



Global Journal on Technology

Vol 5 (2014) 202-207



Selected Paper of 4th World Conference on Information Technology (WCIT-2013)

A New Data Elimination Method Based on Clustering Algorithms for Diagnosis of Diabetes Diseases

Onur Inan*, Selcuk University, Faculty of Engineering Computer Engineering, Konya, Turkey.

Nihat Yilmaz, Selcuk University, Faculty of Engineering Electrical-Electronics Engineering, Konya, Turkey.

Mustafa Serter Uzer, Selcuk University, Faculty of Engineering Electrical-Electronics Engineering, Konya, Turkey.

Suggested Citation:

Inan, O. , Yilmaz, N. & Uzer, M.S. A New Data Elimination Method Based on Clustering Algorithms for Diagnosis of Diabetes Diseases, *Global Journal on Technology* [Online]. 2014, 05, pp 202-207. Available from: www.awer-center.org/pitcs

Received May 04, 2013; revised July 20, 2013; accepted September 23, 2013.

Selection and peer review under responsibility of Prof. Dr. Hafize Keser.

©2014 SPROC - Academic World Education & Research Center. All rights reserved.

Abstract

The most important factors that prevent pattern recognition from functioning rapidly and effectively are the noisy and inconsistent data in databases. This article presents a new data elimination method based on clustering algorithms for diagnosis of diabetes diseases. In this method, K-means Algorithm is used for clustering based data elimination system for the elimination of noisy and inconsistent data and Support Vector Machines is used for classification. This newly developed approach was tested in the diagnosis of diabetes. The data set used in the diagnosis of these diseases is the Pima Indians Diabetes data sets obtained from the UCI database. The proposed system achieved 93.65% classification success rates from this data set. Classification accuracies for these data sets were obtained through using 10-fold cross-validation method. According to the result, the proposed method of performance is successful compared to other results attained, and seems very promising for pattern recognition applications.

Keywords: diabetes diseases, support vector machine, K-means algorithm.

* ADDRESS FOR CORRESPONDENCE: **Onur Inan**, Selcuk University, Faculty of Engineering Computer Engineering, Konya and 42070, Turkey, E-mail address: oinan@selcuk.edu.tr / Tel.: +090-332-223-3333

1. Introduction

Pattern recognition and data mining are used in many aspects of life. These techniques are most frequently used in military, medical and industrial areas. The variety and quantity of data collected used in these applications have considerably increased thanks to the contribution of new measurement systems. It has become nearly impossible for these data sets to be analyzed and evaluated by experts in order to obtain information that would be ultimately useful. For this reason, feature selection and data reduction algorithms are being developed, which increases the performance of analysis systems or recognition systems by filtering and arraying data according to importance and by identifying unnecessary measurements in data sets.

In the algorithm that we develop, k-means algorithm has been used as an instrument that provides for determining the coherent or incoherent datum inside themselves. The data reduction operation is carried out by a heuristic algorithm that runs according to incoherence information gotten from k-means algorithm. Some remarkable practices in the literature that use k-means algorithm for data reduction are like these: [1] have extracted the incorrectly classified datum from original data pattern via simple k-means algorithm on Type-2 diabetic data, and they have classified the rest datum by the k-fold cross-validation method on C4.5 algorithm. [2] have developed a neighborhood-classified-k-means (NC-k-means) algorithm, adding an absolute neighborhood index into the traditional k-means algorithm.

Diabetes is a disease that requires constant monitoring and dramatically decreases the quality of life. As will all diseases, early diagnosis in diabetes is very important both to determine the treatment method and to reduce the damage caused to the body. For the system to be developed for the diagnosis of this disease, Pima Indians Diabetes data set obtained from the UCI database was used again. Some of important studies conducted on this data set are as follows: [3] proposed a new method of biomedical signal classification using complex-valued pseudo autoregressive (CAR) modeling approach. [4] presented a modified version of the Hybrid Multilayer Perceptron (HMLP) network to improve the performance of the conventional HMLP. [5] worked on diabetes disease, which is a very common and important disease using Principal Component Analysis (PCA) and Adaptive Neuro-Fuzzy Inference System (ANFIS). [6] used Generalized Discriminant Analysis (GDA) and Least Square Support Vector Machine (LS-SVM) and they proposed a new cascade learning system based on Generalized Discriminant Analysis and Least Square Support Vector Machine. [7] presented a hybrid neural network that includes Artificial Neural Network (ANN) and Fuzzy Neural Network (FNN) was developed.

Our proposed approach consists of two stages. In the first stage, noisy and inconsistent data in databases were reduced by using K-means algorithm. In the second stage, the rearranged patterns were tested using SVM classifier. The 10-fold cross-validation method was used to demonstrate reliability of the classification. In this study, a hybrid approach was applied to Diabetes datasets. The developed method was compared with studies related to the databases. The method developed as a result of this comparison has been found to have the highest success rate among the results that we attained.

2. K-means Algorithm

The commonly used K-means algorithm was first developed in 1967 by MacQueen [8]. This algorithm is known as one of the most common unsupervised clustering methods and is used to separate observations between a certain numbers of sets according to a predefined distance criterion. Unsupervised data clustering is a task of assigning points to clusters as well as simultaneously estimating cluster location and shape [9]. This algorithm allows each data to belong to only a single set. The aim is to achieve high similarity within the sets, and low similarity between the sets [10] The center of a set is the average value of the elements of a set. At the same time, these centers of sets represent the characters of all elements within the set. Belonging to the sets is determined according

to the center of set that is closest. The most frequently used distance criterion is the Euclidean distance measurement [11].

2.1. Support Vector Machine (SVM)

SVM is an effective method used for pattern recognition, data mining and machine learning. SVM was developed in 1995 by Cortes and Vapnik [12]. SVM is a supervised learning algorithm used in classification and regression analyses. In this algorithm, there are two different categories separated by a linear plane. The training of the algorithm is the process of determining the parameters of this linear plane. In multiclass applications, the problem is categorized into groups as belonging either to one class or to the other classes [13-14].

1.3. Performance evaluation

Four criteria for performance evaluation of Diabetes diagnosis were used. These criteria are classification accuracy, confusion matrix, analysis of sensitivity and specificity, and k-fold cross-validation [15-17].

3. Proposed Data Elimination Method with a K-means Algorithm

Due to various noises, there are data in databases that are inconsistent, duplicate and uncertain in comparison to similar data. Entering these data into the recognition system significantly reduces the performance of the system. Some of the issues in these learning data were resolved in the proposed system by using the k-means algorithm. In this way, the newly obtained data set increase the test performance of the classifier and the stability of the system. The proposed system as it is seen in figure 1 is comprised of two phases. In the first stage, noisy and inconsistent data are removed with the k-means algorithm. In the second stage, the accuracy rates of designed system with the obtained and filtered data are determined by using them in the SVM classifier. In order to increase the reliability of the classifier performance measurement, the 10-fold cross-validation method was used.

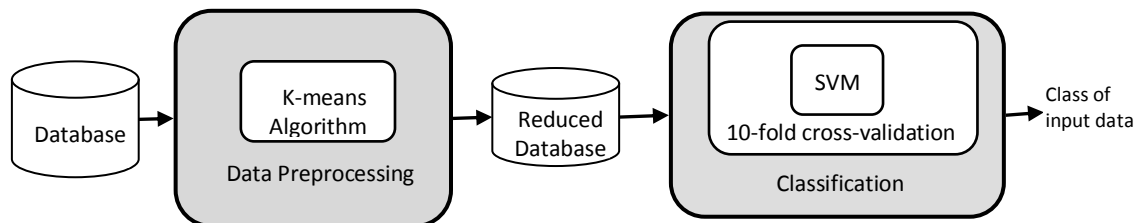


Figure 1. Block diagram of the proposed system

The used database is organized such that the characteristics are found in the columns while the samples are located in the rows. After this, the k-means algorithm was applied for each sample. In the database separated into an N number of sets (N is number of features in dataset), elimination is carried out based on the principle that samples in the same set should also belong to the same class. Hence, the data of the more prevalent class remains within a set. The pseudo-code of this method is given in figure 2. If deemed necessary, a certain rate can be determined during elimination to ensure that all data below this rate are eliminated. In this way, while data elimination is not performed in sets with high inconsistency, disruptive data can be eliminated from highly consistent sets. Because of this approach, the number of data to be eliminated from the database can also be adjusted. This procedure is repeated for all sets. The obtained data set is entered to inputs of the SVM classifier used for 10-fold cross validation.

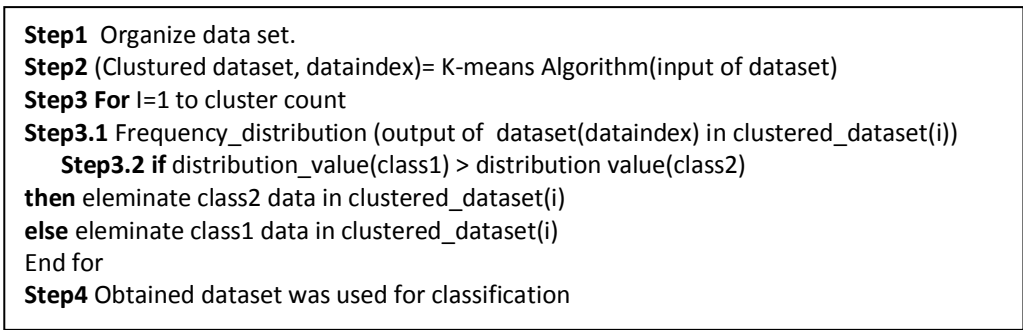


Figure.2 Pseudo-Code presentation of the proposed data elimination method

3.1.Diabetes dataset

In this study, we use the Diabetes diseases dataset introduced by Black C.L. [18]. This dataset contains 768 samples, where each sample has 8 features which are eight clinical findings. All patients in this dataset are Pima Indian women at least 21 years old and living near Phoenix, Arizona, USA. The binary target variable takes the values ‘0’ or ‘1’. While ‘1’ means a positive test for Diabetes, ‘0’ is a negative test. There are 268 cases in class ‘1’ and 500 cases in class ‘0.’⁶

3.2. Experimental Results and Discussion

In this study, the classified that will assist in producing a diagnosis of diabetes as a result of medical attention was selected as the SVM. While using this classifier, training was performed according to the parameters given in Table 1.

Table 1. List of classification parameters

Parameters	Value
Method	SVM
Optimization algorithm	SMO
Validation method	k-fold cross-validation (10-fold CV)
Kernel Function	Linear
Maximum Iteration	15000
The initial value	Random

The test result of the K-means-SVM methods developed for Diabetes data set is given in Table 2.

Table 2. Pre-processing results for the datasets with K-means Algorithm

Database	Number of Classes	Number of Features	Number of Clusters	Samples	Cleared data k-means Samples	Mean Accuracy with 10-fold CV
Diabetes	2	8	8	768	552	93.65

For the Diabetes dataset, classification performance values and the comparisons with the other systems have been given in Table 3 and Table 4 respectively.

Table 3. Performance of classification for Diabetes dataset

Performance Criteria	Raw Data	k-means
Accuracy (%)	77.60	93.65
Sensitivity (%)	88.34	97.75
Specificity (%)	57.32	83.50
Positive predictive value (%)	79.49	93.76
Negative predictive value (%)	74.28	93.15

Table 4. Classification accuracies obtained by our method and other classifiers for Diabetes

Author (Year)	Method	Classification accuracy (%)
Polat and Gunes (2008) ⁶	LS-SVM (10-fold CV)	78.21
	GDA-LS-SVM (10-fold CV)	82.05
Isa and Mamat (2011) ⁴	Clustered-HMLP	80.59
Aibinu et al. (2011) ³	AR1+NN (3-fold CV)	81.28
Patil (2010) ¹	Hybrid Prediction Model (HPM)	92.38
Polat and Gunes (2007) ⁵	Combining PCA and ANFIS	89.47
Kahramanli and Allahverdi (2008) ⁷	Hybrid system(ANN and FNN)	84.2
Our study	k-means (10-fold CV)	93.65

Conclusions

This study was designed for use in the diagnosis of diabetes. The database contain data that are noisy, inconsistent and fragmentary. Such defective data is one of the most significant factors that adversely affect the success of the classifier. The developed systems ensure the elimination of this type of poor quality data. The data elimination process is carried out with the k-means algorithm. Databases that were subject to data elimination are classified by using the SVM. Classification accuracy of the proposed system reached 93.65% for Pima Indians Diabetes dataset. According to the result, the proposed method performance is highly successful compared to other results attained, and seems very promising for pattern recognition applications. We expect that the developed system can be used in all other pattern recognition applications.

Acknowledgment

The authors are grateful to Selcuk University Scientific Research Projects Coordinatorship for support of the manuscript.

References

- [1] Patil, B.M., Joshi, R.C, & Toshniwal D. (2010). Hybrid prediction model for Type-2 diabetic patients. *Expert Systems with Applications* Dec, 37(12), 8102-8108.
- [2] Zhang, B., Li, S.S., Wu, C.S., Gao, L.R., Zhang, W.J., & Peng, M. (2013). A neighbourhood-constrained k-means approach to classify very high spatial resolution hyperspectral imagery. *Remote Sens Lett*, 4(2), 161-170.
- [3] Aibinu, A.M, Salami, MJE, & Shafie, A.A. (2011). A novel signal diagnosis technique using pseudo complex-valued autoregressive technique. *Expert Systems with Applications*, 38(8), 9063-9069.
- [4] Isa NAM, & Mamat WMFW. (2011). Clustered-Hybrid Multilayer Perceptron network for pattern recognition application. *Applied Soft Computing*, 11(1), 1457-1466.

- [5] Polat,K.,& Gunes, S. (2007). An expert system approach based on principal component analysis and adaptive neuro-fuzzy inference system to diagnosis of diabetes disease. *Digital Signal Processing*, 17(4), 702-710.
- [6] Polat, K., Gunes, S.,& Arslan, A.(2008). A cascade learning system for classification of diabetes disease: Generalized discriminant analysis and least square support vector machine. *Expert Systems with Applications*, 34(1), 482-487.
- [7] Kahramanli, H.,& Allahverdi, N. (2008). Design of a hybrid system for the diabetes and heart diseases. *Expert Systems with Applications*, 35(1-2), 82-89.
- [8] MacQueen JB. (1967). Some Methods for classification and analysis of multivariate observations. *In Proceedings of 5th Berkeley symposium on mathematical statistics and probability*. California: Berkeley: University of California, 281–297.
- [9] Zhang, JY, Peng, LQ, Zhao, XX, & Kuruoglu, EE. (2012). Robust data clustering by learning multi-metric Lq-norm distances. *Expert Systems with Applications*, 39(1), 335-349.
- [10] Han, J,& Kamber, M. (2001). *Data Mining Concepts and Techniques*: Morgan Kauffmann Publishers.
- [11] Erisoglu, M., Calis, N., & Sakallioğlu, S. (2011). A new algorithm for initial cluster centers in k-means algorithm. *Pattern Recognition Letters*, 32(14), 1701-1705.
- [12] Cortes, C., & Vapnik, V.(1995). Support-Vector Networks. *Mach Learn*, 20(3), 273-297.
- [13] Ivanciuc, O. (2007). *Reviews in Computational Chemistry*. John Wiley & Sons, Inc., 23.
- [14] Kuan, T.W., Wang, J.F., Wang, J.C., Lin, P.C.,& Gu, G.H. (2012). VLSI Design of an SVM Learning Core on Sequential Minimal Optimization Algorithm. *Ieee Transactions on Very Large Scale Integration (Vlsi) Systems*, 20(4), 673-683.
- [15] Xu, Y., Zhu, Q., Wang, J.H. (2012). Breast cancer diagnosis based on a kernel orthogonal transform. *Neural Comput Appl*, 21(8), 1865-1870.
- [16] Francois, D., Rossi, F., Wertz, V., & Verleysen, M. (2007). Resampling methods for parameter-free and robust feature selection with mutual information. *Neurocomputing*, 70, 1276-1288.
- [17] Diamantidis, N.A.,& Karlis, D., (2000). Giakoumakis EA. Unsupervised stratification of cross-validation for accuracy estimation. *Artificial Intelligence*, 116, 1-16.
- [18] Blake, C.L., M.C.J.(1998). UCI repository of machine learning databases.