of clusters, we need to calculate the CVI for different number of clusters. Fig. 4 shows the applied CVI for case study I. As the figure shows, the optimum number of clusters is equal to four.
Base on Fig. 3 and Fig. 4, we consider fuzziness exponent (m) equal to 2.5 and the number of clusters equal to 4. The result of the clustering is shown in Fig. 5. The result shows that the proposed method can identify the noises from the main data accurately. The expansion of the eclipse shape clusters can be limited by Ω as an input to the model.

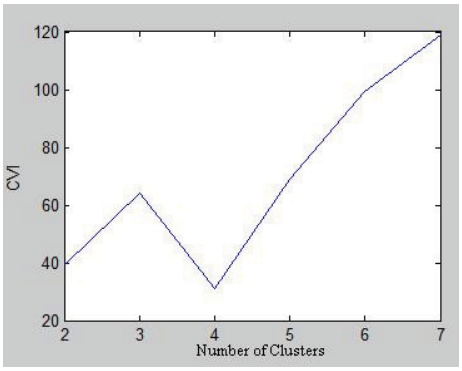

Figure 4: Identification of the optimum number of clusters

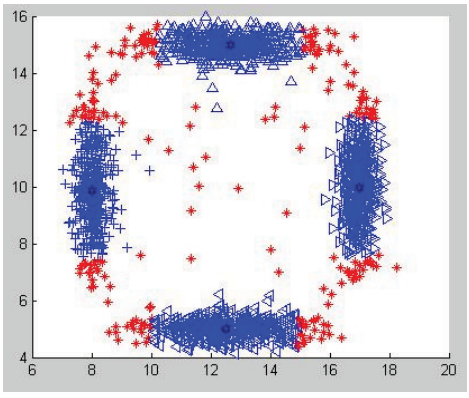
Figure 5: Data after Clustering

It is clear that, the method presented in [1] can not identify the correct shape of the clusters because of the spherical behaviour of the data. In Fig. 5, data which are shown with star "*", have a maximum membership function value lower than 0.2, hence are identified as noise.
In the other cases, each data cloud includes 500 data and 50 data are randomly generated as noise. Main data and noises both have Gaussian distribution. Also only the figures related to clustering and choosing the appropriate number of clusters are presented.

5.2 *Case Study II*

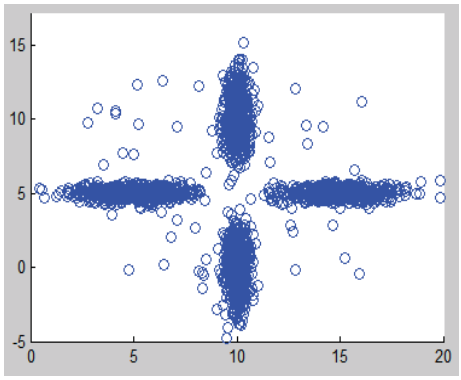
Figure 6: Data before clustering

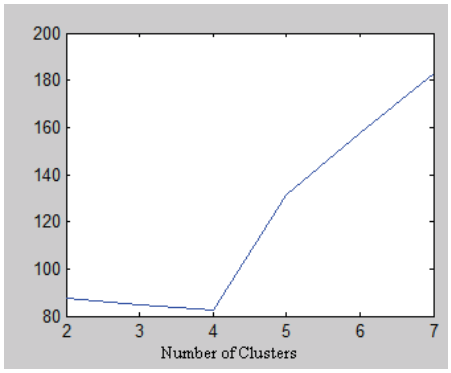

Figure 7: Identification of the optimum number of clusters

This case is presented to show the strength of the used CVI and the effect of Ω. As Figure 8 shows, the combination of these two elements in the proposed model, enable the model to identify the number of clusters, the shapes of the clusters and finally the noises accurately.

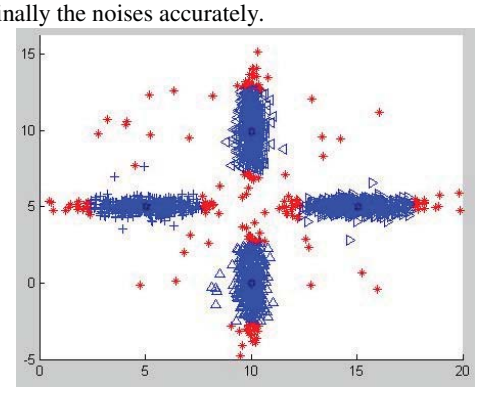
Figure 8: Data after Clustering

5.3 *Case Study III*

Fig 9 shows the data set and noises.

ISBN: 978-989-95079-6-8

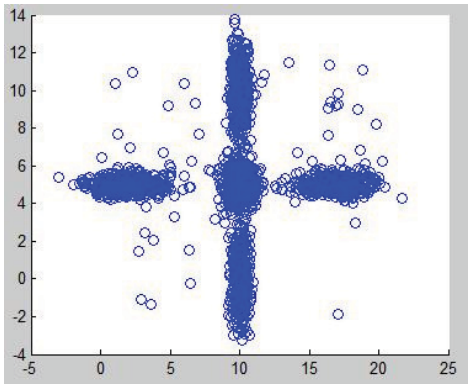

Figure 9: Data before clustering

As Fig. 10 shows, the optimum number of clusters is equal to five. This case is presented to validate that whether the method is able to find the spherical and ellipsoidal shapes simultaneously or is not. As Figure 11 shows, the method could find the circle at the center of the figure and four eclipses around the circle accurately. Also the noises are detected accurately as shown in Figure 11.

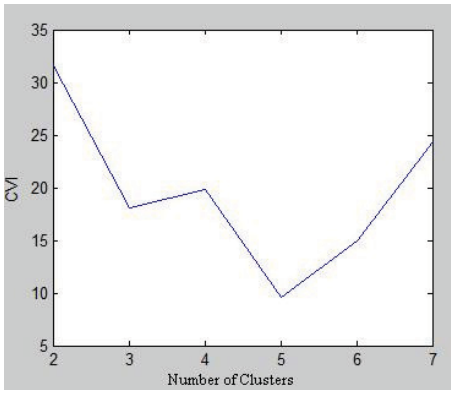

Figure 10: Identification of the optimum number of clusters

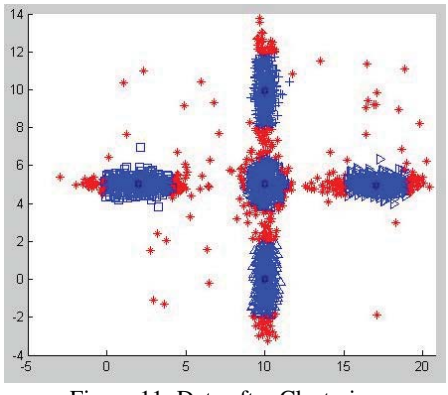

Figure 11: Data after Clustering

5.4 Case Study IV

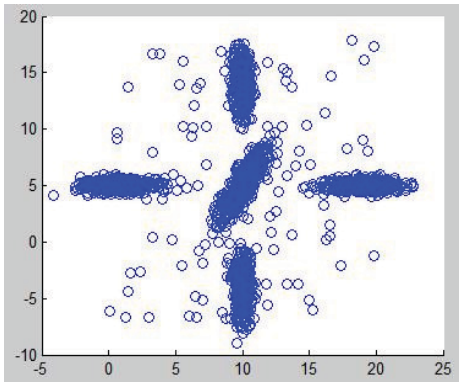

Figure 12: Data before clustering

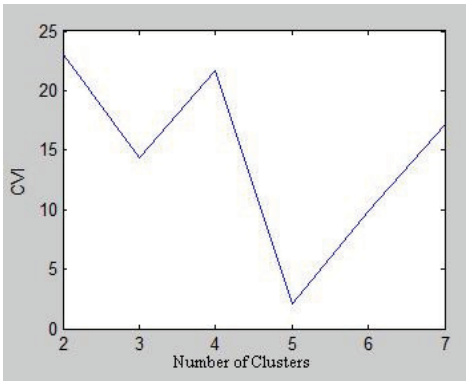

Figure 13: Identification of the optimum number of clusters

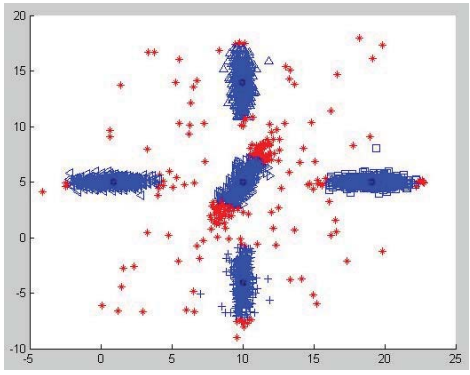

Figure 14: Data after Clustering

In this case, five eclipses are spread in different directions as shown in Figure 12. The clustering results show that the method can find the complex behaviors of data in different directions accurately as well as simple cases.

6 Conclusions

In this paper, a new noise rejection clustering method base on Mahalanobis distance and a different cluster validity index is presented. The new method can be distinguished from the existing methods in the literature from the following points:

ISBN: 978-989-95079-6-8

- The noise rejection clustering methods existing in the literature mainly use Euclidean distance in their clustering method. The proposed method, applies Mahalanobis distance which enables the method to identify the ellipsoidal behavior of data besides spherical behavior.
- In order to find the appropriate number of clusters, the method uses a well-defined cluster validity index which is independent from distance type.
- A new weighting system is attached to the noise rejection method, which helps for better detection of noises.

Method is applied to different cases and the results show that the method is capable of identification of noise and outliers within spherical and ellipsoidal data.

References

[1] W.W. Melek, A.A. Goldenberg, M.R. Emami, A fuzzy noise-rejection data partitioning algorithm, *International Journal of Approximate Reasoning*, 38: 1–17, 2005.

[2] R. Krishnapuram, J.M. Keller, A Possibilistic Approach to Clustering, *IEEE Transactions on Fuzzy Systems,* 2 (1): 98-110, 1993.

[3] M.H. Fazel Zarandi, B. Rezaee, I.B. Turksen, E. Neshat, A type-2 fuzzy rule-based expert system model for stock price analysis. *Expert Systems with Application*, 36:139-154, 2009.

[4] R. Babuska, R.J. van der Veen, U. Kaymak, Improved Covariance Estimation for Gustafson-Kessel Clustering. *IEEE*, 1081-1085, 2002.

[5] X.L. Xie, G. Beni, A validity measure for fuzzy clustering, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(8): 841–847, 1991.

[6] S.H. Kwon, Cluster validity index for fuzzy clustering. *Electronics Letters*, 34(22): 2176–2177, 1998.

[7] S. Tadjudin and D.A. Landgrebe. Covariance estimation with limited training samples. *IEEE Transactions on Geosciences and Remote Sensing*, 37(4): 2113-2118, July 1999.

[8] J.F. Friedman, Regularize discriminate analysis, *J.R. Statist. Soc.*, 84:17-42, 1989.

[9] Kim, Y.I., Kim, D.W., Lee, D., Lee, K.H., A cluster validation index for GK cluster analysis based on relative degree of sharing, *Information Science*, 168:225-242, 2004.

[10] M. Sugeno, T. Yasukawa, A fuzzy-logic-based approach to qualitative modeling, *IEEE Transactions on Fuzzy Systems,* 1(1), 1993.

[11] M.R. Emami, I.B. Turksen, A.A. Goldenberg, Development of a systematic methodology of fuzzy logic modeling, *IEEE Transactions on Fuzzy Systems,* 6 (3): 346–361, 1998.

[12] J.H. Ward, Hierarchical grouping to optimize an objective function, *Journal of American Statistics Association,* 58: 236–244, 1963.

[13] W.W. Melek, Neurofuzzy Control of Modular and Reconfigurable Robots, PhD. Dissertation, University of Toronto, Canada, 2002.

ISBN: 978-989-95079-6-8