Original Article

Healthc Inform Res. 2013 September;19(3):177-185.
http://dx.doi.org/10.4258/hir.2013.19.3.177
pISSN 2093-3681 • eISSN 2093-369X

HIR
Healthcare Informatics Research

Real-Data Comparison of Data Mining Methods in Prediction of Diabetes in Iran

Lily Tapak, MSc[1], Hossein Mahjub, PhD[2], Omid Hamidi, MSc[3], Jalal Poorolajal, PhD[2]

[1]Department of Biostatistics, School of Public Health, Hamadan University of Medical Sciences, Hamadan; [2]Research Center for Health Sciences and Department of Epidemiology & Biostatistics, School of Public Health, Hamadan University of Medical Sciences, Hamadan; [3]Department of Science, Hamadan University of Technology, Hamadan, Iran

Objectives: Diabetes is one of the most common non-communicable diseases in developing countries. Early screening and diagnosis play an important role in effective prevention strategies. This study compared two traditional classification methods (logistic regression and Fisher linear discriminant analysis) and four machine-learning classifiers (neural networks, support vector machines, fuzzy c-mean, and random forests) to classify persons with and without diabetes. **Methods:** The data set used in this study included 6,500 subjects from the Iranian national non-communicable diseases risk factors surveillance obtained through a cross-sectional survey. The obtained sample was based on cluster sampling of the Iran population which was conducted in 2005–2009 to assess the prevalence of major non-communicable disease risk factors. Ten risk factors that are commonly associated with diabetes were selected to compare the performance of six classifiers in terms of sensitivity, specificity, total accuracy, and area under the receiver operating characteristic (ROC) curve criteria. **Results:** Support vector machines showed the highest total accuracy (0.986) as well as area under the ROC (0.979). Also, this method showed high specificity (1.000) and sensitivity (0.820). All other methods produced total accuracy of more than 85%, but for all methods, the sensitivity values were very low (less than 0.350). **Conclusions:** The results of this study indicate that, in terms of sensitivity, specificity, and overall classification accuracy, the support vector machine model ranks first among all the classifiers tested in the prediction of diabetes. Therefore, this approach is a promising classifier for predicting diabetes, and it should be further investigated for the prediction of other diseases.

Keywords: Diabetes, Cluster Sampling, Data Mining, Support Vector Machine, Logistic Regression

Submitted: May 1, 2013
Revised: September 8, 2013
Accepted: September 21, 2013

Corresponding Author
Hossein Mahjub, PhD
Research Center for Health Sciences and Department of Epidemiology & Biostatistics, School of Public Health, Hamadan University of Medical Sciences, Hamadan, Iran. Tel: +98-811-8260661, Fax: +98-811-8255301, E-mail: mahjub@umsha.ac.ir

I. Introduction

Diabetes is one of the most common non-communicable disease (NCDs) that has significantly contributed to increased mortality in patients. It is assuredly one of the most challenging health problems in the 21st century that evidently is epidemic in a large number of developing countries [1]. About 135 million people have been estimated to have diabetes, and it is expected to increase to about 300 million by the year 2025 [2].

Besides diabetes, the condition of impaired glucose tolerance (IGT) or pre-diabetes, with elevated blood glucose

levels that increase the risk of developing diabetes, heart disease, and stroke, is also a major public health problem [1]. People with diabetes are at risk of severe and fatal complications [3]. Cardiovascular disease, stroke, retinopathy and blindness, peripheral neuropathy, end-stage renal disease, and mutilation (amputation) are the most serious complications of diabetes [4]. It has been noted in recent studies that by changes in lifestyle or pharmacotherapy, diabetes can be avoided by pre-diabetic persons [5]. Therefore, early screening and diagnosis of diabetes mellitus plays an important role in effective prevention strategies [6]. Moreover, because of the seriousness of diabetes and its complications, providing an efficient and accurate model to predict persons prone to diabetes, especially based on demographic characteristics, is an important issue that should be investigated.

To achieve this purpose, it is possible to identify people who are at risk for the development of diabetes based on common risk factors, such as body mass index (BMI) and family history of diabetes, through a number of predictive models, such as logistic regression [7]. Ideally, it would be important to amend the predictive power of the models predicting diabetes via learning theory and data mining techniques for classification that require no distributional assumptions. Classical techniques, such as logistic regression (LR) and Fisher linear discriminant analysis (LDA), have been widely used for classification of various problems, especially medical ones where the dependent variable is dichotomous [8].

Recently, the positive performance of data mining methods, with classifiers like neural networks (NN), support vector machines (SVM), fuzzy c-mean (FCM), and random forests (RF), has led to considerable research interest in their application to prediction and classification problems [2,7-9].

Research comparing the accuracy of traditional classifiers and computer intensive data mining methods has been steadily increasing. Three data mining method (NN, SVM, and decision tree) were assessed and compared by Kim et al. [10] in a study with LR for mortality prediction. Maroco et al. [8] evaluated various data mining and traditional classifiers (LDA, LR, NN, SVM, classification tree, and RF) for Alzheimer disease. Son et al. [11] compared various kernel functions in the SVM technique for predicting medication adherence in heart failure patients. In another study conducted by Lee et al. [12], the performance of SVM was evaluated and compared with LR for the classification of chronic disease. In a study conducted by Lehmann et al. [13], the performance of four classification methods (RF, SVM, NN, and LDA) were compared for recognition of Alzheimer disease. Hachesu et al. [14] evaluated and compared the performance of three algorithms, namely, decision tree, SVM, and NN, for the classification of coronary artery disease. However, there has been relatively little research related to the performance of data mining methods and comparison of them in diabetes prediction. Yu et al. [9] compared SVM and LR for the classification of undiagnosed diabetes or pre-diabetes vs. no diabetes. Priya and Aruna [15] compared NN and SVM for the diagnosis of diabetic retinopathy.

Although, some authors maintain that classification based on data mining techniques has higher accuracy and lower error rates than the traditional methods (LDA and LR), this excellence is not apparent with all data sets [16-19]. The results of various studies are inconsistent regarding classification accuracy of data mining classifiers as compared to traditional, less computer demanding methods, and there is disagreement regarding the stability of the findings [16,20]. To our knowledge, there has not been a study comparing various data mining classifiers like NN, FCM, RF, and SVM with traditional classifiers for predicting diabetes.

The aim of this study was to provide a comprehensive comparison of six methods (two classic methods and four commonly used data mining methods), namely, LR, LDA, NN, FCM, RF, and SVM, and apply these methods to distinguish people with either undiagnosed diabetes or pre-diabetes from people without these conditions in the Iranian population.

II. Methods

1. Data Source

This study used a data set obtained from the Iranian national NCDs risk factors surveillance through a cross-sectional study which was conducted in 2005–2009 to assess the prevalence of major NCDs risk factors [21].

A two-stage cluster sampling method was used for data collection. As mentioned in [21], the data collection included three steps:

Step 1: collecting questionnaire-based information about health history and behavioral information;

Step 2: using standardized physical measurements to collect physical and physiological data;

Step 3: taking blood samples for biochemical measurement and laboratory examinations of lipids and glucose status that were performed by trained personnel.

The total sample size was 6,500. Participants were diagnosed with diabetes if they had a measured fasting blood sugar (FBS) ≥126 mg/dL, and those with FBS of 110 to 125 mg/dL were considered to have undiagnosed diabetes or pre-diabetes (as IGT). Participants with FBS <110 mg/dL were considered not to have diabetes. To create a dichotomous

 http://dx.doi.org/10.4258/hir.2013.19.3.177

classification, the diabetic and pre-diabetic subjects were classified as a single group and were compared with those without diabetes.

We selected 10 risk factors that are commonly associated with diabetes, including age, gender, BMI, waist circumference, smoking, job, hypertension, residential region (rural/urban), physical activity, and family history of diabetes.

2. Data Pre-processing and Dealing with Missing Values

Pre-processing of the data set was done in two steps. First, fields with spelling errors, additional tokens, other irregularities and irrelevancies, such as outliers were removed or corrected. Second, because of the missing completely at random (MCAR) mechanism for missingness based on Little MCAR test [22] (p = 0.561), persons with at least one missing variable were removed from analysis. Table 1 shows the demographic and clinical characteristics of the participants.

3. Data Mining Algorithms

1) Neural networks

The NNs method is a flexible mathematical conformation for information processing which is well suited for forecasting, pattern recognition, and classification problems. NNs include multiple input nodes and weighted interconnections [14]. One of the most used NNs is the multilayer perceptron (MLP), in which its neurons apply a nonlinear activation function to calculate their outputs. The activation function includes a sigmoid function ($f(x) = 1 / (1 + \exp(-x))$) in the hidden layer and a linear function ($f_j(x) = \sum_{i=1}^{p} w_{ij}x_i$, where x_i's are predictor variables and w_{ij}'s are input weights) in the output layer. The functional form of the MLP can be written as

$$y_k = f\left(\sum_{i=1}^{N} w_{ji}x_i + b_j\right),$$

where x_i is the i-th nodal value in the previous layer, y_j is the j-th nodal value in the present layer, b_j is the bias of the j-th node in the present layer, w_{ji} is a weight connecting x_i and y_j, N is the number of nodes in the previous layer, and f is the activation function in the present layer [23].

2) Support vector machines

The SVM is a supervised machine learning technique which has wide application in regression and classification problems. The SVM algorithm carries out a classification by mapping a vector of predictors into a higher dimensional plane via maximization of the margin between two data classes [23]. High discriminative power is achieved by using either linear or non-linear kernel functions to alter the input space into a multidimensional space [9].

In the binary classification mode, the equation of the hyperplane segregating two groups, say {-1, +1} in a higher-dimension feature space is given by the relation $y(t)=\sum_{i=1}^{D} w_i\phi_i + b = 0$, where $\{\phi_i(x)\}_{i=1}^{D}$ denotes features, b and $\{w_i\}_{i=1}^{D}$ denote

Table 1. Demographic and clinical characteristics of participants (n = 6,500)

Characteristic	Value
Sex	
Men	3,250 (50.00)
Women	3,250 (50.00)
Smoking	
Yes	716 (17.90)
No	3,284 (82.10)
Job	
Government employee	433 (6.72)
Non-government employee	61 (0.95)
Self-employed	2,036 (31.61)
Unpaid work	45 (0.70)
Student	589 (9.14)
Soldier	10 (0.16)
Housewife	2,641 (41.00)
Retired	220 (3.42)
Jobless able	233 (3.62)
Jobless disabled	85 (1.32)
Others	89 (1.38)
Hypertension	
Yes	1,183 (18.20)
No	5,317 (81.80)
Residential region	
Rural	2,920 (44.92)
Urban	3,580 (55.08)
Physical activity	
Yes	4,296 (66.10)
No	2,204 (33.90)
Family history	
Yes	517 (17.23)
No	2,483 (82.77)
Body mass index	
<20	858 (13.21)
20–24	2,569 (39.55)
25–30	2,088 (32.14)
>30	981 (15.10)
Waist circumference	86.84 ± 0.22
Age (yr)	40.10 ± 14.39

Values are presented as number (%) or mean ± standard deviation.

coefficients that have to be estimated from the data, and $\{(x_i, y_i)\}_{i=1}^{N}$ are a set of samples where $y_i \in \{+1, -1\}$.

In summary, the goal of SVM can be regarded as the solution of the following quadratic optimization problem: $\min_{W,\xi,b} \frac{1}{2}\|w\|^2 + C(\sum_{i=1}^{N}\xi_i)^k$ subject to $y_i(w^T.\Phi(x_i)+b) \geq 1-\xi_i$, where $\xi_i \geq 0$, $i=1$, ..., n, where the training data are mapped to a higher dimensional space by the function Φ, and C is a user-defined penalty parameter on the training error that controls the trade-off between classification errors and the complexity of the model. Therefore, the decision function (predictor) is $f(x)=\text{sign}(w^T\Phi(x)+b)$, where x is any testing vector [11].

In addition, to derive the optimal hyperplane for not linearly separated data, several solutions called kernel functions have been proposed and adopted for SVM. A kernel function is written as $K(x_i, x_j)=\Phi(x_i)^T\Phi(x_j)$, and the most widely used four kernel functions are the linear ($K(x_i, x_j)=x_i^Tx_j$); radial basis function (RBF) ($K(x_i, x_j) = \exp(-\gamma\|x_i-x_j\|^2)$, where γ is the kernel parameter); polynomial ($K(x_i, x_j) = (x_i^Tx_j+1)^d$, $d > 0$ is the degree of the polynomial kernel); and sigmoid ($K(x_i, x_j) = \tanh(x_i^Tx_j+1)$).

3) Random forests

RF is a new "ensemble learning" method in classification which is designed to produce accurate predictions without over fitting the data [8]. This method constructs a series of unpruned classification trees using random bootstrap samples of the original data sample. The outputs of all trees are aggregated to produce one final classification, i.e., the object belongs to a class with the majority of predictions given by the trees in the random forest [8].

4) Fuzzy c-mean

FCM is a method of clustering in which each piece of data may belong to more than one cluster [2] with varying degrees of membership for each cluster. The goal of this method can be regarded as minimizing of the following objective function via an iterative optimization algorithm: $J_m = \sum_{i=1}^{n}\sum_{j=1}^{c} u_{ij}^m \|x_i - c_i\|$, $1<m<a$, where m is any real number greater than 1 which is called the fuzziness index and controls the fuzziness of membership of each observation, x_i is the i-th component of d-dimensional observed data, u_{ij} is the degree of membership of x_i in cluster $j(u_{ij} \in [0,1]$, $\sum_{j=1}^{c} u_{ij}=1 \ \forall i=1,2,...,n, \ \forall j=1,2,...,c)$, c_j is the d-dimension center of the cluster, and $\|\ \|$ is any norm, such as Euclidean distance expressing the similarity between any observed data and the center of cluster [2].

4. Implementation and Performance Criteria

To avoid overfitting due to the use of the same data for the training and testing of different classification methods, a 10-fold cross-validation strategy was used in the training data set. In this regard, the total data set was partitioned into 10 nearly equal subsets. In each of the 10 steps, 9/10 of the sample was used for training and 1/10 for testing.

In this research, all classification methods were implemented on the diabetes data set by using R packages [24] (e1071, nnet, randomForest, glmnet, MASS and pROC) through related functions (e.g.,"cmeans", "svm", "nnet", "randomForest", "lda", and "glm"). The SVM-based model building process was carried out with RBF as kernels. Also, this study employed an MLP with one hidden layer to construct the NN model for the diabetes data set. After fitting the models using the 10-fold cross-validation strategy, predicted values were obtained to evaluate the methods' performance.

To compare the discriminative powers of the six models, receiver operating characteristic (ROC) curves were generated based on the predicted outcome and true outcome, and the area under the curves (AUCs) for the data sets were calculated.

Sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and total accuracy were calculated based on the following formulas:

$$\text{Sensitivity}=\frac{TP}{TP+FN}, \quad \text{Specificity}=\frac{TN}{TN+FP}$$

$$\text{PPV}=\frac{TP}{TP+FP}, \quad \text{NPV}=\frac{TN}{TN+FN}$$

$$\text{Total Accuracy}=\frac{TP+TN}{TP+TF+TN+FN}$$

where TP, FP, TN, and FN represent the number of true positives, false positives, true negatives, and false negatives, respectively [9].

In addition, to find the optimal cut-off point value for calculation of predictive performance in LR, several points were examined, and 0.1 was obtained.

III. Results

The performance of the six classifiers was evaluated in terms of their discriminative accuracy by AUC, sensitivity (the proportion of persons that have diabetes and were correctly diagnosed), specificity (the proportion of persons that did not have diabetes and were correctly diagnosed), PPV, NPV, and total accuracy (Table 2; Figure 1) from the 10-fold cross validation strategy. A graphical comparison of six quantities is shown in Figure 1 to evaluate the performance of the models.

As seen in Table 2, almost all the algorithms generate high

 http://dx.doi.org/10.4258/hir.2013.19.3.177

Table 2. The performance of six classifiers

Method	Sensitivity	Specificity	PPV	NPV	AUC	Total accuracy
Logistic regression	0.133	0.999	0.914	0.935	0.763	0.935
Linear discriminant analysis	0.006	0.998	0.200	0.926	0.710	0.925
Fuzzy c-mean	0.330	0.901	0.210	0.944	0.678	0.859
Support vector machine	0.820	1.000	1.000	0.991	0.979	0.986
Neural network	0.084	0.998	0.750	0.931	0.751	0.931
Random forest	0.081	0.998	0.795	0.932	0.717	0.930

PPV: positive predictive value, NPV: negative predictive value, AUC: area under ROC curve, ROC: receiver operating characteristic.

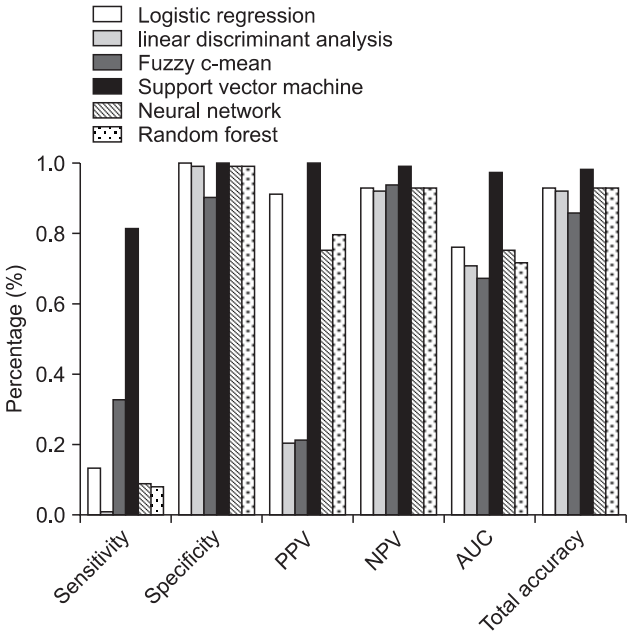

Figure 1. Performance criteria of the six classification methods. PPV: positive predictive value, NPV: negative predictive value, AUC; area under ROC curve, ROC: receiver operating characteristic.

specificity (more than 90%). However, the sensitivity values of the classical methods (LR, 13.3%; LDA, 0.6%), NN (8%) and RF (8%) were very low, and the sensitivity of FCM (33%) was relatively low. In addition, the highest sensitivity in our experiments was obtained with SVM (82%).

Although, the PPV of the LDA (20%) and FCM was relatively low in comparison with the other four methods (SVM, 100%; LR, 91.4%; RF, 79.5%; and NN, 75%), the NPV performance of these two (92.6% for LDA and 94.4% for FCM) is as good as that of the other four technique (SVM, 99.1%; LR, 93.5%; RF, 93.2%; and NN, 93.1%).

The overall discriminative ability of classification schemes is represented by their AUC values, which are 97.9% for SVM with RBF kernel, 76.3% for LR, 75.1% for NN, 71.7% for RF, and 67.8% for FCM (Table 2; Figure 2). Furthermore, all techniques produced a total accuracy of more than 85%. The highest total accuracy was achieved by SVM.

Thus, the SVM approach appears to perform better than the traditional models and the other three data mining methods.

IV. Discussion

It is clear from the sensitivity, specificity, and AUC values presented here that SVM has a distinct advantage over the other methods in terms of predictive capabilities, and it is more effective than LR, LDA, FCM, RF, and NN.

With the exception for FCM (0.67), in terms of AUC, the discriminant power of classification methods was appropriate for most classifiers (more than 0.7). Specificity ranged from a minimum of 0.901 (FCM) to a maximum of 1 (SVM). All the classifiers were quite efficient in predicting group membership in the group with a larger number of elements (the normal group corresponding to 93% of the sample). In terms of total accuracy, SVM outperformed all other classification methods; however, other methods also achieved high total accuracy (more than 0.85).

Judging from the sensitivity of the classification methods, prediction for the group with lower frequency (the diabetic group, 7% of the sample) was quite poor for almost all the used classifiers, with the exception of SVM, despite its high specificity and total accuracy.

The minimum sensitivity value was 0.006 (LDA), and maximum sensitivity was 0.820 (SVM, followed by 0.33 for FCM). Only one of the six tested classifiers showed a sensitivity value higher than 0.5.

Considering that having diabetes is the key prediction in this biomedical application, a classification method with higher sensitivity is desired; therefore, classification methods, such as LR, NN, FCM, RF, and LDA, are inappropriate for this type of binary classification task. However, SVM showed a good sensitivity result; hence, it is an appropriate method for classification. This finding is similar to the results

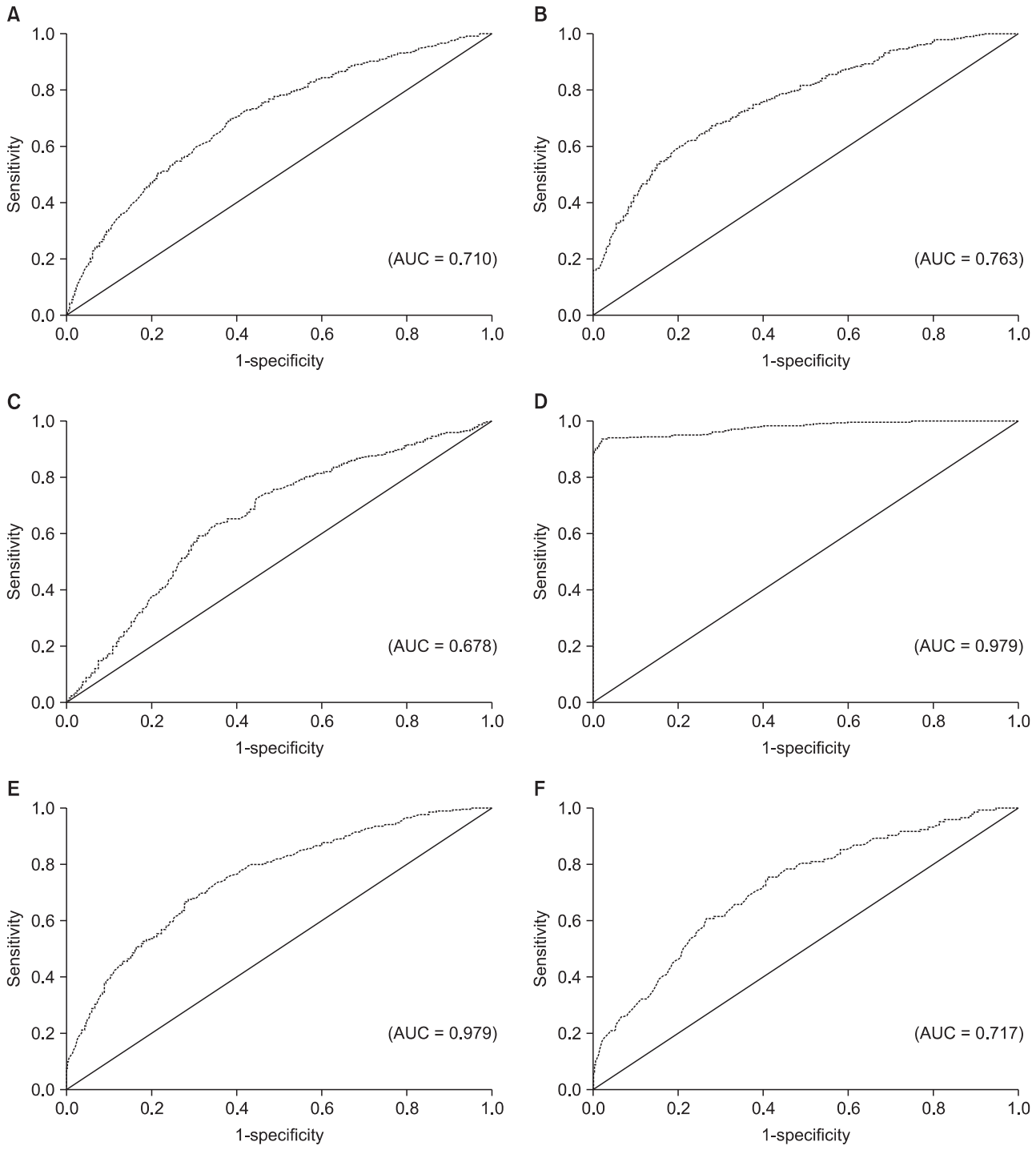

Figure 2. Receiver operating characteristic (ROC) curves for comparison of the six classification methods: (A) linear discriminant analysis, (B) logistic regression, (C) fuzzy c-mean, (D) support vector machine, (E) neural network, and (F) random forest. AUC: area under ROC curve.

of other works comparing various classification methods in other biomedical conditions [2,9].

In a study to compare NN and SVM for diagnosing diabetic retinopathy, Priya and Aruna [15] reported better performance for SVM (accuracy 89.6% for NN and 97.61% for SVM). In a study comparing three data mining methods (NN, SVM, and decision tree) with LR, Kim et al. [10] concluded that the decision tree algorithm slightly outperformed (AUC, 0.892) the other data mining techniques, followed by the artificial neural network (AUC, 0.874) and SVM (AUC, 0.876), which is in contradiction to our results.

In another study [2] titled "Review of automated diagnosis

http://dx.doi.org/10.4258/hir.2013.19.3.177

of diabetic retinopathy" using the support vector machine achieved 99.45% for sensitivity and 100% for specificity, which is similar to our results.

Son et al. [11] demonstrated that SVM with appropriate kernel function can be a promising tool for predicting medication adherence in heart failure patients (sensitivity, 77.6%; specificity, 81.6%; PPV, 77.8%; NPV, 77.6%; and total accuracy, 77.6%)

Lehmann et al. [13] compared several classification methods (RF, SVM, NN, and LDA) for recognition of Alzheimer disease and found that data mining classifiers show a slight superiority compared to classical ones, whereas in our work SVM showed a high performance compare to the other methods. In their study, the sensitivity and specificity of SVM were 89% and 88%, respectively.

In a study conducted by Hachesu et al. [14], three algorithms, namely, decision tree, SVM and NN, were compared for the classification of coronary artery disease. Their findings demonstrated that all three algorithms showed various acceptable degrees of accuracy for prediction, and the SVM was the best fit (96.4% total accuracy and 98.1% sensitivity), which is similar to our results.

Maroco et al. [8] reported in their comparison study of data mining and traditional classifiers (LDA, LR, NN, SVM, classification tree, and RF) the highest total accuracy and specificity for SVM and the lowest sensitivity, while in the present study, SVM had the highest sensitivity. Yu et al. [9], in a comparison between SVM and LR for the classification of undiagnosed diabetes or pre-diabetes vs. no diabetes, showed that the SVM performance based on AUC is as good as that of LR (73.2% for SVM and 73.4 for LR). In contrast, the performance of SVM in our study was better than that of LR.

In another study Lee et al. [12] compared SVM and LR for the classification of chronic disease and showed that SVM achieved higher accuracy with a smaller number of variables than the number of variables used in LR (71.1% for LR and 97.3% for SVM), which is consistent with our results.

Some methods are only good for predicting the larger group membership (high specificity) but quite insufficient in predicting the smaller group membership (low sensitivity); therefore, selecting classification methods only based on total accuracy can be spurious [8]. Some real-data studies have reported unbalanced efficiency for small frequency vs. large frequency groups in LR, NN, and SVM [8,20,25]. However, to our knowledge, such imbalance of RF has not been published elsewhere.

Based on the six performance criteria (total accuracy, specificity, sensitivity, PPV, NPV, and AUC), the traditional classifiers (LDA and LR) appear to perform as well as the FCM, NN, and RF (the newest member of the binary classification family).

It seems that the relatively low observed prevalence of diabetes may limit the performance of some data mining methods evaluated in this study. The present unbalanced sample sizes of two groups did not limit the achievement of acceptable accuracy, specificity, and sensitivity of SVM as reported by other studies [15,26,27]. Furthermore, there have been studies with fairly small samples in which new classification methods such as RF and NN have been applied with high accuracy [8,13,28]. Some studies have reported equivalent or even superior performance of LR and LDA in comparison with NN, SVM, RF, and FCM [9,13,20,29,30]. Since the performance of NN and SVM depends on tuning parameters, these parameters were optimally determined by grid search.

This study focused on the performance of six classification methods in detecting cases of diabetes and pre-diabetes in the Iranian population. Our results demonstrated that the discriminative performance of SVM models was superior to that of other commonly used methods. Therefore, it can be applied successfully for the detection of a common disease with simple clinical measurements. SVM is a nonparametric method that provides efficient solutions to classification problems without any assumption regarding the distribution of data. The SVM method is a learning machine technique in modeling nonlinearity based on minimization of structural risk which avoids finding local minimums instead of general ones because of minimizing structural risk function. Because of the convex optimality problem, SVM gives a unique solution. This is an advantage of SVM compared to other methods, such as NN, which have multiple solutions associated with local minimum and for this reason may not be robust over different samples. In addition, with appropriate choice of kernel function (e.g., RBF kernel) and related parameters, the similarity between individuals is increased. Therefore, when classifying a new subject, it is assigned to the group with the highest similarity [31].

This work demonstrates the predictive power of the SVM with unequal sample sizes. Yu et al. [9] performed a similar study in which they compared LR and SVM performance on diabetes data and concluded that the SVM approach performed as well as the LR model.

Generally, one cannot find a method that always is the best for the classification of different datasets.

Conflict of Interest

No potential conflict of interest relevant to this article was reported.

Acknowledgments

We would like to thank the Ministry of Health and the Iranian Center for Disease Control and Management for financial support of the Iranian non-communicable diseases surveillance system. We also wish to thank the Vice-chancellor of Health of Hamadan University of Medical Sciences and all health workers for their valuable effort and contribution with this survey.

References

1. International Diabetes Federation. IDF Diabetes Atlas: the global burden [Internet]. Brussels, Belgium: International Diabetes Federation; c2013 [cited at 2013 Sep 1]. Available from: http://www.idf.org/diabetesatlas/5e/the-global-burden.
2. Priya R, Aruna P. Review of automated diagnosis of diabetic retinopathy using the support vector machine. Int J Appl Eng Res (Dindigul) 2011;1(4):844-63.
3. Calder R, Alexander C. Cardiovascular disease in people with diabetes mellitus. Pract Diabetol 2000;19(4):7-18.
4. Barr EL, Zimmet PZ, Welborn TA, Jolley D, Magliano DJ, Dunstan DW, et al. Risk of cardiovascular and all-cause mortality in individuals with diabetes mellitus, impaired fasting glucose, and impaired glucose tolerance: the Australian Diabetes, Obesity, and Lifestyle Study (AusDiab). Circulation 2007;116(2):151-7.
5. Pi-Sunyer FX. How effective are lifestyle changes in the prevention of type 2 diabetes mellitus? Nutr Rev 2007;65(3):101-10.
6. IDF Clinical Guidelines Task Force. Global Guideline for Type 2 Diabetes: recommendations for standard, comprehensive, and minimal care. Diabet Med 2006;23(6):579-93.
7. Thomas C, Hypponen E, Power C. Type 2 diabetes mellitus in midlife estimated from the Cambridge Risk Score and body mass index. Arch Intern Med 2006;166(6):682-8.
8. Maroco J, Silva D, Rodrigues A, Guerreiro M, Santana I, de Mendonca A. Data mining methods in the prediction of Dementia: a real-data comparison of the accuracy, sensitivity and specificity of linear discriminant analysis, logistic regression, neural networks, support vector machines, classification trees and random forests. BMC Res Notes 2011;4:299.
9. Yu W, Liu T, Valdez R, Gwinn M, Khoury MJ. Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes. BMC Med Inform Decis Mak 2010;10:16.
10. Kim S, Kim W, Park RW. A comparison of intensive care unit mortality prediction models through the use of data mining techniques. Healthc Inform Res 2011;17(4):232-43.
11. Son YJ, Kim HG, Kim EH, Choi S, Lee SK. Application of support vector machine for prediction of medication adherence in heart failure patients. Healthc Inform Res 2010;16(4):253-9.
12. Lee SK, Kang BY, Kim HG, Son YJ. Predictors of medication adherence in elderly patients with chronic diseases using support vector machine models. Healthc Inform Res 2013;19(1):33-41.
13. Lehmann C, Koenig T, Jelic V, Prichep L, John RE, Wahlund LO, et al. Application and comparison of classification algorithms for recognition of Alzheimer's disease in electrical brain activity (EEG). J Neurosci Methods 2007;161(2):342-50.
14. Hachesu PR, Ahmadi M, Alizadeh S, Sadoughi F. Use of data mining techniques to determine and predict length of stay of cardiac patients. Healthc Inform Res 2013;19(2):121-9.
15. Priya R, Aruna P. SVM and neural network based diagnosis of diabetic retinopathy. Int J Comput Appl 2012;41(1):6-12.
16. Finch H, Schneider MK. Classification accuracy of neural networks vs. discriminant analysis, logistic regression, and classification and regression trees: three- and five-group cases. Methodology (Gott) 2007;3(2):47-57.
17. Austin PC. A comparison of regression trees, logistic regression, generalized additive models, and multivariate adaptive regression splines for predicting AMI mortality. Stat Med 2007;26(15):2937-57.
18. Gelnarova E, Safarik L. Comparison of three statistical classifiers on a prostate cancer data. Neural Netw World 2005;15(4):311-8.
19. Green M, Bjork J, Forberg J, Ekelund U, Edenbrandt L, Ohlsson M. Comparison between neural networks and multiple logistic regression to predict acute coronary syndrome in the emergency room. Artif Intell Med 2006;38(3):305-18.
20. Meyer D, Leisch F, Hornik K. The support vector machine under test. Neurocomputing 2003;55(1-2):169-86.
21. Poorolajal J, Zamani R, Mir-Moeini R, Amiri B, Majzoobi M, Erfani H, et al. Five-year evaluation of chronic diseases in Hamadan, Iran: 2005-2009. Iran J Public Health 2012;41(3):71-81.
22. Little RJ. A test of missing completely at random for multivariate data with missing values. J Am Stat Assoc 1988;83(404):1198-1202.
23. Yoon H, Jun SC, Hyun Y, Bae GO, Lee KK. A compara-

 http://dx.doi.org/10.4258/hir.2013.19.3.177

tive study of artificial neural networks and support vector machines for predicting groundwater levels in a coastal aquifer. J Hydrol 2011;396(1-2):128-38.
24. The Comprehensive R Archive Network (CRAN) package [Internet]. The R Foundation; [cited at 2013 Sep 1]. Available from: http://cran.r-project.org/web/packages/.
25. Schwarzer G, Vach W, Schumacher M. On the misuses of artificial neural networks for prognostic and diagnostic classification in oncology. Stat Med 2000;19(4):541-61.
26. Oliveira PP Jr, Nitrini R, Busatto G, Buchpiguel C, Sato JR, Amaro E Jr. Use of SVM methods with surface-based cortical and volumetric subcortical measurements to detect Alzheimer's disease. J Alzheimers Dis 2010;19(4):1263-72.
27. Zhu Y, Tan Y, Hua Y, Wang M, Zhang G, Zhang J. Feature selection and performance evaluation of support vector machine (SVM)-based classifier for differentiating benign and malignant pulmonary nodules by computed tomography. J Digit Imaging 2010;23(1):51-65.
28. Abbasimehr H, Setak M, Tarokh MJ. A neuro-fuzzy classifier for customer churn prediction. Int J Comput Appl 2011;19(8):35-41.
29. Breiman L. Random forests. Mach Learn 2001;45(1):5-32.
30. Smith A, Sterba-Boatwright B, Mott J. Novel application of a statistical technique, random forests, in a bacterial source tracking study. Water Res 2010;44(14):4067-76.
31. Auria L, Moro RA. Support vector machines (SVM) as a technique for solvency analysis. Berlin, Germany: Deutsches Institut fur Wirtschaftsforschung; 2008.