Study on a New Sparse Rule Algorithm in Liver Disease

Zhangfan Ye
Fuzhou University Zhicheng College
Fuzhou, China
yezhangfan@fzu.edu.cn

Song Chen
School of Physics and Information Engineering, Fuzhou University,
Fuzhou, China
chensongcg@163.com

2020 IEEE 3rd International Conference of Safe Production and Informatization (IICSPI) | 978-1-7281-7738-0/20/$31.00 ©2020 IEEE | DOI: 10.1109/IICSPI51290.2020.9332314

***Abstract*—Most of the current diagnostic methods of liver diseases are black box model which can't fully display the information hidden in the data.In this paper, we propose a new spare rule extraction algorithm on the basis of random forest, combined with elastic L1 and L2 norm regularization to select high precision and interpretable diagnostic rules. The method is compared with the state of art techniques in hepatitis dataset and the experimental results show that our method can extract more accurate, concise and strong explanatory rules. We believe that the proposed approach achieve excellent performance in tradeoff curve and more effectively predict the risk of liver disease in the health care environment.**

Keywords—Machine Learning, Rule Extraction, Random Forest, Norm Regularization, Liver Inflammation

I. Introduction

Substantial progress has been made in the research of liver disease knowledge over the past few decades, for example, hepatitis is an inflammation of the liver, which may deteriorate to fibrosis, cirrhosis or liver cancer. Liver disease can be caused by infection, injury, drug reactions, toxins, autoimmune processes or genetic defects.The diagnosis of liver disease can be treated as a two classification problem. Although at present many diagnostic methods for liver diseases have been successfully applied in practice, they[1-3] are all black box model. The disadvantage of the black box model is that they do not adequately reveal information that is hidden in the data. Although some black box model achieve very high accuracy, but it can't give the reasons for the classification which is very important to doctor.The knowledge representation rules extracted from the data are more popular and easier to be understood than other representations .In this paper we propose a new method to select effective rules using the norm convergence algorithm. This paper design a mixed rule extraction and feature selection method. It is a continuous iteration. The iteration step include selecting the feature from the extracted rules, and extracting important rules from the selected features. A random forest combined with sparse encoding is used to extract important rule. The continuous iteration will not stop until the selected features and rules is changed. In experimental we compare our algorithm with other state of art algorithms in the Pareto curve to exhibit the balance of the concise and accuracy.

II. Related Work

University of California Irvine (UCI) machine learning database [10], it has been widely used as a benchmark for classification algorithms. Hepatitis disease data set is provided by the Joseph Institute in the UCI database.

Many diagnostic methods for hepatitis data sets have been successfully applied to different classification algorithms: Clustering based on attribute weighting [1], learning machine [4], Support Vector Machine [5], Classification and regression tree, Support vector identification [6], principal component analysis [6]. Most of the work is focused on the SVM classifier in order to improve the interpretability of the generated rules. Yoichi Hayashi et al. proposed a new rule extraction algorithm [7] which is a continuous iterative technique combined with the sample selection and the sample data set preprocessing. However, in order to overcome the difficulty of extracting high precision rules, iterative algorithm for continuous sampling is presented. But its cost is less simple at the same time more rules is extracted[7].Inspired by this, this article aims to achieve optimal balance in precision and concise of the rules.

III. Method

We designs a new spare rule extraction algorithm on the basis of random forest, combined with norm regularization, to select high precision and interpretable diagnostic rules. In general, it maps the sample data to the binary coded space defined by the entire leaf node set of a random forest. Secondly, extracts representative rules thought the binary coded space. Then, after rules selection, recalculate the new feature as a sub-feature of the next iteration. The new feature is used to construct a new random forest which will generate a new set of rules. The process is repeated until the stopping condition is met. Finally, the number of features remains stable and the number of rules converge.

A. Binary coding of random forests

Random forests are a kind of classifier which is trained and predicted by many trees. In short, random forests are made up of multiple decision trees. For each tree, the training data are sampled with replacement from the total training set. It means that in the total training set there are some samples which may occur several times in a tree of training, also may never appear in any tree of the training set. In the training of each tree node, the features of sample data extracted accord with a certain proportion of random without replacement. It recursively split the data into subsets. Gini index and information gain [8] is the two common criteria for data splitting.

The path from the root to the leaf node could be interpreted as a decision rule, so a random forest could be

978-1-7281-7738-0/20/$31.00 ©2020 IEEE

regarded as a set of decision rules. Each sample traverses each decision tree of random forest from the root node to only one leaf node, thus we could define a binary eigenvector to represent the random forest leaf node structure. Figure 1 shows the mapping of nodes.

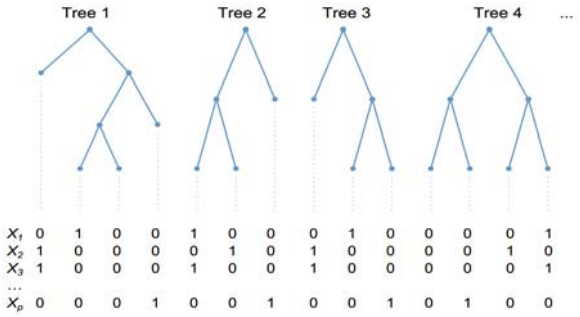

Fig. 1. Binary coding of random forests

X_i represents the leaf node space. In this space, each sample is mapped to the vertices of the hypercube. The dimension of each rule space is defined as a decision rule. So which rules are valid and which are invalid is represented by X_i. The above two binary mapping has been applied to the random forest for edge learning.

B. Rule extraction using sparse coding

A new set of training samples is obtained by the above mapping, $\{(X_1, y_1), (X_2, y_2), ..., (X_p, y_p)\}$ where X_i is a vector of binary attributes and $y \in \{1, 2, ..., K\}$ is the corresponding class label. We define the classifier formula as follows:

$$y = \arg \max_{k \in \{1,...,k\}} \left(W_k^T X + b_k \right) \tag{1}$$

Where the weight vector W_k and scalar b_k define the linear discriminant function of class K. In The formula (1) the weight W_k measures the importance of rules. Obviously, if a weight is 0 for all the class, the rules it represent can be safely removed. So rule extraction is the import problem in learning weight vector.

In this paper, we solve the learning problem using probability (p) norm regularization:

$$\min_{W_k, \xi_{ik}} \left(\lambda \sum_{k=1}^{K} \left\{ p \|W_k\| + (1-p) \|W_k\|_2 \right\} + \sum_{i=1, k=1, ..., K}^{L} \sum_{k \neq y_i} \xi_{ik} \right)$$
$$s.t \left(W_{y_i} - W_k \right)^T X_i + b_{y_i} + b_k + \xi_{ik} \geq 1 \tag{2}$$
$$\xi_i \geq 0, i = 1, ..., p$$

The objective function consists of two parts: the first term $\sum_{k=1}^{K} \left\{ p \|W_k\| + (1-p) \|W_k\|_2 \right\}$ control the number of nonzero weights and rule extraction and the second term is the sum of the relaxation variables. Because the non-zero relaxation variable represents a sample of misclassification, the second term is related to empirical error. The sparsity and empirical error of the results depend on the regularization parameter. L1 and L2 norm sparse coding has been widely used in statistics and machine learning [9]. L1 norm can delete more unimportant features, and L2 norm can prevent data overfitting. So, we choose probability norm that combined L1 with L2 norm, which make our method not only select relatively few features, but also enhance generalization ability.

C. Feature deletion and rule extraction

We design a feature selection method based on random forest. The main step is as follows:

Feature weight of samples represents the importance of a feature of samples. It can be calculated as follows:

- For each decision tree in a random forest, calculate the OOB (Out-of-Bag) data error which is denoted as errOOB1.
- Randomly add the noise to the feature of the OOB data (This can randomly change the value of the feature), then calculate the OOB data error again, which is denoted as errOOB2.
- Suppose that there are N trees in the random forest, then the importance of the feature is defined as:

$$\sum (\text{err}00B1 - \text{err}00B2) / Ntree \tag{3}$$

If the accuracy of the OOB data greatly reduced when adding random noise to a feature, the feature has great impact on the classification results of the sample.

There are two goals for feature selection:

- Find a feature variable that is highly relevant to the dependent variable
- Select a small number of feature variables which can sufficiently predict the result of the dependent variable.

The steps of feature selection:

1) Preliminary estimation and sorting

a) Sort the feature variables in random forests in descending order according importance (Variable Importance, VI).

b) Delete a proportion of the current feature variables according to importance. It will get a new feature set.

c) Builds a new random forest with the new features, then calculates and sorts the VI of each feature in the feature set.

d) Repeat the above steps until the features is unchanged.

2) According to feature set and the random forest which is set up in the step 1, we can calculate the corresponding OOB error rate and select the feature set which has minimum OOB error as the final feature set.

The feature distribution in random forests is determined by the learning process of random forests. Usually, the feature distribution is different those produced by rule extraction in the previous formula. Based on the assumption

that the important features are located in the extracted decision rules, we can use this to select features. If the feature is not in rules extracted by the formula (2), it will be removed because it is not affected by the classifier which is defined by the formula (1). In this way, the rules and features can be selected at the same time.

The regularization parameter λ can be chosen by cross validation of sample set. By selecting the features to reconstruct the random forest, the rules can be further selected to become more concise. In such a process of iteration, the features are selected to construct a new random forest, through the new random forest it can produces new rules. The iteration will not stop until the feature will not change. The generated rules is shown in figure 2.

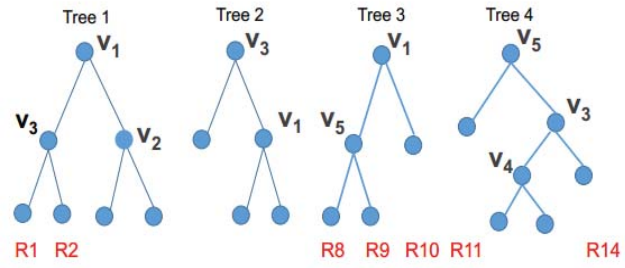

Fig. 2. Generated rules after iteration

IV. Results and Discussion

A. Data set and Preprocessing

Data set: the hepatitis dataset, which consists of 155 instances and contains 19 attributes for each sample. The value of this database include text and numeric value. Also it is a complex and noisy data set because it contains a large number of missing data. There were 32 (20.65%) deaths examples and 123 (79.35%) survived examples. The classification task is to predict the survival or death of a patient with hepatitis[10].

Data preprocessing: if the train data set is imbalanced, it will cause a lot of problems in pattern recognition. For example, if the data set is not balanced, then the classifier prefers to "learn" a largest proportion of samples and cluster them with the highest accuracy. In the practical application, it will not be acceptable with such bias. It can be seen from table I, the data set is seriously imbalanced. In order to achieve the equal distribution of sample data, many methods have been proposed. In this paper, using SMOTE (Synthetic Minority Oversampling Technique) to solve this problem. SMOTE algorithm creates "synthetic" instances for the each minority class with very few samples.

B. Performance Comparison of Rule Extraction Algorithm

In order to verify the validity, this paper uses K-fold (Cross validation, CV) to test the accuracy of classification for the rules. The K-fold CV method is widely used to minimize the bias caused by random sampling. In this experiment, k=5.

In this paper, the number of rules is 4. As shown in Table I, the accuracy of our proposed algorithm is 84.37%. Therefore, the method proposed in this paper has achieved higher accuracy and got more concise rules.

C. Algorithms Comparison for extracting rules

TABLE I. Performance of rule extraction algorithms for the Hepatitis dataset

Rule extraction method [validation method] [Ref.]	**TR ACC (%)**	**TS ACC (%)**	**# Rules**	**Rule set**	**Total #ante.**	**Ave.# ante**
C4.5 [Averaged over 100 runs][11]	-	78.94	5.85	No	-	-
PART [Averaged over 100 runs][12]	-	80.02	6.64	No	-	-
Decision Table [10CV][13]	-	81.93	28	No	-	-
Repeated Incremental Pruning to Produce Error Reduction (RIPPER)[10CV] [14]	-	78.06	4	No	-	-
Partial C4.5 Decision Tree [10CV][15]	-	84.51	8	No	-	-
Ripple Down Rule Learner [10CV][16]	-	78.71	2	No	-	-
Sampling-Continuous Re-RX [5×2CV][117]	89.04	83.24	3.5	Yes	7	1.90
Our Approach	**92.91**	**84.37**	**4**	**Yes**	**8**	**4**

It can be seen that ELSE rules are a black box, which is in R4 extracted by EvoC method in Table III. One of the main goals of rule extraction is to provide a clear understanding and interpretation. In R4 method ELSE is a black box and blackly assigns a class label to the sample. The rules obtained in this study can obtain better accuracy and get more concise rules and each rule can clearly classify survival or death in Table II.

TABLE II. Rules extraction for the Hepatitis dataset using our method

Rule number	**Rule expression**
R1	Malaise > 1.50 AND Varices > 1.50 THEN DIE
R2	Age, years > 32.50 AND Spiders≤1.50 THEN LIVE
R3	Varices > 1.50AND Bilirubin ≤2.250 AND Spleen palpable > 1.50 AND Bilirubin > 0.65 THEN LIVE
R4	Age, years≤64.00AND Spiders > 1.50 AND Liver big > 0.50 AND Varices > 1.50 AND Liver firm > 1.50 THEN LIVE

D. The weight of attributes of sample in rule set

TABLE III. Rules extraction for the Hepatitis dataset using EvoC [11]

Rule number	**Rule expression**
R1	IF Fatigue=Yes AND Age≥30.0 AND ALP≤280.0 AND Albumin 4.3 AND Protime≤46.0 THEN Class=DIE
R2	IF Anorexia=No AND Bilirubin≤1.8 AND AST≤420.0 THEN Class=LIVE
R3	IF Spiders=Yes AND Age≥30.0 AND 62.0≤ALP≤175.0 AND Albumin≤4.3 AND Protime≤85.0 THEN Class=DIE
R4	ELSE Class=LIVE

It can be seen from Figure 3, the probability is relatively

high in the attributes of the rules, indicating that these properties are important factors in hepatitis disease. For example, the attributes (8, 13) have highest probability of diagnosis so their changes should be paid more attention.

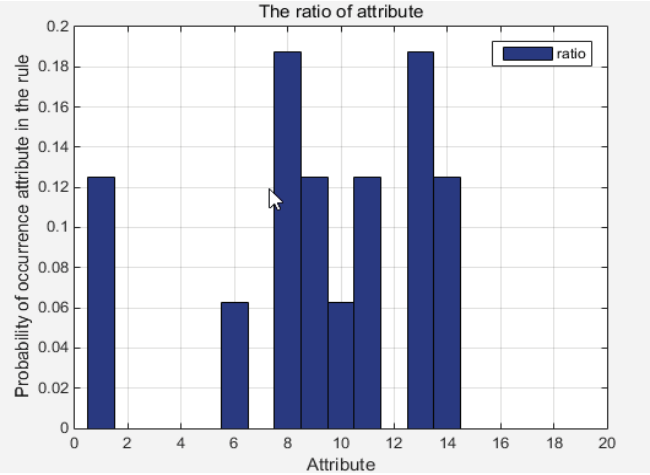

Fig. 3. Probability of occurrence attribute in the rule

E. Accuracy of rule expressions

Rule R can be evaluated using coverage and accuracy. Given a labeled data set D, define n_{covers} the number of coverage of the data and $n_{correct}$ the number of data classified by rule R. The coverage and accuracy is defined as:

$$Coverage(R) = \frac{n_{covers}}{|D|} \tag{4}$$

$$Accuracy(R) = \frac{n_{correct}}{n_{covers}} \tag{5}$$

TABLE IV. RULE VALIDATION

Rule number	Coverage	Accuracy	prediction label
1	0.626	1	2
2	0.594	1	1
3	0.634	1	2
4	0.496	1	2

As shown in Table IV, we can verify the effectiveness of each rule for the hepatitis data set. The higher coverage and the accuracy the rules have, the more confidence the rules will be for auxiliary diagnostics. In this study, the top three rules is used for hepatitis auxiliary diagnosis.

F. The trade-off between classification accuracy and the number of extraction rules

As shown in Table II, researchers have proposed several algorithms for rule extraction for the hepatitis data set. In order to understand the performance of each algorithm better, this paper uses the tradeoff curve to evaluate relationship between precision and rule number, as shown in figure 4.

The reciprocal of the rule number is displayed on the horizontal axis. The red point of our approach located near the tradeoff curve. It shows that optimization between the performance of our algorithm and the number of rules are best.

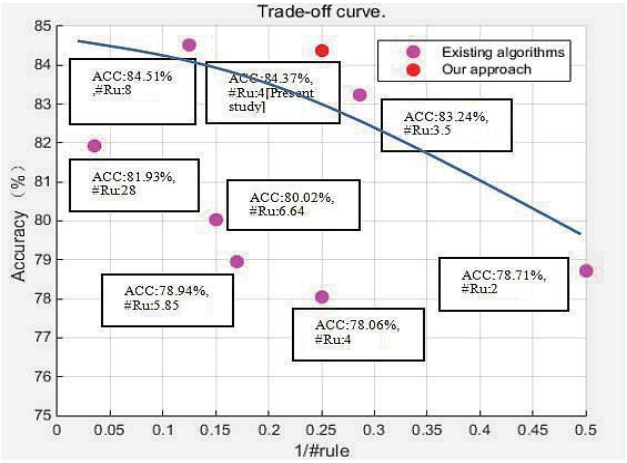

Fig. 4. Trade-off curve between the accuracy and number of rules extracted for the Hepatitis dataset.

V. CONCLUSION

This paper presents a method based on random forest classifier and probability norm regularization to extract more accurate and concise rules. The Pareto curve is used to analyze the optimization of extracted rules optimally and the result demonstrates that our method can achieve good balance between accuracy and interpretability. Also we show that the rules generated by our algorithm can guarantee the accuracy of the diagnosis and is more suitable for medical decision making. In the future, we will focus on how to improve the stability of the algorithm and try to apply it to practical business applications.

ACKNOWLEDGMENT

This work was supported by Fujian Province Young and Middle-aged Teacher Education Research Funding Project (JA15626)

REFERENCES

[1] K.Polat, “Application of Attribute Weighting Method Based on Clustering Centers to Discrimination of Linearly Non-Separable Medical Datasets,” Journal of Medical Systems, 2012, pp. 2657-2673.

[2] M.Seera, CP. Lim, SC.Tan and CK.Loo, “A hybrid FAM–CART model and its application to medical data classification,” Neural Computing & Applications, 2015, pp. 1799-1811.

[3] R.Liu, “A particle swarm optimization based simultaneous learning framework for clustering and classification,” Pattern Recognition, 2014, pp. 2143-2152.

[4] P.Mohapatra, S. Chakravarty and P.K. Dash, “An Improved Cuckoo Search based Extreme Learning Machine for Medical Data Classification,” Swarm & Evolutionary Computation, 2015, pp. 25-49.

[5] J.S.Sartakhti, M.H.Zangooei and K. Mozafari, “Hepatitis disease diagnosis using a novel hybrid method based on support vector machine and simulated annealing (SVM-SA),” Computer Methods & Programs in Biomedicine, 2012, pp. 570-579.

[6] J.Habibi, M.H. Zangooei and R. Alizadehsani, “Disease Diagnosis with a hybrid method SVR using NSGA-II,” Neurocomputing, 2014.

[7] Y.Hayashi, S. Nakano and S. Fujisawa, “Use of the recursive-rule extraction algorithm with continuous attributes to improve diagnostic accuracy in thyroid disease,” Informatics in Medicine Unlocked, 2015.

[8] WY.Loh., “Classification and Regression Trees,” Wiley Interdisciplinary Reviews Data Mining & Knowledge Discovery, 2011, pp. 14-23.

[9] R.Tibshirani, “Regression shrinkage and selection via the lasso: a retrospective,” Journal of the Royal Statistical Society, 2011, pp.58.

[10] K.C.Tan, Q.Yu, C.M. Heng, “Evolutionary computing for knowledge discovery in medical diagnosis,” Artificial Intelligence in Medicine, 2003, pp.129-154.

[11] S.Salzberg,.“Programs for Machine Learning by J. Ross Quinlan.

Morgan Kaufmann Publishers, Inc., 1993," Machine Learning, 1994, pp. 235-240.

[12] J.J.Christopher, H.K.Nehemiah and A. Kannan, "A Swarm Optimization approach for clinical knowledge mining," Computer Methods & Programs in Biomedicine, 2015, pp.137-148.

[13] R.Kurz, "Putting the child first: research as a part of paediatric care The Joseph J Hoet Lecture on Ethics in Paediatric Research given at the European Conference on Clinical Research in Children, 24-25 January 2002, Brussels," International Journal of Pharmaceutical Medicine, 2002, pp. 11-13.

[14] W.W.Cohen,"Fast Effective Rule Induction,"Machine Learning Proceedings, 1995, pp. 115-123.

[15] A.S.Larik, S. Haider, "Rule-Based Behavior Prediction of Opponent Agents Using Robocup 3D Soccer Simulation League Logfiles," Springer Berlin Heidelberg, 2012.

[16] D.Uehara, "Non-invasive prediction of non-alcoholic steatohepatitis in Japanese patients with morbid obesity by artificial intelligence using rule extraction technology," World Journal of Hepatology, 2018, pp. 934-943.

[17] D.Wodzisaw, R. Adamczak, Y. Hayashi, "Neural eliminators and classifiers, 2019.

[18] M. Saravanan and A. Priya (2019). An Algorithm for Security Enhancement in Image Transmission Using Steganography. Journal of the Institute of Electronics and Computer, 1, 1-8.