Rule Extraction From Support Vector Machines Using Ensemble Learning Approach: An Application for Diagnosis of Diabetes

Longfei Han, Senlin Luo, Jianmin Yu, Limin Pan, and Songjing Chen

Abstract—**Diabetes mellitus is a chronic disease and a worldwide public health challenge. It has been shown that 50–80% proportion of T2DM is undiagnosed. In this paper, support vector machines are utilized to screen diabetes, and an ensemble learning module is added, which turns the "black box" of SVM decisions into comprehensible and transparent rules, and it is also useful for solving imbalance problem. Results on China Health and Nutrition Survey data show that the proposed ensemble learning method generates rule sets with weighted average precision 94.2% and weighted average recall 93.9% for all classes. Furthermore, the hybrid system can provide a tool for diagnosis of diabetes, and it supports a second opinion for lay users.**

Index Terms—**diagnosis of diabetes, ensemble learning, random forest (RF), rule extraction, support vector machines (SVMs).**

I. Introduction

Type 2 diabetes mellitus (T2DM) is an increasingly worldwide public health problem. It causes a substantial burden of cost and various complications. The prevalence of diabetes is growing rapidly, and it is estimated that over 550 million people will suffer from diabetes by 2030 [1]. Approximately 50–80% proportion of T2DM is undiagnosed [2]. Especially, the proportion of undiagnosed diabetes is high in developing countries [3], such as China (∼50%), Saharan Africa (∼80%), and Western Pacific (∼60%).

Consequently, considering the clinical impact and healthcare cost of T2DM, it is not surprising that feasible and comprehensive approaches should be proposed to identify people with undiagnosed diabetes. Recently, over 100 risk models for risk detection or risk prediction have been published, and most of them appeared in 2008–2012 [4]. There are two broad approaches for the detection of undiagnosed T2DM. One approach only focuses on assessment of traditional noninvasive factors, such as age, family history, body-mass index, and gender. These models are usually designed as questionnaires, which can be completed by lay individuals. Typical examples of this approach are Diabetes Risk Calculator [5], Cambridge Risk Score [6], MDPPQ [7], ADA [8], Simple Chinese risk score, etc. [9], [10]. The other approach focuses on both the noninvasive and metabolic factors, which should require blood sampling and laboratory measurements, such as high density lipoprotein (HDL), and cholesterol (CHOL). On the other hand, the risk assessment model or risk score can be classified into detective model and predictive model. In this paper, we only focus on the risk models for detection of undiagnosed T2DM here. Other aspect for identifying individuals at risk of developing T2DM was systematically summarized in three recently published reviews [10]–[12].

Over a decade, researchers have applied machine-learning methods for diabetes assessment. Tapak *et al*. compared five machine-learning classifiers [13] (neural network, support vector machines (SVMs), fuzzy c-mean, random forest (RF), and linear discriminant analysis) to classify individuals with or without diabetes; Velu and Kaswan compared three clustering techniques [14] (expectation-maximization algorithm, *h*-means clustering, and genetic algorithm); Akgobek compared six direct rule-extraction approaches [15] (C4.5, NaiveBayes, CN2, PART, CORE, GA-SVM) for assessment on cancer and diabetes; Lee *et al.* applied various classification algorithms including Quest, SVM, C4.5, logistic regression (LR), and K-NN to develop a predictive model using genetic data [16]; Barakat *et al.* [17] proposed intelligible SVMs for diagnosis of diabetes mellitus, etc. (see [18]–[20]).

In this paper, we present an ensemble learning approach for rule extraction from the SVM, which uses RF rule induction technique to develop an affordable and feasible assessment rules for diagnosis of diabetes. In our proposed method, support vectors (SVs) are first extracted from the SVM with acceptable accuracy. Then, new labels of SVs are predicted by the trained SVM model, and original labels of SVs are replaced by predicted labels. Finally, the synthetic data are fed to RF to generate rules.

For extracted rule sets, if the decision tree is large, then each leaf of the tree may have few examples. On the other hand, if the tree is too small, then tree may learn few patterns. All these drawbacks make single decision tree (C4.5) difficult to fit complex models. By utilizing the ensemble learning method, RF can solve the problem mentioned previously. Meanwhile, considering the rule sets are generated from the SVs, the rule sets obtained by SVM + RF are certainly much less and smaller than those of RF, where the large rule sets may make the problem incomprehensible. Moreover, for the skewed classification problem, the proposed method can be a preprocessing technique to reduce the imbalance proportion of skewed data, which can

Manuscript received January 27, 2014; revised April 1, 2014; accepted May 11, 2014. Date of publication May 19, 2014; date of current version March 2, 2015. This work was supported by the 12th five-year National Science and Technology Supporting Project (Approval No: 2012BAI10B01).

L. Han is with the Beijing Institute of Technology, Beijing 100081, China (e-mail : hanlongfei@hotmail.com).

S. Luo, J. Yu, L. Pan, and S. Chen are with the Information System and Security and Countermeasures Experimental Center, Beijing Institute of Technology, Beijing 100081, China (e-mail :luosenlin126@126.com; 945200115@qq.com; 632738491@qq.com; 981886724@qq.com).

Digital Object Identifier 10.1109/JBHI.2014.2325615

improve precision and recall in positive class. The model can assess undiagnosed individuals in an understandable form and give a more comprehensive and transparent representation for end users.

This paper is structured as follows: In Section II, previous rule-extraction techniques from the SVM are reviewed. Section III explains the presented method. Section IV describes dataset and experimental procedure. Finally, Section V presents the results and discussions, and Section VI gives a conclusion for the paper.

II. Rule Extraction From the SVM

SVMs and artificial neural network (ANN) have shown better performance than other machine-learning algorithms in some application areas, such as speech recognition, computer vision, and medical diagnosis. However, the SVM and ANN have an inherent inability to explain models and results, because these algorithms construct black box models [21] and learn patterns with no transparence and comprehensibility to humans. This drawback of these models impedes their application in some areas, especially in medical diagnosis [22].

In recent years, a proliferation of rule-extraction methods for trained SVMs has been proposed. These motifs can be classified into three basic categories: "decompositional" (or transparent), "pedagogical" (or learning based), and "eclectic" (or hybrid) [23]. The decompositional approach focuses on extracting region-based rules by SVs and separating hyperplane. For instance, Núñez *et al.* [24] first proposed the SVM + prototype method, and utilized the defined regions (ellipsoids and hyperrectangles) to refine the rules; Fu *et al.* suggested a RulExSVM method for nonlinear SVMs; Zhang *et al.* [25] proposed the hyperrectangle rule extraction (HRE) algorithm; and Fung *et al.* [26] suggested a linear programming formulation approach for rule extraction from linear SVMs. By contrast, the pedagogical approach treats SVMs model as a black box and uses the generated model to predict the label (class) for an extended data or unlabeled data. The resulting patterns are then used to train a decision tree learning system and to extract the corresponding rule sets [27], such as GEX and G-REX [28]. The eclectic approach incorporates both "decompositional" and "pedagogical" techniques; it only uses the SVs or applied rule-based model to train the artificial data based on SVs. Barakat *et al.* proposed an SQRex-SVM algorithm based on the sequential covering approach. Fuzzy rule extraction [29], Trepan [30], active learning-based approach [31], and other hybrid approaches [32] were suggested as well.

III. Proposed Rule-Extraction Approach

The proposed rule-extraction approach divides the rule-extraction process into three basic steps. During the first step, the training data are applied to build an SVM model with acceptable accuracy by tuning the parameters. In order to obtain a set of rules that can explain the logical workings of SVMs, the trained SVM model should be regarded as an oracle to provide class label y_i. Then, SVs are extracted and predicted by the obtained SVM model, and the predicted labels of SVs will replace the original labels of SVs to generate a synthesis dataset. The motivation of changing labels here is to ensure the future generated rules can mimic the predictions of SVMs as closely as possible. The idea behind this technique is the assumption that the trained SVM model is an oracle, and it can better represent the patterns than the synthesis dataset. By changing the class labels of the data, some noise, that is class overlap in the data, is removed from the data. During the second step, the artificial data are put into RF algorithm, and the best rule sets are generated by tuning the rule induction method's parameters. In the last step, the rules are evaluated by the test data of given classification problem. The proposed approach is depicted in Fig. 1.

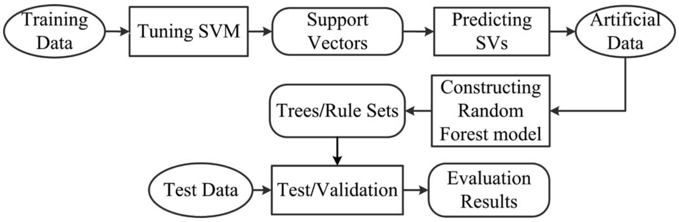


Fig. 1. Block diagram of the proposed rule-extraction approach.

A. Data Preparation

The dataset is firstly divided into two parts, 90% of the data is used for rule extraction, and last 10% of data keeps as test set. In rule-extraction process, tenfold cross validation (CV) is used as the training method to obtain the optimal parameters of models, and tenfold results incorporate together to calculate the averaged accuracy of tenfold CV for the model.

After accomplishing the tenfold validation, the best fold, which has the best precision and recall rate, is regarded as chosen set. Then, the model of best fold is used to generate rule sets and evaluate by the test set.

B. Support Vector Machines

SVM is based on the principle of structural risk minimization [33], and it belong to the supervised learning models for nonlinear classification analysis. The SVM model is achieved by finding the optimal separating hyperplane $(w \cdot x + b = 0)$ with maximizing the margin *d*, which is defined as $d = 2/\,||w||$. This optimal hyperplane can be represented as a convex optimization problem:

$$\text{minimize } \frac{1}{2}\,||w||^2 \text{ subject to } y_i\,(wx_i + b) \geq 1. \tag{1}$$

Then, through introducing the slack variable ε_i to intensify the generalization, the SVM allows for errors on the training set, and (1) is modified as

$$\begin{aligned}&\text{minimize } \frac{1}{2}\,||w||^2 \text{ subject to}\\&y_i\,(wx_i + b) \geq 1 - \varepsilon_i \quad \forall i, \varepsilon_i \geq 0, y_i = +1 \text{ or } -1.\end{aligned} \tag{2}$$

In the case of controlling the tradeoff between overfitting and underfitting, a regularization parameter *C* is defined in the optimization problem, and the (2) is modified to a quadratic programming optimization problem:

$$\begin{aligned}&\text{minimize } \frac{1}{2}\left\|w\right\|^2 + C\sum_{i=1}^{n}\varepsilon_i\\&\text{subject to } y_i\ \left(wx_i + b\right) \geq 1 - \varepsilon_i, \varepsilon_i \geq 0.\end{aligned} \tag{3}$$

For convenience of solving the problem, the primal problem is transformed to its dual optimization problem by introducing dual Lagrange multiplier α. Hence, (3) became

$$\text{maximize } w\left(\alpha\right) = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1,j=1}^{n}\alpha_i y_i \alpha_j y_j \left\langle x_i, x_j\right\rangle$$

$$\text{subject to } C \geq \alpha_i \geq 0 \quad \forall i, \sum_{i=1}^{n}\alpha_i y_i = 0. \tag{4}$$

In the nonlinear classification problem, the SVM uses kernel functions to map the examples into the high-dimensional feature space and separates categories by a clear linear margin [34]. Usually, radial basis function (RBF) is used as the kernel function to map the data:

$$K\left(x, x'\right) = \exp\left(\frac{\left\|\mathrm{x} - x'\right\|_2^2}{2\sigma^2}\right) \tag{5}$$

where $\left\|x - x'\right\|_2^2$ is the squared Euclidean distance between two vectors, and σ^2 is a free parameter, which can involve a parameter gamma:

$$\gamma = -\frac{1}{2\sigma^2}. \tag{6}$$

Gamma has an impact on generalization capability of the SVM, higher value of gamma will make the SVM model have larger number of SVs, and cause the overfitting problem.

Finally, by including kernel functions, (4) now becomes

$$\begin{aligned}&\text{maximize } w\left(\alpha\right) = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1,j=1}^{n}\alpha_i y_i \alpha_j y_j K\left(x_i, x_j\right)\\&\text{subject to } C \geq \alpha_i \geq 0 \quad \forall i, \sum_{i=1}^{n}\alpha_i y_i = 0.\end{aligned} \tag{7}$$

Hence, solving for α by the gradient decent algorithm, the SVs can be obtained by the examples of training data which have nonzero Lagrange multiplier. The hyperplane is completely defined by SVs as

$$f\left(x\right) = \text{sign}\left(\sum_{s=1}^{sv}\alpha_s y_s K\left(x_s, x\right) + b\right). \tag{8}$$

SVs are the only examples that make contribution to the classification of the SVM. Through tuning the parameter *C* and gamma with tenfold CV, we can obtain the SVs from the SVM with best acceptable precision.

C. Rule Generation and Evaluation

The RF [35] is an ensemble learning method for classification. RF constructs a multitude of decision trees and utilizes the mode of individual trees' output to classify the patterns. In the traditional decision tree method, it will be difficult to fit complex models (such as SVMs) if the tree is so large that each only has few examples. Unlike the decision tree, however, RF combines random subspace method and bagging idea to optimize the nonlinear problem, and it is trained based on ensemble learning, which uses multiple models to obtain better performance than any constituent model. In other words, ensemble learning, such as bagging method, can produce a strong learner which has more flexibility and complexity than single model [36], for instance, decision tree. Meanwhile, some ensemble methods, especially bagging, tend to reduce overfitting problems of training data, which also may intensify the generalization of the models. Totally, we utilize RF rather than decision tree to generate rule sets.

The rule generation stage proceeds in two steps: In first step, the SVM model, which is constructed by best fold of CV, is applied to predict the labels of SVs, and the original labels of SVs are discarded. Hence, the artificial synthetic data are generated. During second step, the artificial data are used to train an RF model, and all decision trees of RF are the generated rule sets.

Finally, the performance of the rule sets are evaluated on 10% remained test data, the precision, recall, and F-measure are used to estimate the accuracy of the rule sets.

IV. Experiments

In this study, we proposed an ensemble learning approach (SVM + RF) for rule extraction from SVMs. The method applied the information provided by SVs of the SVM model, and combined ensemble techniques to extract more rules from the complex SVM model. First, C4.5, Naïve Bayes Tree (NBTree), RF, and BP Neural Network were regarded as comparison methods to compare the accuracy with the SVM, which would prove the motivation of rule extraction from the SVM. Then, SVM + C4.5, an eclectic method for rule extraction, was applied to compare the rule sets' learning ability with the proposed method. SVM + C4.5 utilized the C4.5 decision tree to construct rule classifier with the same SVs. The difference between the proposed method and SVM + C4.5 was the difference in the rule induction approach. Finally, all algorithms were tested on the test sets. The flow diagram of experimental setup is shown in Fig. 2.

A. Data Preprocess

The open dataset of Chinese was from China Health and Nutrition Survey (CHNS) [37], an international collaborative project between the University of North Carolina and the National Institute of Nutrition and Food Safety at the Chinese Center. The CHNS collected health data from 228 communities in nine diverse provinces. The 2009 examination surveyed a total of 8597 adult aged 18 or older, and it also included fasting blood collection. After data preprocess, such as vacant data exclusion and noise data canceling, there were total of 7913 individuals with fasting blood. There were no statistically significant differences in the total 2009.

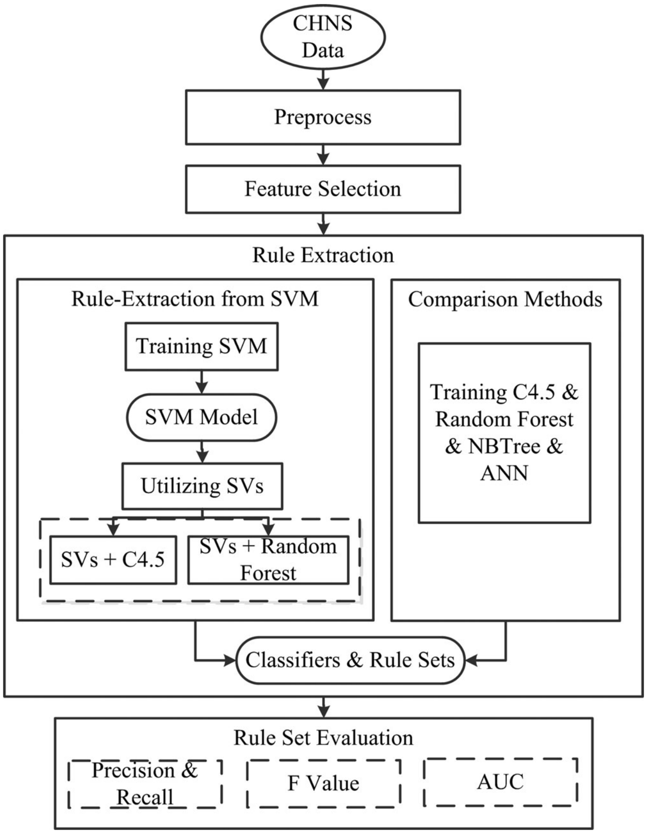


Fig. 2. Flow diagram of the experimental setup.

After excluding the information of individual daily food consumption, there were 56 features remained, which contained both noninvasive factors and metabolic factors. Considering the efficiency of diabetes diagnosis and screening, we proposed a detection model only with few strong relevant and easy available features. Diabetes was defined according to the WHO 2006 guidelines as having a fasting blood glucose measurement ≥7.0 mmol/l. Totally, 646 individuals were marked as diabetic.

B. Feature Selection

As many machine-learning methods have worse performance with large amounts of irrelevant features, feature selection (FS) techniques have become a necessity in all applications [38]. FS can avoid overfitting and gain a deeper insight into the unknown areas, such as occurrence and diagnosis of diseases. As a result, we utilized three filter techniques (univariate LR, chi-square tests, and information gain (IG)-based method) and an embedded technique (RF) to select the strong relevant features. Univariate LR selected the features which were statistical significant with P value < 0.05.

TABLE I
FS RESULTS FOR THE DIAGNOSIS MODEL

Name	LR	Chi-square	IG	RF	Variables (mean ± SD)
HBA1C	0.000	3173.48	0.148	243.2	5.64 ± 0.92
TG	0.000	594.97	0.037	28.09	1.69 ± 1.51
UA	0.000	267.78	0.015	14.65	5.16 ± 1.80
HDL	0.000	215.85	0.015	11.41	1.43 ± 0.51
AGE	0.000	160.66	0.016	10.48	51.82 ± 14.17
DBP	0.000	123.5	0.011	5.44	75.38 ± 11.25
CHOL	0.000	136.52	0.010	9.09	4.88 ± 1.00
WAIST	0.000	285.15	0.025	8.52	81.46 ± 15.11
WEIGHT	0.000	264.14	0.022	8.21	56.99 ± 11.11

In statistics, chi-square test was applied to test the independence of two events. However, in FS procedure, two events represented the occurrence of the feature t and occurrence of the class c_i. The importance of features can be compared to the chi-square distribution with one degree of freedom to judge extremeness [39]. It is defined as

$$\chi^2(t, c_i) = -\frac{N\left[P(t, c_i)\, P(\bar{t}, \overline{c_i}) - P(t, \overline{c_i})\, P(\bar{t}, c_i)\right]^2}{P(t)\, P(\bar{t})\, P(\overline{c_i})\, P(c_i)} \quad (9)$$

where N is the total number of examples in the data. (t, c_i) is the presence of t and category in c_i. $(\bar{t}, \overline{c_i})$ is absence of t and category not in c_i, etc.

IG measured the information obtained for class prediction by knowing the value of a feature; the IG is defined to be the expected reduction in entropy. If features are continuous, IG uses information theoretic binning to discretize the continuous features [40].

The measure of feature importance in RF is the total decrease in node impurities from splitting on the variable, averaged over all trees [41]. The node impurity is measured by the Gini importance. Gini importance is defined as

$$G_k = 2p(1 - p) \quad (10)$$

where p represents the fraction of positive examples assigned to a certain node k and $1 - p$ as the fraction of negative examples.

Taking the efficiency of diabetes diagnosis into account, we decided to choose the number of features no more than 10, and these features also would not impair the accuracy of learning algorithms. For this purpose, we respectively chose the top 15 features from the ranking lists of feature importance in Chi-Square, IG, and RF, so we correspondingly set the thresholds of these three methods as 100, 0.01, and 5. Finally, we selected the same features all occurred in ranking lists, which simultaneously had statistically significant difference in LR ($P < 0.05$). The experiments of FS were accomplished by Weka. The FS results for diagnosis model and the characteristics (mean ± SD) of these features were shown in Table I. According to the result in Table I, hemoglobin A1c (HBA1C), triglyceride (TG), uric acid (UA), HDL, age, diastolic blood pressure (DBP), CHOL, waist, and weight total nine features were selected in the model.

C. Rule Extraction From the SVM

The 2009 CHNS data were highly unbalanced dataset with 8.2% diabetes and 91.8% nondiabetes data, and the unbalanced proportion was about 11:1. To get a fair view of the performance, we conducted ten runs for 2009 CHNS data. First, we randomly shuffled the data ten times, and divided each dataset into a training set and test set with a 9:1 ratio.

TABLE II
AVERAGE RESULTS OF TENFOLD CV FOR POSITIVE CLASS

	Precision (mean ± SD)	Recall (mean ± SD)	F (mean ± SD)	AUC (mean ± SD)
SVM	**88.4% ± 0.01**	40.0% ± 0.01	0.551 ± 0.01	0.852 ± 0.01
RF	81.2% ± 0.01	**49.0% ± 0.01**	**0.612 ± 0.01**	**0.869 ± 0.01**
C4.5	79.8% ± 0.01	46.7% ± 0.01	0.589 ± 0.00	0.779 ± 0.01
NBTree	76.6% ± 0.01	47.3% ± 0.01	0.585 ± 0.01	0.858 ± 0.02
BP NN	81.3% ± 0.01	42.3% ± 0.03	0.556 ± 0.02	0.838 ± 0.02.

Next, in the first run, 90% of the data was used for training SVMs with RBF kernel by tenfold CV. This process was accomplished by R with package "e1071." The data were first normalized to [0, 1]. After the tenfold CV, the optimal hyperparameters (C and gamma) of the SVM obtained by grid-search were 3 and 0.1.Then, the SVM model in the CV was constructed by the best fold, which was defined as the fold gave the best classification rate with the particular fold's test set, and finally the SVM model was used to test on the remained 10% dataset. To ensure the fair performance of the trained model, another nine runs were conducted on remained nine shuffled datasets with the same chosen parameters. Because on any particular randomly drawn test dataset, one classifier may outperform in testing dataset than in tenfold CV. This is a particularly pressing problem for small test datasets.

In addition, if the approaches were applied to the datasets on which rule induction techniques perform better than SVM, the rule extraction from SVM would seem illogical. This aspect was always neglected in this field. In order to illustrate the motivation of rule extraction from SVM, BP neural network (BP NN), RF, C4.5, and NBTree were also implemented in ten runs as the same as SVM, whose optimal parameters were chosen by grid search in first run. The average accuracy of these models was calculated with precision, recall, F score, and AUC. The average results of tenfold CV in ten runs were shown in Table II.

In all of ten runs, the SVs were extracted from the SVM models constructed by best fold. The mean of the number of SVs was 1061 (standard deviation = 14). The unbalanced proportion of dataset was changed from 11:1 to 5.6:1 (standard deviation = 0.2), which proved that the extraction method of SVs is useful to deal with the class imbalance problem.

Finally, the SVs were reloaded to predict the labels by obtained SVM model, and overrode the original labels of SVs to create synthetic artificial data. Then, artificial data generated by SVM were antinormalized, and used to construct rule sets based on RF, whose process was achieved by R with package "randomForest." The optimal parameters of "ntree" (number of trees to grow), "mtry" (number of variables randomly sampled as candidates at each split), and "nodesize" (minimum size of terminal nodes in each tree) in RFs, respectively, were 10, 9, and 1. This process was implemented by grid search in tenfold CV. Generally, the accuracy of rule sets represented the ability of learning from SVM. The average accuracy of tenfold CV in ten runs was calculated, and it was compared with another eclectic method (SVM + C4.5), which utilized C4.5 for rule induction. The results of tenfold CV in ten runs were shown in Table III.

TABLE III
AVERAGE RESULTS OF TENFOLD CV IN TEN RUNS FOR EXTRACTED RULES

	Precision (mean ± SD)	Recall (mean ± SD)	F Value (mean ± SD)
SVM+RF	**81.8% ± 0.02**	**75.6% ± 0.03**	**0.786 ± 0.02**
SVM+C4.5	79.9% ± 0.03	71.2% ± 0.03	0.753 ± 0.02

TABLE IV
AVERAGE RESULTS OF TEST SETS IN TEN RUNS FOR POSITIVE CLASS

	Precision (mean ± SD)	Recall (mean ± SD)	F Value (mean ± SD)
SVM + RF	**89.6% ± 0.05**	44.3% ± 0.04	0.593 ± 0.03
SVM + C4.5	86.3% ± 0.06	41.1% ± 0.06	0.555 ± 0.05
RF	82.3% ± 0.05	48.0% ± 0.05	0.604 ± 0.03
C4.5	81.3% ± 0.07	45.9% ± 0.05	0.584 ± 0.05
NBTree	78.1% ± 0.05	46.8% ± 0.04	0.583 ± 0.03
SVM	87.4% ± 0.07	41.0% ± 0.06	0558 ± 0.06
BP NN	81.6% ± 0.05	**49.1% ± 0.04**	**0.612 ± 0.03**

To compare the performance of rule sets extracted by the proposed hybrid approach, C4.5, RF, NBTree, SVM + C4.5, and the proposed method total five rule induction techniques were tested on the corresponding test sets of ten runs. At the same time, SVM and BP NN were also tested as comparison methods. The best average performance of test set over the ten randomizations was denoted in Table IV with bold face.

V. RESULTS AND DISCUSSIONS

Identifying the potential individuals with undiagnosed diabetes is the basic intention of this study. The quantities employed to measure the quality of the models are precision, recall, and F score. We place high emphasis on precision which contributes to correctly screen the individuals who are undiagnosed diabetic. Consequently, precision is given top priority ahead of others.

By evaluating the proposed diagnosis model, the average results of test sets in ten runs are shown in Table IV.

In 2009 CHNS data, the diagnosis model yields 89.6% precision and 44.3% recall for the positive class, and yields 94.2% weighted average precision and 93.97% weighted average recall for all classes. Weighted average accuracy is computed by weighting the measures of class (precision, recall), and weights of each class are defined as the proportion of instances in that class. Through Table II, it presents that the SVM has highest precision in five methods, which proves that the SVM performs better than rule induction techniques in ten runs. This evidence makes the motivation of rule extraction from SVM seem logical. Through Table IV, it proves that our proposed rule-extraction approach produces comprehensible rule sets with better precision compared to other four rule induction methods (SVM + C4.5, RF, C4.5, and NBTree). In addition, although our proposed method has lower recall and F value than BP neural network and RF, it has highest precision in total seven approaches. Compared with SVM in Table IV, it indicates that the proposed method has better accuracy than SVM, and it intensifies the generalization of SVMs, which is proved by ten runs on same data. All in all, SVM + RF does not impair the capability of SVM. Especially, it learns most of knowledge which is learned by SVM. The rule extracted such as 1) If HBA1C $>$ 7.15 and HDL $>$ 1.57 and CHOL $>$ 5.9 and AGE $>$77, then diabetic. 2) If HBA1C $>$ 7.25, then diabetic. 3) If HBA1C $<$ 7.05 and UA $<$ 4.5 and TG $<$ 3.12 and CHOL $<$ 1.96, then nondiabetic.

Through extracting support vectors from SVM and replacing the labels, the unbalanced proportion declines from 11:1 to 5.6:1, which proves that extraction of SVs can alleviate the imbalance problem just like oversampling and undersampling.

Consequently, through Table III, it shows that SVM + RF has better accuracy and quality compared with SVM + C4.5, which correspondingly has 81.8% precision and 75.6% recall. It demonstrates that our proposed rule-extraction method has better learning ability from SVM than SVM + C4.5. Through Table IV, the rule-extraction result of RF is also better than C4.5, which is proved both by SVM + RF versus SVM + C4.5 and RF versus C4.5. Consequently, as RF actually implements the bagging technique, the experiments verify that the ensemble learning method can intensify the comprehensive ability of models, and bagging method can produce a strong learner which has more flexibility and complexity than single model.

Finally, through Table IV, SVM + RF has higher precision but lower recall than RF. However, the rule sets of SVM + RF is generated only by SVs, so the complexity, time, and sizes of rules are more decreased than RF (the optimal numbers of tree to grow are 10 versus 70). Totally, the larger rule sets may make the learn patterns more transparent but the comprehensibility of the rules is adversely affected. In this aspect, SVM + RF may have superiority than RF.

VI. Conclusion

In this paper, we developed an ensemble system for diabetes diagnosis. In particular, we utilized SVM for diagnosis of diabetes, where a rule-based explanation component was applied to provide comprehensibility and transparent representation. These rule sets can be regarded as a second opinion for diagnosis and a tool to screen the individuals with undiagnosed diabetes by lay users. This will provide an enhanced opportunity for timely and appropriate intervention to apply, which may reduce the incidence of diabetes and its complications.

Results show that our proposed model has high quality in terms of diagnosis with precision, which meant the diagnosis ability of the model.

One of the potential future extensions of this study is how to prune the rule sets of the proposed method, the obtained rule sets are much less and smaller than RF, but still larger than C4.5 and NBTree. Another extension is how to determine the status of diabetes risk not only derived from impaired fasting glucose but also from impaired glucose tolerance. Anyway, the risk factors related to both impaired regulation of glucose metabolism were reported quite similar. But the sensitivity of these risk factors in assessment of two impaired regulations of glucose metabolism was quite different. That is a big challenge to improve the diagnosis model to recognize the risk of impaired glucose tolerance.

References

[1] D. Whiting, L. Guariguata, C. Weil, and J. Shaw, "IDF diabetes atlas: Global estimates of the prevalence of diabetes for 2011 and 2030," *Diabetes Res. Clin. Pract.*, vol. 94, pp. 311–321, 2011.

[2] International Diabetes Federation. (2011). *IDF Diabetes Atlas*, 5th ed. Brussels, Belgium. [Online]. Available: http://www.idf.org/diabetesatlas

[3] N. Brown, J. Critchley, P. Bogowicz, M. Mayige, and N. Unwin, "Risk scores based on self-reported or available clinical data to detect undiagnosed type 2 diabetes: A systematic review," *Diabetes Res. Clin. Pract.*, vol. 98, pp. 369–385, 2012.

[4] D. Noble, R. Mathur, T. Dent, C. Meads, and T. Greenhalgh, "Risk models and scores for type 2 diabetes: Systematic review," *BMJ.*, vol. 343, pp. 1–31, 2011.

[5] K. Heikes, D. Eddy, B. Arondekar, and L. Schlessinger, "Diabetes risk calculator: A simple tool for detecting undiagnosed diabetes and pre-diabetes," *Diabetes Care*, vol. 31, no. 5, pp. 1040–1045, 2008.

[6] S. J. Griffin, p. S. Little, C. N. Hales, A. L. Kinmonth, and N. J. Wareham, "Diabetes risk score: Towards earlier detection of type 2 diabetes in general practice," *Diabetes/Metabolism Res. Rev.*, vol. 16, pp. 164–171, 2000.

[7] M. W. Hanif, G. Valsamakis, A. Dixon, A. Boutsiadis, A. F. Jones, A. H. Barnett, and S. Kumar, , "Detection of impaired glucose tolerance and undiagnosed type 2 diabetes in UK South Asians: An effective screening strategy," *Diabetes, Obesity Metabolism.*, vol. 10, no. 9, pp. 755–762, 2008.

[8] D. B. Rolka, K. M. Narayan, T. J. Thompson, J. Lindenmayer, and D. O. Stuart, "Performance of recommended screening tests for undiagnosed diabetes and dysglycemia," *Diabetes Care*, vol. 24, no. 11, pp. 1899–1903, 2001.

[9] W. G. Gao, Y. H. Dong, Z. C. Pang, H. R. Nan, S. J. Wang, J. Ren, L. Zhang, J. Tuomilehto, and Q. Qiao, "A simple Chinese risk score for undiagnosed diabetes," *Diabetic Med.*, vol. 27, pp. 274–281, 2010.

[10] T. Thitaporn, D. Newby, J. Schneider, and S. C. Li, "Survey of diabetes risk assessment tools: Concepts, structure and performance," *Diabetes/Metabolism Res. Rev.*, vol. 28, pp. 485–498, 2012.

[11] B. Buijsse, R. Simmons, S. Griffin, and M. Schulze, "Risk assessment tools for identifying individuals at risk of developing type 2 diabetes," *Epidemiol. Rev.*, vol. 33, pp. 46–62, 2011.

[12] p. E. H. Schwarz, J. Li, J. Lindstrom, and J. Tuomilehto, "Tools for predicting the risk of type 2 diabetes in daily practice," *Horm. Metab. Res.*, vol. 41, pp. 86–97, 2009.

[13] L. Tapak, H. Mahjub, O. Hamidi, and J. Poorolajal, "Real-data comparison of data mining methods in prediction of diabetes in Iran," *Healthcare Informat. Res.*, vol. 19, no. 3, pp. 177–185, 2013.

[14] C. M. Velu and K. R. Kashwan, "Visual data mining techniques for classification of diabetic patients," in *Proc. IEEE 3rd Int. Adv. Comput. Conf.*, 2013, pp. 1070–1075.

[15] O. Akgobek, "A hybrid approach for improving the accuracy of classification algorithms in data mining," *Energy Edu. Sci. Technol. Part A-Energy Sci. Res.*, vol. 29, no. 2, pp. 1039–1054, 2012.

[16] J. Lee, B. Keam, E. J. Jang, M. S. Park, J. Y. Lee, D. B. Kim, C. H. Lee, T. Kim, B. Oh, H. J. Park, K. B. Kwack, C. Chu, and H. L. Kim, "Development of a predictive model for type 2 diabetes mellitus using genetic and clinical data," *Osong Public Health Res. Perspect.*, vol. 2, no. 2, pp. 75–82, 2011.

[17] N. H. Barakat, A. P. Bradley, and M. N. H. Barakat, "Intelligible support vector machines for diagnosis of diabetes mellitus," *IEEE Trans. Inform. Technol. Biomed.*, vol. 14, no. 4, pp. 1114–1120, Jul. 2010.

[18] M. Marinov and I. Yoo, "Data-mining technologies for diabetes: A systematic review," *J. Diabetes Sci. Technol.*, vol. 5, no. 6, pp. 1549–1556, 2011.

[19] K. Polat, and S. Guenes, "An expert system approach based on principal component analysis and adaptive neuro-fuzzy inference system to diagnosis of diabetes disease," *Digit. Signal Process.*, vol. 17, no. 4, pp. 702–710, 2007.

[20] A. Khan and K. Revett, "Data mining the PIMA dataset using rough set theory with a special emphasis on rule reduction," in *Proc. INMIC 8th Int. Multitopic Conf.*, 2004, pp. 334–339.

[21] N. H. Barakat and A. P. Bradley, "Rule extraction from support vector machines: A sequential covering approach," *IEEE Trans. Knowl. Data Eng.*, vol. 19, no. 6, pp. 729–741, Jun. 2007.

[22] N. Barakat and J. Diederich, "Eclectic rule-extraction from support vector machines," *Int. J. Comput. Intell.*, vol. 2, no. 1, pp. 59–62, 2005.

[23] X. J. Fu, C. J. Ong, S. Keerthit, and G. G. Hung, "Extracting the knowledge embedded in support vector machines," in *Proc. IEEE Int. Conf. Neural Netw.*, 2004, pp. 107–112.

[24] H. Núñez, C. Angulo, and A. Català, "Rule extraction from support vector machines," in *Proc. Eur. Symp. Artif. Neural Netw.*, 2002, pp. 291–296.

[25] Y. Zhang, H. Su, T. Jia, and J. Chu, "Rule extraction from trained support vector machines," in *Proc. 9th Pacific-Asia Conf. Adv. Knowl. Discovery Data Mining*, 2005, pp. 61–70.

[26] G. Fung, S. Sandilya, and R. Rao, "Rule extraction from linear support vector machines," in *Proc. 11th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2005, pp. 32–40.

[27] N. Barakat and J. Diederich, "Learning-based rule-extraction from support vector machines: Performance on benchmark data sets," in *Proc. 14th Int. Conf. Comput. Theory Appl.*, 2004, pp. 178–190.

[28] D. Martens, B. Baesens, T. V. Gestel, and J. Vanthienen, "Comprehensible credit scoring models using rule extraction from support vector machines," *Eur. J. Oper. Res.*, vol. 183, no. 3, pp. 1466–1476, 2006.

[29] A. C. F. Chaves, M. M. B. R. Vellasco, and R. Tanscheit, "Fuzzy rule extraction from support vector machines," in *Proc. 5th Int. Conf. Hybrid Intell. Syst.*, 2005, pp. 6–9.

[30] D. Martens, B. Baesens, and T. V. Gestel, "Decompositional rule extraction from support vector machines by active learning," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 2, pp. 178–191, Feb. 2009.

[31] M. A. H. Farquad, V. Ravi, and R. S. Bapi, "Rule extraction from support vector machine using modified active learning based approach: An application to CRM," *Knowl.-Based Intell. Inform. Eng. Syst.*, vol. 6276, pp. 461–470, 2010.

[32] M. A. H. Farquad, V. Ravi, and R. S. Bapi, "Rule extraction using support vector machine based hybrid classifier," in *Proc. TENCON IEEE Region 10 Conf.*, 2008, pp. 1–6.

[33] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[34] C. C Chang and C. J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 1–27, 2011.

[35] C. Antonio, S. Jamie, and K. Ender, "Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning," *Found. Trends Comput. Vision*, vol. 7, pp. 81–227, 2011.

[36] L. Kuncheva and C. Whitaker, "Measures of diversity in classifier ensembles," *Mach. Learning*, vol. 51, pp. 181–207, 2010.

[37] S. M. Attard, A. H. Herring, E. J. Mayer-Davis, B. M. Popkin, J. B. Meigs, and P. Gordon-Larsen, "Multilevel examination of diabetes in modernising china: What elements of urbanisation are most associated with diabetes?" *Diabetologia*, vol. 55, no. 12, pp. 3182–3192, 2012.

[38] Y Saeys, I. Inza, and P Larrañaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, no. 19, pp. 2507–2517, 2007.

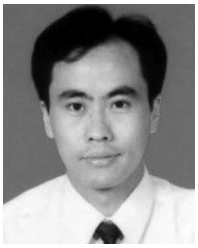
[39] P. Paokanta, "β-thalassemia knowledge elicitation using data engineering: PCA, pearson's chi square and machine learning," *Int. J. Comput. Theory Eng.*, vol. 4, no. 5, pp. 702–706, 2012.

[40] U. Fayyad and K. Irani, "Multi-interval discretization of continuous-valued attributes for classification learning," in *Proc. 10th Proc. 13th Int. Joint Conf. Artif. Intell.*, 1993.

[41] Q. Yanjun, "Random forest for bioinformatics," in *Ensemble Machine Learning*. New York, NY, USA: Springer, 2012, pp. 307–323.

Longfei Han received the Master degree from the School of Information and Electronics, Beijing Institute of Technology, Beijing, China. He is currently working toward the Ph.D. degree at the Information System and Security & Countermeasures Experimental Center, Beijing Institute of Technology.

His current research interests include machine learning and medical data mining.



Senlin Luo received the B.E. and M.E. degrees from the College of Electrical and Electronic Engineering, Harbin University of Science and Technology, Harbin, China, in 1992 and 1995, respectively, and the Ph.D. degree from the School of Information and Electronics, Beijing Institute of Technology, Beijing, China, in 1998.

He is currently a Deputy Director, Laboratory Director, and Professor of Information System and Security & Countermeasures Experimental Center, Beijing Institute of Technology. His current research interests include machine learning, medical data mining, and information security.

Jianmin Yu received the bachelor degree from the School of Mechanical and Electrical Information and Information Engineering, Shandong University. He is currently working toward the master degree at the Information System and Security and Countermeasures Experimental Center of Beijing Institute of Technology.

He current research interests include machine learning, medical data mining.

Limin Pan received B.E. and M.E. degrees from the College of Electrical and Electronic Engineering, Harbin University of Science and Technology, Harbin, China.

She is currently working in Beijing Institute of Technology. Her research interests include data mining and image processing.

Songjing Chen is currently working toward the Ph.D. degree at the Information System and Security and Countermeasures Experimental Center, Beijing Institute of Technology, Beijing, China.

Her research interests include medical data mining, bioinformatics, and machine learning.