FUZZY C-MEANS BASED ADAPTIVE NEURAL NETWORK CLUSTERING

Mübariz EMINOV
Department of Statistics and Computer Science
Mugla University, Mugla, Turkey
E-mail: emubariz@hotmail.com

Keyword: fuzzy clustering, competitive learning, neural network, adaptive algorithm

ABSTRACT

In this paper, in order to cluster an unstructured data set, clustering and unsupervised neural network (connectionist) approaches are together considered. For clustering, the use of competitive learning (CL) based network and train it indirectly using fuzzy c-means (FCM) algorithm is proposed**.**

We deal with two kinds of clustering: a) number of weight vectors is known in advance, b) number of weight vectors is not known, at all. Regarding the former case, FCM based algorithm that can improve a partition performance of existing network trained by CL algorithms is suggested. With respect to the latter, validity criterion V_{sv} based algorithm that can obtain adaptive clustering and maintains results of FCM algorithm by finding optimal number of clusters is proposed. Both proposed algorithms were examined by numerical experiments results of which show their effectiveness.

1. INTRODUCTION

Clustering is a technique to explore a data set for the recognition of structure in them. For this aim, with respect to unstructured large crisp data, mostly a cluster analysis and unsupervised neural network (connectionist) approaches are used. The former approach, due to efficiency in clustering, uses well known probabilistic hard c-means (HCM) algorithm [1] and its fuzzified version the probabilistic fuzzy c-means (FCM) algorithm approaches [2] while the latter utilizes usually competitive learning (CL) algorithm [3]. All these algorithms are appropriate to recognize spherical clouds of high-dimensional points (patterns) in the data set.

The main goal of every clustering algorithm is to obtain optimal partition a data set to clusters. This, as known, means that clustering should get homogeneity within clusters and heterogeneity between them. [4]. For this aim, important cluster parameter such as the optimal values of a cluster centers (a prototypes) and memberships of the data assigned to clusters have to be chosen. Obviously, the cluster centers and memberships in HCM and FCM algorithms and CL algorithm are defined on different manners. Since the former group is algorithmic procedure which maintains alternating optimization minimizing the distance based objective function. Whereas CL algorithm can train cluster centers or weight vectors through heuristic scheme.

As known, CL algorithms yield a relatively appropriate partitioning if both number of clusters and desirable initial values of weight vectors are given during training phase of neural network [5]. However, these values are roughly allotted because, generally, the distribution of a pattern vectors is not known, in advance. Though in order to reduce the dependence of the partitioning results on initial values, numerous studies in which the convergence of weight (reference) vectors and different deleting methods [5,6] are proposed have been done. Still, roughly initialization of weight vectors and probably presentation of pattern vectors during learning phase make a partition performance worse. Another deficiency of the all considered CL algorithms is that they work on the basis of the crispy definition of clusters.

For described reason, we will deal with FCM algorithm that allows the strict assignments pattern vectors regardless of the initial values of cluster centers. Moreover, as mentioned above, using FCM algorithm, due to its fuzzy assignment, a local optimal partition can be fulfilled. Therefore, FCM algorithms are more useful than CL algorithms. Furthermore, in many cases, a number of clusters that requires finding another considerable cluster parameter such as an optimal number of clusters is not given in advance.

In this paper, in order to find the optimal number of clusters and improve the partition performance of traditional CL algorithms based neural network, the adaptive neural network clustering is proposed. The optimal number of clusters here is determined by new algorithm that is based on the validity criterion V_{sv} and maintains partition results of FCM algorithm. This algorithm is preferable especially when there are no compact clusters or they are greatly influenced by "noise". Moreover, training phase of the given network is indirectly executed through FCM algorithm, as well. Such kind of clustering possesses some advantages in comparison with both the pure FCM clustering and neural network trained by CL algorithms in sense both partitioning performance and computational time. The presented algorithms have been tested using artificial and real data sets concerning some attributes of individuals. Numerical experiments show fairly good results.

2. FCM CLUSTERING ALGORITHM

For the recognition of structure in large crisp data set $X= \{x_1, x_2,..., x_n\}\in R^p$, FCM algorithm is mostly used. It is the fuzzified version of HCM or hard ISODATA algorithm introduced by Dunn [1] and improved by Bezdek later [2]. According to this algorithm, a partition of data set $X\in R^p$ into c clusters $1 \le i \le c$ is expressed by (u x c) fuzzy partition matrix U or (U_{ik}) where $u_{ik} \in [0,1]$ is the membership degree of a pattern vector $x_k= \{x_1, x_2,..., x_p\}\in R^p$ to cluster c. In such partition the following conditions have to be satisfied [2,4]:

$$\sum_{k=1}^{n} u_{ik} > 0 \qquad \text{for all } i \in \{1,...,c\} \qquad (2.1)$$

$$\sum_{i=1}^{c} u_{ik} = 1 \qquad \text{for all } k \in \{1,...,n\} \qquad (2.2)$$

where condition (2.1) means that by contrast to HCM algorithm (CL algorithm, as well), a pattern can now belong to several clusters at once with different degrees of memberships.

As known, FCM algorithm recognizes a fuzzy cluster in the form of spherical clouds of patterns and each is represented by a cluster center (a prototype) $V_i \in R^p$. In order to obtain the local optimal partition of a pattern vectors to clusters, the optimal cluster centers for given belongingness of patterns to the clusters have to be chosen. For this aim, this requires the optimization of the distance based objective function [2,4] such as;

$$J(U, V) = \sum_{i=1}^{c} \sum_{k=1}^{n} (u_{ik})^m d^2(v_i, x_k) \qquad (2.3)$$

where d is the Euclidian distance and should be suited the above described constraints and also 1 < m where m is called fuzziness parameter.

FCM algorithm for optimization of the objective function (2,3), as known, uses an alternating optimization technique by which under fixed number of clusters are the randomly generated prototypes $V^{(0)}$ or the partition matrix $U^{(0)}$ and at each optimization step t, their values $U^{(t)}$ and $V^{(t)}$ are updated according to the following equations, respectively:

$$u_{ik} = \frac{1}{\sum_{j=1}^{c} \left(d^2(v_i, x_k) / d^2(v_j, x_k)\right)^{1/(m-1)}}$$

$$v_i = \frac{\sum_{k=1}^{n} (u_{ik})^m x_k}{\sum_{k=1}^{n} (u_{ik})^m}$$

The algorithm proceeds until successive approximation to the prototypes $d(V^{(t)},V^{(t-1)})$ or to the partition $d(U^{(t)},U^{(t-1)})$ have stabilized and the choice of terminating the condition corresponds to the above initialization manner, respectively. In the next section we will use the former manner initialization, i.e. the prototypes $V^{(0)}$ will be randomly generated.

It should be noted that FCM algorithm yields a partition matrix U where each pattern vector x_k is assigned to all given clusters as a fuzzy subset with the different membership degree. For many application problems such as classification, pattern recognition, etc, a unique assignment of a pattern to single cluster is only required. The assignment is based on the mapping the cluster to the relative pattern vector $x_k=\{x_1, x_2,..., x_p\}$. As a result of matching, each pattern will be crisply assigned to corresponding cluster if that pattern possesses highest degree of membership to them or minimum distance to its a cluster center (defuzzification process)[4,7]. This object will be dealt with later during estimation of the clustering quality, too.

3. COMPETITIVE LEARNING ALGORITHM

As mentioned earlier, unsupervised neural networks can be applied to the clustering to recognize the structure in the unstructured data set. In general, these clustering networks comprise two layers: the input and output (cluster) layers and the connections between their node constitute corresponding weight vectors. There are a lot of learning algorithms for unsupervised neural networks [8]. Previously denoted CL algorithms are usually used after which the weight vectors represent centers of the relevant clusters that have to be found during the training phase of the network. CL algorithm under fixed number of nodes (clusters) is similar to the HCM algorithm [1] but the former is different from the latter as it is referred to the heuristic procedure. At present, numerous CL algorithms have been proposed, however, a common basis here is that all the algorithms learn by competitive selection of "winning" neurons and adjusting associated weights that is called winner-take-all strategy. In order to select a winner weight vector, the distances between pattern vector $x(t)=\{x_1, x_2,..., x_p\} \in R^p$ and the weight vectors $w_i(t) \in R^p$ are calculated and the weight vector as a winner index minimizing the distance is selected. The process [3] formulated by

$$q = \arg\min \|x - w_i\| \qquad \text{for all } i$$

where arg(.) gives the index q of the winner. Then with the use of the winner q the weight vector $w_i^{(t)}$ is adjusted according to

$$\triangle w_i = \begin{cases} \alpha(t)(x - w_i) & i \in N_q^{(t)} \\ 0 & \text{otherwise} \end{cases}$$

where $\alpha(t)$ is learning rate and is decreasing function of time ($0<\alpha(t)<1$). For Kohonen's self- organizing algorithm, $N_q^{(t)}$ has set of indexes of topological neighbors for the winner at step t. If $N_q^{(t)}$ has an index of the winner only then the Kohonen's algorithm becomes the standard CL.

As well known, most CL algorithms are highly sensitive to initial values of the weight vectors and consequently, they can minimize the partition error if desirable initial values are given in advance. However, the knowledge about the distribution of the pattern vectors is generally not acquired beforehand, hence the initial values have to be randomly allotted. On the other hand, Kohonen's algorithm has little dependence on initial values [3] but it is not always suitable especially when the patterns do not exist under the influence of neighboring relations [5]. Though, in order to improve a partition performance of the clustering neural networks, numerous studies some of which are based on the generation of the many weight vectors and sequentially deletion to the fixed number according to the different deletion method have been done [5,6]. It should be stressed that learning is processed with CL algorithm. Such kind of algorithm is called the adaptive CL algorithm. We will deal with adaptive CL algorithm [6] where deletion method is based on the distortion error D_i (see Eq.(4.3)) of i^{th} cluster S_i. This method is shortly carried out as follows; with respect to all weight vectors (clusters), D_i is calculated according to (4.3) then s is determined to be deleted if it holds the condition $D_i \geq D_s$ for all i. Learning is proceeded with the remainder amount of clusters at each step.

The performance of conventional and adaptive CL algorithms will be examined in comparison with the FCM based algorithm proposed by us later.

4. Neural Network Clustering By Using Fcm Algorithm

In this section we consider the use of FCM algorithm in the unsupervised neural network in the two cases:

i) number of the weight vectors (clusters) is known in advance;

ii) number of weight vectors(clusters) is not known, at all.

4.1 Clustering With Known Number Of Weight Vectors

As denoted in the previous section, the conventional CL algorithms can get a relatively appropriate partitioning in the data set $X \in R^p$ if both number of weight vectors (clusters) and their desirable initial values are given in advance. Generally, these initial values are not known beforehand and hence weight vectors are roughly allotted in the training phase of network. Therefore, the adaptive CL algorithms one of which is early understood have been introduced. However, both the conventional CL algorithms and the adaptive CL algorithms get training still by roughly initialization of weight vectors and probably presentation of a pattern vectors that cause to converge of weight vectors on the different values. Another deficiency of these CL algorithms is that they work on the basis of crispy partition in the data set. The denoted failings of CL algorithms lead to reduction in the partition performance of clustering.

In order to improve a partition performance of mentioned CL based network, clustering approach after which network is indirectly trained is suggested by us. For this aim, FCM based algorithm from which trained values of the weight vectors are resulted is proposed. On the other hand, fulfilling clustering in the network's constitution instead of pure FCM algorithm becomes additional advantages that are as follows. It will be possible to cluster a whole data set by using the training data only. Furthermore, the crispy assignment of the pattern vectors (defuzzification process) to correspondent cluster using a fuzzy partition matrix U is provided without utilizing an additional classifier module. Since this process can be simply accomplished by selecting the weight vector for which d (x,w_i) is minimum (see (4.2)). It is clear; such kind of approach fairly decreases computational time of clustering.

Now we look at the estimation of the cluster quality of the clustering neural network when number of weight vectors is given but their initial values are not given in advance. As known, a data set $X \in R^p$ composed of c cluster each of which contains a data subset S_i as follows:

$$X= \bigcup_{i=1}^{c} S_i \text{ with } S_i \cap S_j = 0 \quad \forall j \neq i \tag{4.1}$$

After finding a trained values of the weight vectors $W=\{w_1, w_1,..., w_c\} \in R^p$, a data set $X \in R^p$ is divided into the cluster according to the condition.

$$S_i= \{ X \in R^p \mid \| x-w_i \| \leq \| x-w_j \| \quad \forall j \neq i \} \tag{4.2}$$

In this paper, the intra-cluster distance or distortion error in S_i data subset of i^{th} cluster is calculated by the following equitation. [5]:

$$D_i= (1/p) \sum_{x \in Si} \|x-w_i\|^2 \tag{4.3}$$

In addition, the mean square error (MSE) is given as follows:

$$E= (1/n) \sum_{i=1}^{c} D_i \tag{4.4}$$

4.2 Clustering With Unknown Number Of Weight Vectors

Up until now, we have assumed that number of clusters is implicitly known in advance although the distribution of pattern vectors into clusters might not be given. Unfortunately, in many cases information about distribution of whole pattern vectors on the clusters that is the number of clusters is yet not presented. The number of clusters c, as known, is the most significant cluster parameter affecting clustering performance. Different choices of c may lead to different clustering results. Thus, the estimation of the optimal cluster number c* during the clustering process is a prime concern. We refer to clustering that is able to find cluster number c* as adaptive clustering. Such kind of clustering can be executed by neural network approach in which ART networks are representative. These networks suggest some kind of threshold mechanism for new cluster formation. However, ART networks are more sensitive to data noise than HCM algorithm and conventional CL algorithm and hold their disadvantage in the same way [8].

In this paper we consider adaptive clustering neural network approach that uses CL algorithm based network. According to this clustering an optimal number of clusters c* is found by the proposed algorithm which can evaluate the local optimal partition for each number of cluster c using cluster validity or validity criteria. Then both a number of weight vectors and their values for given network are resulted from this algorithm. Thus, the adaptive network clustering which comprises the advantages of the FCM based algorithm described in previous subsection can be provided. Numerous validity criteria of which well known ones are the partition coefficient (V_{pc}) and the partition entropy (V_{pe}) introduced by Bezdek [2] have been proposed. However, V_{pc} and V_{pe} are sensitive to noises or a fuzziness parameter m and give a good result if a data set contains compact and well-separated clusters. In order to overcome this weakness and provide a good performance under a wide range of both c and fuzziness parameter m, some indices have been suggested [9]. One of these is a validity criterion V_{SV} introduced by Kim and Park [9] that can best suit to denoted requirements. This index uses only the structured characteristics and works effectively both HCM and FCM algorithms. Taking into account the advantages of the validity criterion V_{sv}, the V_{sv} based algorithm to determine an optimal number of cluster c* is suggested by us. Before the presentation of this algorithm we shortly describe V_{sv} index below.

The foundation of criteria V_{SV} is that clusters are in the under-partitioned state when c<c* and the over-partitioned state when c>c*. To find the under-partitioned state in the partitioning process, $V_u(c,V,X)$ as an under-partition measure function is used:

$$V_u(c,V,X) = (1/n)\sum_{i=1}^{c} MD_i, \qquad 2 \le c \le c_{max} \qquad (4.5)$$

where MD_i is the mean intra-cluster distance (MICD) of the i^{th} cluster defined as

$$MD_i = \Big(\sum_{x \in Si} \|v_i - x\|^2\Big)/n_i \qquad (4.6)$$

where S_i is a data subset of the i^{th} cluster and n_i is the number of pattern in them. Clearly, this function has very small values for c≥c* (patterns are optimally- or over-partitioned) and relatively large values for c<c* (patterns are under-partitioned).

To determine the over-partitioned state, an over-partition measure function defined as V_o (c,V) is utilized as following:

$$V_o(c,V) = c/d_{min}, \qquad 2 \le c \le c_{max} \qquad (4.7)$$

where d_{min} is the inter-cluster minimum distance which is determined as

$$d_{min} = \min_{i \ne j} \|v_i - v_j\| \qquad (4.8)$$

The function V_o (c,V) has very large values for c≥c* (patterns are over-partitioned that is d_{min} is very small) and relatively small values for c<c* (patterns are optimally- and under-partitioned).Clearly, these functions show opposite properties. After normalization, V_{SV} is formulated by the two partition measure functions as following:

$$V_{sv}(c,V,X) = U_{uN}(c,V,X) + V_{oN}(c,V) \qquad (4.9)$$

Thus, the optimal cluster number c* can be found with the smallest value of V_{sv} for c=2 to c_{max} and V_{sv}=[0,1] will be held.

The algorithm to find c* is presented as follows.

Step 1. Initialization:
give minimum (c=2) and maximum ($c=c_{max}$) numbers of clusters.

Step 2. Partitioning:
provide crisply defined partitioning in data set after FCM algorithm.

Step 3. Calculating:
calculate U_{uN} and V_{oN}
calculate V_{SV}

Step 4. Setting:
set c ← c+1

Step 5. Termination condition:
if $c<c_{max}$ then go to Step 2, otherwise

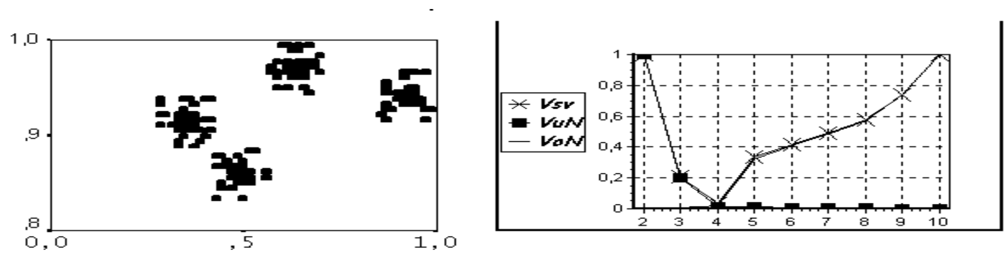
Step 6. Stop.

5. EXPERIMENTAL RESULTS

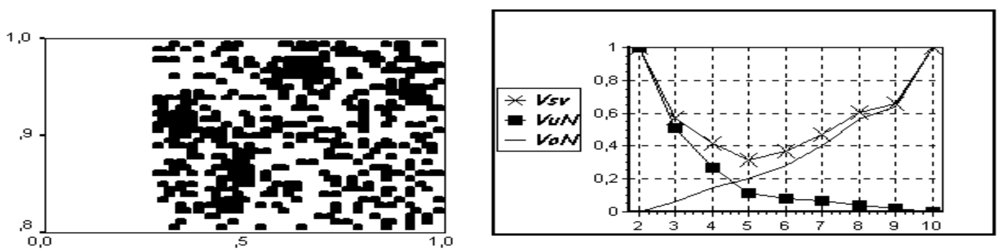
In order to verify both the proposed FCM based partition algorithm and adaptive clustering algorithm, numerical experiments have been performed. In these ones, the four types of the data set related to a population in which patterns represent individuals are used. Each pattern is three-dimensional real-valued vector, which contain some properties of individuals such as Age, Height and Weight. Used types of the data sets correspond different kinds of distributions of the pattern vectors such as: a) mixed Gaussian, b) mixed Gaussian together uniform, c) uniform in the separated spaces and d) uniform in the entire space,. These types of data sets here are called as 'Data 1', 'Data 2', 'Data 3' and 'Data 4', respectively and shown in Fig. Pattern vectors are assigned within [0,1] and in convenience they are presented on x and y axes.

At first, experiments related to determine according to proposed adaptive clustering algorithm, the optimal number of clusters with indicated four data sets were performed. For each of the data sets FCM algorithm was performed with m=2 and termination condition $\varepsilon = 10^{-5}$. The prototypes are chosen as the initial values $V^{(0)}$ instead of the fuzzy partition matrix U. The cluster numbers are varied from 2 to 10 for all data sets. As shown in Fig. (a) and (b), number of clusters is preferably expected to be $c^* = 4$ both for 'Data 1'and 'Data 3' as the optimal cluster numbers. Figure shows partition measure functions V_{uN} and V_{oN} with respect to the cluster number c with respect to all data sets, as well. As clear from this figure, measure functions are fast gradient changed around c = 4,5,4,5, with respect to 'Data 1', 'Data 2', 'Data 3' and 'Data 4', respectively. Consequently, the validity index V_{sv} shows a steep valley at correspondent values of c that is regarded as the optimal number of cluster.

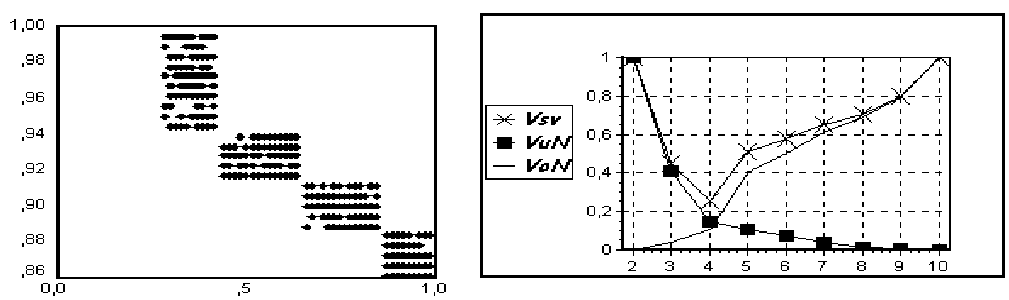
Table 1 comparatively presents the values of the validity indices V_{sv} and V_{pc} with respect to all data sets for c = 2 to 10. There the optimal value of c chosen by each index is highlighted. Table 1 shows that the index V_{sv} correctly defines optimal number of cluster in the all cases. The correctness is missing in V_{pc} when clusters are not compacted and separated. After choosing the optimal number clusters, numerical experiments were proceeded to estimate the partition performance of the FCM based algorithm in comparison with: a) standard CL algorithm, b) adaptive CL algorithm and c) pure FCM algorithm. The partition performance of the four algorithms is examined with above-mentioned types of data sets and it is assumed that optimal number of clusters is chosen for each data set from previous experiments.



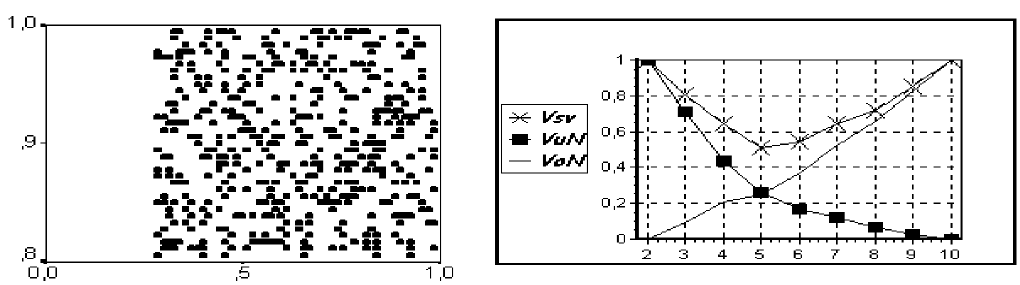
(a) Data1 and results



(b) Data 2 and results



(c) Data 3 and results



(d) Data 4 and results

Figure. Data sets and their V_{sv} index results

Table 1. Performance comparison of the validity indices.

	Data 1		Data 2		Data 3		Data 4	
c	V_{pc}	V_{sv}	V_{pc}	V_{sv}	V_{pc}	V_{sv}	V_{pc}	V_{sv}
2	0,952	1	**0,898**	1	**0,969**	1	**0,887**	1
3	0,984	0,211	0,897	0,567	0,946	0,450	0,816	0,804
4	**0,996**	**0,033**	0,885	0,416	0,967	**0,251**	0,790	0,641
5	0,941	0,334	0,896	**0,315**	0,955	0,511	0,793	**0,509**
6	0,976	0,419	0,887	0,365	0,945	0,577	0,782	0,541
7	0,941	0,486	0,881	0,473	0,932	0,654	0,774	0,641
8	0,906	0,574	0,868	0,603	0,917	0,707	0,779	0,721
9	0,921	0,737	0,868	0,656	0,914	0,796	0,777	0,849
10	0,848	1	0,868	1	0,910	1	0,773	1

The initial parameters of the adaptive CL algorithm are chosen as follows: n = 500, final number of weight vectors l_f is optimal number of clusters in the correspondent data set, initial number of weight vectors $l_o = 10$, termination condition is the convergence of all weight vectors with error $\varepsilon_w = 10^{-4}$ and $\alpha = 0.8$. Both the pure FCM and FCM based algorithms are initialized with same parameters, i.e. $m = 2, \varepsilon = 10^{-5}$ However, the latter is performed by using the some training data subsets each of which contains about 180 patterns.

Table 2 shows MSE calculated according to (4.4) for each algorithm using the 'Data 1', 'Data 2', 'Data 3' and 'Data 4'. The results with respect to CL algorithms are the averages of 10 trials. It is clear, the excellent results are shown by the pure FCM algorithm and FCM based algorithm makes little concessions. If to use the whole data sets, of course, no difference then will exist.

Table 2. MSE values for each algorithm.

Algorithm	Data1 MSE	Data2 MSE	Data3 MSE	Data4 MSE
Standard CL	3,7	3,5	1,5	6,5
Adaptive CL	3,3	2,7	1,3	5,4
FCM based	2,88	2,5	1,1	4,2
Pure FCM	2,81	2,1	1,0	4,0
	$x10^{-4}$	$x10^{-3}$	$X10^{-3}$	$x10^{-3}$

6. CONCLUSION

The considered clustering is performed on the CL based neural network training phase of which can be indirectly executed by using the proposed two algorithms. These algorithms can provide the ordinary and adaptive clustering and were successfully applied to the four types of crisp data sets. The results show enhanced performance in a sense both partitioning correctness and computational time.

REFERENCES

1. R. Duda and H. Hart, Pattern Classification and Scene Analysis, Wiley, New York, 1973.
2. J.C.Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum Press, New York,1981.
3. T.Kohonen, Self-Organization and Associative Memory, Springer-Verlag, Berlin, 1989.
4. F. Hoppner, F. Klawonn, R. Kruse and T. Runkler, Fuzzy Cluster Analysis, John Wiley, New York, 1999.
5. M. Maeda and H. Miyajim and S. Murashima, An Adaptive Learning and Self-Deleting Neural Network for Vector Quantization, IEICE Trans. Fundamentals.vol.E79-A, no 11,1886-1893,1996.
6. M. Maeda and H. Miyajim, Competitive Learning Algorithm Founded on Adaptivity and Sensitivity Deletion Methods, IEICE Trans. Fundamentals. vol. E-83-A,no12,2770-2774,2000.
7. M. Eminov, Rule–Based Fuzzy Classification Using Query Processing, Mathematical &Computational Applications, vol.8, no 2, 253-262,2003.
8. L.M. Fu, Neural Networks in Computer Intelligence, McGraw-Hill, New York, 1994.
9. D. J. Kim, Y. W. Park and D.J. Park, A Novel Validity Index for Clusters, IEICE Trans. Inf.& System., vol.E-84-D,no 2,282-285,2001.