ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

IMPROVING DIAGNOSIS OF DIABETES MELLITUS USING COMBINATION OF PREPROCESSING TECHNIQUES

RAZIEH ASGARNEZHAD[1], MARYAM SHEKOFTEH[2], FARSAD ZAMANI BOROUJENI[3*]

[1, 2, 3] Department of Computer Engineering, Isfahan (Khorasgan) Branch, Islamic Azad University, Isfahan, Iran

E-mail: [1]r.asgarnezhad@khuisf.ac.ir, [2] shekofteh@iausarv.ac.ir, [3]f.zamani@khuisf.ac.ir

ABSTRACT

Diabetes mellitus is one of the most common diseases among people of all age groups, affecting children, adolescents and young adults. There is an increasing interest in using machine learning techniques to diagnose these chronic diseases. However, the poor quality of most medical data sets inhibits construction of efficient models for prediction of diabetes mellitus. Without efficient preprocessing methods, dealing with these kinds of data sets leads to unreliable results. This paper presents an efficient preprocessing technique including a combination of missing value replacement and attribute subset selection methods on a well-known diabetes mellitus data set. The results show that the proposed technique can improve the performance of applied classifier and outperforms the traditional methods in terms of accuracy and precision in diabetes mellitus prediction.

Keywords: *Data Mining, Preprocessing Techniques, Diabetes mellitus*.

1. INTRODUCTION

Nowadays, diabetes is a very common mellitus in the world. The absence of sufficient insulin and imbalance of glucose level in the body imply a diabetes occasion. In the absence of insulin, carbohydrates is changed into glucose, enhancing the sugar level more than normal [1]. The effects of diabetes include chronic malfunction, harm and fracture of several organs. Most of diabetes is divided into two main categories: type-1 & type-2 diabetes mellitus.

The type-2 diabetes accounts for almost 90% of the diabetes cases and commonly called the adult-onset diabetes or the non-insulin dependent diabetes. In this case the various organs of the body become insulin resistant, and this increases the demand for insulin. At this point, pancreas doesn't make the required amount of insulin. To keep this type of diabetes at bay, the patients have to follow a strict diet, exercise routine and keep track of the blood glucose. Obesity, being overweight, being physically inactive can lead to type 2 diabetes. Also with ageing, the risk of developing diabetes is considered to be more. Majority of the Type 2 diabetes patients have border line diabetes or the Pre-Diabetes, a condition where the blood glucose levels are higher than normal but not as high as a diabetic patient.

In the last decade, a large amount of research efforts were conducted to develop machine learning systems for diabetes diagnosis in general, and prediction of diabetes mellitus in particular. As the volume of data in these systems grow larger and larger, the effect of data quality on the process of extracting useful information becomes imperatively important. It is critical for data mining applications in the healthcare industry to investigate how to improve the quality of data as much as possible.

When mining real-world medical data sets, it should be noted that most attributes are highly prone to contain missing values, noise and outliers. There are many possible reasons for dealing with incomplete data such as unavailable information during the data entry, human or computer errors and not recording relevant data due to a misunderstanding or equipment malfunction [2].

In fact, the quality of data determines the performance of the prediction methods and the usefulness of the extracted knowledge. In this regard, the preprocessing techniques can help improve the data quality and lead to obtain more accurate results.

Although several research studies have been conducted to predict diabetes mellitus in recent years, a few of them have employed comprehensive preprocessing techniques which usually involve a combination of missing value replacement and attribute subset selection.

In this study, we investigated the effect of different combination of preprocessing techniques on the performance of the prediction models, in a stepwise manner. In the first step, detected missing values are replaced using Mean, Median, and K-nearest neighbor methods. In the second step, we apply 3 attribute selection methods to improve the classification performance including: combination of forward selection and backward elimination, brute force, and evolutionary technique. The classification is performed by implementing a binary class Support Vector Machines (SVM), i.e. 0 for negative test or 1 for positive test, as it gives higher accuracy than other classification algorithms for the selected data set [1,7,8,9,16]. A brief review of the related methods is reported in the literature review section.

This paper is organized as follows: Section 2, shows a summary of the related works followed by a description of data collection in section 3. The proposed method is presented in Section 4 and evaluated by the experiment explained in section 5. Finally, the paper presents a conclusion in section 6.

2. RELATED WORK

A large number of methods have been proposed for improving the accuracy of diabetes mellitus diagnosis systems. Most of these methods found in literature are focused mainly on improving the performance of the classifier by changing the architecture of the classifier, manipulating the parameters or using the Ensemble Classification methods. The following text presents a brief review on some salient approaches in the literature.

Kayaer and Yıldırım [3] applied several neural network structures like multilayer perceptron (MLP), radial basis function (RBF) and general regression neural network (GRNN) on Pima dataset. This data set educed from UCI Repository of Machine Learning Databases. The data set was pickup from a bigger data set held by the National Institutes of Diabetes mellitus and Digestive and Kidney Diseases. They found that MLP is more accurate than RBF and the best accuracy result (80.21%) is achieved on the test data when an appropriate structure is chosen for GRNN. Although, their method improved the accuracy of the previous ones, they didn’t describe any preprocessing techniques to enhance the quality of data. Koklu and Unal [4] proposed a decision support system for diagnosis of diabetes using.supervised learning techniques. It applied three different classification algorithms namely Multilayer Perceptron, Naive Bayes [5] and Decision Tree Induction (J48). Without using any pre-processing techniques, they attained the accuracy of 75.13%, 73.828%, and 76.302% for MLP, J48, and NB, respectively. They showed that NB outperforms its counterpart algorithms in terms of accuracy of the classifier on Pima dataset.

Diwani and Sam [6] employed NB and J48 to identify diabetic patients via Pima dataset. They observed through analysis of the experimental results that NB outperforms J48 decision tree algorithm in terms of the accuracy of the model. They used knowledge flow for data processing and analysis and WEKA for conducting their experiments. The best algorithms were NB and J48 with 76.30% and 73.83% of correctly classified tuples. A. Parashar et al. [1] implemented and compared two techniques : SVM and Feed Forward Neural Network for classification of patients in Pima dataset. This method aimed at using supervised learning approach to find a set of basis vectors. As a preprocessing step, they used LDA for attribute selection and observed that the performance of LDA-SVM is much higher than the SVM. Their experimental results showed that SVM, with the accuracy of 75.65%, attains better performance versus Feed Forward NN.

Kumar and Govindasamy [7] implemented hybrid model via different classification techniques such as SVM, Regression, Bayes Net, NB and Decision Table for improving the accuracy of diabetes mellitus diagnosis in three phases. Theses phases includes, applying the original diabetes dataset, the use of feature selection, and the comparison of the results with and without feature selection.

In general, these algorithms make use of a particular classification algorithm and run it several times by changing the algorithm’s parameters or input weights to improve the accuracy of the classifier. In the implementation phase, Kumar and Govindasamy used two learning methods that are included in Weka data mining tool. They showed that the process of selecting a subset of relevant features has a great impact on the performance results and leads to increased performance of the classification algorithm. Comparison was made among these classification algorithms such as SVM, Regression, Bayes Net, NB and Decision Table before and after applying feature selection, out of which Decision Table algorithm outperformed all other techniques with 79.81% accuracy after applying feature selection [7].

Kumar and Govindasamy applied various techniques on Pima dataset using ANN, SVM, K-NN Algorithms [8]. Using these techniques, they attempted to make an ensemble model by combining two techniques: Bayesian classification and Multilayer Perceptron. They used feature selection technique to achieve higher accuracy. Various techniques like C4.5, RBF, Bayes Net and MLP are trained and their combination were used after the training phase. C4.5+RF and MLP+Bayes Net achieved the accuracy of 79.31%, 81.89%, respectively. They found that MLP + Bayes Net was a robust model for the classification of data.

Purwar and Singh investigated a summary of the techniques applied for the detect of diabetes [8]. They summarized some of existing works with diverse classification algorithms such as SVM, KNN, NB, ID3, CART, and C5.0. Some of works have applied without pre-processing and it is a disadvantage. They compared the accuracy of these classification models. SVM achieved the accuracy of 81.77% versus others.

In the literature, a variety of useful missing value replacement and imputation techniques were proposed to enhance the quality of data. In particular, missing values for some attributes can be imputed by different techniques such as combination of clustering and neural network [9], K-Nearest Neighbor algorithm [10], decision tree induction [11] and many others. Attribute subset selection is another task which has a great impact on the quality of data, redundant attributes or dimensions can be found and removed by applying numerous techniques such as forward selection, backward elimination [12], rough sets and information entropy [13] and evolutionary techniques [5].

With having reviewed the existing literature, we found that the combination of preprocessing techniques applying on a proper classifier, that didn't investigate in previous work, may be more accurate than the other methods and lead to better results. In fact, this important issue can be considered as a gap in previous work and literature selection criteria.

3. DATA COLLECTION

In this paper, we use Pima Indians diabetes mellitus dataset downloaded from UCI machine learning repository[1]. This data set has about of 768 instances. Each person is recognized in data set by 8 attributes as seen in follow. All attributes are numerical values. This data set educed from UCI Repository of Machine Learning Databases. The data set was pickup from a bigger data set held by the National Institutes of Diabetes mellitus and Digestive and Kidney Diseases. All persons are Indian women at least 21 years old and live in near Phoenix, Arizona, USA. The response variable is binary and takes 0 or 1, where 1 means a positive test and 0 is a negative test for diabetes mellitus. The cases are 268 (34.9%) in class 1 and 500 (65.1%) cases in class 0.

Eight clinical features contained in the Pima dataset are as follows:

1. Number of times pregnant for each person in life era
2. Plasma glucose concentration a 2 hours that is get from an oral glucose tolerance test
3. Diastolic blood pressure for each person (mm Hg)
4. Triceps skin fold thickness for each person (mm)
5. 2-Hour serum insulin for each person (mu U/ml)
6. Body mass index that can be calculated from its formulas
7. Diabetes mellitus pedigree function
8. Age for each person (years)

4. PROPOSED METHOD

The main weakness of PIMA is the lack of completeness. In other words, it suffers from large number of missing values and similar to many other data sets, ignoring important values of its attributes leads to poor classification results. Another problem of this data set is that it has not been prepared for the classification tasks. To overcome these problems, this paper presents the following solutions: The first solution attempts to impute the best values for missing values of the data set. The second solution establishes data reduction. In fact, attribute subset selection techniques are used to improve the performance of the classification algorithm. The proposed method applied SVM classifier to predict of diabetes mellitus. SVM is a supervised learning algorithm [14]. Experiments

[1] http://ftp.ics.uci.edu/pub/ml-repos/machine-learning databases/pima-indians-diabetes

show that SVM is one of the best classifier for diagnosis of diabetes on PIMA data set and it is the cause of choose of it in this research.

The following text explains mentioned techniques in more detail. The proposed method, investigates the effect of three different missing values replacement techniques including Mean, Median, and K-NN on the performance of the classification task. Furthermore, three attribute subset selection techniques including combination of forward selection and backward elimination, optimize selection (brute force) and genetic algorithm were used to reduce the redundancy of attributes and to obtain a set of relevant features in the selected data set. After applying the cleaning and reduction techniques on the data, the SVM classifier was employed to classify the instances. We examined different combinations of missing value replacement and attribute subset selection techniques to find the best case which results in the best values for accuracy and precision measures. The following sections describe the above mentioned techniques and their impact on the accuracy of diabetes mellitus prediction.

4.1. Missing Value Analysis

Missing values in this data set are represented by “0”. In below, we state the most common approaches to fill in missing values examined in this research.

4.1.1. Replace with mean

In the first strategy, the missing values are replaced with the average. This technique replaces all missing values of an attribute by the average of all existing values of that attribute. In most of the cases, this method can approximate values that are close to actual value of the attribute for all missing values [15].

4.1.2. Replace with median

In the second strategy, we replace all missing values with the median. This technique replaces all missing values of an attribute by the median of all existing values of that attribute. In the cases where the outlier exists in the available data, this method can work better than replacing with mean.

4.1.3. K-Nearest Neighbor (KNN)

In the third strategy, different values maybe imputed for each missing value. In fact, for filling missing values, we apply the most probable value using K-nearest neighbor (KNN) algorithm. KNN technique replaces missing values of an attribute with the corresponding value from the closest tuple in terms of Euclidean distance [15]. Depending on the dataset, the K-Nearest Neighbor model can be a regression or classification model [16].

4.2. Attribute Subset Selection

Attribute subset selection is one of the data reduction methods that remove redundant or irrelevant attributes from the data set. To this end, greedy methods are commonly employed that while searching through the attribute space, they always make what looks to be the best selection at the time [17].

There are several techniques for attribute subset selection in the literature [18-20] including brute force, stepwise forward selection, stepwise backward elimination, evolutionary algorithms and many others. For all of the mentioned techniques, it needs a performance measurement which reflects the effectiveness of a subset selection on the given dataset [10]. The following sections describe a set of attribute selection techniques that are applied in this research.

4.2.1. Combination of forward Selection and backward elimination

In this method the most relevant attributes of the related dataset are selected using combination of two greedy attribute subset selection algorithms namely ‘forward selection’ and ‘backward elimination’. In fact, at each stage of this method, the best attribute is selected (forward selection) and the worst attribute among the existing set is removed (backward elimination) [17].

4.2.2. Optimize selection (brute force)

The second approach for attribute subset selection is optimized selection (Brute Force). In this method, the most relevant attributes of the dataset are selected by examining all possible combinations of attribute selections. Due to its comprehensive examination of attribute combinations, the optimize selection (Brute Force) method has a high complexity and may not be applicable to high dimensional data.

4.2.3. Optimize selection (evolutionary) using genetic algorithm

In evolutionary selection method the most relevant attributes of the data set are selected using evolutionary algorithms, e.g. genetic algorithm (GA). GA is a metaheuristic algorithm that typically utilized to create efficient solutions for optimization by reducing the search space [16].

5. EXPERIMENTAL EVALUATIONS

In this study, Rapid Miner tool was used to conduct experiments on Pima dataset and SVM classifier was selected for building a classification model. The experiments aim at evaluating the impact of combination of pre-processing techniques including missing value replacement and attribute subset selection techniques before performing the classification.

For evaluating the performance of classification, two criteria namely accuracy and precision are employed. Accuracy and precision measure the percentage of correctly classified instances of the unseen data.. Accuracy calculates the sum of actual patients that are classified as patient (TP) and the number of healthy persons that are classified as non-patient instances (TN) relative to the total number of classified instances. This includes the number of incorrectly classified cases 1) the number of healthy persons that are recognized as patient (FP) 2) the number of patients that are classified as healthy persons (FN):

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{1}$$

In other words, accuracy is the proximity of classification outcomes to the actual values without considering the class labels. Conversely, precision is the measure of accuracy provided that the patient class has been predicted [14]:

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

Tables 1 and 2 show the comparative results of accuracy and precision, respectively. In conjunction with first three columns with first three rows, the impact of combination of three missing value replacement techniques with three strategy of attribute subset selection is represented. The results of evaluating the classifier without any preprocessing step are also shown in the junction of the forth column and the forth row.

As can be seen in Tables 1 and 2, when the no preprocessing step is performed, the accuracy and precision of the classifier is lower than cases in which missing value replacements and attribute subset selection methods are applied to the data set before the classification.

Table 1: Impact Of Combination Of Preprocessing Techniques On Accuracy Measure (%)

Attribute Subset Selection	Missing value techniques			
	Replace with mean	Replace with median	Replace with k-NN	none
Combination of Backward Elimination & Forward Selection	74.58	77.78	74.14	-----------
Optimize Selection (Brute Force)	75	79.03	77.55	
Optimize Selection (Evolutionary) Using Genetic Algorithm	**83.33**	80.36	76.56	
Without Attribute Subset Selection	-------------------------------			**72.04**

Table 2: Impact Of Combination Of Preprocessing Techniques On Precision Measure(%)

Attribute Subset Selection	Missing value techniques			
	Replace with mean	Replace with median	Replace with k-NN	none
Combination of Backward Elimination & Forward Selection	80	80	79.5	-----------
Optimize Selection (Brute Force)	83.48	84.35	82.17	
Optimize Selection (Evolutionary) Using Genetic Algorithm	**84.35**	82.61	82.17	
Without Attribute Subset Selection	--------------------------------			**76.56**

Besides, as can be seen in both tables, the combination of optimized selection (evolutionary) using genetic algorithm and replacement with mean techniques (the value at the first column of the third row) generate the best results with an accuracy of 84.35% and precision of 83.33% on Pima data set. So that, by using these two techniques, a significant improvement is achieved with respect to non-existence of preprocessing techniques in terms of accuracy and precision.

Furthermore, Table 3 Shows A Comparison Of The Proposed Method With Some Previous Works In Term Of Accuracy. It Is Evident From This Table That The Proposed Method Enhances The Accuracy About 8% With Respect To The Existing Methods.

Table 3: A Comparison Of Our Proposed Method And Others In Term Of Accuracy Measure (%)

Method	Preprocessing Techniques	Classifier	Accuracy (%)
S. A. Diwani and A. Sam (2014)	-----------	Naïve Base	76.3021
D. A. Kumar and R. Govindasamy (2015)	greedy stepwise approach (reduction technique)	SVM	77.73
A.Parashar et al. (2014)	LDA feature selection (reduction technique)	LDA-SVM	75.65%
Proposed Method	Missing value analysis(replace with mean) & optimize selection using genetic algorithm	SVM	84.35

Firstly, experiments show that the combination of preprocessing techniques can lead to a considerable improve of predictive model for diagnosis of diabetes mellitus. Secondly, among several combinations of missing value techniques with attribute subset selection methods, we found that the combination of replacement with mean as a missing value replacement technique with optimize selection (evolutionary) using genetic algorithm as a data reduction technique provides the best results in terms of accuracy and precision of the predictive model for this particular data set. The implications of this finding is that the mentioned combination can improve the diabetes mellitus prediction. But there is no guarantee that this finding holds for other data sets. It should be noted that similar experiments must be conducted on different types of data obtained from other application domains.

6. CONCLUSION

One of the most typical diseases among several age groups is diabetes mellitus. For fast prediction of this hazardous diseases, several machine learning and other computerized techniques has been applied. These techniques are competing on the performance criteria. However, the low quality of most medical data sets leads to unreliable and imprecise results. In fact, the lack of using preprocessing techniques on these data sets inhibits the construction of efficient models for prediction of diabetes mellitus. This paper dwells on the combination of preprocessing techniques including a combination of missing value replacement and attribute subset selection methods during classification on a well-known diabetes mellitus data set namely Pima data set. Several experiments were conducted for evaluating the effect of missing value analysis and attribute subset selection methods to attain the best accuracy and precision results for the predictive model of diabetes diagnosis. Among many alternatives, we found that the combination of replacement with mean (as missing value analysis method) and optimize selection using genetic algorithm (as attribute subset selection method) outperform their counterparts. In fact, the results show that the proposed technique can improve the performance of applied classifier and outperforms the previous methods in diabetes mellitus prediction in terms of accuracy and precision.

For future work, using other techniques of preprocessing and combining with proposed method is suggested. Moreover, in some data sets that are related to diabetes, applying other classifiers like ensemble methods can be effective.

REFRENCES:

[1] A. Parashar, K. Burse, and K. Rawat, "A Comparative Approach for Pima Indians Diabetes Diagnosis using LDA-Support Vector Machine and Feed Forward Neural Network," *International Journal of Advanced Research in Computer Science and Software Engineering,* vol. 4, 2014, pp. 378-383.

[2] J. Han, J. Pei, and M. Kamber, *Data mining: concepts and techniques*: Elsevier, 2011.

[3] K. Kayaer and T. Yıldırım, "Medical diagnosis on Pima Indian diabetes using general regression neural networks," in *Proceedings of the international conference on artificial neural networks and neural information processing (ICANN/ICONIP)*, 2003, pp. 181-184.

[4] M. Koklu and Y. Unal, "Analysis of a Population of Diabetic Patients Databases with Classifiers," *human resources,* vol. 1, 2013, p. 2.

[5] H. H. Inbarani, A. T. Azar, and G. Jothi, "Supervised hybrid feature selection based on PSO and rough sets for medical diagnosis," *Computer methods and programs in biomedicine,* vol. 113, 2014, pp. 175-185.

[6] S. A. Diwani and A. Sam, "Diabetes Forecasting Using Supervised Learning Techniques," *ACSIJ Advances in Computer Science: an International Journal, ,* vol. 3, 2014, pp. 10-18.

[7] D. A. Kumar and R. Govindasamy, "Performance and Evaluation of Classification Data Mining Techniques in Diabetes," *International Journal of Computer Science and Information Technologies,* vol. 6, 2015, pp. 1312-1319.

[8] P. Agrawal and A. kumar Dewangan, "a brief survey on the techniques used for the diagnosis of diabetes-mellitus," *International Research Journal of Engineering and Technology (IRJET),* vol. 02, 2015, pp. 1039-1043.

[9] A. Purwar and S. K. Singh, "Hybrid prediction model with missing value imputation for medical data," *Expert Systems with Applications,* vol. 42, 2015, pp. 5621-5631.

[10] F. Sambo, B. Di Camillo, A. Franzin, A. Facchinetti, L. Hakaste, J. Kravic, *et al.*, "A Bayesian Network analysis of the probabilistic relations between risk factors in the predisposition to type 2 diabetes," in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2015, pp. 2119-2122.

[11] M. G. Rahman and M. Z. Islam, "Missing value imputation using decision trees and decision forests by splitting and merging records: Two novel techniques," *Knowledge-Based Systems,* vol. 53, 2013, pp. 51-65.

[12] E. Gasca, J. S. Sánchez, and R. Alonso, "Eliminating redundancy and irrelevance using a new MLP-based feature selection method," *Pattern Recognition,* vol. 39, 2006, pp. 313-315.

[13] S.-U. Guan, J. Liu, and Y. Qi, "An incremental approach to contribution-based feature selection," *Journal of Intelligent Systems,* vol. 13, 2004, pp. 15-42.

[14] J. Reunanen, "Overfitting in making comparisons between variable selection methods," *Journal of Machine Learning Research,* vol. 3, 2003, pp. 1371-1382.

[15] J. Han, "Feature selection based on rough set and information entropy," in *2005 IEEE International Conference on Granular Computing*, 2005, pp. 153-158.

[16] P. Thirumal and N. Nagarajan, "utilization of data mining techniques for diagnosis of diabetes mellitus-a case study," *Journal of Engineering and Applied Sciences,* vol. 10, 2006, pp. 8-13.

[17] T. Jayalakshmi and A. Santhakumaran, "A novel classification method for diagnosis of diabetes mellitus using artificial neural networks," in *Data Storage and Data Engineering (DSDE), 2010 International Conference on*, 2010, pp. 159-163.

[18] D. Tutorial, "RapidMiner 4.6," 2012.

[19] J. Han, M. Kamber, and J. Pei, *Data mining: concepts and techniques: concepts and techniques*: Elsevier, 2011.