A Novel Method for Disease Prediction: Hybrid of Random Forest and Multivariate Adaptive Regression Splines

Dengju Yao
College of Computer Science and Technology, Harbin Engineering University, Harbin Heilongjiang, China
Email : ydkvictory@163.com

Jing Yang
College of Computer Science and Technology, Harbin Engineering University, Harbin Heilongjiang, China
Email : yangjing@hrbeu.edu.cn

Xiaojuan Zhan
Department of Computer Science and Technology, Heilongjiang Institute of Technology, Harbin Heilongjiang, China
Email : xiaojuanzhan@gmail.com

***Abstract*—Using data mining technology for disease prediction and diagnosis has become the focus of attention. Data mining technology provides an important means for extracting valuable medical rules hidden in medical data and acts as an important role in disease prediction and clinical diagnosis. This paper surveys some kind of popular data mining techniques for disease prediction and diagnosis, such as decision tree, associated rule analysis and clustering analysis. Then, a novel hybrid method of random forest and multivariate adaptive regression splines is proposed for building disease prediction model. Firstly, random forest algorithm is used to perform a preliminary screening of variables and to gain an importance ranks. Then, the new dataset selected by top-k important predictors is input into the MARS procedure, which is responsible for building interpretable models for predicting disease survivability. The capability of this combination method is evaluated using basic performance measurements (e.g., accuracy, sensitivity, and specificity) along with a 10-fold cross-validation. Experimental results show that the proposed method provides a higher accuracy and a relatively simple model.**

***Index Terms*—data mining, medical data, random forest, multivariate adaptive regression splines**

I. INTRODUCTION

With the rapid development of electronic information technology and widely use of digital medical equipment and hospital information systems (HIS), the information capacity in the medical database is constantly expanding. Those data not only include the patient's biological indicators of test values as blood sugar, blood lipids, insulin, etc., but also include the patient's height, weight, age and other natural data, and disease history, symptoms, medication and many other kinds of non-numerical data, and many clinical data will change over time continuously. These data contains a lot of valuable information, and to analyze these historical data can help to identify disease risk factors and the interaction between them, which is useful for disease diagnosis and prediction. However, because of the characteristics of diversity, imperfect, timeliness, and redundancy of medical data, the traditional statistical methods have not been competent enough to solve these complex problems. How to take advantage of these valuable medical information resources for the disease diagnosis and treatment has become a research focus. [1]

With the rapid development of information technology and database technology, data mining technology has received great concern in information industry and the whole society. Data Mining is a process of extracting information and knowledge which is implicit, unknown in advance but potentially useful from a large, incomplete, and noisy, fuzzy and random practical application data. Data mining technology has been widely used in retail, finance, insurance, telecommunications, bioinformatics and medical fields. In medical research field, usually the data mining experts associated with medical experts together to mine medical data aimed at certain diseases (such as breast cancer, lung cancer, tumors, etc.) and extract meaningful medical information hidden in medical data. The result of medical data mining can help doctors make the correct diagnosis and treatment, and this is very significant for human health.

The paper is followed as below: in Section Ⅱ, the characteristic of medical data are briefly introduced. In Section Ⅲ, the related research works of medical data mining are presented. In Section Ⅳ, the basic concepts of RF, MARS, and the combination of RF and MARS are briefly introduced. In part Ⅴ, algorithm evaluation method is given. Finally, experiment result and algorithm discussions are presented.

doi:10.4304/jcp.8.1.170-177

II. THE CHARACTERISTICS OF MEDICAL DATA

Compared with other types of data, medical data has its own uniqueness [1-2].

A. Privacy Characteristic

Medical data is information about people, so inevitably involves patients' some private information. When the private information makes these patients encounter unpredictable intrusion in daily life, it creates a privacy problem. In addition, medical data also involves security and confidentiality. When unauthorized individuals or organizations seek to obtain such private information, it creates a security problem. When researchers that have private information of patients share the information with unauthorized individuals or organizations, it creates a confidential problem. Therefore, it is the obligation and responsibilities of medical data mining to protect the privacy of patients and to ensure that medical data security and confidentiality. This will require research on data mining algorithms and related technologies based on privacy protection for medical data mining.

B. Diversity and Heterogeneity

The diversity and heterogeneity of the data is the most significant features different from other areas. As the medical data are obtained from medical imaging, laboratory data and the exchange of doctors and patients, so the original medical data has many forms, usually including pure data (such as signs of parameters, test results), signals (such as EMG, EEG, etc.), images (such as test results from type-B ultrasonic and other medical imaging equipment), text (such as the identity of the patient records, descriptions of the symptoms, detection and written expression of diagnosis), etc.. There is a big heterogeneity among these data. Because the interpretation for image data or other clinical signals is expressed by unstructured language, which is difficult to standardize, and even different experts in same department can not reach consensus on the state's vague description of the patient. They not only use different names to describe the same disease, but also different grammatical structure to describe the relationship between medicine entries. So its data mining is very difficult, and we need to study specific medical data mining algorithms.

C. Redundancy and Imperfection

Medical database is a huge data resource, in which there are a lot of the same or part same information stored every day, and therefore may contain duplicates, irrelevant, or even contradictory records. For example, for the same disease, the symptoms, test results and treatment measures may be the same. In addition, due to the finiteness of medical case, it is impossible that medical information of any kind of disease can be fully reflected in the medical database, the objective information incomplete and the subjective inaccurate description of disease result in incomplete of medical information. Meanwhile, due to human factors the medical data may lead to bias and incomplete records, and many expressions for medical data record itself has uncertainty and ambiguity. For a large number of vague, incomplete, noisy and redundant information in medical databases, you must clean up and filter the information to ensure data consistency and certainty, and make these data become suitable for mining form before data mining. These characteristics of medical data cause a big difference between the medical data mining and general data mining.

D. Timeliness

Medical data also has timeliness characteristics. Medical testing signals, such as ECG, SPECT images etc., are function about time and have stronger timeliness. There is some static medical information, such as the identity of the patient records, although not with the timing, but they present the patients' medical activity record at some point.

III. RELATED WORK

Data mining technology has been applied to biology and medicine domain since it was proposed. There is a data-mining-related international conference named "Data Mining in Biomedicine", which is specifically for biomedical research. And its purpose is to explore the data mining research and applications in biomedical domain. Currently, Medical data mining study focused on four typical areas, including disease diagnosis, hospital information system, new drug development and genetics. Next, we will introduce the research related with disease diagnosis in detail.

In medical diagnosis domain, the main purpose of data mining is for classification and correlation analysis to find or verify useful medical diagnosis rules. There are a large number of medical record information about the patient's condition and the patient's personal information, including age, gender, residence, occupation, living conditions, etc., association analysis based on the medical database can help to find meaningful relationship or patterns. Not similar to association analysis, classification is a process of analyzing, learning, and establishing the appropriate classification model in specific problem areas based on a large number of clinical data, and its main goal is to classify some objects into a specific category. Feature extraction is the most critical part of classification system, and its' quality is directly related to classification performance.

As a specific application, the medical diagnosis expert system is a good example. Traditional medical diagnostic expert system translate the expert diagnostic experience into the rules, users can make judgments quickly only by entering the patient's symptoms. This can reduce medical errors due to doctors' subjective judgments. However, the diagnostic criteria for expert system are based on one or a few experts' experience, and lack of objectivity and universality. In addition, the expert system's reasoning rules and conclusions are pre-designed, some of the clinical manifestations may not be within this range, so there are some limitations. Along with the development of data mining technique, we can use data mining technology into medical field combination with statistical

methods to analyze and process large amount of historical data in patient information database. Data mining algorithm can explore the disease law and various risk factors for disease and many complications, digging out valuable diagnostic rules. Bases on this, according to the patient's age, gender, the results of laboratory examinations, physiological and biochemical indicators, doctors can make diagnoses. This method can excluding the interference of human factor, and is more objectivity. In addition, because data mining is carried on a large amount of data, the received diagnosis rule has a good universality. There are some successful examples in practice application. University of Southern California Spine Hospital [3] utilized Information Discovery for medical data mining to discover diagnostic rules of the spine; Children's Hospital, Chicago, USA [4] used data mining system based on SPSS to analyze and predict the indicators of brain tumor children for providing a scientific treatment for children's brain tumors cure; Health Trinity Hospital in American [6] established the PPS (Pooling Data, Asking Questions) system based on SAS, which analyzed the medical expenses, physiological indicators and other medical records of millions of patients and discharged from hospital patients from 11 hospitals in the past seven years, and accessed to the law of demand for hospital resources.

Another objective of medical data mining is to predict disease. Forecasts can be made by model building on the historical data or data distribution [2]. By analyzing and comparing the biological data of normal and patient, predicting the precursor of certain diseases, may prevent diseases and save lives of patients. Predictive modeling can be divided into prediction based on classification and prediction based on regression. In fact, the decision tree classification model and nearest neighbor algorithms are used to predict by class labels, and predictive modeling based on regression include linear and nonlinear problems. 2009, Massoud Toussi [7] et al. from France explored the medical decision-making rules to make up for gaps in knowledge of clinical guidelines, and showed the implementation method of data mining with type 2 diabetes example, by utilizing the medical prescription of doctors for medical diagnostic data mining. UK Imperial College and University of Oster [9] used data mining in diabetes research and finally get treatment to classify the relevant disease and rules match; Dr. John from California State University's [18] found that some defective gene can lead to manic depression by mining genetic data, and now many researchers are looking for biological and medical evidence to prove their research. 2010, John H Warner [12] et al. used multivariate adaptive regression splines to analyze the potential biological threat identification in the U.S. adult smokers and non-smoking groups, the study used random forest method to implement initial variable selection, then used multivariate adaptive regression splines to bulid the final statistical model. Through the data mining analysis in the 3585 adult smokers and 1077 non-smokers exposed clinical data, they constructed forecasts model for each potential hazards biomarker BOPH.

Over the last decade, neural networks and fuzzy control techniques have been introduced into this area. R. Setiono [8] proposed a method using artificial neural networks to extract diagnosis classification rules from a breast cancer data; Lee [5] et al. utilized space-time dimension related to predict the onset of the disease in the study of dynamics EEG in schizophrenia patients analysis; Zhao et al. used relevant dimension to predict myocardial ischemia. Cho and Walbot [19] utilized artificial neural networks to match non-redundant gene set that have been known, then they established genetic model based on the matching results; Professor Qian built knowledge bone tumor diagnosis based on rough set [11]; 2009, Liu [21], et al. from China Medical University, applied artificial neural network based on career history to classify and identify coal miners pneumoconiosis high-risk population. The study used longitudinal review data of China Tower coal and Bayesian learning algorithm to build a three-layer artificial neural network based on the coal workers' occupational exposure data. The network includes 6 input variables, 15 hidden layer neurons and 1 output neurons. Sensitivity and ROC analysis are used to explain the importance of input variables and neural network performance, and occupational characteristics and probability values of prediction are used to classify different levels of risk CPW coal workers.

Recently, association rules analysis has become a research hot. There are many researchers try to use association rules analysis method to medical diagnosis. Carlos Ordonez [13-14] et al. proposed an improved Apriori algorithm by adding some constraints to the original Apriori algorithm. The improved algorithm can be used to extract association rules hidden in the medical data quickly and effectively. When the improved algorithm was applied to image data of breast cancer and heart disease data analysis, the result can be comparable with expert analysis; Shah B[15], come from Texas Health Science Center, have analyzed the relations between diabetic patient's age, gender, and pace back and forth bone mineral density. Using the method of time series associated analysis, HarriS ND[16] et al. have discover that QT interval is interrelated with type I diabetes blood glucose concentrations at night, and point out that he revised QTC may be used to alert the incidence of patients sudden death at night. While they used data mining techniques in database of diabetes, Milanzon from Slovenia, Masuda from Japan [17], et al., specifically researched the efficiency of decision trees and association rules algorithm when used in lack of apriori knowledge massive data, they also gave the improved methods.

Another research hot is medical image data mining. In medical image analysis, data mining technology is used to express the characteristics of the target tissue, ie image feature extraction and pattern recognition automatically [2]. In addition, in research of a series of age-related disease and its complications, using data mining techniques for multidimensional analysis of physiological monitoring data is also a new class of research focus.

In a word, data mining techniques are good at discovering and extracting hidden and meaningful knowledge from mass and lack of prior information data predicting future trends and behavior, and making proactive knowledge-based decision-making. It is this advantage makes the data mining analysis is widely used in medical diagnosis domain and made many valuable results.

In this paper, we propose a combination of RF and MARS for predicting breast cancer survivability, using Wisconsin breast cancer data (WBCD) from the UCI Machine Learning Repository. The 10-fold cross-validation [30] method, confusion matrix, accuracy, sensitivity and specificity are used to evaluate the performance of the models for breast cancer survivability prediction.

IV. The Method of combinating RF and MARS

Recently, Random Forest (RF) [23] has become an attractive method in data mining. Random forests (RF) is one of the most successful ensemble learning techniques which have been proven to be very powerful and popular techniques in the pattern recognition and machine learning for high-dimensional classification and skewed problems. As a classifier integration method, RF have the features of classifying fast and training simple and is suitable for feature selection according to variable importance. Moreover, Multivariate Adaptive Regression Splines (MARS) [29] has become particularly popular in the area of data mining because it does not assume or impose any particular type or class of relationship between the predictor and the dependent variable in prior. In this paper, we propose a hybrid of random forest and multivariate adaptive regression splines algorithms for building disease prediction model. Because RF algorithm is based on decision tree algorithm, we firstly introduce the decision tree briefly.

A. Decision Tree

Decision tree [20] is a tree structure, where each internal node represents a test on the attribute, each branch represents a test output, and each leaf node represents a class or class distribution. For a classification problem or rules learning problems, the generation of a decision tree is a process to divide and rule from the top to down. The goal of classification is to build a model by analyzing the training data set, which can be used to classify other data with categories unknown.

A well-known decision tree algorithm is ID3 algorithm proposed by Quinlan in 1986, which is based on information entropy and has been used in many knowledge discovery applications in recent years. Many scholars have conducted for ID3 algorithm improvement.

The following is the general approach constructing ID3 decision tree:

Step one. To select random subset T1 with the window size of w from the training set T

Step two. To construct decision tree in the current window following the rule that regard the speed of information entropy declining as a standard test attribute;

Step tree. To deal with all the training instances by order, and to identify the exception of the current decision tree. If there is no an exception, the training process end; else, go to Step four;

Step four. Combination some training instances in the current window with some exceptions found of in Step three to format a new window, go to Step two.

After a decision tree is build up, we can use this tree to classify new examples. Classification is a process of calculating class label of a case based on its property value. The class label of an example is calculated from the root node of the decision tree, through the internal node of the entire tree, until arriving at a leaf node. In each internal node, the case is tested, and the results of the test will determine the case to reach the following nodes through which branch [11]. The final leaf node class label is considered as the class of the case.

C4.5algorithm is the improvements of ID3 algorithm [20]. It increases some functions based on the ID3 algorithm, including handling continuous attributes and default attribute, using post-pruning techniques to avoid imbalance of the tree, adopting cross-validation, etc. C4.5 algorithm uses the test evaluation function to select optimal test with greatest function value to divide the node at each node. C4.5 algorithm uses two testing evaluation function, namely, information gain function and information gain rate function. This algorithm selects the attributes with the largest Gain value to classify. For continuous descriptive attributes, it takes discrete way.

B. Random Forests

RF is an ensemble classifier that consists of many decision trees and outputs the class that is the mode of the class's output by individual trees [23]. The method combines Breiman's "bagging" idea and the random selection of features in order to construct a collection of decision trees with controlled variation [11]. Bagging RF algorithm can be stated as follows:

Step one: For a given training dataset, extract a new sample set by N times repeated random sampling using bootstrap method. For example, from the data (x_1, y_1),...,(x_n, y_n) to build a sample (x_1^*, y_1^*)....(x_N^*, y_N^*) . Samples which are not being extracted consist of out-of-bag data (OOB).

Step two: Build a decision tree or regression tree based on sample set resulted from step 1;

Step three: Repeat step one to two, result in many trees, composing a forest.

Step four: Let every tree in the forest to vote for x_i.

Step five: Calculate the sum of votes for every class, the class with highest number of votes is the classification label for x_i.

Step six: The percentage of incorrect classification is the classing error ratio of random forest.

Decision tree algorithm has been widely used in medical image data mining and clinical decision analysis or other corresponding application.

C. Multivariate Adaptive Regression Splines

MARS is a nonlinear and nonparametric regression technique which is first proposed by Friedman. Because it

makes no assumption about the underlying functional relationship between the dependent and prediction variables, MARS has become particularly popular in the area of data mining and has been increasingly used in recent years in various scientific fields including disease risk research, human genetics and food sciences [29].

MARS constructs model from a set of coefficients and basis functions of the predictors (x) of the form:

$$(x-t)_+ = \begin{cases} x-t & x>t \\ 0 & otherwise \end{cases} \quad (1)$$

for a dependent variable y, and M terms, the MARS model can be described as the following equation:

$$y = f(x) = \beta_0 + \sum_{m=1}^{M} \beta_m H_{km}(x_{v(km)}) \quad (2)$$

where β_0 andβ_m are model parameters estimated from the training data. Function H is defined as:

$$H_{km}(x_{v(km)}) = \prod_{k=1}^{K} h_{km} \quad (3)$$

where $x_{v(k,m)}$ is the predictor in the k'th of the m'th product.

Constructing a MARS model need three steps:

Step1: A number of basis functions are added to the model by forward stepwise according to a pre-determined maximum which should be considerably larger (twice as much at least) than the optimal (best least-squares fit).

Step2: Over-fitting basis functions are removed in the model by backward. The Generalized Cross Validation error is a measure of the goodness of fit, which takes into account not only the residual error but also the model complexity as well. It is given by

$$GCV = \frac{\sum_{i=1}^{N} (y_i - f(x_i))^2}{(1-\frac{C}{N})^2} \quad (4)$$

with

$$C = 1 + cd \quad (5)$$

where N is the number of cases in the data set, d is the effective degrees of freedom, which is equal to the number of independent basis functions. The quantity c is the penalty for adding a basis function.

Step3: Optimal MARS model is selected by cross validation. The model with minimum mean square error (RMSECV) is the optimal. RMSECV is defined as:

$$RMSECV = \sqrt{\frac{\sum_{i}^{n} (y_i - \hat{y}_{-i})^2}{n}} \quad (6)$$

where y_i is the value of i'th dependent variable, $\hat{y}_{-i}$ is the predicting value of the dependent variable after removing i'th cases.

D. The Combination of RF and MARS

In this paper, we combined the RF and MARS to build the prediction model for breast cancer survivability. RF was used to perform a preliminary screening of prediction variables and to produce variable importance and cross-validated R-squared statistics. The variable importance effectively ranks all variables in dataset with respect to their ability to predict the breast cancer survivability. As a result, 19 predictor-sets together with cross-validated R-squared statistics were kept. The new dataset was generated by these 19 predictors and was input into MARS procedure which was used to find interpretable models for predicting breast cancer survivability. The R statistical package was used for the implementation of RF and MARS.

V. Method of Algorithm evaluation

A. Data Set

The Wisconsin Diagnostic Breast Cancer (WDBC) was obtained from UC Irvine Machine Learning Repository, from Dr. William H. Wolberg [31]. The database contains of 569 samples with 32 attributes (ID, diagnosis, 30 real-valued input features). There are two values in the variable class of breast cancer: benign (non-cancerous) and malignant (cancerous). The database contains 357 benign and 212 malignant.

B. Measures for Performance Evaluation

In this study, the accuracy, sensitivity and specificity were used to evaluate the performance of prediction model resulted from RF, MARS, RF&MARS. They are three commonly used performance measurements and are computed based on the confusion matrix [13]. A confusion matrix is a matrix usually used to represent the relationships between real class attributes and that of predicted classes. In a two-class prediction problem, the upper left cell denotes the number of samples classified as true while they we true (TP), and lower right cell denotes the number of samples classified as false while they were false (TN). The other two cells represent the number of samples misclassified. Specifically, the lower left cell represents the number of samples classified as false while they were true (FN), and the upper right cell represents the number of samples classified as true while they actually were false (FP).

When the confusion matrixes were obtained, the accuracy, sensitivity and specificity could be calculated using the following formulas respectively. The accuracy of classifiers is the percentage of correctness of prediction among the test sets. It is defined in (7). The sensitivity is referred as the true positive rate, and the specificity as the true negative rate. Both sensitivity and specificity are used for measuring the factors that affect the performance, and are computed using (8) and (9), respectively.

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (7)$$

$$sensitivity = \frac{TP}{TP + FN} \quad (8)$$

$$specificity = \frac{TN}{TN + FP} \quad (9)$$

In this study, the sensitivity is the probability of correct tests among "Benign" patients. In contrast, the specificity is the probability of correct tests among "Malignant" patients.

C. k-Fold cross-validation

The WDBC dataset which contains only 569 cases is a smaller samples dataset. In order to minimize the bias associated with the random sampling of the training and holdout data samples in comparing the predictive accuracy of two or more methods, we used a 10-fold cross-validation method. With cross-validation, some of the data is removed before training begins. Then when training is done, the data that was removed can be used to test the performance of the learned model on "new" data [30] [32]. Specifically, we use the following three-step 10-fold cross-validation procedure to estimate the prediction accuracy:

Step one: We randomly divided the dataset (569 records) into 10 disjoint subsets (folds), with each fold containing approximately the same number of records (50-60 records). The sampling is stratified by the class labels to ensure that the subset class proportions are roughly the same as those in the whole dataset.

Step two: For each subset, a MARS classification model is constructed using the nine of the 10 folds and tested on the tenth one to obtain a cross-validation estimate of its prediction accuracy.

Step three: The 10 cross-validation estimates are then averaged to provide an estimate for the classier accuracy constructed from all the data.

VI. Result and Discussion

A. Variable Importance

In our experiment, random forest algorithm is used to discover the variable importance of predictors. RF runs on the initial WDBC dataset and result in the relatively important variables. The RF is implemented in the R software. The parameters *mtry* is set as 5, *ntree* as 1000, and type as "classification". The importance is computed with OOB estimate of error rate: 4.04%. Finally, the order of importance of the total 31 predictor is plotted as Figure. 1.

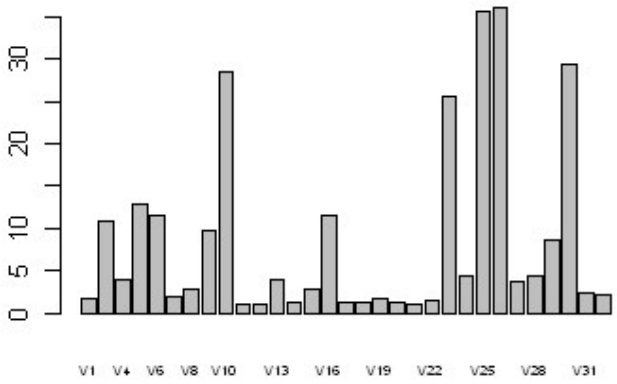

Figure1. Variable importance

According to the Fig. 1, we select top 19 predictors (V3, V4, V5, V6, V8, V9, V10, V13, V15, V16, V23, V24, V25, V26, V27, V28, V29, V30, V31) to construction new dataset from initial WDBC dataset. The result dataset is input into the MARS to build prediction model.

B. Classification Result

In this study, the models were evaluated based on the accuracy measures discussed above (classification accuracy, sensitivity and specificity). The results were achieved using 10 fold cross-validation for each model, and are based on the average results obtained from the test dataset (the 10^{th} fold) for each fold. Table 1 shows the complete set of results in a tabular format. For each fold of each model type, the detailed prediction results of the validation datasets are presented in form of confusion matrixes.

In comparison to the RF, MARS, and RF&MARS method, we found that the RF model achieved a classification accuracy of 0.9626 with a sensitivity of 0.9678 and a specificity of 0.9457. The RF&MARS model achieved a classification accuracy of 0.9629 with a sensitivity of 0.9516 and a specificity of 0.9767. However, the MARS model preformed the best of the three models evaluated. The MARS model achieved a classification accuracy of 0.9670 with a sensitivity of 0.9562 and a specificity of 0.9812. The results indicate that the RF&MARS model has slightly higher classification accuracy than RF model, but slightly lower classification accuracy than MARS model. This result may be because the RF&MARS has selected a smaller number of predictors. However, the RF&MARS model has identical sensitivity with MARS model, and has a simpler regression equation than MARS model.

C. Comparation with Other Algorithms

For discussing the performance of the proposed method, we run C4.5 which is a classical decision tree algorithm and SVMMaj which is a support vector machine algorithm on the WDBC dataset respectively. Similarly, we used 10-fold cross validation to get a more credible result, and computed result of each time is shown by table 1. The result shows that C4.5 algorithm obtains a mean accuracy of 0.9316 with sensitivity of 0.9422 and specificity of 0.9151 and SVMMaj algorithm

obtains a mean accuracy of 0.9585 with sensitivity of 0.9595 and specificity of 0.9553. This result illustrate that the proposed combination method in the paper is superior to the C4.5 algorithm and SVM algorithm.

VII. CONCLUSION

The paper analyzed the characteristic of medical data and proposed a novel method of hybrid of RF and MARS for disease diagnosis and prediction. The proposed method is implemented on R software and is tested on the WDBC dataset. At the end, the performance of the hybrid algorithm of RF and MARS is compared with C4.5 algorithm and SVMMaj algorithm. The result experiment shows that the combination method of RF and MARS is suitable for disease prediction, which has not only good classification accuracy but will result in relatively simple and interpretable model.

ACKNOWLEDGMENT

The authors are grateful to the support of the National Natural Science Foundation of China (61073043, 61073041), the Natural Science Foundation of Heilongjiang Province (F200901), and the Special Fund of academic leaders of Harbin (2011RFXXG015). Thanks to UC Irvine Machine Learning Repository for providing the data. Thanks to Doctor Yang for helpful comments, suggestion and criticisms.

REFERENCES

[1] A. Soltani Sarvestani, A. A. Safavi, N.M. Parandeh, M.Salehi., "Predicting Breast Cancer Survivability Using Data Mining Techniques," IEEE Transl., vol. 2, pp. 227-231, 2010

[2] Zhao HengYu. Application of Data Mining in Medical Field. China Science and Technology Information, , vol.15, pp.129-13, 2009

[3] WANG Hua, J IANG Qicheng, HU Xuegang. Application of data mining to medicine. Anhui Medical and Pharmaceutical Journal, vol. 12, pp. 746-748, 2008

[4] Yue Huang, Paul J. McCullagh, Norman Black,Roy Harper. Evaluation of Outcome Prediction for a Clinical Diabetes Database. KELSI, pp. 181-190, 2004

[5] Jing Yi. Data Mining of Hospital Information and Exploration of Its Practical Implementation [D]. Chongqing Medical University, pp. 25-26, 2007

[6] Lee Yingjie,ZhuYisheng,XuYuhong,et al. The nonlinear dynamical analysis of the EEG in schizophrenia with temporal and spatial embedding dimension. Journal of Medical Engineering & Technology, vol. 25, pp. 79-83, 2001

[7] Shengyue Yang, Zhan Peng, Xiaoping Fan, Yanping Ji. Dynamic BP neural networks based depression diagnosis system. Journal of Railway Science and Engineering, vol.2, pp. 7-74, 2005

[8] Massoud Toussi, Jean-Baptiste Lamy, Philippe Le Toumelin, et. Using data mining techniques to explore physicians' therapeutic decisions when clinical guidelines do not provide recommendations: methods and example for type 2 diabetes. BMC Medical Informatics and Decision Making, vol. 6, pp. 1471-2288, 2009

[9] R.Setiono. Generating concise and accurate classification rules for breast cancer diagnosis,Artificial Intelligence in Medicine, vol. 18, pp. 205-219,2000

[10] Lee Yingjie,ZhuYisheng,XuYuhong,et al. The nonlinear dynamical analysis of the EEG in schizophrenia with temporal and spatial embedding dimension. Journal of Medical Engineering & Technology, vol. 25, pp. 79-83, 2001

[11] Xiaoping Fan, Zhan Peng, Shengyue Yang. Study on Depression Classified System Based on Multilayer Perceptrons Artificial Neural Network. COMPUTER ENGINEERING AND APPLICATIONS. vol. 40, pp. 205-208, 2004

[12] Hui zhang, ZongCai Qian, Jinghui Qu. Application research for building bone tumor assisted diagnostic knowledge database based on rough set. MEDICAL INFORMATICS, vol. 17, pp. 115-118, 2004

[13] John H Warner, Qiwei Liang, Mohamadi Sarkar, et al. Adaptive regression modeling of biomarkers of potential harm in a population of U.S. adult cigarette smokers and nonsmokers. BMC Medical Research Methodology, pp. 1-10, 2010

[14] Carlos Ordonez, CesarA. Santana, Leviende Bral. Discovering interesting association rules in medical data,In ACM SIGNOD Workshop on Research Issues on Data Mining and Knowledge Discovery(DMK02000), pp. 78-85, 2000

[15] Carlos Ordonez. Comparing Association Rules and Decision Trees for Disease Prediction. HIKM06, pp. 17-24, 2006

[16] Shah B. Relationship between diabetes and age in human metatarsal bones. The 17th Southern Biomedical Engineering Conference, pp. 2-32, 1998

[17] Harris ND,Ireland RH,Marques JLB,et al. Can changes in QT interval be used to predict the onset of Hypoglyeemia in type 1 diabetes. Computers in Cardiology, vol. 27, pp. 375-378, 2000

[18] Milan Z,Gou M,Peter K,et al. Mining diabetes database with decision trees and association rules[C]. Proeeedings of the 15th IEEE Symposium on computer-based Medical Systems, pp. 134-139, 2002

[19] Ohrn A, Row land T. Rough sets: a knowledge discovery technique for multi-factorial medical outcomes. Am J Phys Med Rehabil, pp. 79-100, 2002

[20] Cho Y, Walbot V. Computational methods for gene annotation: the arabidopsis genome[J]. Biotechnology, pp. 12:126, 2001

[21] Senlin Luo, Hua Cheng, Yuqing Gu. C4.5 Algorithm in the Construction of the Type 2 Diabetes Classified Rules. Application Research of Computers, vol. 7, pp. 175-177, 2004

[22] Hongbo Liu, Zhifeng Tang, Yongli Yang, Dong Weng, et. Identification and classification of high risk groups for Coal Workers' Pneumoconiosis using an artificial neural network based on occupational histories: a retrospective cohort study. BMC Public Health, Accepted: 29 September 2009

[23] L. Breiman, Random Forests, J. Machine Learning, vol. 45, pp. 5–32, 2001.

[24] Jaree Thongkam, Guandong Xu and Yanchun Zhang, "AdaBoost Algorithm with Random Forests for Predicting Breast Cance Survivability," International Joint Conference on Neural Networks (IJCNN 2008), pp. 3062-3069, 2008

[25] Yue Huang,Paul J. McCullagh, Norman Black,Roy Harper. Feature Selection and Classification Model Construction

on Type 2 Diabetic Patient Data. Industrial Conference on Data Mining, pp. 153-162, 2004
[26] Agrawal, Imielinski and Swami. Mining Association Rules between Sets of Items in Large Databases, SIGMOD, 1993
[27] A. K. Jain, M. N. Murty, P. J. Flynn. Data clustering: a review. ACM Computing Surveys (CSUR), vol. 31, pp. 264-323, 1999
[28] PCharu C. Aggarwal, PJoel L. Wolf, Philip S. Yu, Cecilia Procopiuc, Jong Soo Park. Spectral clustering for multi-type relational data.Proceedings of the 1999 ACM SIGMOD international conference on Management of data, vol. 5, pp. 585 – 592, 1999
[29] Friedman, "Multivariate Adaptive Regression Splines (with discussion) ," Annals of Statistics, vol. 19, pp.1–141, 1991
[30] N. Meinshausen, "Quantile Regression Forests, " Machine Learning Research, vol. 7, pp. 983–999, 2006
[31] http://archive.ics.uci.edu/ml/
[32] D. Delen, G. Walker and A. Kadam, "Predicting breast cancer survivability: a comparison of three data mining methods," J.Artificial Intelligence in Medicine, vol. 34, pp. 113-127, 2005.

Dengju Yao, born in 1980, Ph.D. candidate. His research interests include database, data mining, and privacy preservation.

Jing Yang, born in 1962, Ph.D. supervisor. Hers research interests are database, knowledge engineering and knowledge discovery, and privacy preservation.

Xiaojuan Zhan, born in 1978, lecturer. Hers research interests include data mining, machine learning, and privacy preservation.

TABLE 1
TABULAR RESULTS FOR 10-FOLD CROSS-VALIDATION FOR ALL FOLDS AND ALL MODEL TYPES

Method \ Fold		1	2	3	4	5	6	7	8	9	10	Mean	St.Dev
RF	Accuracy	0.9677	0.8837	0.9783	1	0.9744	0.9804	0.9444	0.9841	0.95	0.9630	0.9626	0.0321
	Sensitivity	0.9411	0.9677	0.9655	1	0.9787	0.9714	0.9677	1	0.9429	0.9429	0.9678	0.0321
	Specificity	1	0.6667	1	1	0.9677	1	0.9130	0.95	0.96	1	0.9457	0.1024
MARS	Accuracy	0.9516	0.9535	0.9783	1	0.9615	0.9804	0.9815	0.9841	0.9167	0.9630	0.9670	0.0233
	Sensitivity	0.9143	0.9706	0.9655	1	0.9583	0.9714	0.9697	0.9778	0.8919	0.9429	0.9562	0.0319
	Specificity	1	0.8889	1	1	0.9667	1	1	1	0.9565	1	0.9812	0.0362
RF & MARS	Accuracy	0.9516	0.9302	0.9783	0.9655	0.9615	0.9804	0.9815	0.9841	0.9333	0.9630	0.9629	0.0195
	Sensitivity	0.9143	0.9697	0.9655	0.9512	0.9583	0.9714	0.9697	0.9778	0.8947	0.9429	0.9516	0.0272
	Specificity	1	0.8	1	1	0.9667	1	1	1	1	1	0.9767	0.0630
C4.5	Accuracy	0.92	0.9492	0.9348	0.92	0.9298	0.9344	0.9219	0.96	0.9184	0.9277	0.9316	0.0137
	Sensitivity	0.9688	0.9118	0.9697	0.9118	0.9024	0.9744	0.9318	0.9375	1	0.9149	0.9422	0.0336
	Specificity	0.8333	1	0.8462	0.9375	1	0.8636	0.9	1	0.8261	0.9444	0.9151	0.0709
SVM	Accuracy	0.95	0.9474	0.9412	0.9677	0.9792	0.9455	0.9677	0.9455	0.975	0.9661	0.9585	0.0140
	Sensitivity	0.9643	0.9487	0.9231	0.9556	1	0.9667	0.95	0.9310	0.9818	0.9737	0.9595	0.0230
	Specificity	0.9167	0.9444	0.96	1	0.9357	0.92	1	0.9615	0.96	1	0.9553	0.0284