International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:10, No:4, 2016

# Intelligent Recognition of Diabetes Disease via FCM Based Attribute Weighting

Kemal Polat

***Abstract*—**In this paper, an attribute weighting method called fuzzy C-means clustering based attribute weighting (FCMAW) for classification of Diabetes disease dataset has been used. The aims of this study are to reduce the variance within attributes of diabetes dataset and to improve the classification accuracy of classifier algorithm transforming from non-linear separable datasets to linearly separable datasets. Pima Indians Diabetes dataset has two classes including normal subjects (500 instances) and diabetes subjects (268 instances). Fuzzy C-means clustering is an improved version of K-means clustering method and is one of most used clustering methods in data mining and machine learning applications. In this study, as the first stage, fuzzy C-means clustering process has been used for finding the centers of attributes in Pima Indians diabetes dataset and then weighted the dataset according to the ratios of the means of attributes to centers of theirs. Secondly, after weighting process, the classifier algorithms including support vector machine (SVM) and k-NN (k- nearest neighbor) classifiers have been used for classifying weighted Pima Indians diabetes dataset. Experimental results show that the proposed attribute weighting method (FCMAW) has obtained very promising results in the classification of Pima Indians diabetes dataset.

***Keywords*—**Fuzzy C-means clustering, Fuzzy C-means clustering based attribute weighting, Pima Indians diabetes dataset, SVM.

## I. Introduction

IN real world datasets, the data distribution of datasets cannot always have a linearly separable dataset due to noise. Therefore, the data preprocessing methods should be used prior to classification process to increase the classification performance. In this paper, a new data preprocessing method called subtractive clustering based attribute weighting has been proposed to classify the Pima Indians diabetes disease dataset that is to be real world application.

Clustering is a process appointing the patterns into a set of groups, where such that patterns within a group are more similar to each other than are patterns belonging to different clusters. Clustering has been applied in a wide variety of fields, ranging from engineering, computer sciences, life and medical, to earth sciences, social sciences, and economics [1]-[3].

Data preprocessing methods are used for applications including noise removing, outlier detection, data normalization, missing value finding etc. These methods are employed prior to classification process in classification of datasets. In literature, there are a lot of data pre-processing methods. Among these, Polat et. al. proposed a new attribute weighting method based on fuzzy logic and called fuzzy weighting method in classification of ECG dataset [4]. Polat et al. have proposed k-NN (k-nearest neighbor) based attribute weighting method to reduce the variance within attribute s in dataset and applied to medical datasets [5]-[7]. Polat et al. suggested a novel attribute weighting method based on similarity measure between attribute s and applied to classification of Doppler signals to diagnose Atherosclerosis disease [8]. Dua et al. present a method based on the connected component theory to remove the labels from the image and crop the image to a relevant reduced size as data preprocessing method [9], [10].

As far as we know, there are a lot works related with classification of Pima Indians Diabetes dataset in literature. When the studies in the literature related with this classification application are examined, it can be seen that a great variety of methods were used which reached high classification accuracies. Among these, Statlog obtained 77.7%, 77.6%, 76.8%, and 75.7% classification accuracies using Logdisc, DIPOL92, SMART, and RBF with 10-fold cross validation [11]. Norbert Jankowski achieved 77.6% classification accuracy with 10-fold cross validation [11]. Ster and Dobnikar accomplished 77.5%, 76.6%, 76.5%, 76.4%, 75.8%, and 75.8% using Linear Discriminant Analysis, ASI, Fisher Discriminant Analysis, MLP+BP, LVQ, and LFC with 10-fold cross validation [11]. 76.8% classification accuracy was achieved using GTO DT with 10-fold cross validation by Bennet and Blue [11]. Zarndt obtained 75.8% classification accuracy using MLP+BP by way of 10-fold cross validation [11]. Ster and Dobnikar obtained 75.5% classification accuracy using MLP+BP by way of 10-fold cross validation [11]. Karol Grudziński obtained 75.5% classification accuracy using k-NN (k=22) by way of 10-fold cross validation [11]. AIRS classifier algorithm obtained 79.22% classification accuracy with 10-fold cross validation, Fuzzy-AIRS achieved 84.42% accuracy for classification of Pima Indians Diabetes dataset conducted by Polat et al. [12].

In this study, a clustering based attribute weighting method with fuzzy c-means clustering have been proposed for classification of diabetes disease dataset that has a non-linearly separable distribution. In this study, firstly, fuzzy c-means clustering based attribute weighting has been employed to diabetes disease dataset and then weighted diabetes disease dataset according to cluster centers of attributes. Secondly, after weighting process, the classification algorithms have

Kemal Polat (Professor) is with the Abant Izzet Baysal University, Department of Electrical and Electronic Engineering, 14280, Bolu, Turkey (e-mail: kpolat@ibu.edu.tr).

been applied to weighted diabetes disease dataset. As classifier algorithms, support vector machine and k-NN classifier have been used. While only SVM and k-NN classifiers obtained 73.18% and 72.92% classification accuracies, the combinations of FCMAW-SVM and FCMAW - k-NN achieved the success rates of 91.41% and 84.38% on the classification of Pima Indians diabetes dataset, respectively.

## II. Material

Diabetes is a metabolic disease in which there is a deficiency or absence of insulin secretion by the pancreas. Diabetes occurs in two major forms: type I, or insulin-dependent diabetes, and type II, or non-insulin-dependent diabetes. Most studies of the age of onset of type I diabetes have been restricted to children and adults under 30 years old. The age distributions in females and males show small differences which have been most clearly demonstrated in surveys in children and young adults. Both genders are affected, but in many communities the majorities with type II diabetes are female. The prevalence of type II diabetes rises with increasing age and in many populations the majority of those with diabetes are either middle-aged or elderly. Diabetes may be considered as a disorder of the metabolic disposal of food. The interaction of food and diabetic state must be assessed from two aspects: first, whether food precipitates the diabetic condition and, second, the type of food that is most appropriate for the person with established diabetes, whether it is insulin-dependent or non-insulin dependent. Diabetes shows considerable familial aggregation which may result from either the inheritance of disease susceptibility or the sharing of a common environment by members of the same family. Therefore, it is important to determine the existence of diabetes in families of the subjects. The occurrence of gestational diabetes, which either first appears or is first recognized during pregnancy, is associated with increased risk for the development of diabetes in subsequent years. Thus, women who experience gestational diabetes should be considered a high risk group for the development of type II diabetes [13]. In this study, the Pima Indians diabetes database [14] was analyzed. The data consist of 768 records and according to the examination results 268 of them are diabetics and the rest of them are non-diabetics. Each record has eight attributes and these are detailed in Table I. Eight independent input parameters, essentially risk factors for diabetes, are incorporated in the classifiers [15]. This dataset was taken from UCI machine learning database [16].

TABLE I
Pima Indians Diabetes Dataset: Description of Attributes [16]

| Attribute description | Minimum Value | Maximum Value | Average | Standard Deviation |
|---|---|---|---|---|
| **1.**Number of times pregnant | 0 | 17 | 3,845 | 3,3695 |
| **2.**Plasma glucose concentration a 2 hours in an oral glucose tolerance test | 0 | 199 | 120,89 | 31,97 |
| **3.**Diastolic blood pressure (mm Hg) | 0 | 122 | 69,10 | 19,3558 |
| **4.**Triceps skin fold thickness (mm) | 0 | 99 | 20,53 | 15,9522 |
| **5.**2-Hour serum insulin (mu U/ml) | 0 | 846 | 79,79 | 115,244 |
| **6.**Body mass index | 0 | 67,1 | 31,99 | 7,88416 |
| **7.**Diabetes pedigree function | 0,078 | 2,42 | 0,471 | 0,33132 |
| **8.**Age | 21 | 81 | 33,24 | 11,760 |
| **9.**Number of times pregnant | 0 | 17 | 3,845 | 3,3695 |

*Note: 768 observations comprising 500 normal and 268 patient (diabetes) subjects

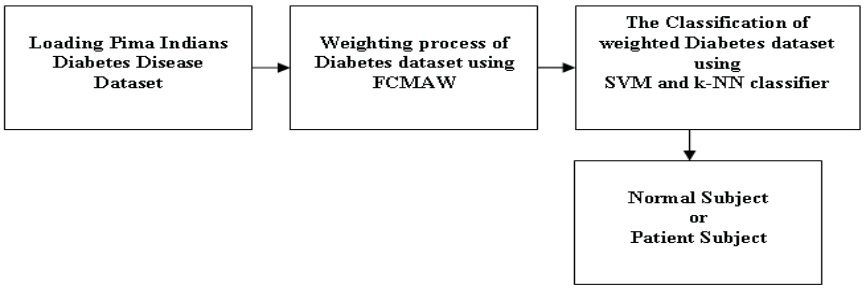


Fig. 1 The block diagram of proposed method

## III. Method

### A. The Overall System

In this study, a new attribute weighting method called FCM based attribute weighting has been proposed for classifying Pima Indians diabetes disease. As classifier algorithm, support vector machine (SVM) and k-NN (k-nearest neighbor) classifiers have been used to classify the weighted diabetes disease dataset. The proposed hybrid method is demonstrated in Fig. 1.

*B. Fuzzy C-Means Clustering Based Attribute Weighting (FCMFW)*

Clustering algorithms are used widely not only to collect similar or dissimilar data, but also useful for data compression and data reduction. The most used clustering algorithms are K-means clustering [17], fuzzy C-means clustering [18], the mountain clustering [19], and subtractive clustering [20]. Among these clustering methods, the fuzzy C-means clustering (FCM) is chosen in weighting process since FCM is the extend version of K-means clustering and is mostly used in applications and literature [21].

The goal of attribute weighting method is to map the attribute s according to their distributions in a dataset and also transform from non-linearly separable dataset to linearly separable dataset. Attribute weighting method works based upon principle that decreasing the variance in attributes forming dataset. Thanks to this weighting method, the similar data in same attribute are gathered and the discrimination ability of classifier is increased. In this study, a new weighting method called FCM clustering based attribute weighting is proposed. The fuzzy C-means is briefly explained and then explained the proposed weighting method.

Fuzzy C-means is one of the most popular fuzzy clustering algorithms [22]. FCM was realized by [18]. FCM attempts to find a partition for a set of data points $x_j \in R^d$, $j = 1,...N$ while minimizing the cost function (1) [21];

$$J(U,M) = \sum_{i=1}^{c} \sum_{j=1}^{N} (u_{i,j})^m D_{ij} \quad (1)$$

where $U = [u_{i,j}]_{cxN}$ is the fuzzy partition matrix and $u_{i,j} \in [0,1]$ is the membership coefficient of the jth object in the ith cluster [21]; $M = [m_1,...,m_c]$ is the cluster prototype matrix; $m \in [1,\infty]$ is the fuzzification parameter; $D_{ij} = D(x_j, m_i)$ is the distance measure between $x_j$ and $m_i$.

The FCM can be briefly summarized the FCM as follows, in which the Euclidean or $L_2$ norm distance function is used [21], [23]:

a) Select the cluster centers $c_i, i = 1,2,...,c$ randomly from the n points $\{X_1, X_2, X_3,...., X_n\}$.

b) Compute the membership matrix U using (2);

$$u_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{d_{ij}}{d_{kj}}\right)^{2/(m-1)}} \quad (2)$$

where $d_{ij} = \left\| c_i - x_j \right\|$ is the Euclidean distance between ith cluster center and jth data point, m is the fuzziness index.

c) Compute the cost function according to the following equation. Stop the process if it is below a certain threshold;

$$J(U, c_1,..., c_c) = \sum_{i=1}^{c} J_i = \sum_{i=1}^{c} \sum_{j=1}^{n} \mu_{ij}^m d_{ij}^2 \quad (3)$$

d) Compute new c fuzzy cluster centers $c_i, i = 1,2,...,c$ using;

$$c_i = \frac{\sum_{j=1}^{n} \mu_{ij}^m \ X_j}{\sum_{j=1}^{n} \mu_{ij}^m} \quad (4)$$

Otherwise go to (b).

This weighting method works as follows: firstly, the cluster centers are calculated using FCM method. After computing the centers of attribute s, the ratios of means of attribute s to their centers are calculated and these ratios are multiplied with data of each attribute.

*C. Used Classifier Algorithms: Support Vector Machine (SVM) and k-NN Classifier*

After FCMAW process, the classifier algorithms including SVM and k-NN classifiers are used. These algorithms have been explained in the following subsections. In the training and testing of classifiers, 50-50% training-testing split is used.

SVM is a reliable classification technique, which is based on the statistical learning theory. This technique was firstly proposed for classification and regression tasks by [24].

It is a method for creating functions from a set of labeled training data. The function can be a classification function (the output is binary: is the input in a category) or the function can be a general regression function. For classification, SVMs operate by finding a hypersurface in the space of possible inputs. This hypersurface will attempt to split the positive examples from the negative examples. The split will be chosen to have the largest distance from the hypersurface to the nearest of the positive and negative examples. Intuitively, this makes the classification correct for testing data that is near, but not identical to the training data [25].

The k-nearest neighbor (k-NN) algorithm is one of the simplest algorithms among all machine learning algorithms. In k-nearest-neighbor classification, the training dataset is used to classify each data of a "target" dataset. The structure of the data is that there is a classification variable of interest and a number of additional predictor variables. k-NN algorithm works as follows [26]-[28]:

a) For each row in the target dataset (the set to be classified), locate the k closest members (the k nearest neighbors) of the training dataset. A Euclidean Distance measure is used to calculate how close each member of the training set is to the target row that is being examined.

b) Examine the k nearest neighbors - which classification do most of them belong to? Assign this category to the row being examined.
c) Repeat this procedure for the remaining rows in the target set.
d) The user selects a maximum value for k, builds models parallel on all values of k up to the maximum specified value and scoring is done on the best of these models.

## IV. Results and Discussion

In this paper, fuzzy C-means clustering attribute weighting process has been suggested and applied to classification of Pima Indians diabetes dataset that is a real world medical application. The aim of this weighting method is to transform from non-linearly separable dataset to linearly separable dataset. The working principle of FCMAW methods is as follows: (i) the centers of attributes in dataset were found with fuzzy c-means clustering (ii) the ratios of means of attributes to their centers were calculated and these ratios were multiplied to each attribute in dataset. Two stage classification including data preprocessing and classification has been used. In the first step, fuzzy c-means clustering based attribute weighting has been separately employed to Pima Indians diabetes dataset and then weighted Pima Indians diabetes dataset according to cluster centers of attributes. In the second step, later weighting process, the classifier algorithms have been applied to weighted Pima Indians diabetes dataset and then classified the Pima Indians diabetes dataset as normal or patient. As classifier algorithm, SVM and k-NN classifiers have been used. In classification of Pima Indians diabetes dataset, the 50-50% training–testing split has been used. In order to see the efficiency of FCMAW, the data distributions of raw and weighted Pima Indians diabetes dataset were given. Fig. 2 shows the data distribution of Pima Indians diabetes dataset according to first three attribute s (1st, 2nd, and 3rd attributes) and the distribution of weighted Pima Indians diabetes dataset with FCM based attribute weighting method according to first three attribute s (1st, 2nd, and 3rd attributes). As it can be seen from this Fig. 2, the linearity of raw Pima Indians diabetes dataset was increased by using of FCMAW.

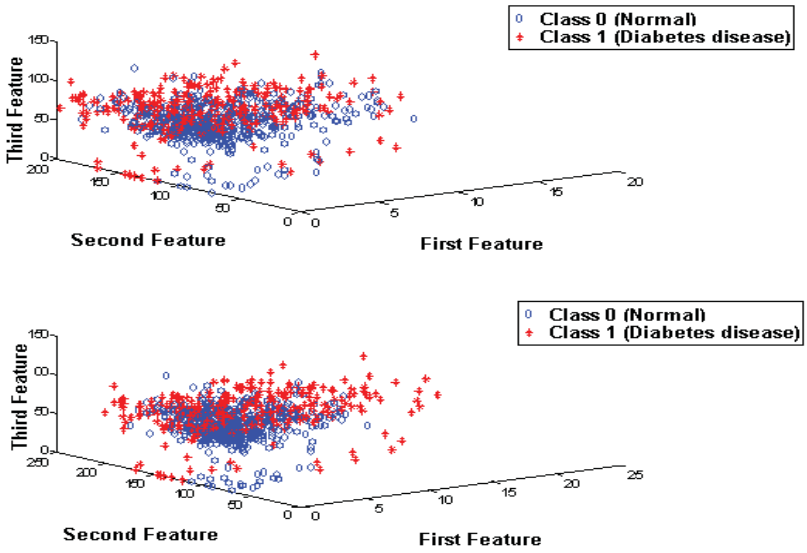


Fig. 2 (a) The distribution of Pima Indians diabetes dataset according to first three attributes (1st,2nd, and 3rd), (b) The distribution of weighted Pima Indians diabetes dataset with FCM based attribute weighting method according to first attributes attribute s (1st, 2nd, and 3rd)

TABLE II
The Obtained Results from SVM, k-NN, Combination of FCMAW and SVM, and Combination of FCMAW and k-NN (for k=50) Using 50-50% Training-Testing Dataset Split of Diabetes Dataset

| Used Method | Classification Accuracy (%) | Sensitivity (%) | Specificity (%) | AUC value |
|---|---|---|---|---|
| SVM classifier | 73.18 | 64.22 | 73.90 | 0.683 |
| k-NN classifier (for k of 50 value) | 7.92 | 77.78 | 72.12 | 0.632 |
| Combination of FCMAW and SVM | **91.41** | 89.76 | 92.21 | 0.899 |
| Combination of FCMAW and k-NN (for k of 50 value) | **84.38** | 87.75 | 83.21 | 0.796 |

The obtained results show that the adding of FCMAW to classifier algorithms is an effective and robust solution for classification of Pima Indians diabetes disease.

In order to test the success of the proposed method, the classification accuracy, sensitivity, specificity, and AUC (area under the ROC curve) values have been used on recognition of Pima Indians diabetes dataset. Table II presents the

obtained results the from SVM, k-NN, combination of FCMAW and SVM, and combination of FCMAW and k-NN using 50-50% training-testing dataset split of diabetes dataset.

## V. Conclusions

Classification of real world medical datasets is a difficult task due to these datasets having noise, nonlinearly separable datasets, outlier point, missing data etc. Therefore, the data preprocessing methods should be used prior to classifier algorithms on classification of medical datasets. As medical application, Pima Indians diabetes dataset has been classified using proposed method based on combining FCMAW and classifier algorithms (SVM and k-NN). In this paper, for data preprocessing method, fuzzy C-means clustering based attribute weighting method has been proposed and combined with classifiers to classify the diabetes disease dataset. The purpose of this method was to transform from non-linearly separable medical datasets to linearly separable medical datasets. By this way, the superior results have been obtained for diabetes disease dataset. The proposed attribute weighting method could be confidently used as real time medical diagnosis applications.

## References

[1] R. Hathaway and J. Bezdek, Fuzzy c-means clustering of incomplete data, *IEEE Trans. Syst., Man, Cybern.*, 31(5), 2001, 735–744.
[2] B. Everitt, S. Landau, and M. Leese, Cluster Analysis. London: Arnold, 2001.
[3] J. Hartigan, Clustering Algorithms. New York: Wiley, 1975.
[4] Polat, K., Şahan, S., Güneş, S., A new method to medical diagnosis: Artificial immune recognition system (AIRS) with fuzzy weighted pre-processing and application to ECG arrhythmia, *Expert Systems with Applications*, 31(2), 2006, 264-269.
[5] Polat, K., Şahan, S., Güneş, S., Automatic detection of heart disease using an artificial immune recognition system (AIRS) with fuzzy resource allocation mechanism and k-nn (nearest neighbour) based weighting preprocessing, *Expert Systems with Applications*, 32(2), 2007, 625-631.
[6] Polat, K., Güneş, S., A hybrid medical decision making system based on principles component analysis, k-NN based weighted pre-processing and adaptive neuro-fuzzy inference system, *Digital Signal Processing*, 16(6), 2006, 913-921.
[7] Polat, K., Güneş, S., The effect to diagnostic accuracy of decision tree classifier of fuzzy and k-NN based weighted pre-processing methods to diagnosis of erythemato-squamous diseases, *Digital Signal Processing*, 16(6), 2006, 922-930.
[8] Polat K, Latifoğlu F, Kara S, Güneş S, Usage of Novel Similarity Based Weighting Method to Diagnose the Atherosclerosis from Carotid Artery Doppler Signals, *Medical & Biological Eng. & Computing*, 46, 2008, 353-362.
[9] Haralick and Shapiro, 1992 R.M. Haralick and L.G. Shapiro, Computer and robot vision Vol. 1, Addison-Wesley (1992).
[10] Dua, S., Singh, H., Thompson, H.W., Associative classification of mammograms using weighted rules, *Expert Systems with Applications*, 36(5), 2009, 9250-9259.
[11] Datasets used for classification comparison of results. http://www.phys.uni.torun.pl/kmk/projects/datasets.html (last accessed: 2016).
[12] Polat, K., Güneş, S., An improved approach to medical data sets classification: artificial immune recognition system with fuzzy resource allocation mechanism, *Expert Systems*, 24(4), 252-270 (2007).
[13] Besser, G.M., H.J. Bodansky and A.G. Cudworth (1988) Clinical Diabetes: An Illustrated Text, London: Gower Medical.
[14] www.cormactech.com/neunet, (last accessed: 2016).
[15] Ubeyli, E.D. Comparison of different classification algorithms in clinical decision-making, *Expert Systems*, 24(1), 2007, 17-31.
[16] UCI machine learning database, ftp://ftp.ics.uci.edu/pub/machine-learning-databases (last accessed: 2016).
[17] MacQueen, J. B., Some Methods for classification and Analysis of Multivariate Observations, Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, University of California Press, 1967, 1:281-297.
[18] Bezdek, J. C., Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum Press, 1981, New York
[19] Yager, R. R., Filev, D. P., Generation of fuzzy rules by mountain clustering, *IEEE Transactions on Systems, Man and Cybernetics*, (1994), 24, 209–219.
[20] Chiu, S. L., Fuzzy model identification based on cluster estimation, Journal of Intelligent and Fuzzy Systems, (1994), 2.
[21] Xu R., Wunsch II D., Survey of Clustering Algorithms, *IEEE Transactions on Neural Networks*, 2005, 16(3) 645:678.
[22] Höppner, F., Klawonn, F., and Kruse, R., Fuzzy Cluster Analysis: Methods for Classification, Data Analysis, and Image Recognition. New York: Wiley, 1999.
[23] Guldemır, H., Sengur, A., Comparison of clustering algorithms for analog modulation classification, *Expert Systems with Applications*, 30(4), 2006, 642-649.
[24] V. Vapnik, 1995. The Nature of Statistical Learning Theory, Springer, New York.
[25] http://research.microsoft.com/~jplatt/svm.html (last arrived: 2016).
[26] Dasarathy, B. V., editor (1991) Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques, ISBN 0-8186-8930-7.
[27] Shakhnarovich, G., Darrell, T., Indyk, P., Nearest-Neighbor Methods in Learning and Vision, The MIT Press, 2005, ISBN 0-262-19547-X
[28] http://www.resample.com/xlminer/help/k-NN/knn_intro.htm, (last accessed: 2016).