fuzzy clustering, robust method,
hybrid method, weighted myriad.,

Tomasz PRZYBYŁA

HYBRID FUZZY CLUSTERING METHOD

A new hybrid clustering method based on fuzzy myriad is presented. Proposed method could be treated as generalization of well known fuzzy c-means method (FCM) proposed by Bezdek. The form of objective function of proposed method allows applying existing modification of the FCM method such as conditional clustering or partial supervised clustering.

1. INTRODUCTION

The clustering aims at assigning a set of objects to clusters in such a way that objects within the same cluster have a high degree of similarity, while objects belonging to different clusters are dissimilar. The clustering methods can be divided into hierarchical and nonhierarchical (partitioning) methods. In this paper, clustering by minimization of criterion function will be considered. The most traditional clustering methods are "hard" partitioning i.e. every object belongs to one group. Such kind of partition finds sharp boundaries among clusters. However, in practice, the boundaries are not strict but ambiguous. Thus soft partitioning is more suitable in this case. Fortunately, the fuzzy set theory proposed by Zadeh [1] allows to describe soft partitioning. The most popular method of fuzzy clustering is the fuzzy c-means (FCM) method proposed by Bezdek [2]. The main disadvantage of the FCM method is its sensitivity to presence of outliers and noise in clustered data. In real applications, the data are corrupted by noise and assumed models such a Gaussian distribution are never exact. The FCM method is a prototype-based method, where the prototypes are weighted (fuzzy) means. The performance of linear estimation of prototypes is optimal for the Gaussian model of data statistics. The Gaussian model is inadequate in an impulsive environment. Impulsive signals are more accurately modeled by distributions which density functions have heavier tails than the Gaussian distribution [3, 4].

2. HYBRID FUZZY CLUSTERING METHOD (HFCMyr)

2.1. WEIGHTED MYRIAD

Let us consider a set of *N* independent and identically distributed observations (iid), *X={x1,x2, ...,xN}*, and a set of assigned weights *U={u1,u2,...,uN}*. A weighted myriad is a value, , that minimizes the weighted myriad objective function defined as follows

$$\hat{\Theta} = \arg\min_{\Theta\in\Re} \sum_{k=1} \ln\left[K^2 + u_k\left(x_k - \Theta\right)^2\right] \tag{1}$$

The value of weighted myriad depends on the data set ***X***, assigned weights ***U*** and the parameter *K,* called a linearity parameter. Two interesting cases may occur: first, when the *K* value tends to infinity (i.e. *K*→∝), then value of weighted myriad converges with the weighted mean, that is

$$\lim_{K\to\infty} \hat{\Theta}_K = \frac{\sum_{k=1}^{N} u_k x_k}{\sum_{k=1}^{N} u_k}, \tag{2}$$

where $\hat{\Theta}_K = myriad\left\{u_k \Diamond x_k\right\}_{k=1}^{N}$. Second case occurs when the value of *K* parameter tends to zero (i.e. *K*→0), then value of weighted myriad is always equal to one of the most frequent values in the input data set.

2.2. OBJECTIVE FUNCTION

Fuzzy myriad selectivity depends on value of *K* parameter. Assuming, that for each cluster a different value of *K* is assigned, the objective function of proposed method can be described as follows:

$$J_m(\mathbf{U},\mathbf{V}) = \sum_{i=1}^{c}\sum_{k=1}^{N}\sum_{l=1}^{p} \ln\left[K_i^2 + u_{ik}^m\left(x_k(l) - v_i(l)\right)^2\right], \tag{3}$$

where K_i^2 is the myriad linear parameter assigned to *i*-th cluster and $1 \le i \le c$, $u_{ik} \in \mathbf{U}$ is fuzzy partition matrix, $x_k(l)$ is l-th feature of k-th input data set, $v_i(l)$ is l-th feature of i-th cluster prototype and *m* is a fuzzyfier.

Using Lagrange multipliers technique, the minimization of (3) can be done only if: for fixed number of clusters *c* and parameters *m* and *K*, are sets defined as

$$\forall_{1\le k\le N} \Im_k = \left\{i \mid 1 \le i \le c; \left\|\mathbf{x}_k - \mathbf{v}_i\right\|^2 = 0\right\}$$

$$\tilde{\Im} = \left\{1,2,\ldots,c\right\} - \Im_k,$$

the values of partition matrix are described by

$$\underset{1\le i\le c}{\forall}\ \underset{1\le k\le N}{\forall}\ u_{ik}=\begin{cases}\left[\sum_{j=1}^{c}\left(\dfrac{\left\|\mathbf{x}_k-\mathbf{v}_i\right\|}{\left\|\mathbf{x}_k-\mathbf{v}_j\right\|}\right)^{2/(m-1)}\right]^{-1} & if\quad \Im_k=\varnothing\\ 0 & if\quad \underset{i\in\Im_k}{\forall}\\ 1 & if\quad \Im_k\neq\varnothing\end{cases}\tag{4}$$

where $\|\bullet\|$ is an Euclidean norm, and $\mathbf{v}_i$ are prototypes $1\le i\le c$.

For fixed number of clusters *c*, parameters *m* and *K* and fixed partition matrix **U**, the prototype value minimizing (3) are fuzzy myriad described as follows

$$v_i(l)=\arg\min_{\Theta\in\Re}\sum_{k=1}^{N}\ln\left[K^2+u_{ik}^{m}\left(x_k(l)-\Theta\right)^2\right],\tag{5}$$

where: *i* is the cluster number ($1\le i\le c$), and *l* is the component (feature) number ($1\le l\le p$).

2.3. *K* VALUE ESTIMATION

The α-stable distribution is a generalization of Gaussian distribution (α=2) or a Cauchy distribution (α=1). So, methods for evaluating parameters of α-stable distribution can be applied for Gaussian or Cauchy distributions.

Assuming, that ***x*** is α-stable random variable, and $\mathbf{y}=\ln|\mathbf{x}|$, the following dependency can be proofed [5]

$$Var(\boldsymbol{y})=\frac{\pi^2}{6}\left(\frac{1}{\alpha^2}+\frac{1}{2}\right),\tag{6}$$

where $0\le\alpha\le 2$.

For Gaussian distribution (α=2), the *K* value should tends to infinity. In spite of fact that for *K*>50, differences between fuzzy myriad and fuzzy mean can be omitted. For the α<1, the fuzzy myriad estimator should as selective as possible (*K* should tends to 0). The following relation between α and *K* value has been proposed [6]

$$K=\sqrt{\frac{\alpha}{2-\alpha}}\tag{7}$$

For the data set ***X***, the estimation for each cluster can be done in the following way:

1. For *i*-th cluster, for each feature $1\le l\le p$ of feature vectors belonging to *i*-th cluster compute the K_l parameter.
2. Finally, value of K_i is compute as $K_i=\min\limits_{1\le l\le p}K_l$.

2.4. CLUSTERING DATA WITH HFCMYR CLUSTERING METHOD

The hybrid clustering method can be described as follows:

1. given the data set ***X****={**x**$_1$,…**x**$_N$}* where $\boldsymbol{x} \in \Re^p$, fix the number of clusters *c∈{2,…,N-1},* the fuzzyfier *m*∈[1,∝) and the tolerance limit ε. Initialize randomly the partition matrix **U** and fix initial values of *K* parameter for each cluster, fix *l*=0,
2. calculate the prototype values ***V***, as weighted myriads. A weighted myriad has to be calculated for each feature of $\boldsymbol{v}_i$ using (5),
3. update the partition matrix ***U*** using (4),
4. Update $K_i, 1 \le i \le c$, based on (6) and (7),
5. if $\left\| \mathbf{U}^{(l+1)} - \mathbf{U}^{(l)} \right\| < \varepsilon$ stop the clustering algorithm, otherwise *l=l+1* and go to (2°).

3. NUMERICAL EXPERIMENTS

Two sets have been chosen as test data. The first set is a synthetic containing two realization of random variables with Gaussian distribution and Cauchy distribution. Each realization has 100 points in 2D space. The first data set has been presented on figure (1).

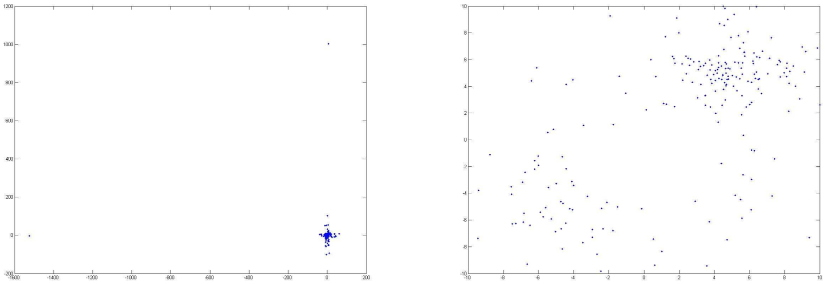

Fig.1 The first data set

The centers of generated clusters are: $\boldsymbol{v}_1 = [-5.0 \ -5.0]^{\mathrm{T}}$, $\boldsymbol{v}_2 = [5.0 \ 5.0]^{\mathrm{T}}$.

As a reference method has be chosen a familiar FCM method.

The following parameters have been fixed:

- number of clusters: c=2,
- fuzzyfier: m=2.0,
- tolerance limit: $\varepsilon = 10^{-5}$.

The obtaining results are presented in table 1.

Table 1**.** The value of prototypes from proposed and reference method

	FCM	HFCMyr	K
V$_1$	[-7.6006 –6.7342]$^{\mathrm{T}}$	[-3.1394 –8.9157]$^{\mathrm{T}}$	2
V$_2$	[4.5958 4.1041]$^{\mathrm{T}}$	[4.8655 4.9354]$^{\mathrm{T}}$	50

Obtained results from proposed have been presented on figure 2.

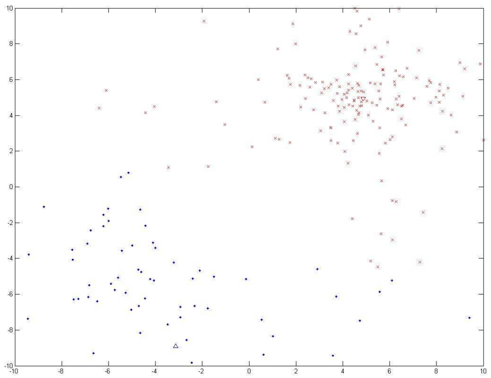

Fig.2 Two clusters found in the test data set

Results from proposed method and reference method are very similar, hence only from proposed method the test data set separation has been presented.

As the second data Fisher's Iris data set has been chosen. For this data set the following values have been fixed:

- the number of clusters *c=3*,
- the fuzzyfier *m=2*,
- the tolerance limit $\varepsilon=10^{-6}$.

Obtained results are presented in table 2 and 3 for the proposed and the reference methods, respectively.

Table 2**.** Results for IRIS data set and proposed method

	I	II	III	K
X_1	50			2.0
X_2		45	5	2.0
X_3		7	43	2.0

Table 3. Results for IRIS data set and reference method

	I	II	III
X_1	50		
X_2		47	3
X_3		13	37

4. CONCLUSIONS

This paper has dealt with clustering of data corrupted by noise and outliers. Well known methods such as Bezdek's FCM are sensitive on outliers, hence the obtained groups can be different than primary expected. Therefore methods which results are intuitively correct (the same or very similar to expected) are worth searched for.

Results of the proposed method are more accurate than the reference method outputs. A nonlinear estimation of group prototypes has increased robustness of the clustering method. The cost of increased robustness and flexibility is longer computational time.

BIBLIOGRAPHY

[1] Zadeh L., A., *Fuzzy sets*, Information and Control 8 (1965), 338-353.

[2] Bezdek, J. C., *Pattern recognition with fuzzy objective function algorithms*, Plenum Press, New York, (1981)

[3] Arce G. R. Kalluri S., *Fast algorithm for weighted myriad computation by fixed{point search*, IEEE Transactions on Signal Processing 48 (2000), no. 1, 159-171.

[4] Arce, G. R. Kalluri S., *Robust frequency-selective filtering using weighted myriad filters admitting real-valued weights*, IEEE Transactions on Signal Processing 49 (2001), no. 11, 2721-2733.

[5] Georgiu, P. G., Tsakalides, P., Kyriakakis C., *Alpha-Stable Modeling of Noise and Robust Time-Delay Estimation in the Presence of Impulsive Noise*, IEEE Transactions on Multimedia 1 (1999), no. 3, 291-301.

[6] Gonzales, J. G., Arce, G. R., *Optimality of the Myriad Filter in Practical Impulsive-Noise Environments.* IEEE Transactions on Signal Processing 49 (2001), no. 2, 438-441.