Journal of Engineering and Applied Sciences 9 (10-12): 372-377, 2014
ISSN: 1816-949X

A New Fuzzy Clustering by Outliers

[1]Amina Dik, [1]Khalid Jebari, [1,2]Abdelaziz Bouroumi and [1]Aziz Ettouhami
[1]LCS Laboratory, Faculty of Sciences, Mohammed V-Agdal University, UM5A Rabat, Morocco
[2]LMI Laboratory, Ben M'sik Faculty of Sciences, University Hassan II Mohammedia (UH2M)
Casablanca, Morocco

Abstract: This study presents a new approach for partitioning data sets affected by outliers. The proposed scheme consists of two main stages. The first stage is a preprocessing technique that aims to detect data value to be outliers by introducing the notion of object's proximity degree. The second stage is a new procedure based on the Fuzzy C-Means (FCM) algorithm and the concept of outliers clusters. It consists to introduce clusters for outliers in addition to regular clusters. The proposed algorithm initializes their centers by the detected possible outliers. Final and accurate decision is made about these possible outliers during the process. The performance of this approach is also illustrated through real and artificial examples.

Key words: Similarity measure, outlier detection, FCM, proximity degree, illustrated

INTRODUCTION

The goal of data clustering is to find a structure in dataset (Jain, 2010). It aims to organize a set of objects into homogeneous clusters such as objects in the same cluster should be more similar to each other than are those belonging to different clusters (Bouroumi *et al*., 2000). Clustering has been widely applied in several different fields and various disciplines.

Several clustering algorithms are proposed in the literature. The most widely used clustering algorithm is Fuzzy C-Means (FCM) originally proposed by Bezdek (1981). FCM has been widely used and adapted (Krishnapuram and Keller, 1993; Bezdek *et al*., 1999; Hathaway and Bezdek, 2001). However, FCM is sensitive to outliers. They lead FCM to have difficulties in extracting the clusters correctly (Jain, 2010; Jolion and Rosenfeld, 1989).

Several methods have been proposed to detect outliers (Dave and Sen, 1997; Dave and Krishnapuram, 1997) a new concept of noise cluster was introduced (Dave, 1991; Ohashi, 1984). Unfortunately, these methods require some parameters that are not trivial to estimate.

This study presents an approach of identifying possible outliers and partitioning data sets containing outliers by an adapted FCM algorithm. The proposed approach deals with the outliers problem by introducing two concepts: object's degree of proximity and outliers clusters. The first reflects the closeness of an object to other considered objects. The second signifies that instead of considering a single noise cluster containing all outliers as proposed in noise clustering, each outlier is considered as center to an outlier cluster. The proposed approach offers the possibility to remove or not such points and the adapted FCM algorithm called Possible Outliers FCM (POFCM) allows reducing the influence of outliers on the regular clusters.

Related work: Several clustering algorithms are proposed in the literature. The most widely used clustering algorithm is FCM originally proposed by Bezdek (1981). Based on fuzzy set theory, this algorithm allows each point to have a degree of belonging to all clusters instead of belonging to one cluster. It partitions the considered dataset $X = \{x_1, x_2,\ldots, x_n\}$ d U^p where $x_i 0 U^p$ represents an object and x_{ij} its jth feature. Similar, objects are in the same cluster and dissimilar objects belong to different clusters. FCM optimizes an objective function J_m defined by:

$$J_m(U,V;X) = \sum_{k=1}^{n}\sum_{i=1}^{c}(u_{ik})^m d^2(x_k, v_i) \quad (1)$$

Where:

- m (1<m<4) = Weighting exponent used to control the relative contribution of each object vector x_i and the fuzziness degree of the final partition
- u_{ik} = Degree to which the object x_k belongs to the ith cluster (1#i#c and 1#k#n)
- V (v_1, v_2,…, v_c) = c-tuple of prototypes, each prototype characterizes one of the c clusters
- d (x_k,v_i) = Distance between the ith prototype and the kth object

Corresponding Author: Amina Dik, LCS Laboratory, Faculty of Sciences, Mohammed V-Agdal University, UM5A Rabat, Morocco

Bezdek proved that FCM converges to an approximate solution under two conditions (Bouroum *et al*., 2000): the pseudo-code of FCM algorithm is given in Algorithm A (Bouroumi *et al*., 2000).

$$u_{ik} = \left[\sum_{j=1}^{c} \left(\frac{d(x_k, v_i)}{d(x_k, v_j)} \right)^{2/m-1} \right]^{-1} ; 1 \le i \le c; 1 \le k \le n \quad (2)$$

$$v_i = \frac{\sum_{k=1}^{n} (u_{ik})^m x_k}{\sum_{k=1}^{n} (u_{ik})^m}; 1 \le i \le c \quad (3)$$

Algorithm A; FCM algorithm:

Store unlabeled Dataset X={x_1, x_2, …, x_n} d U^p;
Choose
    1<c<n; m>1; t_{max} (iteration limit); the , (tolerance bound);
    norm for clustering criterion J_m;
    norm for termination error E_t=‖V_t-V_{t-1}‖$_{err}$;
Initialize
    prototypes V_0 = ($v_{1,0}$, $v_{2,0}$, …,$v_{c,0}$) 0 U^{cxp}
    t = 0; (iteration index)
do { t++;
    Calculate U_t using V_{t-1} and (Eq. 2);
    Calculate V_t using U_t and (Eq. 3);
} while (‖V_t-V_{t-1}‖$_{err}$> ,) and (t<t_{max}));
U* = U_t; V* = V_t;
Use U* and/or V*;

FCM optimizes the function J_m that depends on the distances of the objects to the cluster centers weighted by the membership degrees. Thus, an outlier influences on the estimates of the cluster means (Jolion and Rosenfeld, 1989) and cluster centers can be placed away from the real centers. FCM is not robust against outliers.

An outlier is an item considerably dissimilar from the remainder of the data (Han and Kamber, 2006). Generally, outliers are far away from all the other items without neighbors. Recently, some approaches have been proposed on outlier detection (Knorr and Ng, 1998; Ramaswamy *et al*., 2000) and the outliers themselves become the focus in outlier mining tasks (Tang *et al*., 2012).

These approaches can be classified into distribution-based and proximity-based approaches. Distribution-based approaches where outliers are defined based on the probability distribution (Hawkins, 1980; Barnett and Lewis, 1994), develop statistical models. Items that have low probability to belong to the statistical model are declared as outliers (Al-Zoubi *et al*., 2008).

In proximity-based methods, outlier is an isolated point that is far away from the remaining data. This modeling contains specifically three methods: distance-based, clustering-based or density-based approaches.

Distance-based approaches consider a point x as an outlier if there are no more than M points in the dataset at a distance d from x (Knorr and Ng, 1998). However as the values of M and d are decided by the user, it is difficult to determine their values (Knorr and Ng, 1998; Ramaswamy *et al*., 2000). To overcome this limit, another algorithm was proposed (Angiulli *et al*., 2006). This algorithm computes outlier factor of each point as the sum of distances from its k nearest neighbors.

Others approaches are based on density. Density-based approaches compute the region's density in the data and consider items in low dense regions as outliers (Breunig *et al*., 2000). They assign an outlying degree to each data point. This degree represents how much this data point is an outlier.

Clustering-based approaches use the size of the resulting clusters as indicators of the presence of outliers. These approaches argue that outliers form small clusters whereas normal objects belong to dense clusters (Loureiro *et al*., 2004).

Solving both clustering and outlier detection is highly desired. Some fuzzy clustering algorithms have been proposed to partition data sets containing outliers. The most known algorithm is robust fuzzy C-Means (robust-FCM) (Dave, 1991). In this algorithm, the notion of noise cluster is introduced. This noise cluster is characterized by a fictitious prototype that has a constant distance * from all data points. Hence, the importance of the distance * that is a critical parameter of the algorithm (Cimino *et al*., 2007).

In the following, we propose an intuitive pre-processing approach that determines the possible outliers without requiring a preliminary knowledge of the data. The proposed preprocessing approach is a hybrid approach between distance-based and density-based approaches. An adapted FCM is also proposed to partition dataset containing outliers.

MATERIALS AND METHODS

The proposed preprocessing technique to detect outliers is intuitively and based on the notion of proximity degree. This notion reflects the closeness of an object to other considered objects. Here, the point's closeness is determined by the sum of its similarity to each other object. This degree of proximity can be considered as an opposite of isolation degree or outlier factor that characterizes outliers. However, instead of assigning an outlier factor to any object depending on its distance from its local neighborhood (Breunig *et al*., 2000), the proposed degree of proximity depends on all the data, since the local neighborhood is not determinate in preprocessing phase.

The key idea is that a normal object has more neighbors with which it has similar characteristics. Therefore, the object has a high degree of proximity when its neighbors are several and more this degree is high the object is not an outlier.

The proposed preprocessing technique does not require any notion of clusters. It just indicates if the object is more likely to be outlier.

To detect outliers among these objects, researchers determine the proximity degree for each object by using the following formula:

$$D(x_i) = \left(\sum_{\substack{j=0 \\ j \neq 1}}^{n} sim(x_i, x_j) \right) \quad (4)$$

Where:

$$Sim(x_i, x_k) = 1 - \frac{\|x_i - x_k\|_A^2}{p} \quad (5)$$

Sim (x_i,x_k) = Similarity between the objects x_i and x_k

A = Positive definite pxp matrix defined by (Bouroumi *et al*., 2000)

$$A_{jt} = \begin{cases} (r_j)^{-2}, j = t \\ 0, otherwise \end{cases} \quad (6)$$

The factor r_j represents the difference between the upper and the lower limits of the attribute's values. It is defined by:

$$r_j = \max_{1 \le i \le n} \{x_{ij}\} - \min_{1 \le i \le n} \{x_{ij}\}, 1 \le j \le p \quad (7)$$

Denote D^1_{min}, D^2_{min}, D^3_{min} and D^4_{min} the 4 less measure of proximity degree and $D_{range} = \max_{1 \le i \le n}(D(x_i)) - \min_{1 \le i \le n}(D(x_i))$ the difference between the upper and the lower degree of proximity. Researchers compute the follows values:

$$D^1_{min} / D_{range}, D^2_{min} / D_{range}, D^3_{min} / D_{range} \text{ and } D^4_{min} / D_{range}$$

If the value of D^1_{min}/D_{range} is lower than the others values, the vector corresponding to D^1_{min} is a possible outlier. Otherwise no outliers are in the dataset.

The approach does not require the minimal distance d that the user should define in the distance-based approaches. It does not require segmenting the space or the points. Moreover, it allows to determine the top M outliers, M chosen by the user within the (M+2) small proximity degree.

Some outliers should be removed from data sets when they result from error. However, some outliers contain important information and should be kept. The proposed algorithm allows deciding about this and two cases are considered: the possible outliers are removed from dataset and then FCM is executed normally.

The data set is clustered by using the M possible outliers as centers in addition to c random centers. Thereby, POFCM is executed with M+c centers initialized with the M outliers and others points. At the end of processing, we verify if outliers clusters contain more than the outliers. If so, N possible outliers (N = M) are not true outliers and POFCM is executed again with c+M–N centers. The pseudo-code of the POFCM is given in algorithm:

Algorithm; proposed POFCM algorithm:

Store unlabeled Dataset $X = \{x_1, x_2, \ldots, x_n\}$ d U^p;
Step 1: Determine the possible M outliers y_i by using the proposed preprocessing technique.
Step 2: Choose
$1<c<n$; $m>1$; t_{max} (iteration limit); the , _(tolerance bound);
norm for clustering criterion J_m;
norm for termination error $E_t=\|V_t-V_{t-1}\|_{err}$;

Step 3: do {
$M_{init} = M$;
Initialize prototypes $V_0 = (y_1, y_2, \ldots y_M, v_{1,0}, v_{2,0}, \ldots, v_{c,0})$ 0 $U^{(c+M)xp}$
do {Calculate U_t using V_{t-1} and (Eq.2);
Calculate V_t using U_t and (Eq.3);
} while ($\|V_t-V_{t-1}\|_{err}>$,) and ($t<t_{max}$));
For each outlier cluster C_i {
if (card $(C_i)>1$) *M*--;} //the possible outlier is not a true outlier
} while ($M != M_{init}$)
Step 4: $U^* = U_t$; $V^* = V_t$;

RESULTS AND DISCUSSION

To evaluate the performance of the method, experiments are conducted on an artificial dataset X1 (Bouroumi *et al*., 2000) and four real-world datasets available from the UCI Machine Learning Repository (Blake *et al*., 1998): Wine, breast cancer, spect heart and breast tissue (Table 1).

The artificial data set (X1) is an artificial example derived from (Bouroumi *et al*., 2000). It contains two well separated clusters in the plane and two outliers.

Wine dataset is a result of a chemical analysis of wines from three different cultivars. There are 13 attributes and 178 samples from three classes corresponding to three different cultivars with respectively 59, 79 and 48 samples per variety.

Table 1: Description of the studied datasets

Dataset	No. of samples	No. of attributes	No. of classes
X1	42	1	2
Wine	178	13	3
BCW	699	9	2
SPECT heart	267	22	2
Breast tissue	106	9	6

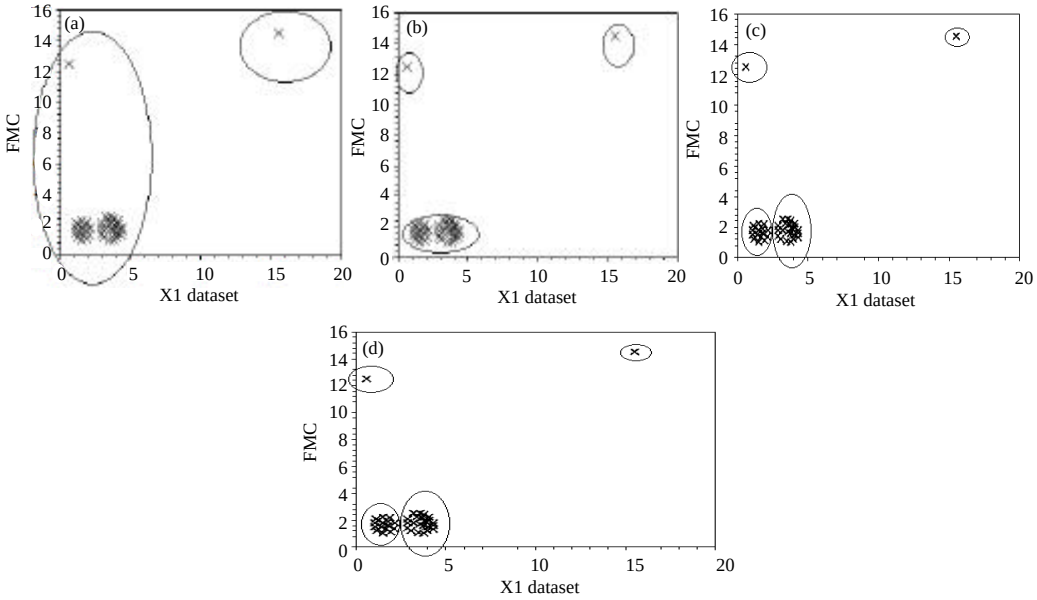

Fig. 1: Representation of results on the X1 dataset for FCM with: a) c = 2; b) c = 3; c) c = 4; d) POFCM with c = 2, M = 2

Table 2: Outlier detection

Dataset	D^1_{min}	D^2_{min}	D^3_{min}	D^4_{min}	D_{range}	D^1_{mi}/D_{range}	D^2_{mi}/D_{range}	D^3_{mi}/D_{range}	D^4_{mi}/D_{range}
X1	4.19	17.17	35 .96	35.99	33.23	0.12	0.51	1.08	0.92
BCW	165.73	166.20	172.03	174.66	352.07	0.47	0.472	0.48	0.49
Wine	109.67	113.80	114.88	115.60	27.90	3.93	4.07	4.11	4.14
Heart	51.52	53.12	56.15	57.04	79.40	0.65	0.67	0.71	0.72
Breast tissue	27.00	54.91	55.99	57.53	60.04	0.44	0.91	0.93	0.95

Breast cancer dataset is a 9-dimensional pattern classification problem with 699 samples from malignant (cancerous) class and benign (non-cancerous) class. The two classes contain, respectively 458 and 241 points.

The third dataset describes diagnosing of cardiac Single Proton Emission Computed Tomography (SPECT) images. There are 22 attributes and 267 samples from two classes corresponding to normal and abnormal patients with 55 and 212, respectively samples per category.

The last example is breast tissue recognition dataset that is the result of a measure of breast tissue by electrical impedance spectroscopy. It is a 9-dimensional pattern classification problem with 106 samples from six classes. Table 1 describes the type of data and gives information about attributes, size and number of classes.

At first, researchers search if there are possible outliers in the considered dataset. For this, researchers compute proximity degree for the objects and search the four small values.

For X1 dataset, $D^1_{min}/D_{range} = 0.12$ and $D^2_{min}/D_{range} = 0.51$ whereas D^3_{min}/D_{range} and D^4_{min}/D_{range} have a higher values (1.08 and 0.92, respectively). For the case of wine dataset, $D^1_{min}/D_{range} = 3.93$ whereas D^2_{min}/D_{range}, D^3_{min}/D_{range} and D^4_{min}/D_{range} have a higher values (4.07, 4.11 and 4.14, respectively).

For the breast tissue dataset, $D^1_{min}/D_{range} = 0.44$ whereas D^2_{min}/D_{range}, D^3_{min}/D_{range} and D^4_{min}/D_{range} have almost the same value (0.91, 0.93 and 0.95, respectively).

The results in Table 2 show that there are possible outliers for X1, wine and breast tissue datasets. Once the possible outliers are determined for the dataset (Table 3), the algorithm POFCM is executed. The results of this algorithm are presented in Table 4.

Table 3: Indexes of possible outliers

Dataset	Index of object 1 (outlier)	Index of object 2	Index of object 3	Index of object 4
X1	0	1 (outlier)	10	11
Wine	121	158	146	59
Breast tissue	102	86	97	105

Table 4: Recognition rate for FCM with and without possible outliers

Dataset	c	M	Recognition rate with outliers (%)	Recognition rate without outliers
X1	2	2	61.91	100.00%
Wine	3	1	69.67	77.97%
Breast tissue	6	1	30.13	31.43

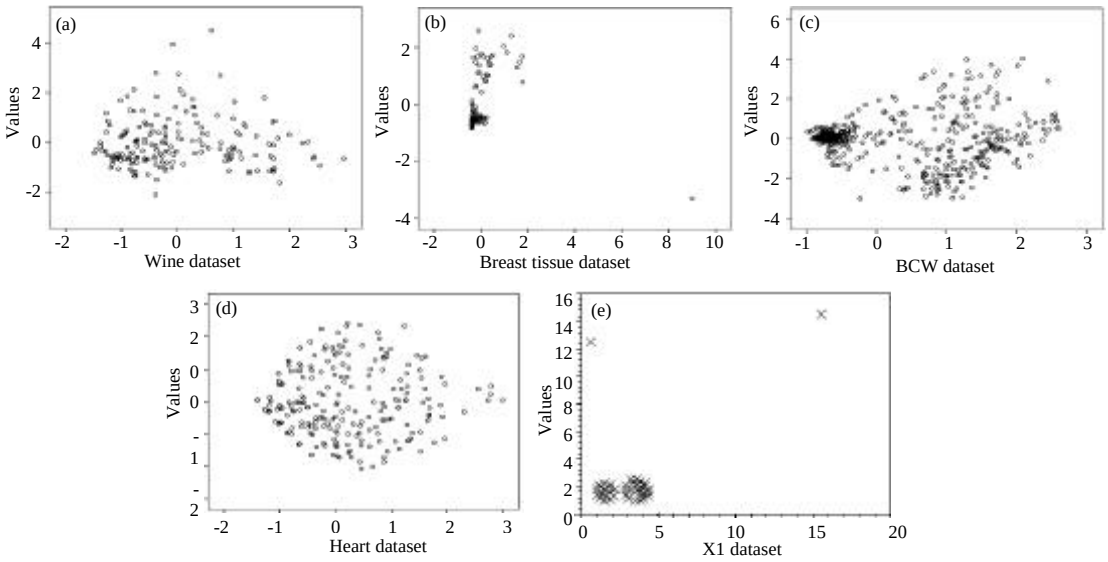

Fig. 2: Representation of the dataset in the plane

Table 5: Recognition rate for POFCM with outliers

Dataset	c	M	FCM (%)	Adapted FCM (%)
X1	2	2	61.91	100.00
Wine	3	1	69.67	69.67
Breast tissue	6	1	30.13	32.08

Table 5 shows that X1 and breast tissue were performed by the adapted FCM (Fig. 1). It means that they have true outliers. However, the possible outlier in the wine dataset is not a true outlier. These results are also confirmed by the representation of the dataset in the plane. Indeed, Fig. 2 shows that X1 and breast tissue contain outliers.

CONCLUSION

An efficient adapted method for outlier detection and clustering dataset is proposed in this study. The proposed method consists of two main stages. The first stage is an intuitive pre-processing method that identifies some M points which can be considered as possible outliers by using the concept of proximity degree.

In the second stage, the new algorithm POFCM based on FCM is executed with the M outliers as centers in addition to c random centers. POFCM verifies the presence of objects in outliers clusters others than the possible outliers and decided if so they are not true outliers. The experimental results show that the approach out performed FCM on clustering dataset that contain outliers.

REFERENCES

Al-Zoubi, M.D.B., A.D. Ali and A.A. Yahya, 2008. Fuzzy clustering-based approach for outlier detection. Proceedings of the 9th WSEAS International Conference on Applications of Computer Engineering, March 23-25, 2010, Penang, Malaysia, pp: 192-197.

Angiulli, F., S. Basta and C. Pizzuti, 2006. Distance-based detection and prediction of outliers. IEEE Trans. Knowl. Data Eng., 18: 145-160.

Barnett, V. and T. Lewis, 1994. Outliers in Statistical Data. 3rd Edn., John Wiley and Sons, New York, ISBN: 0-471-93094-6, Pages: 584.

Bezdek, J., J.Keller and R. Krishnapuram, 1999. Fuzzy Models and Algorithms for Pattern Recognition and Image Processing. 1st Edn., Kluwer Academy Publishers, Norwell, MA., USA., ISBN: 0792385217, pp: 792.

Bezdek, J.C., 1981. Pattern Recognition with Fuzzy Objective Function Algoritms. 1st Edn., Plenum Press, New York, USA.

Blake, C., E. Keogh and C. Merz, 1998. UCI Repository of Machine Learning Database. University of California, Irvine, CA., USA.

Bouroumi, A., M. Limouri and A. Essaid, 2000. Unsupervised fuzzy learning and cluster seeking. Intell. Data Anal., 4: 241-253.

Breunig, M.M., H.P. Kriegel, R.T. Ng and J. Sander, 2000. LOF: Identifying density-based local outliers. Proceedings of the International Conference on Management of Data, May 15-18, 2000, Dallas, TX., USA., pp: 93-104.

Cimino, M.G.C.A., G. Frosini, B. Lazzerini and F. Marcelloni, 2007. On the noise distance in robust fuzzy c-means. Int. J. Comput. Sci. Eng., Vol. 1.

Dave, R.N. and R. Krishnapuram, 1997. Robust clustering methods. A unified view. IEEE Trans. Fuzzy Syst., 5: 270-293.

Dave, R.N. and S. Sen, 1997. Noise clustering algorithm revisited. Proceedings of the Annual Meeting of the North American Fuzzy Information Processing Society, September 21-24, 1997, Syracuse, NY., USA., pp: 199-204.

Dave, R.N., 1991. Characterization and detection of noise in clustering. Pattern Recogn. Lett., 12: 657-664.

Han, J. and M. Kamber, 2006. Data Mining: Concepts and Techniques. 2nd Edn., Morgan Kaufmann Publisher, San Fransisco, USA., ISBN: 1-55860-901-6.

Hathaway, R.J. and J.C. Bezdek, 2001. Fuzzy c-means clustering of incomplete data. IEEE Trans. Syst. Man Cybernet. B: Cybernet., 31: 735-744.

Hawkins, D.M., 1980. Identifications of Outliers. Chapman and Hall, London, UK., ISBN-13: 9780412219009.

Jain, A.K., 2010. Data clustering: 50 years beyond K-means. Pattern Recogn. Lett., 31: 651-666.

Jolion, J.M. and A. Rosenfeld, 1989. Cluster detection in background noise. Pattern Recogn., 22: 603-607.

Knorr, E.M. and R.T. Ng, 1998. Algorithms for mining distance-based outliers in large dataset. Proceedings of the 24rd International Conference on Very Large Data Bases, August 24-27, 1998, San Francisco, CA., USA., pp: 392-403.

Krishnapuram, R. and J.M. Keller, 1993. A possibilistic approach to clustering. IEEE Trans. Fuzzy Syst., 1: 98-110.

Loureiro, A., L. Torgo and C. Soares, 2004. Outlier detection using clustering methods: A data cleaning application. Proceedings of KDNet Symposium on Knowledge-Based Systems for the Public Sector, June 3-4, 2004, Bonn, Germany.

Ohashi, Y., 1984. Fuzzy clustering and robust estimation. Proceedings of the 9th International SAS Users Group Meeting, March 18-21, 1984, Hollywood Beach, USA.

Ramaswamy, S., R. Rastogi and K. Shim, 2000. Efficient algorithms for mining outliers from large data sets. Proceedings of the International Conference on Management of Data, May 15-18, 2000, Dallas, TX., USA., pp: 427-438.

Tang, C., S. Wang and Y. Chen, 2012. Clustering of steel strip sectional profiles based on robust adaptive fuzzy clustering algorithm. Comput. Inform., 30: 357-380.