An Effective Sample Preparation Method for Diabetes Prediction

Shima Afzali and Oktay Yildiz
Computer Engineering Department, Gazi University, Turkey

Abstract: *Diabetes is a chronic disorder caused by metabolic malfunction in carbohydrate metabolism and it has become a serious health problem worldwide. Early and correct detection of diabetes can significantly influence the treatment process of diabetic patients and thus eliminate the associated side effects. Machine learning is an emerging field of high importance for providing prognosis and a deeper understanding of the classification of diseases such as diabetes. This study proposed a high precision diagnostic system by modifying k-means clustering technique. In the first place, noisy, uncertain and inconsistent data was detected by new clustering method and removed from data set. Then, diabetes prediction model was generated by using Support Vector Machine (SVM). Employing the proposed diagnostic system to classify Pima Indians Diabetes data set (PID) resulted in 99.64% classification accuracy with 10-fold cross validation. The results from our analysis show the new system is highly successful compared to SVM and the classical k-means algorithm & SVM regarding classification performance and time consumption. Experimental results indicate that the proposed approach outperforms previous methods.*

Keywords: *Diabetes, clustering, classification, K-means, SVM, sample preparation.*

Received November 28, 2015; accepted February 3, 2016

1. Introduction

Diabetes is a disorder of metabolism and a chronic condition in which patient's body fails to produce enough insulin. Insulin is a hormone made by the pancreas that helps your body to extract and store glucose from carbohydrates in the food. Diabetes is a common disease that affects every part of the body and may lead to disability and even death in many cases [5].

Physicians have to analyse several critical factors and conduct time consuming tests before diagnosing diabetes. Nowadays, significance of the problem and the need to reduce the accompanied cost and ensure accurate and meaningful decisions, have widened the use of machine learning algorithms to perform automatic diagnosis and classification of such diseases.

To this end, medical data sets are fed into machine learning algorithms. These datasets are negatively affected by the noise in which accompanies them.

Noise also is defined as "irrelevant or meaningless data" may be the unfavourable result of an imperfect data collection process. However, the problem is that the system will learn to base its decisions on the set that includes both correct and noise-induced incorrect data. Consequently, the decisions might be fallacious. The case in point is that the focus of k-means algorithm is on detection and removal of noise data [16, 20].

This study analyses a decision support system capable of automatically detecting diabetes status of a person. The main goal is to improve the diagnosis system by removing duplicate, uncertain and inconsistent data in shorter computational time. Here, an effective modified k-means algorithm is proposed to cluster Pima Indians Diabetes (PID) data set by incorporating Fisher Discriminant Ratio (FDR) to distance function. The reason of employing FDR is to involve importance and influence of each feature separately. In other words, features with high separation ability and importance have higher priority in the assigning process. In this way, noisy and inconsistent data are detected and removed from each cluster. Finally, diabetes prediction model is obtained from filtered data set and classified by Support Vector Machine (SVM) method. The k-fold cross validation method is used to demonstrate reliability of the classification. To show superiority of the proposed algorithm, a set of experiments such as SVM, k-means and SVM, modified k-means with FDR & SVM are conducted on the data set. The obtained results show an accuracy of 99.64% indicating that the proposed system outperforms in comparison with other methods regarding classification performance and time consumption.

2. Related Works

Use of machine learning algorithms to detect diseases has been the subject of several research works. In 2015, Iyer *et al*. [10] developed a model for diagnosis of diabetes by employing Decision Tree and Naïve Bayes algorithms. In this work, J48 algorithm and Naïve Bayes gave 76.95% and 79.56% respectively by using the percentage split of 70:30 of the data set. In 2015, Durajiraj and Kalaiselvi provided a survey of

different soft computing techniques for the prediction of diabetes. This survey presented that the Artificial Neural Networks (ANNs) gives more accurate result than other classification techniques such as SVM, KNN and C4.5 [7]. Seera and Lim [18] presented a hybrid intelligent model, the Fuzzy Min-Max neural network, the classification and Regression Tree, and the Random Forest model for undertaking medical decision support tasks. They applied FMM-CART-RF method over PID data set and reported 78.39% accuracy with 10-fold cross validation. In 2014, Keerthana and Srividhya [13] employed the Density Based Clustering Algorithm by using Naïve Bayes on PID data set and obtained 96.35% performance. In 2014, Sanakal and Jayakumari [17] performed Fuzzy C-means clustering (FCM) algorithm and SVM on PID data set and found 94.30% classification accuracy. In 2014, Yilmaz *et al*. [20] proposed a modified k-means algorithm by incorporating a weight factor to the Euclidean distance.

They applied this method by using SVM on Statlog (Heart), SPECT images and PID data sets and obtained 97.87 %, 98.18 % and 96.71 % classification accuracy respectively.

In 2013, Christobel and Sivaprakasam [6] designed a new Class wise k-Nearest Neighbor (CkNN) method to improve the classification accuracy of the standard kNN. This method resulted in a classification accuracy of 78.18% over PID data set with 10-fold cross validation. In 2013, Anand *et al*. [2] performed Principle Component Analysis (PCA) to reduce feature dimension of diabetes disease and then applied higher order neural network. They attained a lower mean square error and faster convergence with PCA pre-processing. In 2013, Koklu and Unal [14] achieved prediction accuracy ratios of 75.13%, 73.82%, and 76.30% for Multi-Layer Perceptron (MLP), J48, and Naive Bayes, respectively over PID data set.

In 2013, Anuja and Chitra [15] executed Radial Basis Function (RBF) as kernel of SVM for classifying PID data set and achieved 78% accuracy.

In 2012, Karegowda *et al*. [12] used k-means clustering to detect and remove incorrectly classified samples and Decision tree C4.5 to classify PID data set. They reported 93.33% accuracy for 10-fold cross validation. In 2011, Ganji and Abadeh [9] created a new algorithm called FCS-ANTMINER by combining ant colony optimization and fuzzy logic. This method used 10-fold cross validation and achieved 84.24% accuracy for PID data set [9]. In 2011, Çalişir and Doğantekin [4] applied Linear Discriminant Analysis (LDA) method for feature extraction and Morlet Wavelet Support Vector Machine (MWSVM) classifier for classifying data. They concluded 89.74% classification accuracy by using PID data set.

3. Materials and Methods

3.1. Pima Indian Diabetes Data set

PID data set contains of 768 samples among which 500 are healthy and 268 are diabetes. Each sample has eight features as shown in Table 1. In this study, PID data set taken from the University machine learning repository of California at Irvine (UCI) is used for training and testing experiments [8].

Table 1. The features of pima indians diabetes data set.

N	Brief Description
1	Number of times pregnant
2	Plasma glucose concentration a 2 hours in an oral glucose tolerance test
3	Diastolic blood pressure (mm Hg)
4	Triceps skin fold thickness (mm)
5	2-Hour serum insulin (mu U/ml)
6	Body mass index (weight in kg/(height in m)^2)
7	Diabetes pedigree function
8	Age (years)
9	Class variable (0 or 1)

3.2. Methodology

- *Fisher discriminant ratio*: FDR tries to rank features with regard to their class-discriminatory power and can be independent of the type of the underlying class distribution. This content may be illustrated by Equation (1).

$$FDR = \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2} \quad (1)$$

Where μ_1 and μ_2 represents the means of class 1 and class 2 respectively, and σ_1^2, σ_2^2 are the corresponding variances. The goal of this method is to maximize the distance between the means of the classes and minimize the variance within each class [19].

- *K-means*: K-means is the simplest unsupervised learning algorithm that divides a data set into k groups. This method tries to create k clusters and determine k centroids for each of them. The centroid is defined as the average value of the elements of a set. This algorithm can act better when centroids are far away from each other and there is high similarity within clusters, and low similarity between separate clusters. After specifying centroids, each element of data set is assigned to the closest cluster. To this end, a proximity measure such as Euclidean distance quantifies the notion of the closest. Euclidean distance between two points *x* and *y* for *n* features is shown in Equation (2). After assigning all elements to their proper clusters, each cluster centre is recomputed and position of each centroid is changed to new points. Then, distance of elements to new centroids are computed and assigned to the closest clusters. This process is

repeated until all data points are settled into the appropriate clusters [1, 11].

$$Euclidean\,distance = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (2)$$

- *Support vector machine*: SVM has been proposed as an efficient statistical learning method for pattern recognition. SVM allows the use of kernel method to match the high-dimensional input data that is not separated as linear. SVM tries to find a hyper-plane maximizing the distance between the nearest points of two different classes of two class data set. Data points which are the closest to the hyperplane are called support vectors and margin is the distance between the support vectors and the class boundary hyperplanes. The aim of SVM is to maximize the margin. SVM can produce different decision boundaries with different kernel functions or approaches [3].

4. Experimental Study and Results

Noisy, duplicate, uncertain, and inconsistent data could negatively affect the performance of a system. One of the effective ways to address this problem is clustering method and many clustering approaches can be found in the literature [11]. But in the most of these studies, the importance and separability criterion of features have not been considered. The purpose of this system is to involve this matter which increases the performance of the classifier and the stability of the system and also reduce the time consumption rate. To fulfil this goal, the present work intends to create a high precision diagnostic system by modifying k-means clustering technique. In this research, we apply this technique on PID data set to remove irrelevant or meaningless data and discriminate diabetic people from healthy one with high accuracy and reliability.

The proposed system consists of two phases (Figure 1):

1. Phase 1: Clustering data set by the modified k-means.
2. Phase 2: Classifying and obtaining diabetes prediction model by using SVM.

In the first phase, the modified k-means algorithm follows an effective procedure to assign data to clusters. In this way, the distance between sample and cluster centre is measured by incorporating FDR to the Euclidean distance measurement which is defined by Equation (3).

$$Distance(x,c) = \sqrt{\sum_{i=1}^{n}((x_i - c_i)FDR_i)^2} \quad (3)$$

Where Distance (x, c) denotes the distance between each sample and cluster centre, n denotes the total number of features. Moreover, x_i is i^{th} feature value of sample data, c_i is i^{th} feature value of the centroid, and FDR_i is FDR value of i^{th} feature. FDR values of features are shown in Table 2, it can be seen that feature 2 with 0.5743 having the highest score and feature 3 with 0.0088 having the lowest score amongst all. Here, discrimination effectiveness of FDR vector is factored in. Accordingly, important and effective features are allowed to have more contribution in assigning points to appropriate clusters. After clustering diabetes data set by the modified k-means, noisy and inconsistent data is removed and thus the new reduced data set is produced by this phase Figure 2. In the second phase, SVM classification method is used to classify data given by the first phase; consequently, the diagnosis model is generated.

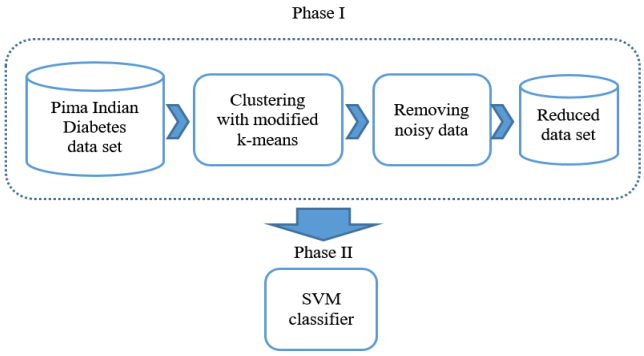

Figure 1. The proposed model.

An overview of the proposed diagnostic system as pseudocode is given below. Here, PID data set is used as an input data. In the first stage, FDR values of features are computed. In the second stage, the modified k-means clustering method is executed to partition data to k clusters. In the third stage, the number of healthy and diabetes samples in each cluster are evaluated. Then, the group with smaller members in each cluster is eliminated. Finally, SVM classifier is applied on the reduced data set.

Algorithm 1: Classification and Diagnosis of Diabetes

Output: Classification of PID
1: Compute FDR vector
2: Compute the modified k-means clustering
3: Count the number of healthy and diabetes samples in each cluster
4: Remove the group with small size in each cluster
5: Reduced data set
6: Run SVM classifier (reduced data set as input data)
Input: PID data set

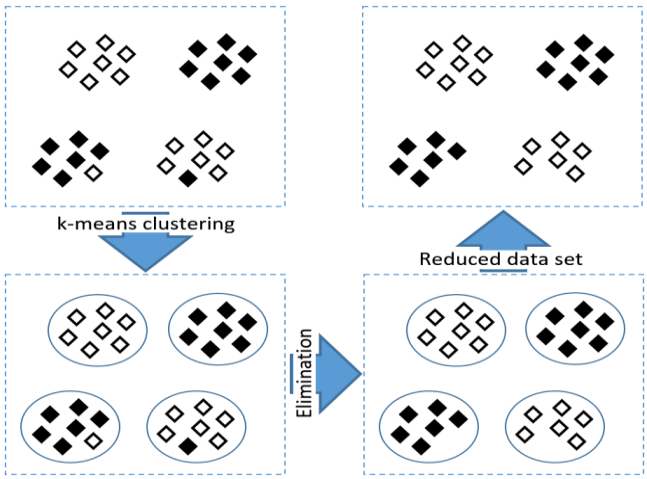

Figure 2. Diagram of phase 1.

Table 2. FDR value of features.

Features	1	2	3	4
FDR	0.1064	0.5743	0.0088	0.0117

Features	5	6	7	8
FDR	0.0343	0.2092	0.0639	0.1347

Table 3. Number of healthy and diabetes samples in each cluster

Clusters	1	2	3	4	5	6	7	8	9	10
Healthy	129	22	8	48	67	3	35	70	114	4
Diabetic	15	48	39	45	36	2	21	4	41	17

As shown in Table 3, data set is partitioned to 10 clusters to obtain meaningful results and a more accurate separation. Each cluster is comprised of samples which belong to first group (healthy people) or second group (diabetic people) or both of them. The group having more samples than the other is kept in the cluster and the other group is removed. For example, in Table 3, in cluster 1 there are 129 samples of healthy people and 15 samples of diabetic people. Here, diabetic samples are inconsistent and should be removed from the cluster to make an accurate decision. Totally, in this experiment, 570 samples out of 768 samples are taken as input data for SVM classifier and 198 samples as noisy and inconsistent data is removed.

One of the more popular performance evaluation methods is k-fold cross validation (k fc) method in which the data set is separated into k mutually exclusive sets of data and the performance of final result is the average performance of the k sets of test data. In this study, 10-fold cross-validation method is performed to ensure that the performance of the applied classifier is reliable. The training set has 513 cases and the test set consists of 57 cases. The obtained classification accuracy by SVM is 99.64% with 10-fold cross validation method.

In order to demonstrate the importance of this work, the performance of the modified k-means and SVM is compared to the results of SVM and k-means and SVM on PID data set. The performance values of the mentioned solutions are shown in Table 4. It reveals that the modified k-means with SVM considerably improved classification ratio and reduced the computational time. In this table, PPV is Positive predictive value and NPV is Negative predictive value.

Table 4. Performance of classification for SVM, k-means and SVM and the modified k-means and SVM.

Performance Criteria	SVM	K-means & SVM	Modified K-means with FDR & SVM
Sensitivity (%)	0.79	0.95	0.99
Specificity (%)	0.71	0.96	0.99
PPV (%)	0.84	0.99	0.99
NPV (%)	0.65	0.76	0.99
Accuracy (%)	76.30	95.12	99.64
Time Consumption (Second)	5.10	1.52	1.37

5. Conclusions

Diabetes can damage human body's system and lead to many other diseases. Thus, early detection and treatment of this kind of disease is crucial. Machine learning methods have become an easier and efficient way of diagnosing diseases by biomedical science experts. In this study, we tried to introduce a high precision diagnostic system to make it different from the existing methods. We proposed the modified k-means clustering algorithm by incorporating FDR to the Euclidean distance which performs better than [20]. This technique was employed to divide PID data set into 10 clusters which consist of diabetes and/or healthy samples. After clustering step, irrelevant and inconsistent data were eliminated from all clusters.

Then, diabetes prediction model was generated by SVM.

The results of experiments demonstrate an improvement in the classification ratio of PID data set after applying the modified k-means algorithm. In order to verify it, Table 4 gives a comparison of our new approach with SVM and k-means & SVM on PID data set regarding sensitivity, specificity, positive predictive value, negative predictive value, accuracy and time consumption. Apparently, Figure 3 gives graphical information about our experiments and allows us to conclude that the new method is a significant improvement over the mentioned two previous methods.

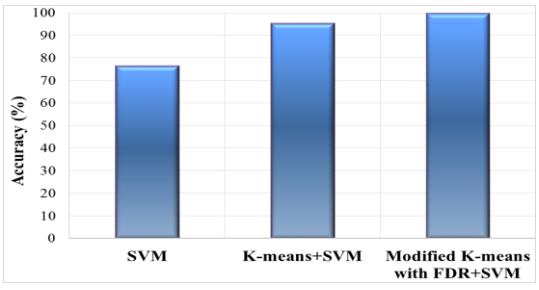

Figure 3. A bar chart of performance comparison among three methods.

Table 5. Classification accuracy comparison with previous studies.

Method	Classification Accuracy (%)
Naive Bayes [10]	79.56
FMM-CART-RF [18]	78.39
Density Based Clustering Naïve Bayes [13]	96.35
Fuzzy C-means clustering (SVM) [17]	94.30
Modified k-means + SVM [20]	96.71
Cass wise k Nearest Neighbour (CkNN) [6]	78.18
Naive Bayes [14]	76.30
SVM (RBF) [15]	78.00
K-means clustering-Decision tree C4.5 [12]	93.33
FCS-ANTMINER [9]	84.24
(LDA)-MWSVM [4]	89.74
Proposed method	**99.64**

As it appears from the Table 5, the new method provides superior results in comparison with previous diabetes disease diagnosis methods. In conclusion, the results from our analysis show the proposed method increases the classification performance as well as reduce the consumption time. Moreover, it can be effectively applied in real time applications. In the future, we will conduct further case studies.

References

[1] Abbas O., "Comparisons Between Data Clustering Algorithms," *The International Arab Journal of Information Technology*, vol. 5, no. 3, pp. 320-325, 2008.

[2] Anand R., Kirar V., and Burse K., "K-fold Cross Validation and Classification Accuracy of PIMA Indian Diabetes Data set using Higher Order Neural Network and PCA," *International Journal of Soft Computing and Engineering*, vol. 2, no. 6, pp. 2231-2307, 2013.

[3] Burgers J., "A Tutorial on Support Vector Machines for Pattern Recognition," *Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 121-167, 1998.

[4] Çalişir D. and Doğantekin E., "An Automatic Diabetes Diagnosis System based on LDA-Wavelet Support Vector Machine classifier," *Expert Systems with Applications*, vol. 38, no. 7, pp. 8311-8315, 2011.

[5] Causes of Diabetes, National Institute of Diabetes and Digestive and Kidney Diseases, https://www.niddk.nih.gov/health-information/diabetes/causes, Last Visited, 2014.

[6] Christobel Y. and Sivaprakasam P., "A new Class wise k Nearest Neighbor Method for the Classification of Diabetes Dataset," *International Journal of Engineering and Advanced Technology*, vol. 2, no. 3, pp. 396-400, 2013.

[7] Durairaj M. and Kalaiselvi G., "Prediction of Diabetes using Soft Computing Techniques-A Survey," *International Journal of Scientific and Technology Research*, vol. 4, no. 3, pp. 190-192, 2015.

[8] Frank A. and Asuncion A., UCI Machine Learning Repository, Irvine, CA: University of California, School of Information and Computer Science 2010.

[9] Ganji M. and Abadeh M., "A Fuzzy Classification System based on Ant Colony Optimization for Diabetes Disease Diagnosis," *Expert Systems with Applications*, vol. 38, no. 12, pp. 14650-14659, 2011.

[10] Iyer A., Jeyalatha S., and Sumbaly R., "Diagnosis of Diabetes using Classification Mining Techniques," *International Journal of Data Mining and Knowledge Management Process*, vol. 5, no. 1, pp. 1-14, 2015.

[11] Jain A., Murty M., and Flynn P., "Data Clustering: A Review," *ACM Computing Surveys*, vol. 31, no. 3, pp. 264-323, 1999.

[12] Karegowda A., Punya V., Jayaram M., and Manjunath A., "Rule based Classification for Diabetic Patients using Cascaded K-Means and Decision Tree C4.5," *International Journal of Computer Applications*, vol. 45, no. 12, pp. 45-50, 2012.

[13] Keerthana G. and Srividhya V., "Performance Enhancement of Classifiers using Integration of Clustering and Classification Techniques," *International Journal of Computer Science Engineering*, vol. 3, no.3, pp. 200-203, 2014.

[14] Koklu M. and Unal Y., "Analysis of A Population of Diabetic Patients Databases with Classifiers," *International Journal of Biomedical and Biological Engineering*, vol. 7, no. 8, pp. 481-483, 2013.

[15] Kumari A. and Chitra R., "Classification of Diabetes Disease using Support Vector Machine," *International Journal of Engineering Research and Applications*, vol. 3, no. 2, pp. 1797-1801, 2013.

[16] Lowongtrakool C. and Hiransakolwong N., "Noise Filtering in Unsupervised Clustering using Computation Intelligence," *International Journal of Math. Analysis*, vol. 6, no. 59, pp. 2911-2920, 2012.

[17] Sanakal S. and Jayakumari S., "Prognosis of Diabetes using Data Mining Approach-Fuzzy C Means Clustering and Support Vector Machine," *International Journal of Computer Trends and Technology*, vol. 11, no. 2, pp. 94-98, 2014.

[18] Seera M. and Lim C., "A hybrid Intelligent System for Medical Data Classification," *Expert Systems with Applications*, vol. 41, no. 5, pp. 2239-2249, 2014.

[19] Theodoridis S. and Koutroumbas K., *Pattern Recognition*, Academic Press, 1999.

[20] Yilmaz N., Inan O., and Uzer M., "A new Data Preparation Method based on Clustering Algorithms for Diagnosis Systems of Heart and Diabetes Diseases," *Journal of Medical Systems*, vol. 38, no. 5, 2014.

Shima Afzali received her B.Sc. degree in computer engineering (software engineering) from the University of Zanjan, Iran, in 2009 and her M.Sc. degree in computer engineering from Gazi University, Turkey, in 2014. She has been working toward the Ph.D. degree in computer science, Victoria University of Wellington, New Zealand, since March 2016. She has been awarded a Victoria Doctoral Scholarship. Her main area of research is machine learning, bioinformatics, evolutionary computation.

Oktay Yıldız received his M.Sc. degree in Institute of Science from Gazi University, in 2004 and Ph.D. degree in Institute of Information Sciences from Gazi University, in 2012. He has been with the Computer Engineering Department at Gazi University, Ankara, Turkey since 2009. His research interests include machine learning, and data mining.