Int. J. Appl. Math. Comput. Sci., 2001, Vol.11, No.4, 993–1007

AN ε-INSENSITIVE APPROACH TO FUZZY CLUSTERING

JACEK ŁĘSKI*

Fuzzy clustering can be helpful in finding natural vague boundaries in data. The fuzzy c-means method is one of the most popular clustering methods based on minimization of a criterion function. However, one of the greatest disadvantages of this method is its sensitivity to the presence of noise and outliers in the data. The present paper introduces a new ε-insensitive Fuzzy C-Means (εFCM) clustering algorithm. As a special case, this algorithm includes the well-known Fuzzy C-Medians method (FCMED). The performance of the new clustering algorithm is experimentally compared with the Fuzzy C-Means (FCM) method using synthetic data with outliers and heavy-tailed, overlapped groups of the data.

Keywords: fuzzy clustering, fuzzy c-means, robust methods, ε-insensitivity, fuzzy c-medians

1. Introduction

Clustering plays an important role in many engineering fields such as pattern recognition, system modelling, image processing, communication, data mining, etc. The clustering methods divide a set of N observations (input vectors) $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$ into c groups denoted by $\Omega_1, \Omega_2, \ldots, \Omega_c$ in such a way that the members of the same group are more similar to one another than to the members of other groups. The number of clusters may be pre-defined or it may be set by the method.

Generally, clustering methods can be divided into (Duda and Hart, 1973; Fukunaga, 1990; Tou and Gonzalez, 1974) the following kinds: hierarchical, graph theoretic, decomposing a density function, minimizing a criterion function. In this paper, clustering by minimization of a criterion function will be considered. Usually, the clustering methods assume that each data vector belongs to one and only one class. This method can be natural for clustering of compact and well-separated groups of data. However, in practice, clusters overlap, and some data vectors belong partially to several clusters. Fuzzy set theory (Zadeh, 1965) is a natural way of describing this situation. In this case, a membership degree of the vector $\mathbf{x}_k$ to the i-th cluster (u_{ik}) is a value from the $[0, 1]$ interval. This idea was first introduced by Ruspini (1969) and used by Dunn (1973) to construct a fuzzy clustering method based on minimization of a

* Institute of Electronics, Silesian University of Technology, Akademicka 16, 44–100 Gliwice, Poland, e-mail: jl@boss.iele.polsl.gliwice.pl

criterion function. Bezdek (1982) generalized this approach to an infinite family of fuzzy c-means algorithms using a weighted exponent on the fuzzy memberships.

Fuzzy c-means clustering algorithm has been successfully applied to a wide variety of problems (Bezdek, 1982). However, one of the greatest disadvantages of this method is its sensitivity to noise and outliers in data. In this case, computed clusters centres can be placed away from the true values. In the literature, there are a number of approaches to reduce the effect of outliers, including the possibilistic clustering (Krishnapuram and Keller, 1993), fuzzy noise clustering (Davé, 1991), L_p norm clustering ($0 < p < 1$) (Hathaway and Bezdek, 2000), and L_1 norm clustering (Jajuga, 1991; Kersten, 1999). In this paper, the last approach is of special interest.

In real applications, the data are corrupted by noise and outliers, and assumed simplified models are only approximators to the reality. For example, if we assume that the distribution of the data in clusters is Gaussian, then using the weighted (by a membership degree) mean should not cause a bias. In this case, the L_2 norm is used as a dissimilarity measure.

The noise and outliers existing in data imply that the clustering methods need to be robust. According to Huber (1981), a robust method should possess the following properties: (i) it should have a reasonably good accuracy at the assumed model, (ii) small deviations arising from the model assumptions should impair the performance only by a small amount, (iii) larger deviations arising from the model assumptions should not cause a catastrophe. In the literature, there are many robust estimators (Davé and Krishnapuram, 1997; Huber, 1981). In this paper, Vapnik's ε-insensitive estimator is of special interest (Vapnik, 1998).

The goal of the present paper is to establish a connection between fuzzy c-means clustering and robust statistics using Vapnik's ε-insensitive estimator. The paper presents a new fuzzy clustering method based on a robust approach. The fuzzy c-medians can be obtained as a special case of the introduced clustering method.

The paper is organized as follows. Section 2 presents a short description of clustering methods based on minimization of a criterion function. A novel clustering algorithm is described in Section 3. Section 4 presents simulation results of clustering for synthetic data with outliers and heavy-tailed groups of data, as well as a comparative study with the fuzzy c-means method. Finally, conclusions are drawn in Section 5.

2. Clustering by Minimization of Criterion Function

A very popular way of data clustering is to define a criterion function (a scalar index) that measures the quality of the partition. In the fuzzy approach (Bezdek, 1982), the set of all possible fuzzy partitions of N p-dimensional vectors into c clusters is defined by

$$\Im_{fc} = \left\{ \mathbf{U} \in \Re_{cN} \;\middle|\; \underset{\substack{1\le i\le c\\ 1\le k\le N}}{\forall} \; u_{ik} \in [0,1],\ \sum_{i=1}^{c} u_{ik} = 1,\ 0 < \sum_{k=1}^{N} u_{ik} < N \right\}. \quad (1)$$

$\Re_{cN}$ denotes a space of all real $(c \times N)$-dimensional matrices. The fuzzy c-means criterion function has the form (Bezdek, 1982)

$$J_m(\mathbf{U}, \mathbf{V}) = \sum_{i=1}^{c} \sum_{k=1}^{N} (u_{ik})^m \, d_{ik}^2, \tag{2}$$

where $\mathbf{U} \in \Im_{fc}$, $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_c] \in \Re_{pc}$ is a prototypes matrix and m is a weighting exponent in $[1, \infty)$. Furthermore, d_{ik} is the inner product induced norm

$$d_{ik}^2 = \|\mathbf{x}_k - \mathbf{v}_i\|_{\mathbf{A}}^2 = (\mathbf{x}_k - \mathbf{v}_i)^T \, \mathbf{A} \, (\mathbf{x}_k - \mathbf{v}_i), \tag{3}$$

where $\mathbf{A}$ is a positive definite matrix. Criterion (2) for $m = 2$ was introduced by Dunn (1973). An infinite family of fuzzy c-means criteria for $m \in [1, \infty)$ was introduced by Bezdek. Using the Lagrange multipliers technique, the following theorem can be proved, via obtaining necessary conditions for minimization of (2) (Bezdek, 1982):

Theorem 1. *If m and c are fixed parameters, and I_k is the set defined as*

$$\underset{1 \le k \le N}{\forall} \quad I_k = \big\{ i \mid 1 \le i \le c;\ d_{ik} = 0 \big\}, \tag{4}$$

then $(\mathbf{U}, \mathbf{V}) \in (\Im_{fc} \times \Re_{pc})$ may be globally minimal for $J_m(\mathbf{U}, \mathbf{V})$ only if

$$\underset{\substack{1 \le i \le c \\ 1 \le k \le N}}{\forall} \quad u_{ik} = \begin{cases} (d_{ik})^{\frac{2}{1-m}} \Big[\sum\limits_{j=1}^{c} (d_{jk})^{\frac{2}{1-m}} \Big]^{-1}, & I_k = \emptyset, \\ \begin{cases} 0, & i \notin I_k, \\ \sum\limits_{i \in I_k} u_{ik} = 1, & i \in I_k, \end{cases} & I_k \neq \emptyset, \end{cases} \tag{5}$$

and

$$\underset{1 \le i \le c}{\forall} \quad \mathbf{v}_i = \frac{\sum\limits_{k=1}^{N} (u_{ik})^m \, \mathbf{x}_k}{\sum\limits_{k=1}^{N} (u_{ik})^m}. \tag{6}$$

The optimal partition is determined as a fixed point of (5) and (6) (Pal and Bezdek, 1995):

1° Fix c $(1 < c < N)$, $m \in (1, \infty)$. Initialize $\mathbf{V}^{(0)} \in \Re_{pc}$, $j = 1$,

2° Calculate the fuzzy partition matrix $\mathbf{U}^{(j)}$ for the j-th iteration using (5),

3° Update the centres for the j-th iteration $\mathbf{V}^{(j)} = [\mathbf{v}_1^{(j)} \mathbf{v}_2^{(j)} \cdots \mathbf{v}_c^{(j)}]$ using (6) and by $\mathbf{U}^{(j)}$,

4° If $\big\|\mathbf{U}^{(j+1)} - \mathbf{U}^{(j)}\big\|_F > \xi$, then $j \leftarrow j + 1$, and go to 2°.

Here $\|\cdot\|_F$ denotes the Frobenius norm ($\|\mathbf{U}\|_F^2 = \mathrm{Tr}\,(\mathbf{U}\mathbf{U}^T) = \sum_i \sum_k u_{ik}^2$), and ξ is a pre-set parameter. The above algorithm is called the fuzzy ISODATA or fuzzy c-means. Note that the original FCM algorithm proposed in (Bezdek, 1982) does not use the Frobenius norm, but rather a convenient matrix norm (usually the maximum norm). The FCM algorithm is often called the alternating optimization one as it loops through a cycle of estimates for $\mathbf{V}^{(j-1)} \to \mathbf{U}^{(j)} \to \mathbf{V}^{(j)}$. Equivalently, the procedure can be shifted by one-half cycle, so that initialization is done on $\mathbf{U}^{(0)}$, and a cycle of estimates is $\mathbf{U}^{(j-1)} \to \mathbf{V}^{(j)} \to \mathbf{U}^{(j)}$. The convergence theory is the same in both the cases (Pal and Bezdek, 1995). In the FCM algorithm, the parameter m influences the fuzziness of the clusters; a larger m results in fuzzier clusters. For $m \longrightarrow 1^+$, the fuzzy c-means solution becomes a hard one, and for $m \longrightarrow \infty$ the solution is as fuzzy as possible: $u_{ik} = 1/c$, for all i, k. Because there is no theoretical basis for the optimal selection of m, usually $m = 2$ is chosen.

3. A New Clustering Algorithm

The clustering algorithm described in the previous section uses a quadratic loss function as a dissimilarity measure between the data and the cluster centres. The reason for using this measure is mathematical, i.e. it is employed owing to the simplicity and low computational burden. However, this approach is sensitive to noise and outliers. In the literature, there are many robust loss functions (Huber, 1981), but due to its simplicity, Vapnik's ε-insensitive loss function (Vapnik, 1998) is of special interest. If we denote an error by t, then the ε-insensitive loss function has the form

$$|t|_\varepsilon = \begin{cases} 0, & |t| \le \varepsilon, \\ |t| - \varepsilon, & |t| > \varepsilon, \end{cases} \tag{7}$$

where ε stands for the insensitivity parameter. The well-known absolute error loss function is a special case of (7) for $\varepsilon = 0$.

Using the ε-insensitive loss function, the fuzzy c-means criterion function (2) takes the form

$$J_{m\varepsilon}(\mathbf{U}, \mathbf{V}) = \sum_{i=1}^{c} \sum_{k=1}^{N} (u_{ik})^m \; |\mathbf{x}_k - \mathbf{v}_i|_\varepsilon \,, \tag{8}$$

where

$$|\mathbf{x}_k - \mathbf{v}_i|_\varepsilon = \sum_{l=1}^{p} |x_{kl} - v_{il}|_\varepsilon \,. \tag{9}$$

Write λ^+ or λ^- as $\lambda^{(\pm)}$, and the cardinality of a set A as $\mathrm{card}(A)$.

Theorem 2. *If m, c and ε are fixed parameters, then $(\mathbf{U},\mathbf{V}) \in (\Im_{fc} \times \Re_{pc})$ may be globally minimal for $J_{m\varepsilon}(\mathbf{U},\mathbf{V})$ only if*

$$\underset{\substack{1\le i\le c\\ 1\le k\le N}}{\forall} \quad u_{ik} = \begin{cases} \left(|\mathbf{x}_k - \mathbf{v}_i|_\varepsilon\right)^{\frac{1}{1-m}} \left[\sum_{j=1}^{c} \left(|\mathbf{x}_k - \mathbf{v}_j|_\varepsilon\right)^{\frac{1}{1-m}}\right]^{-1}, & I_k = \emptyset, \\ \begin{cases} 0, & i \notin I_k, \\ \sum_{i\in I_k} u_{ik} = 1, & i \in I_k, \end{cases} & I_k \neq \emptyset, \end{cases} \tag{10}$$

and

$$\underset{1\le i\le c}{\forall} \quad \underset{1\le l\le p}{\forall} \quad v_{il} = \begin{cases} (x_{kl} + \varepsilon), & \underset{\left\{k \mid \lambda_k^+ \in \Lambda_i^+\right\}}{\forall}, \\ (x_{kl} - \varepsilon), & \underset{\left\{k \mid \lambda_k^- \in \Lambda_i^-\right\}}{\forall}, \end{cases} \tag{11}$$

where

$$\Lambda_i^{(\pm)} = \Big\{ \lambda_k^{(\pm)} \in \left(0, (u_{ik})^m\right) \Big| \min_{\{\lambda_k^+\},\{\lambda_k^-\}} \sum_{k=1}^{N} \left(\lambda_k^+ - \lambda_k^-\right) x_{kl} + \varepsilon \sum_{k=1}^{N} \left(\lambda_k^+ + \lambda_k^-\right),$$

$$\textit{subject to} \quad \sum_{k=1}^{N} \lambda_k^+ = \sum_{k=1}^{N} \lambda_k^- \quad \textit{and} \quad \lambda_k^+, \lambda_k^- \in [0, (u_{ik})^m] \Big\}, \tag{12}$$

and the set I_k is defined as $I_k = \{i \mid 1 \le i \le c;\ |\mathbf{x}_k - \mathbf{v}_i|_\varepsilon = 0\}$, $k = 1, 2, \ldots, N$.

Proof. If $\mathbf{V} \in \Re_{cp}$ is fixed, then the columns of $\mathbf{U}$ are independent, and the minimization of (8) can be performed term by term:

$$J_{m\varepsilon}(\mathbf{U},\mathbf{V}) = \sum_{k=1}^{N} g_k(\mathbf{U}), \tag{13}$$

where

$$\underset{1\le k\le N}{\forall} \quad g_k(\mathbf{U}) = \sum_{i=1}^{c} (u_{ik})^m \ |\mathbf{x}_k - \mathbf{v}_i|_\varepsilon . \tag{14}$$

The Lagrangian of (14) with constraints from (1) is as follows:

$$\underset{1\le k\le N}{\forall} \quad G_k(\mathbf{U},\lambda) = \sum_{i=1}^{c} (u_{ik})^m \ |\mathbf{x}_k - \mathbf{v}_i|_\varepsilon - \lambda \left[\sum_{i=1}^{c} u_{ik} - 1\right], \tag{15}$$

where λ is the Lagrange multiplier. Setting the Lagrangian's gradient to zero, we obtain

$$\underset{1\le k\le N}{\forall} \quad \frac{\partial G_k(\mathbf{U},\lambda)}{\partial \lambda} = \sum_{i=1}^{c} u_{ik} - 1 = 0 \tag{16}$$

and

$$\underset{\substack{1\leq s\leq c\\ 1\leq k\leq N}}{\forall} \quad \frac{\partial G_k\left(\mathbf{U},\lambda\right)}{\partial u_{sk}} = m\left(u_{sk}\right)^{m-1}\left|\mathbf{x}_k-\mathbf{v}_s\right|_\varepsilon - \lambda = 0. \tag{17}$$

From (17) we get

$$u_{sk} = \left(\frac{\lambda}{m}\right)^{\frac{1}{m-1}}\left(\left|\mathbf{x}_k-\mathbf{v}_s\right|_\varepsilon\right)^{\frac{1}{1-m}}. \tag{18}$$

From (18) and (16) it follows that

$$\left(\frac{\lambda}{m}\right)^{\frac{1}{m-1}}\sum_{j=1}^{c}\left(\left|\mathbf{x}_k-\mathbf{v}_j\right|_\varepsilon\right)^{\frac{1}{1-m}} = 1. \tag{19}$$

Combining (18) and (19) yields

$$\underset{\substack{1\leq s\leq c\\ 1\leq k\leq N}}{\forall} \quad u_{sk} = \frac{\left(\left|\mathbf{x}_k-\mathbf{v}_s\right|_\varepsilon\right)^{\frac{1}{1-m}}}{\sum\limits_{j=1}^{c}\left(\left|\mathbf{x}_k-\mathbf{v}_j\right|_\varepsilon\right)^{\frac{1}{1-m}}}. \tag{20}$$

If $I_k \neq \emptyset$, then the choice of $u_{ik}=0$ for $i \notin I_k$ and $\sum_{i\in I_k} u_{ik}=1$ for $i \in I_k$ results in minimization of the criterion (8), because the elements of the partition matrix are zero for non-zero distances, and non-zero for zero distances. Finally, the necessary conditions for minimization of (8) with respect to $\mathbf{U}$ can be written as in (10).

A more difficult problem is to obtain necessary conditions for the prototype matrix $\mathbf{V}$. Combining (8) and (9) yields

$$J_{m\varepsilon}\left(\mathbf{U},\mathbf{V}\right) = \sum_{i=1}^{c}\sum_{k=1}^{N}\left(u_{ik}\right)^m \sum_{l=1}^{p}\left|x_{kl}-v_{il}\right|_\epsilon = \sum_{i=1}^{c}\sum_{l=1}^{p} g_{il}\left(v_{il}\right), \tag{21}$$

where

$$g_{il}\left(v_{il}\right) = \sum_{k=1}^{N}\left(u_{ik}\right)^m \left|x_{kl}-v_{il}\right|_\varepsilon \tag{22}$$

can be called the weighted ε-insensitive (or fuzzy ε-insensitive) estimator. For $\varepsilon=0$ we obtain the fuzzy median estimator (Kersten, 1999).

Our problem of minimizing the criterion (8) with respect to the prototypes can be decomposed to $c\cdot p$ minimization problems of (22) for $i=1,\ldots,c$ and $l=1,\ldots,p$. In a general case, the inequalities $x_{kl}-v_{il}\leq\varepsilon$ and $v_{il}-x_{kl}\leq\varepsilon$ are not satisfied for all data. If we introduce slack variables $\xi_k^+,\xi_k^-\geq 0$, then for all the data x_{kl} we can write

$$\begin{cases} v_{il}-x_{kl}\leq\varepsilon+\xi_k^+, \\ x_{kl}-v_{il}\leq\varepsilon+\xi_k^-. \end{cases} \tag{23}$$

Now, the criterion (22) can be written in the form

$$g_{il}\left(v_{il}\right)=\sum_{k=1}^{N}\left(u_{ik}\right)^{m}\ \left(\xi_{k}^{+}+\xi_{k}^{-}\right), \tag{24}$$

and minimized subject to the constraints (23) and $\xi_{k}^{+}\geq 0,\ \xi_{k}^{-}\geq 0$. The Lagrangian of (24) with the above constraints is

$$\begin{aligned} G_{il}\left(v_{il}\right)\ =\ &\sum_{k=1}^{N}\left(u_{ik}\right)^{m}\ \left(\xi_{k}^{+}+\xi_{k}^{-}\right)-\sum_{k=1}^{N}\lambda_{k}^{+}\left(\varepsilon+\xi_{k}^{+}-v_{il}+x_{kl}\right)\\ &-\sum_{k=1}^{N}\lambda_{k}^{-}\left(\varepsilon+\xi_{k}^{-}+v_{il}-x_{kl}\right)-\sum_{k=1}^{N}\left(\mu_{k}^{+}\xi_{k}^{+}+\mu_{k}^{-}\xi_{k}^{-}\right), \end{aligned} \tag{25}$$

where $\lambda_{k}^{+},\lambda_{k}^{-},\mu_{k}^{+},\mu_{k}^{-}\geq 0$ are the Lagrange multipliers. The objective is to minimize the Lagrangian with respect to v_{il},ξ_{k}^{+} and ξ_{k}^{-}. It must be also maximized with respect to the Lagrange multipliers. The following optimality conditions (for the saddle point of the Lagrangian) are obtained by differentiating (25) with respect to $v_{il},\ \xi_{k}^{+},\ \xi_{k}^{-}$ and setting the result to zero:

$$\begin{cases} \dfrac{\partial G_{il}\left(v_{il}\right)}{\partial v_{il}}=\displaystyle\sum_{k=1}^{N}\left(\lambda_{k}^{+}-\lambda_{k}^{-}\right)=0,\\ \dfrac{\partial G_{il}\left(v_{il}\right)}{\partial \xi_{k}^{+}}=\left(u_{ik}\right)^{m}-\lambda_{k}^{+}-\mu_{k}^{+}=0,\\ \dfrac{\partial G_{il}\left(v_{il}\right)}{\partial \xi_{k}^{-}}=\left(u_{ik}\right)^{m}-\lambda_{k}^{-}-\mu_{k}^{-}=0. \end{cases} \tag{26}$$

The last two conditions in (26) and the requirements $\mu_{k}^{+},\mu_{k}^{-}\geq 0$ imply $\lambda_{k}^{+},\lambda_{k}^{-}\in[0,\left(u_{ik}\right)^{m}]$. Imposing conditions (26) on the Lagrangian (25), we get

$$G_{il}\left(v_{il}\right)=-\sum_{k=1}^{N}\left(\lambda_{k}^{+}-\lambda_{k}^{-}\right)x_{kl}-\varepsilon\sum_{k=1}^{N}\left(\lambda_{k}^{+}+\lambda_{k}^{-}\right). \tag{27}$$

Maximization of (27) subject to constraints

$$\begin{cases} \displaystyle\sum_{k=1}^{N}\left(\lambda_{k}^{+}-\lambda_{k}^{-}\right)=0,\\ \lambda_{k}^{+},\lambda_{k}^{-}\in[0,\left(u_{ik}\right)^{m}] \end{cases} \tag{28}$$

constitutes the so-called Wolfe dual formulation (problem). It is well-known from optimization theory that at the saddle point, for each Lagrange multiplier, the following

Karush-Kühn-Tucker conditions must be satisfied:

$$\begin{cases} \lambda_k^+ \left(\varepsilon + \xi_k^+ - v_{il} + x_{kl}\right) = 0, \\ \lambda_k^- \left(\varepsilon + \xi_k^- + v_{il} - x_{kl}\right) = 0, \\ \left((u_{ik})^m - \lambda_k^+\right) \xi_k^+ = 0, \\ \left((u_{ik})^m - \lambda_k^-\right) \xi_k^- = 0. \end{cases} \tag{29}$$

The last two conditions in (29) show that $\lambda_k^+ \in (0, (u_{ik})^m)$ leads to $\xi_k^+ = 0$ and $\lambda_k^- \in (0, (u_{ik})^m)$ implies $\xi_k^- = 0$. In this case, the first two conditions in (29) yield

$$\begin{cases} v_{il} = x_{kl} + \varepsilon, & \text{for } \lambda_k^+ \in \left(0, (u_{ik})^m\right), \\ v_{il} = x_{kl} - \varepsilon, & \text{for } \lambda_k^- \in \left(0, (u_{ik})^m\right). \end{cases} \tag{30}$$

Thus we can determine the cluster centre v_{il} from (30) by taking arbitrarily any x_{kl} for which the corresponding Lagrange multipliers are within the open interval $(0, (u_{ik})^m)$. Equations (27), (28) and (30) can be written with elegance as (11) and (12). ■

The freedom to choose x_{kl} in (30) results from the fact that we may have many x_{kl} at a distance $\pm\varepsilon$ from the cluster centre v_{il}. Each datum at a distance less than ε has $\lambda_k^{(\pm)} = 0$, and at a distance greater than ε it has $\lambda_k^{(\pm)} = (u_{ik})^m$. So, taking any x_{kl} for which $\lambda_k^{(\pm)} \in (0, (u_{ik})^m)$, we must obtain the same value of the cluster centre. But, from the numerical point of view, it is better to take the mean value of v_{il} obtained for all the data for which the conditions (30) are satisfied, i.e.

$$\underset{1\le i\le c}{\forall} \; \underset{1\le l\le p}{\forall} \quad v_{il} = \frac{1}{\operatorname{card}\left(\Lambda_i^+ \cup \Lambda_i^-\right)} \left[\sum_{\left\{ k \mid \lambda_k^+ \in \Lambda_i^+ \right\}} (x_{kl} + \varepsilon) + \sum_{\left\{ k \mid \lambda_k^- \in \Lambda_i^- \right\}} (x_{kl} - \varepsilon) \right]. \tag{31}$$

On the basis of Theorem 2 and (31), we obtain an algorithm that can be called ε-insensitive Fuzzy C-Means (εFCM):

1° Fix $c\ (1 < c < N)$, $m \in (1, \infty)$ and $\varepsilon \ge 0$. Initialize $\mathbf{V}^{(0)} \in \Re_{pc}$, set $j = 1$.

2° Calculate the fuzzy partition matrix $\mathbf{U}^{(j)}$ for the j-th iteration using (10).

3° Update the centres for the j-th iteration $\mathbf{V}^{(j)} = [\mathbf{v}_1^{(j)} \mathbf{v}_2^{(j)} \cdots \mathbf{v}_c^{(j)}]$ using (31), (12) and $\mathbf{U}^{(j)}$.

4° If $\left\|\mathbf{U}^{(j+1)} - \mathbf{U}^{(j)}\right\|_F > \xi$, then $j \leftarrow j + 1$, and go to 2°.

4. Numerical Experiments

In all the experiments for both FCM and εFCM the weighted exponent $m = 2$ was used. The iterations were stopped as soon as the Frobenius norm in a successive pair of $\mathbf{U}$ matrices was less than 10^{-5} for the FCM and 10^{-2} for the εFCM. For a computed terminal prototypes, we measured the performance of clustering methods by the maximal (on vectors components) absolute difference between the true centres and terminal prototypes. The Frobenius norm distances between the true centres and terminal prototype matrices were computed as well. All the experiments were run in the MATLAB environment. The linear optimization with constraints was performed using the `linprog` procedure.

4.1. Synthetic Data with a Varying Number of Outliers

The purpose of this experiment was to investigate the sensitivity of the FCM and εFCM methods to outliers. The two-dimensional (two features) data set presented in Fig. 1 consists of three well-separated groups and a varying number of outliers located at point (9,9). The number of outliers varies from 0 (no outliers) to 20 (the number of outliers equal to the cardinality of the upper-right cluster marked by crosses). The true cluster centres, calculated without outliers, are marked by triangles. Both the tested methods were initialized using prototypes $(4, 4)$, $(5, 5)$ and $(6, 6)$, marked with squares in Fig. 1. The εFCM method was tested for the parameter ε equal to 0.2 (less than the cluster radius), 2.0 (approximately equal to the cluster radius), and 3.0 (greater than the cluster radius).

The results for the FCM and εFCM methods with the data from Fig. 1 are presented in Tabs. 1 and 2. They show that for both the methods there is no deterioration

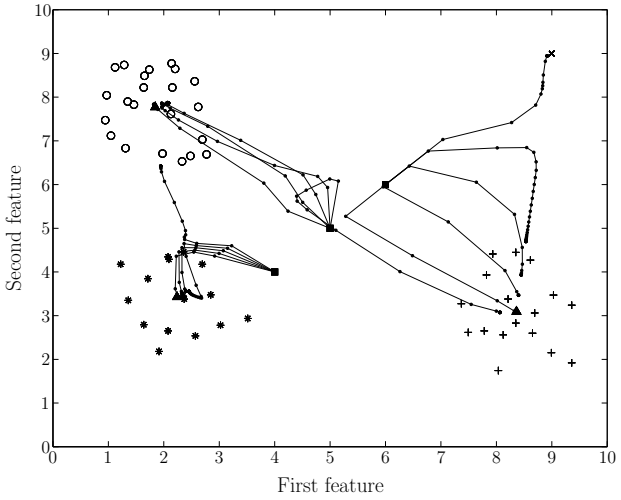

Fig. 1. Performance of the FCM method for the following numbers of outliers: 0, 4, 8, 12 and 16.

Table 1. Maximum cluster centre errors for the synthetic data with outliers.

Number of outliers	FCM	εFCM		
	—	$\varepsilon = 0.2$	$\varepsilon = 2.0$	$\varepsilon = 3.0$
0	0.0188	0.2307	0.1712	0.1047
	0.0652	0.1047	0.1294	0.1294
	0.0101	0.0314	0.3282	0.0126
2	0.0581	0.0314	0.1047	0.1047
	0.0780	0.1047	0.1831	0.1334
	0.1764	0.2718	0.0314	0.0126
4	0.0922	0.1110	0.1047	0.1047
	0.1171	0.1465	0.1891	0.2463
	0.3739	0.2797	0.0314	0.1350
6	0.1263	0.2180	0.1110	0.1047
	0.1766	0.4497	0.1334	0.2710
	0.5898	0.1465	0.1821	0.0126
8	0.1585	0.1047	0.1173	0.1047
	0.2236	0.1467	0.1334	0.3277
	0.8375	0.2039	0.2889	0.0126
10	0.1870	0.6085	0.1047	0.1753
	0.2514	0.1334	0.1334	2.2363
	1.1453	5.9081	0.3779	0.0126
12	0.2169	0.1722	0.1251	0.1047
	0.2435	2.6480	0.1334	0.4338
	1.6039	5.9081	0.8405	0.0784
14	5.8203	0.6085	0.1047	0.1753
	1.3452	0.1334	0.1334	0.4780
	5.8417	5.9081	0.2390	0.2519
16	2.9990	0.6085	0.1047	0.0336
	6.2144	0.1334	0.1294	1.0494
	5.8503	5.9081	0.2519	6.7621
18	2.9988	0.9190	0.1173	1.2760
	6.2143	0.2313	0.1334	0.1294
	5.8569	7.1561	3.0385	5.9081
20	2.9987	0.9190	0.1173	0.1173
	6.2142	0.1511	0.1334	0.1334
	5.8621	7.1561	3.1936	5.9081

Table 2. Frobenius norm of clusters centre errors for the synthetic data with outliers.

Number of outliers	FCM	εFCM		
	—	$\varepsilon = 0.2$	$\varepsilon = 2.0$	$\varepsilon = 3.0$
0	0.0055	0.0821	0.1546	0.0281
2	0.0456	0.1241	0.0629	0.0461
4	0.1794	0.1298	0.0661	0.1133
6	0.4226	0.2831	0.0807	0.1023
8	0.8141	0.1194	0.1957	0.1363
10	1.4568	35.7255	0.2739	5.1902
12	2.7336	42.3886	0.8917	0.2237
14	70.2390	35.7264	0.1326	0.3784
16	104.3053	35.7264	0.1273	77.9249
18	104.3905	53.6137	9.6860	36.9925
20	104.4583	53.5580	10.6388	35.3492

in the quality of the terminal prototype vectors for as many as 8 outliers. For more than 8 outliers, in the case of the εFCM method for $\varepsilon = 0.2$ (the approximately fuzzy median method), the deviation of the computed prototypes from the true centres increases with the number of outliers. In the case of the FCM method, for more than 12 outliers, the errors of the prototype calculation suddenly increase with the number of outliers. However, for the εFCM and $\varepsilon = 0.2$ the errors are smaller with respect to the FCM method for more than 12 outliers.

In the case of the εFCM method, for $\varepsilon = 2$ and as many as 16 outliers, the errors of the prototype calculation are small. When a number of outliers is comparable with the cardinality of the upper-right cluster, one prototype is located between the centre of this cluster and outliers. Other prototypes are placed correctly. In the case of the equal number of bad (outliers) and good points in the data, there is no method to distinguish between good and bad points. In other words, the εFCM for $\varepsilon = 2$ has a very high (nearly the highest possible) breakdown point. For the εFCM and $\varepsilon = 3$ the breakdown point is at 16 outliers, but the errors are smaller than for the FCM method. We can conclude that the best performance of the εFCM method is obtained for the parameter ε approximately equal to the radius of groups in the data. Figure 1 illustrates the performance of the FCM method for a varying number of outliers (equal to 0, 4, 8, 12 and 16). In this figure, the traces of the prototypes calculated in consecutive iterations are also shown. A larger number of outliers results in a greater attraction of the terminal prototypes towards outliers. Figure 2 illustrates the performance of the εFCM method with $\varepsilon = 2$ for 6, 14 and 20 outliers. From this figure we can observe that for 6 outliers the true centres are reached. For 14 outliers the trace of one prototype passes through outliers, but terminates close to the true cluster centre. For 20 outliers the trace of one prototype also passes through outliers and terminates halfway between the true centre and outliers.

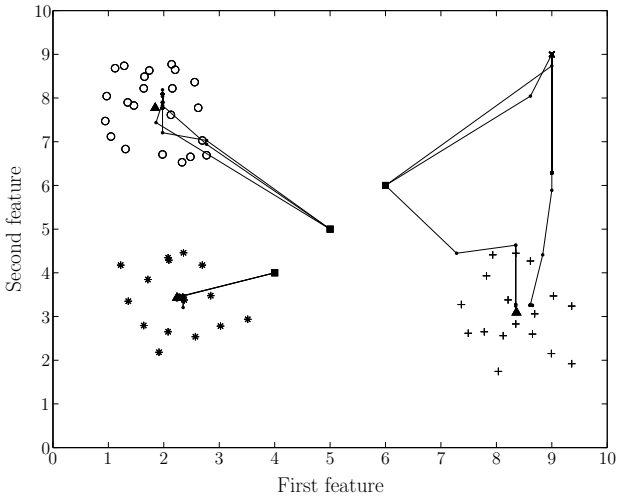

Fig. 2. Performance of the εFCM method with $\varepsilon = 2$ for the following numbers of outliers: 6, 14 and 20.

4.2. Heavy-Tailed and Overlapped Groups of Data

The purpose of this experiment was to compare the usefulness of the FCM and εFCM methods for clustering heavy-tailed and overlapped groups of the data. The two-dimensional data set, presented in Fig. 3, consists of three overlapped groups. Each group was generated by a pseudorandom generator with t-distribution. The true clus-

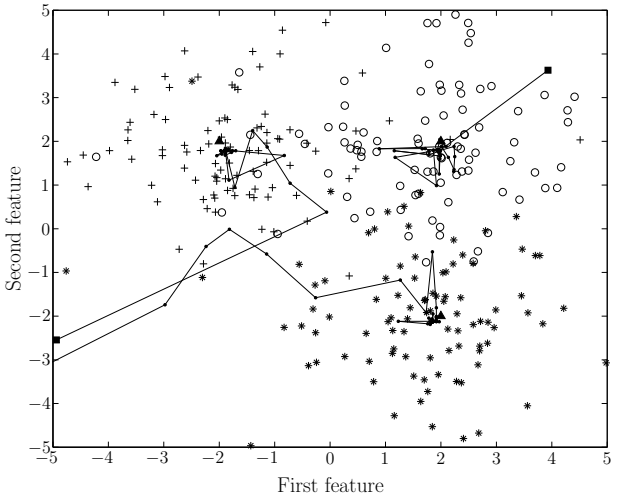

Fig. 3. Performance of the εFCM method with $\varepsilon = 2$ on heavy-tailed and overlapped clusters.

ter centres are $(2,2)$, $(-2,2)$ and $(2,-2)$. These centres are marked with triangles. For all calculations performed in this subsection, the k-th component of the j-th initial prototype was obtained as

$$\underset{1\le j\le c}{\forall} \quad v_{jk} = m_k + j\frac{M_k - m_k}{c+1}, \tag{32}$$

where c is the number of clusters, $m_k = \min\limits_i x_{ik}$, $M_k = \max\limits_i x_{ik}$.

The εFCM algorithm was tested with the parameter ε varying from 0 to 4.0 by 0.5. In Tab. 3, for these parameter values, the maximal absolute deviation of the cluster centres from the true centres is presented. Also, the Frobenius norm of this deviation is shown in Tab. 4. It can be seen from these tables that the εFCM method terminated very close to the true values. The method is not very sensitive to the choice of the ε parameter, but the best result was obtained for $\varepsilon = 2$. We can also see that the FCM method terminated far away from the true centres. Figure 3 illustrates the performance of the εFCM method for $\varepsilon = 2$. This figure presents only a part of the data with the true cluster centres. In this figure, the traces of prototypes calculated in sequential iterations are shown as well. The performance of the FCM method is illustrated in Fig. 4. It should be noted that the range of this plot is wider than in the previous figure. We can see that one prototype terminates in a wrong place (an erroneous local minimum of the criterion function) and hence the other prototypes must represent the data (the first in the cluster centre and the second between the remaining cluster centres).

Table 3. Maximal of cluster centre errors for the data with heavy-tails.

εFCM				
$\varepsilon = 0$	$\varepsilon = 0.5$	$\varepsilon = 1.0$	$\varepsilon = 1.5$	$\varepsilon = 2.0$
1.2774	0.2186	1.0891	0.0349	0.1188
0.3591	1.4384	0.9446	0.6262	0.2173
0.4361	0.1706	1.1458	1.4908	0.2358

εFCM				FCM
$\varepsilon = 2.5$	$\varepsilon = 3.0$	$\varepsilon = 3.5$	$\varepsilon = 4.0$	—
0.2161	0.2943	0.6850	0.6233	15.7694
0.2124	0.4546	0.4765	0.0501	1.1294
0.1772	0.6164	0.8026	0.4453	0.5398

5. Conclusions

Noise and outliers in the clustered data imply that the clustering methods need to be robust. This paper has established a connection between the fuzzy c-means method

Table 4. Frobenius norm of clusters centre errors for the data with heavy-tails.

εFCM				
$\varepsilon = 0$	$\varepsilon = 0.5$	$\varepsilon = 1.0$	$\varepsilon = 1.5$	$\varepsilon = 2.0$
2.0700	2.1618	3.4484	2.6981	0.1372

εFCM				FCM
$\varepsilon = 2.5$	$\varepsilon = 3.0$	$\varepsilon = 3.5$	$\varepsilon = 4.0$	–
0.1414	0.7422	1.3814	0.6587	250.6377

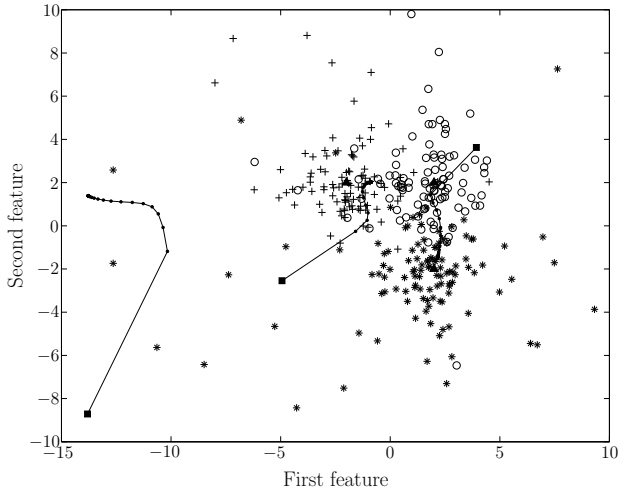

Fig. 4. Performance of the FCM method on heavy-tailed and overlapped clusters (the range of this plot is wider than in Fig. 3).

and robust statistics. The introduced ε-insensitive fuzzy clustering method is based on Vapnik's ε-insensitive loss function. The new method was introduced by solving a constrained minimization problem of the criterion function. The necessary conditions for obtaining a local minimum of the criterion function were proved. The existing fuzzy c-median method can be obtained as a special case of the method introduced in this paper. A comparative study of the ε-insensitive fuzzy c-means with traditional fuzzy c-means was also included. The numerical examples show the usefulness of the proposed method in clustering the data with outliers and with heavy-tailed and overlapped groups of the data. The proposed clustering method requives an insensitivity interval ε. This parameter depends upon the data structure, i.e. the cluster size and shape, outliers location and cardinality. In application of this method to high-dimensional real-word data, selection of the parameter ε can be made by means of a validity index.

Interesting tasks for future research include: (1) studying the use of the εFCM for high-dimensional real-word databases, (2) determining the insensitivity parameter ε, (3) establishing a computationally effective method for the cluster centre determination without using linear programming.

References

Bezdek J.C. (1982): *Pattern Recognition with Fuzzy Objective Function Algorithms.* — New York: Plenum Press.

Davé R.N. (1991): *Characterization and detection of noise in clustering.* — Pattern Recogn. Lett., Vol.12, No.11, pp.657–664.

Davé R.N. and Krishnapuram R. (1997): *Robust clustering methods: A unified view.* — IEEE Trans. Fuzzy Syst., Vol.5, No.2, pp.270–293.

Duda R.O. and Hart P.E. (1973): *Pattern Classification and Scene Analysis.* — New York: Wiley.

Dunn J.C. (1973): *A fuzzy relative of the ISODATA process and its use in detecting compact well-separated cluster.* — J. Cybern., Vol.3, No.3, pp.32–57.

Fukunaga K. (1990): *Introduction to Statistical Pattern Recognition.* — San Diego: Academic Press.

Hathaway R.J. and Bezdek J.C. (2000): *Generalized fuzzy c-means clustering strategies using L_p norm distances.* — IEEE Trans. Fuzzy Syst., Vol.8, No.5, pp.576–582.

Huber P.J. (1981): *Robust statistics.* — New York: Wiley.

Jajuga K. (1991): *L_1-norm based fuzzy clustering.* — Fuzzy Sets Syst., Vol.39, No.1, pp.43–50.

Kersten P.R. (1999): *Fuzzy order statistics and their application to fuzzy clustering.* — IEEE Trans. Fuzzy Syst., Vol.7, No.6, pp.708–712.

Krishnapuram R. and Keller J.M. (1993): *A possibilistic approach to clustering.* — IEEE Trans. Fuzzy Syst., Vol.1, No.1, pp.98–110.

Pal N.R. and J.C. Bezdek (1995): *On cluster validity for the fuzzy c-means model.* — IEEE Trans. Fuzzy Syst., Vol.3, No.3, pp.370–379.

Ruspini E.H. (1969): *A new approach to clustering.* — Inf. Contr., Vol.15, No.1, pp.22–32.

Tou J.T. and Gonzalez R.C. (1974): *Pattern Recognition Principles.* — London: Addison-Wesley.

Vapnik V. (1998): *Statistical Learning Theory.* — New York: Wiley.

Zadeh L.A. (1965): *Fuzzy sets.* — Inf. Contr., Vol.8, pp.338–353.

Received: 30 March 2001
Revised: 4 July 2001