ISSN: 2278-4632
Vol-10 Issue-6 No. 1 June 2020

Machine Learning Algorithm for Analysis and Prediction of Seer Cancer

Vikram Pareek[1], Shankar Sinha Saubhagya[2], Sorathia Nawaz[3], Dr. Marlene Grace[4]

[1,2,3]UG Student, [4]Professor & HOD, Department of CSE
[1,2,3,4]Kommuri Pratap Reddy Institute of Technology, Ghatkesar, Hyderabad, India.

ABSTRACT

The recurrence of breast cancer is a prevailing problem that decreases the quality of patients' lives, creates high burdens on the healthcare system, and impacts the wellbeing of society. Advanced sensing provides an unprecedented opportunity to increase information visibility and characterize patterns of event occurrences. However, few, if any, of previous works have investigated survival analysis of breast cancer recurrences based on large amount of data readily available in the health system. There is a dire need to leverage data to decipher important factors that play a role in the recurrence of breast cancer. This paper presents an ensemble method of random survival forest for time-to-event analysis of breast cancer recurrences in the surveillance, epidemiology, and end results (SEER) data. Our model characterizes the survival function among patients with and without recurrences of breast cancer. Ensemble models are constructed via sampling and bootstrapping into the big data. Experimental results show that the age when cancer recurrence happens and time-between-recurrences approximately follow the Gaussian and exponential distributions with the means of 61.35 ± 14.03 and 2.61 years, respectively. In addition, the results show age, surgery status, stage of tumors, and histological grade are significant factors that influence the probability of breast cancer recurrences. The proposed survival analysis approach shows strong potentials to help healthcare practitioners in prognosis, treatment, and decision-making of breast cancer recurrences.

Keywords: Seer cancer, machine learning, random forest.

1. INTRODUCTION

Cancer incidence and mortality have been increasing at an accelerated pace over the past 3 decades globally, making cancer the major public health problem [1]. Among females, breast cancer is known as the most diagnosed cancer and the main cause of cancer deaths in more than 100 countries. In 2018, there are about 2.1 million newly diagnosed breast cancer cases around the world, responsible for nearly 1 in 4 cancer cases among females [2]. However, the causes of breast cancer are still not clearly known to doctors. Early diagnosis of breast cancer can make the disease easier to treat [3,4]. Several diagnosis techniques are commonly used to distinguish malignant breast tumors from benign ones. Fine Needle Aspiration (FNA) is a well-known procedure used to diagnose breast cancer, but it suffers from a lack of satisfactory diagnosis performance. For FNA, radiologist, oncologist, and pathologist are required to render final judgment together in breast cancer

ISSN: 2278-4632
Vol-10 Issue-6 No. 1 June 2020

diagnosis, which is time-consuming. Also, there is higher possibility to give rise to errors due to exhaustion or inexperience, which panic patients when false-positive result happens or miss optimum treatment time when false-negative result appears. Therefore, developing an efficient diagnosis support system to assist doctors' diagnosis of breast cancer has great significance for medical diagnosis process. Machine learning based interpretable diagnosis systems can enhance the cancer diagnosis ability and decrease the diagnosis misjudgment. What is more, diagnosis support system can provide early breast cancer diagnosis support for the doctors and patients. Cancer research is commonly clinical in practice, and hence data driven research method has been becoming popular. Data mining assists humans in utilizing available data to discover previously unknown and potentially useful knowledge with the help of machine learning [5]. Existing research demonstrates that introducing data mining and machine learning methods into breast cancer diagnosis can gain lots of benefits including increasing diagnosis accuracy, cutting down costs and reducing medical resources [4]. Among the methods for cancer diagnosis, breast cancer diagnosis is often treated as the classification problem of distinguishing benign breast tumors from malignant ones [6]. However, how to get an idea performance for general classification problems is still difficult until now [7]. Particularly, breast cancer diagnosis has become more and more important since the rules for classification and the classification results have a great effect on patients' treatment, which requires both high accuracy and strong interpretability.

2. LITERATURE REVIEW

In machine learning, many single learning methods, such as LR, NB, KNN, SVM, ANN and DT, have been widely employed for breast cancer diagnosis for their technical maturity, stability, and cost saving. Karabatak [11] proposed a weighted NB to predict breast cancer status with high sensitivity, specificity, and accuracy, and use a heuristic search algorithm for allocating different weights for each NB classifier as input. Bagui et al. [12] introduced a rank nearest neighbors' rule (RNN) method to improve KNN and maintain the prediction accuracy of 97%. Akbulut [13] devised a SVM-based method incorporating with feature selection that obtained 99.51% classification accuracy. Ahmad et al. [14] applied ANN to the breast cancer detection and compared the results on the Wisconsin breast cancer dataset. ANNs are usually regarded as a black box because of the complex nonlinear mapping of artificial neural network to data, and it is difficult to clearly describe the decision-making process of a well-trained network. Also, producing rules from multilayer perceptron artificial neural network is an NP-hard problem [9]. To surmount this limitation, Jafari-Marandi et al. [15] developed a decision-oriented ANN classification method, referred to as Life-Sensitive Self-Organizing Error-Driven (LS-SOED), which focused on improving decision making instead of improving accuracy only. Dancey et al. insisted that DT, because of the combination of symbolic information and graphical representation, had become one of the easiest forms of representation for pattern recognition knowledge. Sumbaly et al. [6] discussed various data mining methods for breast cancer diagnosis and proposed a DT-based data mining method for early breast cancer diagnosis. Delen et al. used ANN, DT along with LR to develop the prediction models using SEER dataset and found that DT performed the best. Sivakumar et al.

ISSN: 2278-4632
Vol-10 Issue-6 No. 1 June 2020

concluded many research about breast cancer and found that different classification methods such as Best first (BF) Tree, Alternative Decision (AD) Tree, Functional Trees (FT) Tree, J48, Random Tree (RT), Random Forest (RF) tree, and Classification And Regression Tree (CART) have been widely applied to breast cancer diagnosis, because they could develop a decision tree to select features that were significantly more valuable than others.

3. IRFRE-BASED INTERPRETABLE DIAGNOSIS SYSTEM

In this section, we devise an interpretable diagnosis system for breast cancer detection. To be specific, an improved random forest-based rule extraction (IRFRE) method is developed to acquire accurate and functional decision rules from the previous diagnostic records, which are then available for joint decision making when dealing with new cases to be diagnosed. The process of cancer diagnosis procedure based on IRFRE can be illustrated in Fig. 1. After patients are examined by doctors using different facilities, their data is transferred to system for preprocessing. The transformed data is then brought into the proposed IRFRE method to generate, optimize, and derive the most effective rule ensemble model, followed by identifying the well-performed rules which will be provided to doctors for examining and further validating. Thereafter, the remaining validated rules are stored in the rule base for doctors to predict the diagnosis result of patients. The diagnostic instances that have clear and definite result can be used to enrich the database and update the existing prediction model and the rule base. The core part of the diagnosis system tends to be the proposed IRFRE method, which provides fundamental and irreplaceable supports for the diagnostic model learning and decision-making.

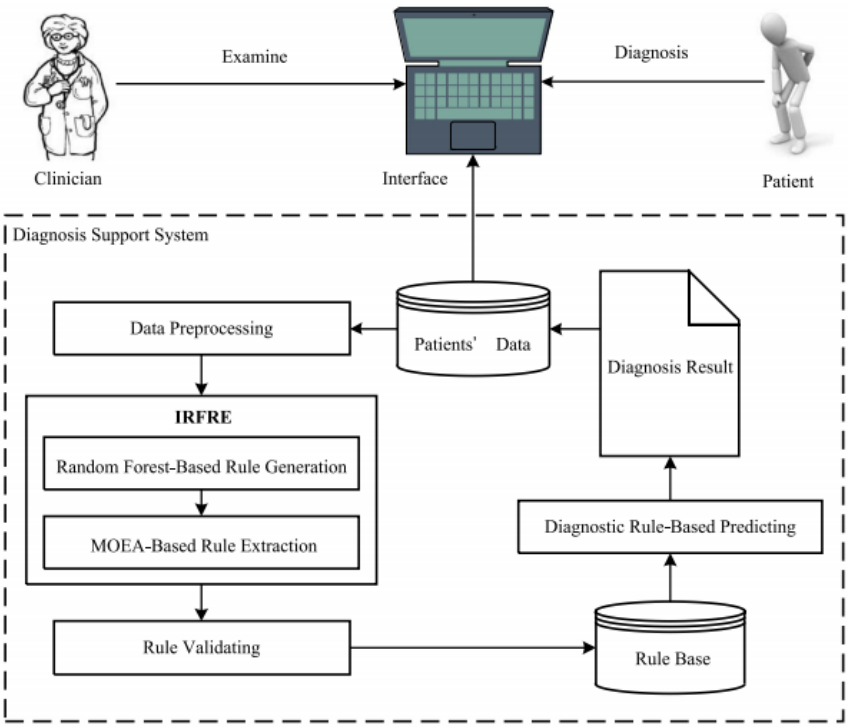

Fig. 1: Flow chart of IRFRE-based patients' cancer diagnosis process.

In this section, we mainly elaborate the mechanism of IRFRE. IRFRE is a kind of rule extraction method for classification. The structure of IRFRE comprises two parts: Random Forest based rule generation and multi-objective evolutionary algorithm (MOEA)-based rule extraction, where the second part is naturally executed after the results of the first part are obtained without human intervention. The main idea of the two parts of IRFRE structure can be briefly described as follows, respectively.

(1) As for the Random Forest-based rule generation, Random Forest is first constructed with Classification and Regression Tree (CART) as base learner for generating amounts of decision trees. IF-THEN rules are then detached from the trained trees by tracking the path from the root node to each leaf node in each tree, which are combined into a rule set.

(2) As for the MOEA-based rule extraction, a multi-objective evolutionary algorithm is employed to identify optimal combination of rules. As far as we know, this is the first time that a MOEA is introduced to rule extraction from ensembles of CART.

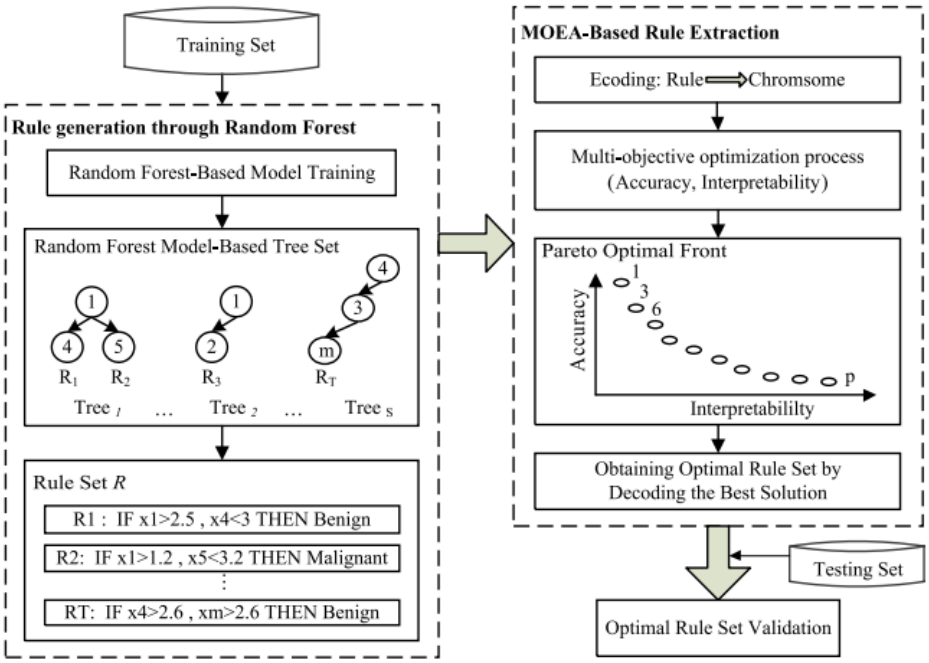

Fig. 2: The framework of IRFRE.

To be precise, there are two mains steps of our MOEA-based rule extraction. First, formalized description of multi-objective rule extraction incorporating chromosome representation, rule matching and objectives settings is provided, where the two objectives, accuracy and interpretability of each rule set are optimized simultaneously. Second, the rule extraction evolutionary process involving non-dominated sorting-based selection, uniform crossover and flip bit mutation problems is carried out to find the Pareto optimal front comprising a series of accuracy interpretability trade-offs corresponding to the evolved optimal rule sets. Fig. 2 illustrates the framework of IRFRE,

ISSN: 2278-4632
Vol-10 Issue-6 No. 1 June 2020

where m and n denote the numbers of features and training samples, respectively, S is the number of CARTs of Random Forest, T represents the total number of rules derived from CART, and p stands for the number of extracted rule sets obtained from MOEA. In the following part, we formally illustrate the implementing details for the two main parts of IRFRE, i.e., Random Forest-based rule generation and MOEA-based rule extraction.

This subsection focuses on describing the process of rule generation and developing a formal algorithm to extract rule sets from Random Forest. The generation of effective rule sets relies on the certain rule generation mechanism. Random Forest-based rule generation model is a kind of efficient and reliable approach among different rule generation methods. The rules obtained from these models can clearly reflect the entire decision-making process. As an outstanding ensemble learning method, Random Forest (RF) has good performance in generalization and still has room for improvement in interpretability. It is notable and well recognized that RF is generally more robust to noise compared with other decision tree ensembles. RF, a collection of classifiers (H) is constructed based on Gini index through bootstrap sampling and feature randomizing. It considers randomly selecting splitting features based on decision trees and each tree is constructed independently by bootstrap sampling. To deal with categorical or continuous values of training set efficiently, Classification and Regression Tree (CART), initially introduced by Berk for either regression or classification problems, is successfully used as the base classifier for feature splitting. CART is a non-parametric procedure with categorical or continuous features, where the data and labels are divided into nodes and subsets with binary tree recursively, so it is a white-box algorithm and easy to extract rules from it. CART of RF can be transformed into classification IF-THEN rules by tracing the path from the root node to each leaf node in the search tree. In each path, the features of node correspond to the rule conditions, and the class of leaf node corresponds to the conclusions of the rules. According to Nguyen, the IF-THEN rules can be expressed as: IF THEN . For instance, a rule structure of breast cancer prediction can be expressed as: Rule 1: IF X1 < A, X4 < B, X8 < C THEN Class = Benign where X1, X4 and X8 are the features used to classify whether the breast tumors of patients are benign or malignant. When the feature values are less than the threshold value and the conditions are satisfied, then the tumor can be treated as benign. Each rule is exclusive and complete, indicating that each instance is covered by only a path or a rule.

4. EXPERMRNTAL RESULTS

We leveraged the SEER dataset from various regions and over different time periods to study and characterize recurrent events among the patients who are initially diagnosed with breast cancer.

ISSN: 2278-4632
Vol-10 Issue-6 No. 1 June 2020

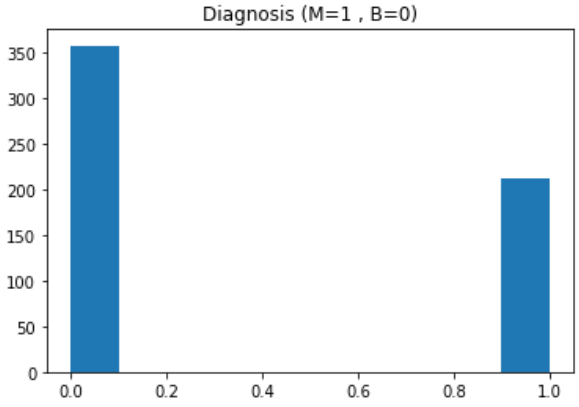

Fig. 3: SEER diagnosis analysis.

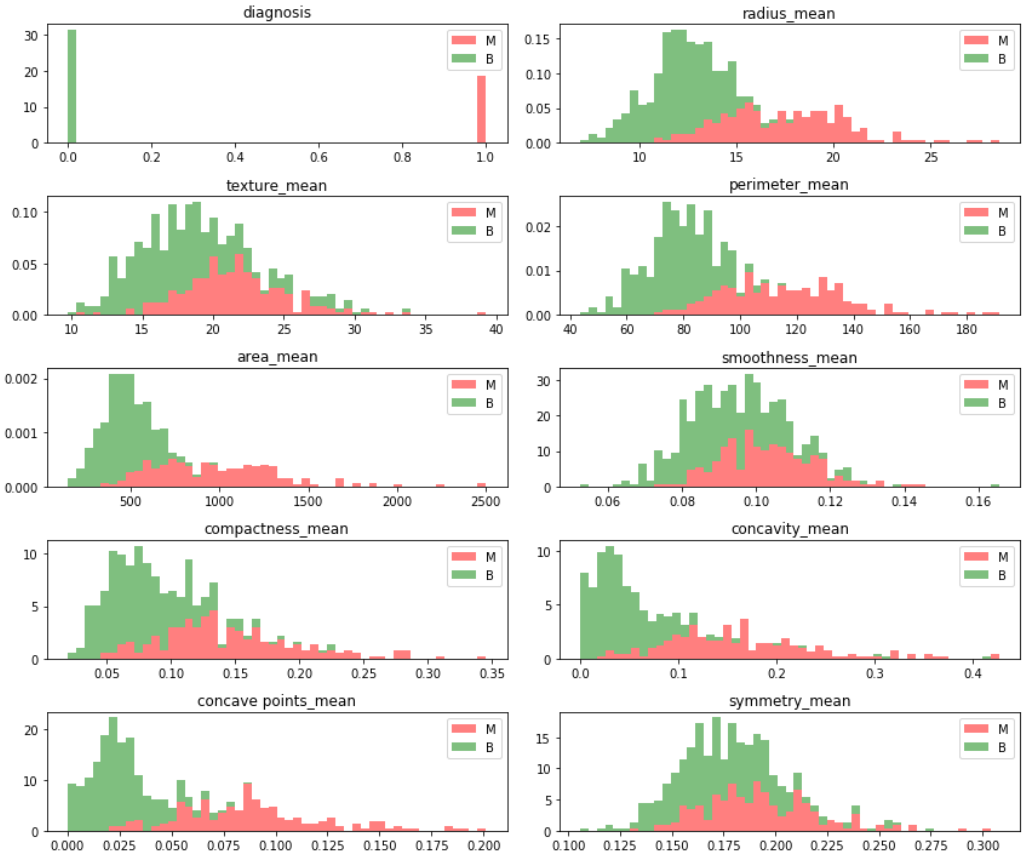

Fig. 4: Various metrics analysis.

From the above fig 3, 4 we can make the following observations. For SEER dataset, the proposed method IRFRE performs better than NB, k-NN, and significantly ANN in terms of accuracy, sensitivity, and Kappa with a comparable specificity. SVM and RF achieve higher accuracy, sensitivity, and Kappa values compared with IRFRE on this dataset. Therefore, the proposed method shows primary advantages in interpretability and good classification ability on this dataset. As for WOBC dataset, the prediction ability of IRFRE is statistically superior to NB and comparable to k-NN, SVM, ANN, and RF. This result shows that IRFRE can achieve a higher or close prediction performance in

ISSN: 2278-4632
Vol-10 Issue-6 No. 1 June 2020

comparison to popular black-box models when seeking the best interpretability. For SEER dataset, IRFRE outperforms all the compared black-box models in terms of all the three indicators on accuracy and interpretability. In the aspect of sensitivity, the performance of IRFRE is like that of Random Forest and superior to those of the other black-box methods statistically. The specificity of SVM is extraordinarily high with a sacrifice of sensitivity. ANN has a good specificity with big variance and the lowest sensitivity. Note that in diagnosis practice, reducing Type I error, i.e., improving sensitivity, is of vital importance for making sure the patients could be identified correctly without delay. Basically, IRFRE can perfectly balance the sensitivity and specificity on the three datasets. It is worthwhile to notice that the sensitivity performance of IRFRE shows apparent superiority on each dataset.

5. CONCLUSIONS

In this paper, we develop an improved Random Forest-based rule extraction (IRFRE) method for breast cancer diagnosis, where the black-box Random Forest is first opened with a rule extraction mechanism and then an improved multi-objective evolutionary algorithm is developed to support the accuracy and interpretability-oriented rule optimization process. To ensure the diversity of solutions, we propose an improved nondominated sorting mechanism for discarding the duplicate individuals and ensure the diversity of solutions. All the experiments are performed on three breast cancer datasets (WDBC, WOBC, and SEER dataset) from the perspectives of method accuracy and interpretability. We also conduct the independent T-test for validating the significance of IRFRE. The experimental results show that the developed methods can primely explain the black box methods and outperform several popular single methods, ensemble learning methods, and rule extraction methods in terms of accuracy and interpretability, which can dramatically improve cancer diagnosis performance. It can also improve interpretability while keep accuracy compared to original Random Forest on three datasets.

REFERENCES

[1] E.M. Ward, C.E. DeSantis, C.C. Lin, J.L. Kramer, A. Jemal, B. Kohler, O.W. Brawley, T. Gansler, Cancer statistics: Breast cancer in situ, CA. Cancer J. Clin. 65 (2015) 481–495.

[2] F. Bray, J. Ferlay, I. Soerjomataram, R.L. Siegel, L.A. Torre, A. Jemal, Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, CA. Cancer J. Clin. 68 (2018) 394–424.

[3] R.L. Siegel, K.D. Miller, A. Jemal, Cancer statistics, 2018, CA. Cancer J. Clin. 68 (2018) 7–30.

[4] L. Peng, W. Chen, W. Zhou, F. Li, J. Yang, J. Zhang, An immune-inspired semi-supervised algorithm for breast cancer diagnosis, Comput. Methods Programs Biomed. 134 (2016) 259–265.

[5] M. Kantardzic, Data Mining: Concepts, Models, Methods, and Algorithms, second ed., John Wiley & Sons, Inc, 2011, pp. 1–25.
[6] R. Sumbaly, N. Vishnusri, S. Jeyalatha, Diagnosis of breast cancer using decision tree data mining technique, Int. J. Comput. Appl. 98 (2014) 16–24.
[7] H. Wang, B. Zheng, S.W. Yoon, H.S. Ko, A support vector machine-based ensemble algorithm for breast cancer diagnosis, European J. Oper. Res. 267 (2018) 687–699.
[8] G. Fung, S. Sandilya, R.B. Rao, Rule extraction from linear support vector machines, in: Proceeding Elev. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. - KDD '05, ACM Press, New York, USA, 2005, p. 32.
[9] M. Mashayekhi, R. Gras, Rule extraction from random forest: the RF+HC methods, in: Advances in Artificial Intelligence, Barbosa, D. and Milios, E., Springer International Publishing, Canada, 2015, pp. 223–237.
[10] M.B. Gorzałczany, F. Rudziński, A multi-objective genetic optimization for fast, fuzzy rule-based credit classification with balanced accuracy and interpretability, Appl. Soft Comput. 40 (2016) 206–220.
[11] M. Karabatak, A new classifier for breast cancer detection based on Naïve Bayesian, Measurement 72 (2015) 32–36.
[12] S.C. Bagui, S. Bagui, K. Pal, N.R. Pal, Breast cancer detection using rank nearest neighbor classification rules, Pattern Recognit. 36 (2003) 25–34.
[13] M.F. Akay, Support vector machines combined with feature selection for breast cancer diagnosis, Expert Syst. Appl. 36 (2009) 3240–3247.
[14] F. Ahmad, N.A. Mat Isa, Z. Hussain, M.K. Osman, S.N. Sulaiman, A GA-based feature selection and parameter optimization of an ANN in diagnosing breast cancer, Pattern Anal. Appl. 18 (2015) 861–870.
[15] R. Jafari-Marandi, S. Davarzani, M. Soltanpour Gharibdousti, B.K. Smith, An optimum ANN-based breast cancer diagnosis: Bridging gaps between ANN learning and decision-making goals, Appl. Soft Comput. 72 (2018) 108–120.
income and middle-income countries. The Lancet
370,1929-38.

