Mathematical Biosciences and Engineering

http://www.aimspress.com/journal/MBE

MBE, 19(3): 2193–2205.
DOI: 10.3934/mbe.2022102
Received: 16 September 2021
Revised: 28 November 2021
Accepted: 14 December 2021
Published: 04 January 2022

Research article

Breast cancer diagnosis using feature extraction and boosted C5.0 decision tree algorithm with penalty factor

Jian-xue Tian and Jue Zhang *

School of Information Engineer, Yulin University, Road chongwen, Yulin 719000, China

* **Correspondence:** Email: zhangjue@stumail.nwu.edu.cn; Tel: +8618991097062.

Abstract: To overcome the two class imbalance problem among breast cancer diagnosis, a hybrid method by combining principal component analysis (PCA) and boosted C5.0 decision tree algorithm with penalty factor is proposed to address this issue. PCA is used to reduce the dimension of feature subset. The boosted C5.0 decision tree algorithm is utilized as an ensemble classifier for classification. Penalty factor is used to optimize the classification result. To demonstrate the efficiency of the proposed method, it is implemented on biased-representative breast cancer datasets from the University of California Irvine(UCI) machine learning repository. Given the experimental results and further analysis, our proposal is a promising method for breast cancer and can be used as an alternative method in class imbalance learning. Indeed, we observe that the feature extraction process has helped us improve diagnostic accuracy. We also demonstrate that the extracted features considering breast cancer issues are essential to high diagnostic accuracy.

Keywords: breast cancer diagnosis; boosted C5.0 algorithm; principle component analysis

1. Introduction

Breast cancer is one of the most deadly diseases for women around the globe [1]. The issue of breast cancer was predicted to be doubled to 1.6 million by the year 2025 [2]. Until now, the cause of breast cancer is still not known to doctors. Early diagnosis of breast cancer is the only way to ensure a long survival of the patients [3, 4]. Hence, if the earlier the tumor detects before spreads, the greater hope it cures. Therefore, accurate diagnosis of breast cancer has become one of the important and urgent problems in medical science fields.

Several machine learning studies are commonly used to improve classification accuracy. Chen et al. [5] used rough set-based SVM classifier and improved the accuracy to appropriate 97% with different features combination which had a clue to physicians. Li et al. [6] proposed a novel supervised dimensionality reduction method that can preserve the relationship of the data, and the parameters

were conducted which obtained 96.98% accuracy. Zheng et al. [7] proposed a method that combined K-means and SVM algorithms, the K-means algorithm was utilized for dimensionality reduction, and SVM algorithm was used as classifier. This method improved the accuracy to 97.38%. Gorunescu and Belciug et al. [8] proposed an evolutionary-based method and tested it on five datasets to show its effectiveness. Karabatak et al. [9] proposed a weighted Naïve Bayes classifier and applied it in breast cancer diagnosis with 5-fold cross-validation sampling method. R.Sheikhpoure et al. [10] hybridized the particle swarm optimization with non-parametric kernel density estimation method in the breast cancer diagnosis and obtained high performance. An immune-inspired semi-supervised algorithm for breast cancer diagnosis proposed by Peng et al. [4] was a modification of artificial intelligence inspired by biological systems and very effective in experiments.

All the methods mentioned above [4–10] have achieved high accuracy. However, all these methods aim to improve the overall classification accuracy but ignore the minority class accuracy. But in practice, for breast cancer prediction, the cancer cases are pretty rare compared with the healthy populations, and the accuracy of data classification in the minority class is critical. Thus, the breast cancer diagnosis problem should be handled from class imbalance perspective.

Many variants of data mining algorithms are designed to solve the class imbalance problem effectively. Typically, Ijaz et al. [11] proposed a cervical cancer prediction model (CCPM) which combined the outlier detection methods, DBSCAN and iForest, the data oversampling methods, SMOTE and SMOTETomek, and random forest classifier for cancer prediction to improve the performance. Mandal et al. [12] presented a new tri-stage feature selection framework for disease classification which use ensemble of four filter methods in the first phase ,Correlation in the second phase and classifer in next stage. Experiments show effeciveness of these proposed algorithms, and these two methods are regarded to be as the best method, but the use on different combinations of imbalanced algorithm and feature selection algorithm has not been investigated.

In general, four types of methods have been used to tackle the class imbalanced problem. They are data-level method, algorithm-level method, cost-sensitive method, and ensemble learning method [13]. The data-level methods are simple, undersampling or oversampling may alter the original class distribution of data [14]. The goal of algorithms method is to propose novel algorithms or modify existing algorithms to directly handle data sets. High cost is assigned to minority class in cost-sensitive methods to improve the classification performance [15]. Among the popular algorithms for binary imbalanced classification, the ensemble of classifiers with penalty factor has attracted significant attention. In addition, considering that each method has its own shortcoming relative to others, we come up with an ensemble strategy to make use of the advantages of the multiple methods and avoid the shortcomings [16]. C5.0 decision tree algorithm is an improved algorithm of C4.5, as the most fundamental and widely used classification method in various fields. It has noticeably low error rates, less memory and can easily support for boosting technqiue. Research [17, 18] showed the benefit of tree-based ensemble approaches for classifying imbalanced data. Thus, in this paper, we adopt as the boosted C5.0 ensemble algorithm with penality factor to solve the classification problem.

Furthermore, feature preprocessing has become an essential preprocessing step that cannot ignore to get a more accurate result in expert system [19]. A variety of feature selection approaches have been proposed, but most of them need to exhaustive examine all possible feature subsets and select the smallest feature [12, 16, 20, 21] . Therefore, these algorithms are computationally inefficient since the exhuastive search. Recently, applications of machine learning algorithms for the feature extraction

have become increasingly popular with techniques [22,23], such as PCA and LDA. Thus, in this paper, we combined feture extraction and boosted C5.0 ensemble algorithm with penalty factor to improve performance. To do so, we can improve the classification performance.

The proposed method is called the PCA and boosted C5.0 with penalty factor (P-Boosted C5.0), where it refers to the three stages in this study. First, PCA is used to transform the original feature subset into a new smaller feature subset. To the best of our knowledge, PCA is popular, simplicity and traditional algorithm for feature extraction. Second, boosted C5.0 algorithm is used as classifier, as boosted C5.0 refers to a general and practical is an effective solution to deal with class imbalance problem, is a reasonable approach to leverage the strength of individual classifiers. Third, penalty factor matrix is employed to impact the classification results, since it represents a balance between maximizing the classification interval and minimizing the classification error. The proposed algorithm is evaluated on famous UCI breast cancer datasests, and the experimental results show its effectivenss and efficieny.

The main contributions of this paper can be summarized as follows: 1) PCA as feature extraction algorithm is used for extract the optimal feature subset; 2) Boosted C5.0 is used as ensemble learning approach for classification to further improve the performance; 3) penalty factor matrix is used to adjust the result by adding a high misclassification cost to the minority class; 4) The empirical results on WDBC datasets reveal the effectiveness of PCA-Boosted C5.0.

The remainder of this paper is structured as follows: Section 2 describes the proposed PCA-Boosted C5.0 algorithm. In Section 3, we present the experiment results and compare them with several other traditional algorithms. The discussion and conclusions are presented in Section 4.

2. The proposed boosting C5.0 decision tree algorithm with penalty factor

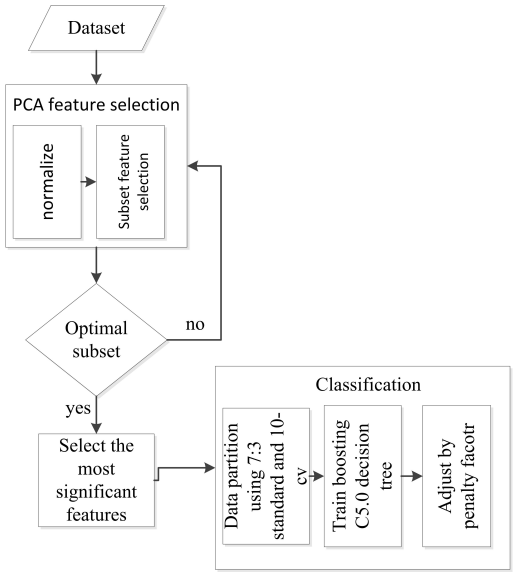

Figure 1. Block diagram of the proposed classification algorithm.

Figure 1 shows the block diagram of the proposed PCA-Boosted C5.0 algorithm. This method consists of three stages. The first step is based on feature extraction method. For instance, PCA is used on the dataset to extract optimal feature subset, and thereby good feature subset leads to high classification accuracy. For instance, in order to reduce dimensionality of features, we use PCA to extract features, as the resulting contribution of estimated principal components is calculated, and those whose contribution less than 10% to the total are eliminated to improve the accuracy of the breast cancer predicion. The second stage is to perform the ensemble classification algorithm on the subset. The obtained subset which obtained in the first step is given as input to a second-stage learning model. Cost sensitive matrix is employed to adjust the classification result in the third stage. Specifically, cost sensitive methods consider high-cost weights for minority class.

2.1. PCA for feature extraction

Feature extraction is a crucial factor for computational systems applied to diagnosis [24]. Improving the feature selection performance could improve the classification performance. We employ PCA in our search as a preprocessing step for enhancing classification effectiveness. PCA is a popular unsupervised linear technique which attempts to transform the original feature sets which include a large number of features to a new smaller feature space, so that the current data can be experssed with a few number of features variable. First of all, we use a normalize function Eq (2.1) to rescale the features' values to a standard range between 0 and 1 since different intervals of features' value were present, so can be measured in a single standard.

$$x = (x - min(x))/(max(x) - min(x)) \tag{2.1}$$

The details of PCA are shown in Algorithm 1.

Algorithm 1: The feature extraction algorithm procedure

Input: a set of 30 dimension feature vectors of original dataset $\langle D \rangle$ $D = \{x_i\}, i = 1, 2, \ldots, n$,

Output: Projection matrix $w = \{w_d\}, d = 1, 2, \ldots, n$,

First, let $u = \frac{1}{m} \sum_{t=1}^{M} x_t$

Then, the convariance matrix of samples is

$C = \frac{1}{m} \sum_{t=1}^{M} (x_t - \mu)(x_t - \mu)^T$

The pricipal components (PCs) are computed by solving the eigen value problem of covariance matrix C, $Cv_i = \lambda_i v_i$

Where $\lambda_i (i = 1, 2, \ldots, n)$ are the eigen values and they are sorted in descending order,$v_i (i = 1, 2, \ldots, n)$ are the corresponding eigenvectors.

To represent the raw feature vectors with low-dimensional ones, what needs to be done is to compute the first $k(k <= n)$ eigenvectors which correspond to the k largest eigen values. To select the number k, a threshold ϑ is introduced to denote the approximation precision of k largest eigenvectors.

$$\frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{n} \lambda_i} >= \theta \tag{2.2}$$

Given the precision parameter θ, the number of k eigenvectors can be decided.

$V = [v_1, v_2, \ldots, v_k]$, Λ=diag$[\lambda_1, \lambda_2, \ldots, \lambda_k$

After the matrix V is decided, the low-dimensional feature vetor of raw ones are determined as follows:

$$P = V^T x_t \tag{2.3}$$

Thus, with PCA the maximum variance is explained by the first principal compont, after that the second variant is calculated, it orders the principle components so that those with the largest variation come first, and eliminate the features which contribute least to variation.

In our datasset, The screen plot of the main components by feature extraction algorithm of PCA is shown in Figure 2. The red color line changes in Figure 2 tend to be stable after the 9th principle component, which indicates 9th principle component has reached most of the original data. Therefore, we can transform the original feature data into a quantitative structure for training convenience. In practice, the first 9th principal components are chosen as inputs to a second-stage classification algorithm. Here, the benefit of feature extraction is that the information can be maintained as the original data avoiding iteration combination.

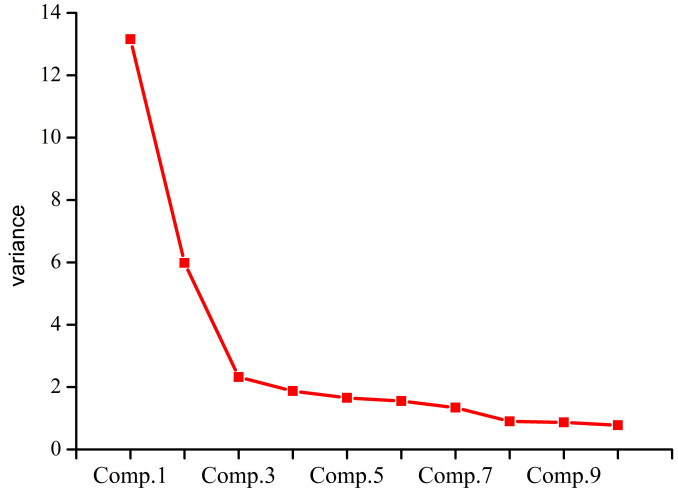

Figure 2. Principle component analysis for tumor.

2.2. Improved C5.0 decision tree for classification

Decision tree, which is the most fundamental and widely used classification method in machine learning field. C5.0 decision tree is an improved top-down algorithm of C4.5, and it uses information gain as splitting criteria to build a decision tree. The criteria of C5.0 is Gain Ration which is a modification of the information gain. The benefit of C5.0 is noticeably low error rates, less memory, and high optimization. Therefore, C5.0 algorithm is more accurate and much faster. C5.0 has tree like structures, prunes the original decision tree, and creates decision tree in the way of "divide and rule". In addition, the most improvement in C5.0 is boosting technique.

Boosting is a simple and effective ensemble learning method for producing accurate classifiers. The principle of boosting algorithm is repeatedly calling weak learners and giving these weak learners high weight vote value. By doing so, the training process can focus more on the cases that caused error, which tends to reduce bias. With respect to C5.0, the most critical feature of C5.0 is boosting technique, and another is the construction of a cost-sensitive. As mentioned above, the boosting and

cost-sensitive technique can provide superior in accuracy of the overall performance. As for this, in this work, we propose a novel breast cancer automated diagnosis method, which employs PCA for feature extraction, boosted C5.0 for classification, and cost sensitive matrix for adjusting classification results. In our proposed approach, we not only consider the classification performance but also the unequal misclassification costs of tumors. In our experiment, PCA feature extraction algorithm was employed to achieve the optimal feature subset that leads to the optimal classification performance to improve the overall performance. Then, boosted C5.0 was used as classification algorithm. Lastly, cost matrix was used to adjust the classification results.

2.3. Decision making trade-off with cost sensitive matrix

In order to solve the imbalance problem, the paper adds a misclassification cost into the weight of instance. To best of our knowledge, the cost associate with missing a cancer case (false negative) is much higher than those of mislabeling a benign one (false negative). Specifically, false negative cases may spend more cost associated with unnecessary biopsies for pathological analyses, but in false positive cases, the patient may miss timely treatment and lead to death. In other words, Cost sensitive methods consider different cost weights for majority and minority classes. This attempt is more beneficial for the final classification boundary away from the minority class, then enhances the absolute classification accuracy, especially for the minority class. Consequently, we aim to use a matrix of costs associated with possible errors to adjust classification results. In this paper, a cost matrix formed by $C \times C$ where C is the number of classes is used. In this paper, cost sensitive matrix is provided in Table 1. A value of 4 in the matrix indicates that the cost of predicting a patient as healthy (false negative) is four times the cost of predicting health as patient (false positive). This value is suggested by research.

In this paper, cost sensitive matrix is provided in Table 1. A value of 4 in the matrix indicates that the cost of predicting a patient as healthy (false negative) is four times the cost of predicting health as patient (false positive). This value is suggested by research.

Table 1. Cost sensitive matrix.

	Actual positive	Actual negative
predict positive	0	1
predict negative	4	0

3. Experimental results and discussion

In order to evaluate the performance of hybrid approach on imbalanced datasets, we test the proposed algorithm on WDBC and WBCD datasets. The experiments are performed on *R* version *x*64 3.2.5 on a PC with an Intel(R) Core(TM) *i*3 – 4130 CPU (3.40 GHz) with 4 GB of RAM, using Windows 10 operating systems. The P-Boosted C5.0 algorithm was implemented with C50, caret, e1071, kernlab, ROCR, gplot and gmodels packages of R. Note that packages with default setting were used.

To test the effectiveness of the proposed P-Boosted C5.0 for breast cancer diagnosis, two standard breast cancer datasets are applied. In addition, to assert the contribution and significance of the proposed algorithm, the proposed algorithm was compared with some of the previous results reported by

earlier methods in literature. Meanwhile, to evaluate the effectiveness of our proposed method, we compare the result of P-Boosted C5.0 with two well-known classifiers on two standard breast cancer datasets. In addition, in order to make the observation more convincing, we conduct 10 independent runs of experiments for each partition, and the average classification performance results are computed, respectively.

3.1. Dataset

We used real-world Wisconsin breast cancer dataset (WDBC taken from the UCI machine learning repository) in our experiment. This dataset is commly used among researchers who used for breast cancer classification, so it can provide us easily to compare the performance of our method with that of literature methods. The WDBC include 569 observations and 32 patient attributes, which include 30 tumor feature, and an ID and one class label. Tumor features were collected from a digitized image of a fine needle aspirate. The ten main variables used to predict benign or maligant cases were 1) radius, 2) texture, 3) perimeter, 4) area, 5) smoothness, 6) compactness, 7) concavity, 8) concave points, 9) symmetry and 10) fractal dimension. 212 samples of the dataset belong to malignant class and 357 are of the dataset are of malignant class. Specifically, the information of each dataset is summarized in Table 2.

Table 2. Specification of breast cancer imbalanced dataset.

Dataset	No.of data sample	No. of feature	Imbalance ratio
WDBC	569	30	1.800

3.2. Evaluation metrics

According to Raeder [25], evaluation measures play an important role in assessing classification performance. Generally, the class of methods usually adopts accuracy as the performance evaluation index. But in class-imbalanced scenario, the overall accuracy as evaluation criteria is not so meaningful since the classification interest is often the minority class. Actually, for two-class imbalanced problems, the class success is typically measured by the geometric mean of true positive and true negative rates [26] which G-mean represents. Thus, in this paper the G-mean are adopted as the performance metric for evaluating imbalanced learning classifier. It is better indicators to show the performance trade-off between classes than overall accuracy for their imbalanced class distribution.

The evaluation indicators are computed based on evaluation metrics derived from the binary confusion matrix presented in Table 3.

Table 3. Confusion matrix.

	Predicted positive	Predicted negative
Actual positive	True positive(TP)	False negative(FN)
Actual negative	False positive(FP)	True negative(TN)

Where TP, TN represents correctly classified the instances as benign or malignant, and FP, FN represents incorrectly classified the instances as benign or malignant. The calculation formulas are

defined as follows:

$$accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{3.1}$$

$$sensitivity = \frac{TP}{TP + FN} \tag{3.2}$$

$$specificity = \frac{TN}{FP + TN} \tag{3.3}$$

$$G - mean = \sqrt{sensitivity \times specificity} \tag{3.4}$$

3.3. Experiment results

In order to perform a comprehensive comparison of proposed algorithm in handling breast cancer problem, we conducted experiments on 1 UCI data sets. They are well-chosen and concluded in Table 2. In our experiment, it includes two parts (i) the classification performance is compared, revealing the importance of extracting features, (ii) we perform a comprehensive comparison of all algorithms in each dataset.

Figure 2 reports the result of PCA for WDBC dataset. In the figure, we observe that the 9th feature subset has the same discernibility as the original set of fetures. Therefore, utilizing feature extraction algorithm is the key to simplifying the part of the data processing phases and improving the performance by choosing significant features.

In contrast, to evaluate the performance of the proposed ensemble approach, we compare the results P-Boosted C5.0 with P-SVM, P-NB, RUSBoost [27] and SMOTE-Boosted C5.0. First, performance comparison of P-Boosted C5.0, P-SVM and P-NB to show the superior performance of Boosted C5.0 since NB and SVM have been considered as the most effective and common algorithm for breast cancer; Second, the comparison between P-Boosted C5.0 and SMOTE-Boosted C5.0 shows that the benefits of PCA algorithm; Third, P-Boosted C5.0 is compared with RUSBoost which are the state-of-the-art approach for imbalanced data to show the benefit of the proposed hybrid P-Boosted C5.0 algorithm. Herein, SMOTE-Boosted C5.0 is a classical hybrid algorithm, SMOTE algorithm as sampling method is used to imbalanced the class distribution, Boosted C5.0 algorithm is used as the ensemble classifier. In contrast, the parameter of over and under in SMOTE is set to 100 and 300. Also, trails parameters values of boosting algorithm are sets as 25, an empirical value is suggested by literature.

Moreover, in order to obtain statistically meaningful conclusion 10-fold cross validation is repeated ten times, and average results are presented in SMOTE-Boosted C5.0, RUSBoost, P-SVM and P-NB. Among these algorithms the best classification G-mean is highlighted in bold typeface.

Table 4. Confusion matrix of proposed method.

	Predicted benign	Predicted malignant
Actual benign	128	3
Actual malignant	1	38

Table 5. Performance comparison(%).

Method	Accuracy	Sensitivity	Specifictiy	G-mean
P-Boosted C5.0	97.65	92.68	99.22	**95.89**
P-SVM	93.53	97.44	92.37	94.87
P-NB	93.57	94.87	93.18	94.02
RUSBoost	94.40	93.00	95.40	94.20
SMOTE-Boosted C5.0	92.50	93.90	91.10	92.48

The confusion matrix of P-Boosted C5.0 is listed in Table 4. Table 5 reports the accuracy, specificity, sensitivity, and G-mean of P-Boosted C5.0 and different classification methods for WDBC dataset. As shown in Table 5, P-Boosted C5.0 outperforms other methods where 70–30 partition is performed in terms of G-mean. As it can be observed from the results listed, 95.89% G-mean with nine features is obtained by proposed P-Boosted C5.0 which gets the best performance among all methods. Both theoretical and experimental results show that the combination of hybrid P-Boosted C5.0 is a promising system.

To further validate the performance of the proposed P-Boosted C5.0 algorithm, the comparisons are also conducted with literature methods and several base classifiers, such as naïve Bayes NB. It is noticeable that, for fair comparison, NB are directly reported as benchamark binary classification method without any feature extraction prior actions. To introduce some more novel and advanced strategies for comparison, we adopted some recent methods, such as the IGSAGAW-CSSVM [28], RIPPER [29] and MaxE [30].

Finally, Table 6 illustrates the performance of the comparison methods mentioned earlier. The symbol is given as "-" which means we do not get data from literature. From the results of Table 6, the proposed P-Boosted C5.0 obtained the highest performance among the classifier reported in the literature [22–25]. The best G-mean achieved by the Aisl method is 97.28%. There may be two main reasons. First, feature selection is employed in literature method, which can identify the significant features and eliminate the irrelevant to improve the classification performance; However, our method of P-Boosted C5.0 uses feature extraction which tranform the original feature sets a new smaller feature space. Thus, this method disturbs the original data distribution, in some content it brings some noisy data; Second, performance of learning algorithm can be impacted by different factors, such as feature space characteristics and parameters. Nevertheless, the value of trial in P-Boosted C5.0 is suggested by the research which is not appropriate for specific issues. In addition, parameter setting and feature extraction play an essential role in the performance of breast cancer diagnosis.

As it can be observed from the result listed, the classification model performs well for diagnosis of breast cancer, the performance is significantly affected by the feature extaction algorithm and ensemble learning algorithm with penalty factor. However, the deep learning methods have shown promising results in cancer prediction [20, 21, 32], but it need more time and hyper-parameters. According to the aformentioned analysis, P-Boosted C5.0 is a promising and effective approach with imbalanced daaset with large number of features.

Table 6. Performance comparison(%).

ML method	Accuracy	Sensitivity	Specificity	G-mean
QKCLDA [6]	97.26	-	-	-
K-SVM [7]	97.38	-	-	-
Aisl [4]	98.00	95.90	98.70	**97.28**
BCT [31]	94.00	90.50	96.10	93.26
MaxE [30]	89.70	98.60	84.20	91.12
IGSAGAW-CSSVM [28]	95.70	-	-	93.60
IGSAGAW-KNN [28]	95.40	-	-	92.90
RIPPER [29]	94.40	91.10	95.78	93.40
NB	93.53	92.31	93.89	93.1
P-Boosted C5.0	97.65	92.68	99.22	95.89

4. Conclusion and future work

Biological data often consist of redundant and irrelevant feature, especially for breast cancer data. As the tumor features can be described as much detail as possible, the redundant information leads to large computation time for tediou calculation but without significant contribution time to the final results. Also, as the number of descriptive tumor features increases, the computational time increase rapidly as well. In this case, feature extraction which can remove irrelevant information into a new smaller feature subset, has becoming a crucial preprocessing step for classification system. Meanwhile, the issue of dealing with imbalanced data sets in breast cancer prediction is still unsolved.

To overcome the class imbalance problem in breast cancer classification and meanwhile keep the optimal new feature subset, a P-Boosted C5.0 algorithm is proposed. P-Boosted C5.0 is a three-step approach that first uses PCA for feature extraction to obtain the new optimal feature subset. Next, the Boosted C5.0 algorithm with fix value of trial is performed for classification. Third, cost senstive matrix is suggested for the penalty factor parameter, which was determined according to literature. Experiments were conducted on WDBC dataset with 569 samples. The experimental results demonstrated the advantages of the proposed P-Boosted C5.0 for solving the imbalance problem.

Future studies shall involve the setting of parameter according to special issues. Also, a deep learning method can be applied with a high-dimensional dataset since the deep learening methods have superiority in performance most time, yet not stable due to the impact of parameters. Thus, in future work we aim to create an adaptive method for setting parameter values in the deep learning method, where the value will be dependent on the minority class.

Acknowledgments

This work was supported by National Natural Science Foundation of China (61966029, 51868076), Shaanxi Science and Technology Research and Development Project (2018GY-024, 2020NY-163) and the Science and Technology Research of Yulin High-tech Zone (CXY-2021-44, CXY-2021-30, CXY-2020-09).

Conflict of interest

The authors declare there is no conflict of interest.

References

1. L. A. Torre, F. Bray, R. L. Siegel, J. Ferlay, J. Lortet-Tieulent, A. Jemal, Global cancer statistics, 2012, *CA Cancer J. Clin.*, **65** (2015), 87–108. https://doi.org/10.3322/caac.21262
2. M. F. Akay, Support vector machines combined with feature selection for breast cancer diagnosis, *Expert Syst. Appl.*, **36** (2009), 3240–3247. https://doi.org/10.1016/j.eswa.2008.01.009
3. R. L. Siegel, K. D. Miller, A. Jemal, Cancer statistics, 2018, *CA Cancer J. Clin.*, **68** (2018), 7–30. https://doi.org/10.3322/caac.21442
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. https://doi.org/10.1016/j.cmpb.2016.07.020
5. H. L. Chen, B. Yang, J. Liu, D. Y. Liu, A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis, *Expert Syst. Appl.*, **38** (2011), 9014–9022. https://doi.org/10.1016/j.eswa.2011.01.120
6. J. B. Li, Y. Peng, D. Liu, Quasiconformal kernel common locality discriminant analysis with application to breast cancer diagnosis, *Inf. Sci.*, **223** (2013), 256–269. https://doi.org/10.1016/j.ins.2012.10.016
7. B. Zheng, S. W. Yoon, S. S. Lam, Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms, *Expert Syst. Appl.*, **4** (2014), 1476–1482. https://doi.org/10.1016/j.eswa.2013.08.044
8. F. Gorunescu, S. Belciug, Evolutionary strategy to develop learning-based decision systems. Application to breast cancer and liver fibrosis stadialization, *J. Biomed. Inform.*, **49** (2014), 112–118. https://doi.org/10.1016/j.jbi.2014.02.001
9. M. Karabatak, A new classifier for breast cancer detection based on Naive Bayesian, *Meas.*, **72** (2015), 32–36. https://doi.org/10.1016/j.measurement.2015.04.028
10. R. Sheikhpour, M. A. Sarram, R. Sheikhpour, Particle swarm optimization for bandwidth determination and feature selection of kernel density estimation based classifiers in diagnosis of breast cancer, *Appl. Soft Comput.*, **40** (2016), 113–131. https://doi.org/10.1016/j.asoc.2015.10.005
11. M. F. Ijaz, M. Attique, Y. Son, Data-driven cervical cancer prediction model with outlier detection and over-sampling methods, *Sensors*, **20** (2020), 2809. https://doi.org/10.3390/s20102809
12. M. Mandal, P. K. Singh, M. F. Ijaz, J. Shafi, R. Sarkar, A Tri-Stage Wrapper-Filter Feature Selection Framework for Disease Classification, *Sensors*, **21** (2021), 5571. https://doi.org/10.3390/s21165571
13. H. Patel, G. S. Thakur, Classification of imbalanced data using a modified fuzzy-neighbor weighted approach, *Int. J. Intell. Eng. Syst.*, **10** (2017), 56–64. https://doi.org/10.22266/ijies2017.0228.07

14. W. C. Lin, C. F. Tsai, Y. H. Hu, J. S. Jhang, Clustering-based undersampling in class-imbalanced data, *Inf. Sci.*, **409** (2017), 17–26. https://doi.org/10.1016/j.ins.2017.05.008

15. P. D. Turney, Cost-sensitive classification: Empirical evaluation of a hybrid genetic decision tree induction algorithm, *J. Artif. Intell. Res.*, **2** (1994), 369–409. https://doi.org/10.1613/jair.120

16. H. E. Kiziloz, Classifier ensemble methods in feature selection, *Neurocomputing*, **419** (2021), 97–107. https://doi.org/10.1016/j.neucom.2020.07.113

17. M. Galar, A. Fernández, E. Barrenechea, H. Bustince, F. Herrera, Ordering-based pruning for improving the performance of ensembles of classifiers in the framework of imbalanced datasets, *Inf. Sci.*, **354** (2016), 178–196. https://doi.org/10.1016/j.ins.2016.02.056

18. J. Zhang, L. Chen, J. Tian, F. Abid, W. Yang, X. Tang, Breast cancer diagnosis using cluster-based undersampling and boosted C5. 0 algorithm, *Int. J. Control Autom. Syst.*, **19** (2021), 1998–2008. https://doi.org/10.1007/s12555-019-1061-x

19. Z. Zheng, X. Wu, R. Srihari, Feature selection for text categorization on imbalanced data, *ACM Sigkdd Explor. Newsl.*, **6** (2004), 80–89. https://doi.org/10.1145/1007730.1007741

20. S. Punitha, F. Al-Turjman, T. Stephan, An automated breast cancer diagnosis using feature selection and parameter optimization in ANN, *Comput. Electr. Eng.*, **90** (2021), 106958. https://doi.org/10.1016/j.compeleceng.2020.106958

21. P. N. Srinivasu, J. G. SivaSai, M. F. Ijaz, A. K. Bhoi, W. Kim, J. J. Kang, Classification of skin disease using deep learning neural networks with MobileNet V2 and LSTM, *Sensors*, **21** (2021), 2852. https://doi.org/10.3390/s21082852

22. H. Naeem, A. A. Bin-Salem, A CNN-LSTM network with multi-level feature extraction-based approach for automated detection of coronavirus from CT scan and X-ray images, *Appl. Soft Comput.*, **113** (2021), 107918. https://doi.org/10.1016/j.asoc.2021.107918

23. P. Huang, Q. Ye, F. Zhang, G. Yang, W. Zhu, Z. Yang, Double L2, p-norm based PCA for feature extraction, *Inf. Sci.*, **573** (2021), 345–359. https://doi.org/10.1016/j.ins.2021.05.079

24. H. D. Cheng, X. J. Shi, R. Min, L. M. Hu, X. P. Cai, H. N. Du, Approaches for automated detection and classification of masses in mammograms, *Pattern Recognit.*, **4** (2006), 646–668. https://doi.org/10.1016/j.patcog.2005.07.006

25. T. Raeder, G. Forman, N. V. Chawla, Learning from imbalanced data: Evaluation matters, in *Data mining: Foundations and intelligent paradigms*, Springer, (2012), 315–331. https://doi.org/10.1007/978-3-641-23166-7_12

26. S. Piri, D. Delen, T. Liu, A synthetic informative minority over-sampling (SIMO) algorithm leveraging support vector machine to enhance learning from imbalanced datasets, *Decis. Support Syst.*, **106** (2018), 15–29. https://doi.org/10.1016/j.dss.2017.11.006

27. C. Seiffert, T. M. Khoshgoftaar, J. Van. Hulse, A. Napolitano, RUSBoost: A hybrid approach to alleviating class imbalance, *IEEE Trans. Syst. Man Cybern. Part A: Syst. Hum.*, **40** (2009), 185–197. https://doi.org/10.1109/tsmca.2009.2029559

28. N. Liu, E. S. Qi, M. Xu, B. Gao, G. Q. Liu, A novel intelligent classification model for breast cancer diagnosis, *Inf. Process. Manage.*, **56** (2019), 609–623. https://doi.org/10.1016/j.ipm.2018.10.014

29. S. Wang, Y. Wang, D. Wang, Y. Yin, Y. Wang, Y. Jin, An improved random forest-based rule extraction method for breast cancer diagnosis, *Appl. Soft Comput.*, **86** (2020), 105941. https://doi.org/10.1016/j.asoc.2019.105941

30. H. Wang, B. Zheng, S. W. Yoon, H. S. Ko, A support vector machine-based ensemble algorithm for breast cancer diagnosis, *Eur. J. Oper. Res.*, **267** (Year), 687–699. https://doi.org/10.1016/j.ejor.2017.12.001

31. L. Breiman, Bagging predictors, *Mach. Learn.*, **24** (1996), 123–140. https://doi.org/10.1007/BF00058655

32. A. Taherkhani, G. Cosma, T. M. McGinnity, AdaBoost-CNN: An adaptive boosting algorithm for convolutional neural networks to classify multi-class imbalanced datasets using transfer learning, *Neurocomputing*, **404** (2020), 351–366. https://doi.org/10.1016/j.neucom.2020.03.064

AIMS Press