© 2016 IJSRSET | Volume 2 | Issue 2 | Print ISSN : 2395-1990 | Online ISSN : 2394-4099
Themed Section: Engineering and Technology

Ensemble Learning Approach based Rule Extraction from Support Vector Machine

Chitra A*, Anto S
Computer Science and Engineering, Sri Krishna College of Technology, Coimbatore, Tamilnadu, India

ABSTRACT

In recent years, support vector machines (SVMs) have shown good performance in a number of application areas. The existing system is concentrated on the discovery of risk of having pre-diabetes or undiagnosed diabetes and to facilitate people decide whether they should see a physician for further evaluation. However the existing system has issue with prediction results by using C4.5, naïve bayes tree and neural network algorithms. To avoid the above mentioned issue we go for proposed system. In proposed scenario, we introduced an efficient algorithm named as Support Vector Machine (SVM) which is utilized to screen diabetes, and an ensemble learning module is added. The proposed system is used to develop an ensemble system for diabetes diagnosis. Specifically, the rules are extracted from the SVM algorithm and it is 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. From the experimental result, we can conclude that the proposed system is better than the existing scenario in terms of reduction of the incidence of diabetes and its complications.

Keywords: Diagnosis of Diabetes, Ensemble Learning, Random Forest (RF), Rule Extraction, Support Vector Machines (SVMs)

I. INTRODUCTION

Diabetes is a disease in which the body does not generate or correctly use insulin, the hormone that unlocks the cells of the body, allowing glucose to go into and fuel. Diabetes increases the risks of initial kidney disease, loss of sight, nerve injure, blood vessel damage and it contributes to heart disease. The cause of diabetes continues to be an ambiguity, although both genetics and ecological factors such as obesity and be short of of exercise come out to take part in roles. Some of the most accepted classification techniques are based on the formation of propositional if-then rules from prelabeled training data. These methods are in principle that can provide an entirely transparent classification decision, but, in fact, their performance and comprehensibility frequently bear in cases of high-dimensional data and continuously valued attributes. Another trendy family of classifiers exemplified by support vector machines (SVMs) and artificial neural networks (ANNs) builds a mathematical form of the data that often performs much better in these situations.However, these methods construct black box models with little or no explanation capacity. In application areas such as medical analysis, there is a obvious need for an description component to be coupled with classification decisions in order to aid the approval of these methods by users[1]. One may disagree that in spite of all the hard work in the grassland of rule extraction from ANNs, there is no clear proof that this area of study was victorious. This dispute is valid to a great extent and can be mostly attributed to the refuse in the use of ANNs in the late 1990s, as they were largely outdated by SVMs because of their superior performance in a number of ordinary applications. Another motivation is that the popular of rule extraction algorithms from ANNs were narrow to a specific network type or architecture. However, in the near future at least, we can estimate an augment in the development and use of SVM rule extraction techniques corresponding with the developments and use of SVMs in a diversity of applications. Furthermore, a number of capable SVM rule extraction algorithms published to date are both simple and largely applicable.

In this paper, we focus an ensemble learning approach for rule extraction from the SVM, which uses RF rule induction technique to develop an inexpensive and possible assessment rules for diagnosis of diabetes. In our proposed method, support vectors (SVs) are primary extracted from the SVM with adequate accuracy. Then, new labels of SVs are predicted by the trained SVMmodel, and unique labels of SVs are replaced by predicted labels. At last, the fake 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 little examples. On the other hand,if the tree is too small, then tree may find out few patterns. All these drawbacks make single decision tree (C4.5) difficult to in shape complex models. By utilizing the ensemble learning method, RF can answer the problem mentioned previously. Meanwhile, In view of the rule sets are generated from the SVs, the rule sets obtained by SVM + RF is definitely much less and smaller than those of RF, where the large rule sets may create the problem unintelligible. Moreover, for the skewed classification trouble the proposed method can be a preprocessing technique to decrease the imbalance proportion of skewed data, which can develop precision and recall in positive class. The model can measure undiagnosed individuals in a clear form and give a more comprehensive and obvious representation for end users.

A. The significance of Rule-Extraction Algorithms

The capability of representative AI systems to present a declarative demonstration of knowledge about the complexity domain offers a natural motivation capability for the decisions made by the system. Reference [3] argues that even limited explanation can absolutely influence the system's reception by the user. This capability is important, mainly in the case of medical applications. A motivation capability can also offer a check on the interior logic of the system as well as being able to give a learner nearby into the problem [4]. In addition, the explanations given by rule-extraction algorithms extensively enhance the capabilities of AI systems to discover data and support the initiation and construction of new theories ANN‘s & SVMs have no such declarative knowledge structures, and hence, are limited in providing explanations.

B. The Classification of Rule-Extraction Algorithms

One possible method for classifying rule-extraction algorithms is in terms of the "translucency" of the sight taken within the rule-extraction method of the fundamental classifier. This pattern yields two crucial categories of rule-extraction techniques: “translucent " and "instructive". The distinctive feature of the "translucent" approach is that the focal point is on extracting rules at the level of entity components of the fundamental machine learning method. But in the feedforward neural networks, these are hidden and output units. Such methods obviously are used in combination with a learning algorithm that consist of rule-based explanations and the basic pattern is to use the trained classifier for generating examples for a second learning algorithm that generates rules as output [5],[6],[7]. This is the "hybrid" or ―eclectic‖ group [1], [2], [8].Clearly, this classification scheme, initially developed for rule-extraction from neural networks, is appropriate to support vector machines as fit. Decompositional system can be based on the investigation of support vectors generated by the SVM even as learning-based classification learns come again? the SVM has learned. An example for learning-based rule-extraction from SVMs is [10].

Related Works

An amount of methods have been proposed for rule extraction from SVMs. Broadly speaking, these methods can be regarded as into three major families—learning based, decompositional, and eclectic method—as recommended by Andrews et al. [2] for ANNs. Learning-based method ensures the model (classifier) as a black box describing only the relationship between the inputs and the outputs. In general, learning-based approaches use another machine learning technique, which has an account capability, to study what the classifier has learned. Not like learning-based, decompositional approaches open the model, glance into its individual components, and then try to extract rules at the level of these components. Therefore, in principle, this is the most obvious approach. The eclectic approach slander in between the learning-based and decompositional approaches. The following sections review these methods.B. Buijss et al[2] proposed “Risk assessment tools for identifying individuals at risk of developing type 2 diabetes” .Type 2 diabetes is

associated with increased risk of cardiovascular disease and premature mortality and is the leading cause of blindness, kidney failure, and non traumatic amputations resulting from micro vascular complications. P. Paokanta et al[6] proposed "β-thalassemia knowledge elicitation using data engineering: PCA, pearson's chi square and machine learning. Data Engineering is one of the knowledge elicitation and Analysis methods, among several techniques; Feature Selection methods play an important role for these processes which are the processes in data mining technique especially classification tasks. In this scenario, the Thalassemia knowledge[15] is extracted using Data engineering techniques (PCA, Pearson's Chi square and Machine Learning). This knowledge presented in form of the comparison of classification performance of machine learning techniques between using Principal Components Analysis (PCA) and Pearson's Chi square for screening the genotypes of β-Thalassemia patients. According to using PCA, the classification results show that the Multi-Layer Perceptron (MLP) is the best algorithm, providing that the percentage of accuracy reaches 86.61, K- Nearest Neighbors (KNN), NaiveBayes, Bayesian Networks (BNs) and Multinomial Logistic Regression with the percentage of accuracy 85.83, 85.04, 85.04 and 82.68.

Q. Yanjun et al[8] "Random forest for bioinformatics," in Ensemble Machine Learning. Modern biology has experienced an increasing use of machine learning techniques for large scale and complex biological data analysis. In the area of Bioinformatics, the Random Forest (RF) technique, which includes an ensemble of decision trees and incorporates feature selection and interactions naturally in the learning process, is a popular choice. K. Heikes et al[17] proposed "Diabetes risk calculator: A simple tool for detecting undiagnosed diabetes and prediabetes". The objective of this study was to develop a simple, self-administered, paper-based screening tool that could be used by the public to determine their risk of having pre-diabetes or undiagnosed diabetes and to help people decide whether they should see a physician for further evaluation. L. Tapak et al[18] propoded "Real-data comparison of data mining methods in prediction of diabetes in Iran". Diabetes is one of the most common non-communicable disease (NCDs) that has significantly contributed to in-creased mortality in patients. 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. 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.

O. Akgobek et al[19] proposed "A hybrid approach for improving the accuracy of classification algorithms in data mining". Data Mining is the discovery of previously unknown, potentially useful and hidden knowledge in databases. The classification task can be carried out by various techniques such as: Decision Tress, Bayesian classify and Bayesian networks (Belief Networks), Neural Networks, Rule induction, K-nearest neighbor, Genetic algorithms, Rough sets, Fuzzy logic and so on. By merging some classification techniques new techniques also have been developed (ex: Fuzzy rule induction, Fuzzy decision trees, Neuro-fuzzy networks, etc.). J. Lee et al[20] propoded "Development of a predictive model for type 2 diabetes mellitus using genetic and clinical data". Recent genetic association studies have provided convincing evidence that several novel loci and specific single nucleotide polymorphisms (SNPs) are associated with an increased risk of T2DM. The aims of this study are: 1) to develop a predictive model of T2DM using genetic and clinical data; and 2) to compare misclassification rates of different models.

Diabetes is a major health problem all over the world. Many classification algorithms have been applied for its diagnoses and treatment. An Ensemble Learning Approach and Support Vector Machine is proposed for the classification of diabetes patients. In medical diagnosis field accuracy is a major factor.By using SVM+C4.5 algorithm ,the accuracy and performance decreases.So ,an attempt has been made in study to improve accuracy by using new ensemble learning approach called "Random Forest".It obtains better accuracy,improve the prediction quality and also reduce time-consuming.The proposed Algorithm is implemented and evaluated using Pima Indians Diabetes Data set from UCI repository of machine learning databases. Hence the above limitation can be overcome by using the Random Forest algorithm for diagnosis of diabetes .

II. METHODS AND MATERIAL

A. Data Preprocessing

Data preprocessing is a data mining technique[11] that involves transforming raw data into an understandable format. Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors. Data preprocessing is a proven method of resolving such issues. Data preprocessing prepares raw data for further processing. Data preprocessing is used database-driven applications such as customer relationship management and rule-based applications (like neural networks).
Data goes through a series of steps during preprocessing:

Data Cleaning: Data is cleansed through processes such as filling in missing values, smoothing the noisy data, or resolving the inconsistencies in the data.

Data Integration: Data with different representations are put together and conflicts within the data are resolved.

What can be wrong with data? There is a hierarchy of problems that are often encountered in data preparation and pre-processing:

1. Impossible values have been inputted.
2. Unlikely values have been inputted.
3. No values have been inputted (missing values).
4. Irrelevant input features are present in the data at hand.

Impossible values should be checked for by the data handling software, ideally at the point of input so that they can be re-entered[13]. These errors are generally straightforward, such as coming across negative prices when positive ones are expected. If correct values cannot be entered, the problem is converted into missing value category, by simply removing the data.Incomplete data is an unavoidable problem in dealing with most real world data sources. Generally, there are some important factors to be taken into account when processing unknown feature values. One of the most important ones is the source of unknowingness[14]:

(i) A value is missing because it was forgotten or lost;
(ii) A certain feature is not applicable for a given instance (e.g., it does not exist for a given instance);
(iii) For a given observation, the designer of a training set does not care about the value of a certain feature.

B. Feature Selection

Feature selection (FS) techniques have become a necessity in all applications. FS can avoid over fitting and gain a deeper insight into the unknown areas, such as occurrence and diagnosis of diseases. As a result, we utilized two filter techniques (univariate LR, chi-square tests) select the relevant features. Univariate LR selected the features which were statistical significant with P value < 0.05.

In statistics, chi-square test was applied to test the independence of two events[7]. However, in FS procedure, two events represented the occurrence of the feature t and occurrence of the class ci .

$$X^2 = \frac{N[P(t,c_i)P(\bar{t},\bar{c_i}) - P(t,\bar{c_i})P(\bar{t},c_i)]^2}{P(t)P(\bar{t})P(\bar{c_i})P(c_i)} \quad (1)$$

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

IG is defined to be the expected reduction in entropy[16]. If features are continuous, IG uses information theoretic binning to discretize the continuous features.The measure of feature importance in RF is the total decrease in node impurities from splitting on the variable, averaged over all trees.

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

Where p represents the fraction of positive examples assigned to a certain node k and 1 – p as the fraction of negative examples.

C. Rule Extraction From SVM

In this module, the unbalanced dataset is handled and data is used for training SVMs with RBF kernel[9]. SVM is based on the rule of structural risk minimization and it belongs to the supervised learning models for nonlinear classification analysis[12]. The SVM model is achieved by finding the optimal separating hyperplane

(w.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:

$$\text{Minimize } \tfrac{1}{2} \|w\|^2 \text{ subject to } y_i (wx_i + b) \geq 1 \quad (3)$$

$$\text{Minimize } \tfrac{1}{2} \|w\|^2 + C \sum_{i=1}^{n} \varepsilon_i \quad (4)$$

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. Usually, radial basis function (RBF) is used as the kernel function to map the data

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

Where $||x - x'||^2$ the squared Euclidean distance between two is vectors and σ^2 is a free parameter.

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.SVs are the only examples that make contribution to the classification of the SVM. 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.

Additionally, 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. 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 Proposed system is depicted as Fig1.

D. Rule Generation and Evaluation

The RF is an ensemble learning method for classification [10]. 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, for instance, decision tree. Meanwhile, some ensemble methods, especially ba gging, 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.

III. RESULTS AND DISCUSSION

PERFORMANCE EVALUATION

Confusion Matrix: Confusion matrix shows predicted and actual classifications. A confusion matrix for a classification problem with two classes is of size 2×2, and it is given in Table 1.

Table 1. Confusion Matrix

Predicted	**Actual**	
	Positive	**Negative**
Positive	TP (true positive)	FP (false positive)
Negative	FN (false negative)	TN (true negative)

- TP represents an instance, which is actually positive and predicted by the model as positive.
- FN represents an instance, which is actually positive but predicted by the model as negative.
- TN represents an instance, which is actually negative and predicted by the model as negative.
- FP represents an instance, which is actually negative but predicted by the model as positive.

Sensitivity and Specificity: Sensitivity is the true positive rate, and specificity is the true negative rate. They are defined as follows:-

$$\text{Sensitivity}= \frac{TP}{TP+FN}$$

$$\text{Specificity} = \frac{TN}{FP+TN}$$

$$\text{Accuracy} = \frac{\text{Number of true records predicted}}{\text{Number of total records}}$$

The precision is calculated as follows:

$$\text{Precision} = \frac{\text{True positive}}{\text{True positive+False positive}}$$

The calculation of the recall value is done as follows: and Comparison of accuracy is shown in Table 2.

$$\text{Recall} = \frac{\text{True positive}}{\text{True positive+False negative}}$$

The Fidility is calculated as

$$\text{F-Measure} = \frac{\text{Precision .Recall}}{\text{Precision + Recall}}$$

Table 2. Comparison of Accuracy for PID

Diabetes	SVM	GA	GA-SVM	SVM+ C4.5	SVM +RF
Accuracy %	77.73	82.98	78.64	81.87	89.02

Table 3. shows that accuracy,precision,recall and F-measure value of the proposed system.

Table 3. Accuracy, Precision, Recall, F-Measure of the Proposed System

Medical Dataset	**Proposed SVM+Random Forest**			
	Accuracy	Precision	Recall	F-Measure
PIMA	89.02	0.89	0.90	0.90

Figure 2. Shows that Comparison of accuracy with existing methods**.**

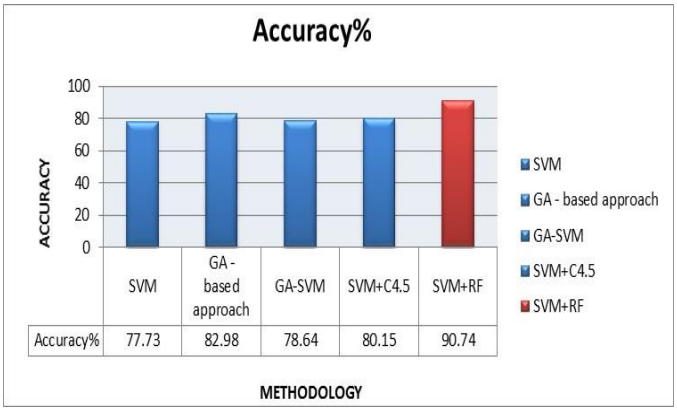

Figure 3. Comparison of Accuracy

IV. CONCLUSION

An expert system based on Ensemble Learning is proposed for medical diagnosis of diseases such as Diabetes. The SVM+C4.5 model can be used to obtain better accuracy and improve the prediction quality. The proposed system's performance is evaluated using a data set from UCI repository with respect to classification accuracy. Precision,recall and F-measure rates are presented for further analysis of the system. Chi-Square is used for the selection of the most significant feature set of the dataset. The disease classification process is based on Support Vector Machine. SVM+Random Forest has provided a classification accuracy of 89.02% for diabetes. In future work,prune the rule sets of the proposed system, the obtained rule sets are much less and smaller than RF, but still larger than C4.5 and NBTree.

V. REFERENCES

[1] Heikes, Kenneth E., et al. "A Simple Screening Tool for Detecting Undiagnosed Diabetes and Prediabetes." Diabetes 56 (2007).

[2] Buijsse, Brian, et al. "Risk assessment tools for identifying individuals at risk of developing type 2 diabetes." Epidemiologic reviews (2011): mxq019.

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

[4] 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 modernizing china: What elements of urbanisation are most associated with diabetes?" Diabetologia, vol. 55, no. 12, pp. 3182–3192, 2012.

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

[6] 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.

[7] 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.

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

[9] H. N´u˜nez, C. Angulo, and A. Catal`a, "Rule extraction from support vector machines," in Proc. Eur. Symp. Artif. Neural Netw., 2002, pp. 291–296.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

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

[16] X. J. Fu, C. J.Ong, S.Keerthit, andG.G.Hung, "Extracting the knowledge embedded in support vector machines," in Proc. IEEE Int. Conf. Neural Netw., 2004, pp.107112.

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

[18] 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.

[19] 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.

[20] 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.