The Positive Effects of Fuzzy C-Means Clustering on Supervised Learning Classifiers

Bekir Karlik *bkarlik@selcuk.edu.tr*
Department of Computer Engineering
Selçuk University
Konya, 42075, Turkey

Abstract

Selection of inputs is one of the most substantial components of classification algorithms for data mining and pattern recognition problems since even the best classifier will perform badly if the inputs are not selected very well. Big data and computational complexity are main cause of bad performance and low accuracy for classical classifiers. In other words, the complexity of classifier method is inversely proportional with its classification efficiency. For this purpose, two hybrid classifiers have been developed by using both type-1 and type-2 fuzzy c-means clustering with cascaded a classifier. In this proposed classifier, a large number of data points are reduced by using fuzzy c-means clustering before applied to a classifier algorithm as inputs. The aim of this study is to investigate the effect of fuzzy clustering on well-known and useful classifiers such as artificial neural networks (ANN) and support vector machines (SVM). Then the role of positive effects of these proposed algorithms were investigated on applied different data sets.

Keywords: Fuzzy C-Means Clustering, ANN, SVM, Learning, Classifier.

1. INTRODUCTION

Data mining is also known as knowledge discovery in database which is the process of investigating knowledge such as patterns, anomalies, changes or significant structures from complicated data, data warehouse, storage data in database etc. Many scientists have proposed different classifiers that provide significant speedup and make the algorithms more practical to solve complicated pattern classification and data mining problems [1-4]. Cluster analysis is a structural approach to solve the problem of the pattern classification without training samples.

Last decades, Fuzzy clustering methods have been provided solution stability to distortion of the source data has become quite noticeable. The source data is considered to inform about classes, particularly number of classes and density of their distribution in the investigated region of the feature spaces. For example, the data sets can contain objects of unrepresentative (or unlabeled) classes. Fuzzy classification rules can be obtained from fuzzy clustering results. Generally classifier is trained training dataset obtained using different fuzzifier parameters to analyze the effect of fuzzifier parameters for training period. Fuzzy c-means clustering (FCM) algorithm is successfully used for solving various problems. However, the algorithm has two major problems. Firstly, the algorithm and the second is highly dependent on the starting block centers, the algorithm is attached to local minima. These various hybrid methods for preventing the problem for the performance of FCM algorithm is used in the literature to improve. In literature, various hybrid studies have been showed clearly that combination supervised artificial neural network and unsupervised fuzzy c-mean algorithms which have better accuracies than the classical clustering and statistical classifiers [5-8]. Both Type-1 FCM [9-12] and type-2 FCM clustering [13-18] are used to select the best patterns related alike label in developed methods in previous studies.

This study presents two hybrid classification systems which are combination of both type-1 and type-2 FCM clustering algorithms and artificial neural networks. Then cluster sets are entered as input to neural networks, which has well-known back propagation algorithm. For this purpose, three data sets are clustered for three different examples. The first one it is clustered for ECG data set, the second one is clustered for the EMG data set. The last one is clustered for perfume data set. The results of hybrid fuzzy clustering based classifiers accuracies are shown to be better than the ordinary neural networks and support vector machines classification algorithms.

The remainder of this paper is arranged as follows: In the following section, fuzzy c-means clustering (type-1 and type-2) methodologies are briefly reviewed. Section 3 explains the structure of fuzzy clustering neural networks and its applications. Section 4 describes the structure of fuzzy c-means clustering based SVM and its applications. The results and some concluding remarks are provided in the last section.

2. FUZZY C-MEANS CLUSTERING (TYPE-1 AND TYPE-2)

Clustering is a task of grouping objects into classes (or labels) of similar objects. It is an unsupervised classification of partitioning of patterns into groups (or clusters) depending on their locality and connectivity within an m-dimensional feature space [19]. In the clustering process, it should be more similarity between sets of elements in the cluster, but the similarities should be less. FCM algorithms form a basis for clustering techniques that dependent on objective functions. The essential of fuzzy clustering is to divide the data into fuzzy partitions that overlap with one another. For this reason, the subsumption of data in a cluster is described with a membership grade between 0 and 1. Unsupervised FCM algorithm is a technique based on the objective function. The algorithm, which is a generalization of the least squares method works by translating the following objective function to minimize.

$$\min\left\{ J_m(U,V,X) = \sum_{k=1}^{N} \sum_{i=1}^{c} (u_{ik})^m \left\| x_k - v_i \right\|_A^2 \right\} \tag{1}$$

where

$$U \in M_{fcn} = \left\{ U \in \Re^{cN} \middle| \begin{array}{c} 0 \le u_{ik} \le 1 \,\forall\, ik \,\&\, \forall k, u_{ik} > 0 \,\exists\, i \\ 0 < \sum_{k=1}^{N} u_{ik} > n \,\forall\, i \,\&\, \sum_{i=1}^{c} u_{ik} = 1 \,\forall\, k \end{array} \right\} \tag{2}$$

where $V=\{v_1,v_2,...,v_c\}$ is defined the vector of cluster centers, and has $\|x\|$ *A=(xTAx)1/2* an interior product norm. *A* is an *h x h* positive matrix, that describes the form of the clusters. The matrix *A* is generally chosen as the identity matrix, leading to Euclidean distance and, as a result, to spherical clusters [20]. The *U* membership matrix randomly assigning algorithm starts. In the second step centers vectors are calculated. According to the calculated cluster centers recalculated using the matrix equation.

$$u_j(x_i) = \frac{1}{\sum_{k=1}^{C} \left(\frac{d_{ji}}{d_{ki}}\right)^{2/(m_1-1)}} \tag{3}$$

where $d_{ji}(d_{ki})$ is defined the distance between cluster prototype $v_j(v_k)$ and pattern x_i [21]. FCM algorithm is an important parameter affecting the behavior of m coefficient used blur algorithm. The value of this parameter specifies the maximum turbidity limit. The value of the defocus coefficient 1 closer algorithm solid (hard) starts to cluster up. There is no rule or method to determine the value of the blur factor. Optimum values for this coefficient can only be found by testing. Definition of the distance of prime membership for each pattern represents both the lower and the upper distance memberships using two dissimilar rates of *m*. A type-2 FCM algorithm is an extension of the conventional type-1 FCM algorithm. The prime memberships that expand pattern x_i by distance type-2 fuzzy sets are formed [22].

$$\bar{u}_j(x_i)=\begin{cases}\dfrac{1}{\sum_{k=1}^{C}(d_{ji}/d_{ki})^{2/(m_1-1)}} & \text{if } \dfrac{1}{\sum_{k=1}^{C}(d_{ji}/d_{ki})^{2/(m_1-1)}}>\dfrac{1}{\sum_{k=1}^{C}(d_{ji}/d_{ki})^{2/(m_2-1)}}\\ \dfrac{1}{\sum_{k=1}^{C}(d_{ji}/d_{ki})^{2/(m_2-1)}} & \text{otherwise}\end{cases}$$

and (4)

$$\underline{u}_j(x_i)=\begin{cases}\dfrac{1}{\sum_{k=1}^{C}(d_{ji}/d_{ki})^{2/(m_1-1)}} & \text{if } \dfrac{1}{\sum_{k=1}^{C}(d_{ji}/d_{ki})^{2/(m_1-1)}}\le\dfrac{1}{\sum_{k=1}^{C}(d_{ji}/d_{ki})^{2/(m_2-1)}}\\ \dfrac{1}{\sum_{k=1}^{C}(d_{ji}/d_{ki})^{2/(m_2-1)}} & \text{otherwise}\end{cases}$$

where *m₁* and *m₂* symbolize dissimilar fuzzy degrees of fuzzifiers. When we describe the distance of prime membership for each pattern, we can use the highest and lowest prime membership of the distance for each pattern. These rates are defined by upper and lower membership for a pattern, respectively. The uses of fuzzifiers are given dissimilar objective functions for minimizing in FCM in Eq. (5) [22].

$$J(U,V)=\sum_{j=1}^{C}\sum_{i=1}^{N}(u_j(x_i))^{m}\,d_{ji}^{2}\quad\text{subject to }\sum_{j=1}^{C}u_{ji}=1\ \text{ for all } i \tag{5}$$

It needs to reduce for an interval type-2 fuzzy set before hard-partitioning. The type reduction before hard-partition must be workable attentively since upper and lower membership $\bar{u}(x_i)$ and $\underline{u}(x_i)$ is usable. This type reduction is performed in order to calculate cluster centers. For this, left memberships $u_j^L(x_i)$ and right memberships $u_j^R(x_i)$ for whole patterns are calculated to arrange left (v_L) and right (v_L) cluster center (Eq. (6).

$$v_x=\frac{\sum_{i=1}^{N}x_i u(x_i)}{\sum_{i=1}^{N}u(x_i)} \tag{6}$$

For this reason, type-reduction can be reached using $u_j^L(x_i)$ and $u_j^R(x_i)$ to partition a pattern set into clusters. Taking the average of the upper and lower membership value of block located in the center of the matrix to be used in calculating membership. Then with FCM algorithm for heap centers are calculated. Condition of end of algorithm is satisfied, solid-classification process should be made for the segmentation process. However, because the data set is now type-2 does not apply, such as the type-1 algorithm. In this case, there is need for such reduction. Type reduction is performed as follows;

$$\text{Type-reduction: } u_j(x_i) = \frac{u_j^R(x_i) + u_j^L(x_i)}{2}, \qquad j = 1, \ldots, C \tag{7}$$

In Eq (7), memberships $u_j^L(x_i)$ and $u_j^R(x_i)$ are dissimilar in respect of each features for a pattern. For this reason, a sample rate for left and right prime membership is required to calculate for each feature such as [22];

$$u_j^R(x_i) = \frac{\sum_{l=1}^{M} u_{jl}(x_i)}{M} \qquad \text{where } u_{jl}(x_i) = \begin{cases} \overline{u}_j & \text{if } x_{il} \text{ uses } \overline{u}_j(x_i) \text{ for } v_j^R \\ \underline{u}_j & \text{otherwise} \end{cases}$$

$$\text{and} \tag{8}$$

$$u_j^L(x_i) = \frac{\sum_{l=1}^{M} u_{jl}(x_i)}{M} \qquad \text{where } u_{jl}(x_i) = \begin{cases} \overline{u}_j & \text{if } x_{il} \text{ uses } \overline{u}_j(x_i) \text{ for } v_j^L \\ \underline{u}_j & \text{otherwise} \end{cases}$$

where *M* is described the number of features for each pattern of x_i.

3. THE ALGORITHM OF FUZZY CLUSTERING NEURAL NETWORKS

The hybrid algorithm which is called as fuzzy c-means clustering neural networks (FCNN) has been developed by Karlık [23]. As seen in Fig.1, It consists of an unsupervised fuzzy clustering and a supervised artificial neural network (ANN) algorithm. Selection of the ANN inputs is one of the most significant components of designing ANN based upon pattern recognition since even the best classifier will perform badly if the inputs are not chosen very well. FCM suitable number of clusters for given input data is found beforehand. The aim of fuzzy clustering is to divide the data into fuzzy partitions, which overlap with each other. A number of data points are reduced by FCM clustering before inputs are applied to an ANN. For this reason, the subsumption of each data to each cluster is represented by a membership rate in [0, 1]. Then, proposed hybrid FCNN architecture is used for the training.

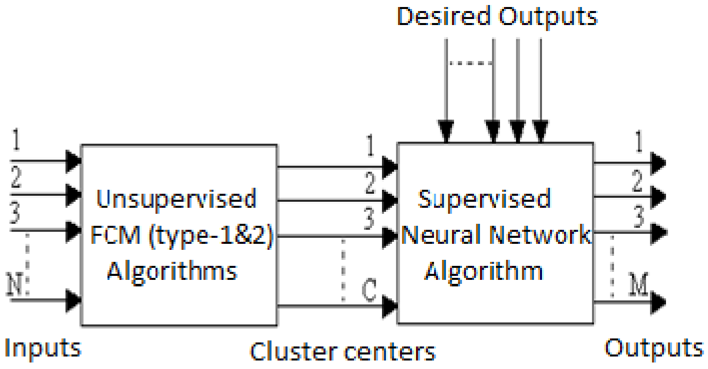
Desired Outputs
1
2
3
N
Unsupervised FCM (type-1&2) Algorithms
1
2
3
C
Supervised Neural Network Algorithm
1
2
3
M
Inputs
Cluster centers
Outputs

FIGURE 1: The architecture of FCNN Algorithm.

In this paper, the two hybrid fuzzy clustering neural network algorithms (type-1 and type-2) have been developed and provided for classification biomedical signals (ECG and EMG) using different training sets. Here, each cluster center obtained from each class provides a new attribute for describing different biomedical signals such as type of arrhythmia from ECG, or arm movements from EMG. The training and testing of fuzzy clustering neural network and/or ordinary neural network have been performed by these training set (feature sets). Then comparative assessment of the performance between proposed both type-1 and type-2 fuzzy clustering neural networks (FCNN algorithms and classical ANN, more reliable results are found by using both hybrid FCNN algorithms than ordinary ANN classifier for the classification of biomedical signals.

No	Architecture	Size of training set	Accuracy (%)
1	Ordinary NN	200*106	78
2	Type-1 FCNN	200*67	97,8
3	Type-2 FCNN	200*75	98,84

TABLE 1: Test results of classification accuracies for ECG data set.

No	Architecture	Size of training set	Feature ext.	Accuracy (%)
1	Ordinary NN	12*6	AR method	96,1
2	Type-1 FCNN	6*6	AR method	98,3
1	Ordinary NN	470*100	DWT	97,02
2	Type-1 FCNN	470*20	DWT	99,58

TABLE 2: Test results of classification accuracies for EMG data set.

In this new approach, the number of instance in training set is reduced by using FCM algorithm and process of reducing is performed on each class of biomedical data set separately. As seen in Table 1 and Table 2, according to ECG application, the size of training sets are 200 samples*67 patterns for type-1 FCNN, and 200 samples*75 patterns for type-2 FCNN respectively [10, 13, 17]. For the EMG, the size of training sets are 6 samples*6 patterns for Type-1 FCNN using Autoregressive (AR) parametric feature extraction method, and 470 samples*20 patterns for Type-1 FCNN using discrete wavelet transform (DWT) feature extraction method [3, 24]. These training feature data sets are defined the FCNN with corresponding class (or label) of outputs.

4. THE ALGORITHM OF FUZZY CLUSTERING BASED SVM CLASSIFIER

Support Vector Machines (SVM) classifier is aimed to be done through the aid of a linear or a nonlinear function. This technique is based on the estimation of the most suitable function to separate the data from each other. The hyperplane geometry is for a generalization of the plane into a different number of dimensions. The cases where the data cannot be separated in linear or linear classifiers are not complex enough sometimes, then the nonlinear classifiers can be used instead of the linear ones to map the data into a richer feature space including nonlinear ones by constructing a hyperplane in that space. The problem of the non-seperability of the data can be resolved through the addition of nonnegative and error-indicative loose variables into the optimization model [25]. So, an effective hybrid algorithm which is called as fuzzy c-means clustering based Support Vector Machines (SVM) has been developed by Esme & Karlik [26]. As shown in Fig.2, This hybrid algorithm consists of parallel combination of unsupervised fuzzy clustering and supervised SVM algorithms. A number of data points are reduced by FCM clustering before inputs are applied to SVM classifier.

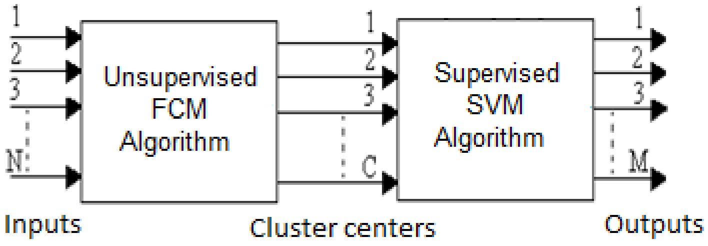

FIGURE 2: The architecture of FCM based SVM Algorithm.

Clustering is performed separately for each perfume class. For each perfume different cluster centers are obtained. In this way, FCM clustering is used to choose representative instances as well features of each perfume classes in original training set [27]. Table 3 shows the positive effect of Fuzzy C-Means (type-1) Clustering on two well-known classifiers (MLP and SVM).

Number of Training Sample	Number of Test Sample	Using ANN Accuracy (%)	Using FCNN Accuracy (%)	Using SVM Accuracy (%)	Using FCM based SVM Accuracy (%)
400	160	79,375	92,250	83,125	93,175

TABLE 3: Test results of classification accuracies for perfume data set.

5. CONCLUSION

For both in large numbers features and a great number of instances, classification applications are improved by multiple classes (or labels). The ever-increasing growth in data quantity and computational difficulty causes of bad the performance and accuracy of classification methods. In other words, the difficulty of classifier model is generally non proportional to its classification efficiency. This study has presented two hybrid classifier method for biomedical and odor applications which has a successful and better accuracy. This method is acquired with incorporating the algorithms of fuzzy c-means clustering technique and supervised classifiers such as back-propagation learning and SVM statistical learning, and comparison of their advantages and disadvantages. So, the results of the hybrid proposed algorithms, which is more beneficial structure than ordinary ANN which has Multi-layered perceptron (MLP) structure and Back-propagation training algorithm to classify both ECG and EMG biomedical data sets, are obtained. Ordinary ANN is still able to use for finding acceptable recognition accuracy. However, training period is quite long especially for big data. The goal in developing FCNN was to achieve more optimal results with relatively few data sets. The test results represent that FCNN hybrid algorithms are better than the ordinary ANN classifiers without requiring extra computational effort. In doing so, test results show that the type-2 fuzzy c-means was less responsive to noise than type-1 fuzzy c-means for application of biomedical data sets. Moreover, the second hybrid system developed using FCM based SVM has been proposed to increase the classification success of application of the perfume data set. It is also given better accuracy (93.175%) than the other hybrid (FCNN) classifier which was able to correctly classify as 92.25%.

In the future, this proposed method can also be used to solve the other applications such as industrials, environmental, agriculture, and medicine.

6. REFERENCES

[1] I.O. Bucak, B. Karlık, “Detection of Drinking Water Quality Using CMAC Based Artificial Neural Networks.” Ekoloji, vol. 20(78), pp. 75-81, 2011.

[2] B. Karlık, “Hepatitis Disease Diagnosis Using Backpropagation and the Naive Bayes Classifiers.” BURCH Journal of Science and Technology, vol. 1(1), pp. 49-62, 2011.

[3] B. Karlık, “Machine Learning Algorithms for Characterization of EMG Signals.” International Journal of Information and Electronics Engineering, vol. 4, no. 3, pp. 189-194, May 2014.

[4] B. Karlık, “Soft Computing Methods in Bioinformatics: A Comprehensive Review.” Mathematical & Computational Applications, vol.18, no.3, pp. 176-197, 2013.

[5] B. Karlık, H. Torpi, M. Alcı, “A Fuzzy-Neural Approach for the Characterization of the Active Microwave Devices.” Proceeding of CriMiCo’02, Sevastopol, Ukraine, September 9-14, 2002, pp. 9-14.

[6] R.H. Abiyev, O. Kaynak, “Identification and Control of Dynamic Plants Using Fuzzy Wavelet Neural Networks.” IEEE International Symposium on Intelligent Control (ISIC 2008), San Antonio, TX, USA, September 3-5, 2008, pp.1295-1301.

[7] M. Korürek, B. Doğan, “ECG Beat Classification Using Particle Swarm Optimization and Radial Basis Function Neural Network.” Expert Systems with Applications, vol. 37, no. 12, pp. 7563-7569, 2010.

[8] B. Doğan, M. Korürek, “A new ECG Beat Clustering Method Based on Kernelized Fuzzy C-Means and Hybrid Ant Colony Optimization for Continuous Domains.” Applied Soft Computing, vol. 12, no. 11, pp. 3442-3451, 2012.

[9] B. Karlık, O. Tokhi, M. Alcı, “A Fuzzy Clustering Neural Network Architecture for Multi-Function Upper-Limb Prosthesis.” IEEE Transactions on Biomedical Engineering, vol. 50, no. 11, pp. 1255-1261, 2003.

[10] Y. Özbay, R. Pektatlı, B. Karlık, “A Fuzzy Clustering Neural Network Architecture for Classification of ECG Arrhythmias.”, Computers in Biology and Medicine, vol. 36, pp.376–388, 2006.

[11] B. Karlık, K. Yüksek, “Fuzzy Clustering Neural Networks for Real Time Odor Recognition System.” Journal of Automated Methods and Management in Chemistry, Article ID 38405, doi:10.1155/2007/38405, Dec. 2007.

[12] B. Karlık, M. Korürek, Y. Koçyiğit, “Differentiating Types of Muscle Movements using Wavelet Based Fuzzy Clustering Neural Network.” Expert Systems, vol. 26(1), pp. 49-59, 2009.

[13] R. Ceylan, Y. Özbay, B. Karlık, “A Novel Approach for Classification of ECG Arrhythmias: Type-2 Fuzzy Clustering Neural Network.” Expert Systems with Applications, vol. 36, issue. 3, part. 2, pp. 6721-6726, 2009.

[14] R.H. Abiyev, O. Kaynak, “Type-2 Fuzzy Neural Structure for Identification and Control of Time-varying Plants.” IEEE Trans.on Industrial Electronics, vol.57(12), pp. 4147-4159, 2010.

[15] R. Ceylan, Y. Özbay, B. Karlık, “Telecardiology and Teletreatment System Design for Heart Failures Using Type-2 Fuzzy Clustering Neural Networks.” International Journal of Artificial Intelligence and Expert Systems vol. 1(4), pp. 100-110, 2011.

[16] R.H. Abiyev, O. Kaynak, T. Alshanableh, F. Mamedov, “A type-2 Neuro-Fuzzy System Based on Clustering and Gradient Techniques Applied to System Identification and Channel Equalization.” Applied Soft Computing, vol. 11(1), pp. 1396-1406, 2011.

[17] Y. Özbay, R. Ceylan, B. Karlık, “Integration of Type-2 Fuzzy Clustering and Wavelet Transform in a Neural Network Based ECG Classifier.” Expert Systems with Applications, vol. 38, pp. 1004-1010, 2011.

[18] R. Ceylan Rahime, Y. Özbay, B. Karlık, “Comparison of Type-2 Fuzzy Clustering Based Cascade Classifier Models for ECG Arrhythmias.” Biomedical Engineering: Applications, Basis and Communications (BME), vol. 26, no. 6, 2014-1450075, 2014.

[19] O. Ornek, A. Subasi, “Clustering Marketing Datasets with Data Mining Techniques.” The 2nd Inter. Symposium on Sustainable Development, Sarajevo, Bosnia and Herzegovina, June 8-9, 2010, vol. 3, pp: 408-412.

[20] J.C. Bezdek, R. Ehrlich, W. Full, “FCM: The Fuzzy C-Means Clustering Algorithm.” Computers & Geosciences, vol. 10, pp. 191–203, 1984.

[21] J. Fan, W. Zhen, and W. Xie, “Supervised Fuzzy C-Means Clustering Algorithm.” Elsevier Science Pattern Recognition Letters, vol. 24, pp. 1607-1612, 2003.

[22] F.C.H. Rhee, C. Hwang, “A Type-2 Fuzzy C-Means Clustering Algorithm.” IEEE Transaction on Neural Networks, vol. 9, no. 1, pp. 83-105, 2001.

[23] B. Karlık, “The Effects of Fuzzy Clustering on the Back-Propagation Algorithm.” International Conference on Computational and Applied Mathematics, Ukraine, Abstract Book, pp. 9-10 September, 2002, Kiev, Ukraine.

[24] B. Karlik, “Differentiating Type of Muscle Movement via AR Modeling and Neural Networks Classification." Turk J Elec Eng & Comp Sci, vol. 7, pp. 45-52, 1999.

[25] I.O. Bucak, “Performance Evaluation of Neural Classifiers through Confusion Matrices to Diagnose Skin Conditions.” International Journal of Artificial Intelligence and Expert Systems (IJAE), vol.5, Issue: 2, pp. 15–27, 2014.

[26] E. Esme and B. Karlık, “FCM Based SVM Classifier for Perfume Recognition.” Applied Soft Computing (re-submitted).

[27] B. Karlık, Y. Bastaki, “Real Time Monitoring Odor Sensing System Using OMX-GR Sensor and Neural Network.” WSEAS Transactions on Electronics, issue 2, vol.1, pp.337-342, 2004.