International Journal of
Mathematical Modelling & Computations
Vol. 03, No. 02, 2013, 109- 114

Fuzzy K-Nearest Neighbor Method to Classify Data in a Closed Area

M. Amirfakhrian and S. Sajadi*

Department of Mathematics, Islamic Azad University, Central Tehran Branch, PO. Code 14168-94351, Iran.

Received: 25 February 2013; Accepted: 3 June 2013.

Abstract. Clustering of objects is an important area of research and application in variety of fields. In this paper we present a good technique for data clustering and application of this technique for data clustering in a closed area. We compared this method with K-nearest neighbor and K-means.

Keywords: Artifical Neural Networks, Clustering, Fuzzy K-nearest Neighbor, K-nearest neighbor, K-means.

Index to information contained in this paper

1. Introduction

Clustering of objects is an important area of research and of practical applications in variety of fields, including pattern recognition and artifical intelligence, statistics, vision analysis and medicine. The K-nearest neighbor (K-NN) algorithm is used to perform the classification [13]. This decision rule provides a simple nonparametric procedure for the assignment of a class label to the input pattern based on the class labels represented by the K-closest [14]. But this method has some problems. One of the problems encountered in using the K-NN classifier is that normally each of the sample vectors is considered equally important in the assignment of the class label to the input vector [13]. This frequently causes difficulty in those places where the sample sets overlap. A typical vectors representing the weight of each cluster is given. Another difficulty is that once an input vector is assigned to a class, there is no indication of its strength of membership in that class. To solve

*Corresponding author. Email: ssajadi29@gmail.com

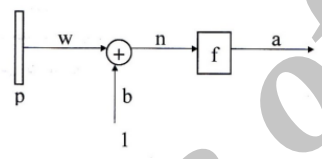
this problems, fuzzy K-NN applied [3, 7].
A fuzzy K-NN algorithm is developed utilizing fuzzy class memberships of the sample sets and thus, producing a fuzzy classification rule [11].
The paper is organized as follows: In Section 2 and 3, we recall artificial neural network and clustering respectively. In Section 4 we apply F-KNN for data clustering in a closed area and compare this technique with K-nearest neighbor classification and K-means [15].

2. Artificial neural network

Artificial neural network (ANN) is a network of interrelated elements that are inspired by natural neural systems. The smallest unit of information processing is a neuron, that will form the basis of function neural networks. Neural networks provide an appropriate output according to the input patterns fed to the network [1, 4].
Characteristics of neural networks are: Ability to learn, distribution of information, generalization, parallel processing, [9].

2.1 *Mathematical structure of neurons and neural network architecture*

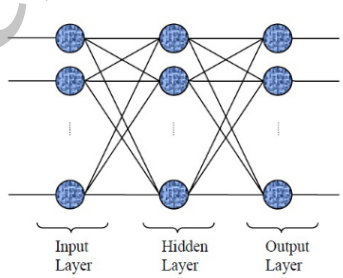
A simple mathematical model of a neuron is shown in Figure 1.



(a)

Figure 1. Neuron

Different structures of neural networks can be made through various combination of the neurons together [2, 12]. A multilayer neural network is shown in Figure 2.



(a)

Figure 2. multilayer network

Typically, models of neural networks are divided into categories in terms of signal transmission manner: Feed-forward neural networks and recurrent neural networks. They are built up using different frameworks, which give rise to different fields of applications [10].

3. Clustering

Clustering is considered as a pre-processing phase design of neural networks.
"Clustering" is a process to obtain a partition P of a set E of N objects x_i $(i = 1, 2, \ldots, N)$, using the resemblance or disemblance measure, such as a distance measure d. A partition P is a set of disjoint subsets of E and the element P_s of P is called a *cluster* and the centers of the clusters are called *centroids* or *prototypes* [8]. The goal of a clustering analysis is to divide a given set of data or objects into a cluster, which represents subsets or a group. The partition should have two properties:

- Homogeneity inside clusters: the data, which belongs to one cluster should be as similar as possible.
- Heterogeneity among the clusters: the data, which belongs to different clusters should be as different as possible.

Many techniques have been developed for clustering data. In this paper, Fuzzy K-nearest neighbor (F-KNN) clustering is used [8, 13].

3.1 *Fuzzy K-nearest neighbor method*

According to what stated in Section 1, to resolve the defects of K-nearest neighbors, we use fuzzy K-nearest neighbor.

In this technique, we consider that membership function is as a distance function. In fact, the fuzzy K-nearest neighbor algorithm assigns class membership to a sample vector rather than assigning the vector to a particular class. The advantage is that no arbitrary assignments are made by the algorithm. In addition, the vectors membership values should provide a level of assurance to accompany the resultant classification.

The basis of this algorithm is to assign membership as a function of the vectors distance from its K-nearest neighbors and those neighbors memberships in the possible classes [7, 16]. Let $W = \{x_1, x_2, \ldots, x_N\}$ be a set of N labeled samples. Also, let $u_i(x)$ be the assigned membership of the vector x, and u_{ij} be the membership in the ith class of the jth vector of the labeled sample set.
$u_i(x)$ is computed through statment 1.

$$u_i(x) = \frac{\sum_{j=1}^{K} u_{ij}\left(\frac{1}{\| x - x_j \|}\right)^{\frac{2}{m-1}}}{\sum_{j=1}^{K}\left(\frac{1}{\|x - x_j\|}\right)^{\frac{2}{m-1}}}, \tag{1}$$

where variable m determines how heavily the distance is weighted when calculating each neighbor's contribution to the membership value. As seen by (1), the assigned memberships of x are influenced by both the inverse of the distances from the nearest neighbors and their class memberships. The inverse distance serves to weight a vector's membership more if it is closer and less if it is farther from the vector under consideration [6, 7, 14]. u_{ij} is an optimal solution that is obtained by solving the following nonlinear programming [5].

$$\min Q = \sum_{i=1}^{N}\sum_{k=1}^{M} u_{ik}^{m}(\| x_k - v_i \|)^2 \tag{2}$$

$$subject\ to \quad \sum_{i=1}^{N} u_{ik} = f_k,$$

$$0 < \sum_{k=1}^{M} u_{ik} < M,$$

$$0 < u_{ik} < 1.$$

Then,

$$u_{ij} = \frac{f_k}{\sum_{j=1}^{C}(\frac{\|x_k - v_i\|}{\|x_k - v_j\|})^{\frac{2}{m-1}}}. \tag{3}$$

We arrive at the (3) by transforming (2) to a standard unconstrained optimization by making use of Lagrange multipliers and determining a critical point of the resulting function. In fact, u_{ij} is a local minimum or a saddle point of this function [11].

4. Data clustering in a closed area

In this Section, we supposed that data set is a closed area such as a square $[0, 2] \times [0, 2]$. Then, we are clustering the data randomly with both of K-means and K-nearest neighbor techniques. Finally, we compare the obtained results of these methods with fuzzy K-nearest neighbor.

K-means clustering is a method commonly used to automatically partition a data set into K groups. It proceeds by selecting K initial cluster centers and then iteratively refining them as follows:

1 Each instance d_i is assigned to its closest cluster center.
2 Each cluster center C_j is updated to be the mean of its constituent instances.

The algorithm converges when there is no further change in assignment of instances to clusters. We initialize the clusters using instances chosen at random from the data set [15].
The K-nearest neighbor is conceptually simple. We compute the distance from an observation y_i to all other points y_j using the distance function

$$(y_i - y_j)' S_{pl}^{-1}(y_i - y_j), \quad i \neq j$$

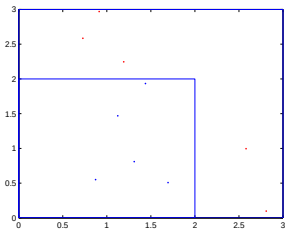
where S_{pl} is the mixed variance [13].

To classify y_i into one of two groups, the K points nearest to y_i are examined, and if the majority of the K points belong to G_1, assign y_i to G_1; otherwise assign y_i to G_2. If we denote the number of points from G_1 as K_1, with the remaining K_2

points from G_2, where $K = K_1 + K_2$, then the rule can be expressed as: Assign y_i to G_1 if

$$K_1 > K_2$$

and to G_2 otherwise [3, 13]. These rules are easily extended to more than two groups. So, we are clustering data with these methods.
Now, we use the fuzzy K-nearest neighbor method to classify points inside a closed area [17]. Data set is shown in Figure 3.



(a)

Figure 3. Data set in a closed area

After comparing these methods, we will see with choosing $K = 1$, all data will allocate to a just one cluster. In the case $K = 2n$, $n \in \mathbb{N}$, data can't classify, because the amount of assignment to each cluster is equal. In the case $K = 2n + 1$, $n \in \mathbb{N}$, data can classify easily and due to each method, clustering will have done. Clusters created through these methods and their comparison are shown in Table 1. In this table $M.\ V.\ F\text{-}KNN$, $F\text{-}KNN$, KNN, K-means denote Membership values of $F\text{-}KNN$, Predicted classes of $F\text{-}KNN$, Predicted classes of KNN and Predicted classes of K-means respectively.

Table 1. Data set in a square

Data	K	$M.\ V.\ F\text{-}KNN$	$F\text{-}KNN$	KNN	K-means
0.3193	3	0.6807	2	1	1
0.7400	3	0.2600	1	1	1
0.3801	3	0.6199	2	1	1
0.3705	3	0.6295	2	1	1
0.1189	3	0.8811	2	1	1
0.5317	3	0.4683	1	1	1
0.3705	3	0.6295	2	1	1
0.1788	3	0.8212	2	1	1
0.5643	3	0.4357	1	1	1
0.8113	3	0.1887	1	1	1

Now, for example we are clustering data in a circle. We choose data randomly in a circle with radius 2 and center $(0, 0)$. The equation of this circle is $x^2 + y^2 \leqslant 9$. We are clustering this data with $F\text{-}KNN$, KNN and K-means methods. At the end of this work, we compare the output.
The result of comparison is shown in Table 2.
According to this table, data classified to 3 clusters with own membership value by $F\text{-}KNN$ method and clssified to 2 classes by KNN and K-means.

Table 2. Data set in a circle

Data	K	$M.V.F\text{-}KNN$	$F\text{-}KNN$	KNN	K-means
0.3093	3	0.7107	3	1	1
0.5678	3	0.4807	2	2	1
0.8201	3	0.1599	1	2	1
0.4255	3	0.6239	2	1	2
0.1009	3	0.9081	3	1	1
0.6377	3	0.4853	2	2	1
0.6410	3	0.3240	1	2	1
0.1387	3	0.8812	3	1	1
0.3915	3	0.7995	3	1	2
0.8196	3	0.1917	1	2	1

5. Conclusion

In this work, we introduced fuzzy K-nearest neighbor and used it to clustering data in a closed area, then we compared this method with K-means and K-nearest neighbor. The advantage of this method is that no arbitrary assignments are made. Moreover, the clusters are created during this process to maintain its homogeneity property.

References

[1] J. A. Anderson, *An introduction to neural networks*, Cambridge, MA: Mit press, 1995.
[2] C.M. Bishop, *Neural Networks for pattern Recognition*, Oxford University Press, New York, USA, 1995.
[3] T.M. Cover, *Estimation by the nearest neighbor rule*, IEEE Trans. Inf. Theory IT, **14 (1)** (1968) 50-55.
[4] L. Fausett, *Fundamental of Neural Networks: Architecture, Algorithms and Appliction*, Prentice Hall, 1994.
[5] R. Fletcher, *Practical methods of optimization*, Department of mathemtacs University of Dundee, Scatland. UK. Wiley, New York, second edition.
[6] A. Jozwik, *A 'learning scheme for fuzzy $K-NN$ rule' Pattern Recognition Letters*, **1** (1983) 287-289.
[7] J. M. Keller, M. R. Gray, and J. A. Givens, *A fuzzy K- nearest algorithm*, IEEE, Trans. Syst. Man Sybern, SMC-**15 (4)** (1985) 580-585.
[8] R. Krasteva, *Bulgarian hand- printed character recognition using fuzzy C-means clustering*, Central laboratory of mechatranics and instrumentation, Bulgarian Academy of Sciences, 1113 Sofia, 112-117
[9] K. Mehrotra, C. K. Mohan, S. Ranka, *Elements of Artificial Neural networks*, Bradford books.
[10] P. Pikthin, *Base and application of neural networks*, Macmillan Publishing Co., Inc. Indianapolis, IN, USA. 2000.
[11] W. Pedrycz, *Conditional fuzzy clustering in the design of radial basis function neural networks*, IEEE, Trans, Neural Networks, **9** (4) (1998) 601-612.
[12] P. Peretto, *An introduction to the modelling of neural networks*, Cambridge, New York: Cambridge University Press, 1992.
[13] A. C. Rencher, *Methods of multivariate analysis*, Brigham Young University, Wiley, New york, second edition, 1934.
[14] S. B. Roh, T. C. Ahn, W. Pedrycz, *The design methodology of radial basis function neural networks based on fuzzy K- nearest neighbors approach*, Fuzzy Set and Systems, **161** (2010) 1803-1822.
[15] S. Rogers, S. Schroedl, *Constrained K-means clustering with Bachground Knowledge*, proceedings of the Eighteenth Conference on machine learning, (2001) 577-584.
[16] A. Staiano, R. Tagliaferri, W. Pedrycz, *Improving RBF networks performance in regression tasks by means of a supervised fuzzy clustering*, Neuro Computing, **69** (2006) 1570-1581.
[17] C. Zhu, T. Jing, *Multicontext Fuzzy Clustering For Sepration Of Brain Tissues In Magnetic Resonance Images*, Science Direct Neuroimage, **18** (2003) 685-696.