Iranian Journal of Medical Physics

ijmp.mums.ac.ir

Breast Cancer Diagnosis from Perspective of Class Imbalance

Jue Zhang[1,2], Li Chen [1*]

1. School of Information and Technology, Northwest University, Xi'an, China
2. School of Information Engineer, Yulin University, Yulin, China

ARTICLE INFO

Article type:
Original Article

Article history:
Received: Jun 04, 2018
Accepted: Sep 16, 2018

Keywords:
Breast Cancer
Classification
Imbalance
Computer Aided Diagnosis

ABSTRACT

Introduction: Breast cancer is the second cause of mortality among women. Early detection is the only rescue to reduce the risk of breast cancer mortality. Traditional methods cannot effectively diagnose tumor since they are based on the assumption of well-balanced dataset.. However, a hybrid method can help to alleviate the two-class imbalance problem existing in the diagnosis of breast cancer and establish a more accurate diagnosis.

Material and Methods: The proposed hybrid approach was based on improved Laplacian score (LS) and K-nearest neighbor (KNN) algorithms called LS-KNN. An improved LS algorithm was used for obtaining the optimal feature subset. The KNN with automatic K was utilized for classifying the data which guaranteed the effectiveness of the proposed method by reducing the computational effort and making the classification more faster. The effectiveness of LS-KNN was also examined on two biased-representative breast cancer datasets using classification accuracy, sensitivity, specificity, G-mean, and Matthews correlation coefficient.

Results: Applying the proposed algorithm on two breast cancer datasets indicated that the efficiency of the new method was higher than the previously introduced methods. The obtained values of accuracy, sensitivity, specificity, G-mean, and Matthews correlation coefficient were 99.27%, 99.12%, 99.51%, 99.42%, respectively.

Conclusion: Experimental results showed that the proposed approach worked well with breast cancer datasets and could be a good alternative to the well-known machine learning methods.

►*Please cite this article as:*
Zhang J, Chen L. Imbalance Class of Perspective the from Diagnosis Breast cancer. Iran J Med Phys 2019; 16: 241-249. 10.22038/ijmp.2018.31600.1373.

Introduction

Breast cancer is one of the most threatening type of cancer in women. The main cause of breast cancer is still unknown and there is no early symptoms in most patients. Research efforts have reported that an early and accurate diagnosis is of utmost importance in the field of medicine [1] in order to enhance the chance of survival in an effective way [2].

It has been widely accepted that applying machine learning techniques in breast cancer diagnosis can be beneficial. A large number of studies [3-9] have been conducted to gain a deep understanding of accurate breast cancer diagnosis based on the breast cancer datasets taken from University of California at Irvine (UCI) machine learning repository [10]. Akay [3] proposed a support vector machine (SVM) combined with F-score for breast cancer diagnosis. The SVM was a classifier which used F-score to evaluate the importance of features compared to the last obtained optimal feature subset. This method was improved to the accuracy level of 99.51%. In another study, Chen [4] proposed SVM with rough set based feature selection method for breast cancer diagnosis. The rough set was employed as a feature selection algorithm, and SVM was used for classification. This method improved the accuracy to 96.55%, 96.72%, and 96.87% in 50-50%, 70-30%, and 80-20% of partition, respectively. El-Baz [5] presented hybrid intelligent system-based rough set and ensemble classifier for breast cancer diagnosis, in which rough set theory was used for feature selection, and KNN was used as a classifier with the reported accuracy of 99.41%. The above-mentioned methods were evaluated on Wisconsin Diagnostic Breast Cancer (WDBC) dataset.

Zheng [6] introduced a hybrid of K-means and SVM algorithms for breast cancer diagnosis, where K-means algorithm referred to feature selection, and SVM was used as a classifier with the obtained accuracy of 97.38%. Pashaei [7] presented an improving medical diagnosis reliability using boosted C5.0 decision tree classifiers by particle swarm optimization method (PSO). The PSO was used for feature selection and boosted C5.0 was employed as a classifier, which achieved 96.38% accuracy. All these methods were evaluated on Breast Cancer Wisconsin (original, BCWO) dataset.

In a study conducted by Peng [8], an immune-inspired semi-supervised algorithm was proposed for breast cancer diagnosis. It achieved an accuracy of 98% and 98.3% on WDBC and BCWO datasets,

*Corresponding Author: Tel: +8613709262269; E-mail: chenli@nwu.edu.cn

respectively. Sheikhpour [9] also developed a particle swarm optimization for bandwidth determination and feature selection of kernel density estimation based on the classifiers in the diagnosis of breast cancer on WDBC and BCWO datasets. The obtained accuracies were 98.45% and 98.53%, respectively.

The obtained results of these methods reported an improvement in the accuracy of cancer diagnosis. However, the main disadvantage of these methods referred to the fact that training accuracy in these methods was used as the only criterion to evaluate their performance based on the assumption of balanced dataset. It should be noted that the balanced data assumption for medical diagnosis is frequently violated because the primary class of interest in medical diagnosis is usually the minority class.

Imbalanced classification problem in breast cancer diagnosis should be carefully addressed since the existing methods maximize the classification accuracy by correctly classifying the majority class, but misclassifying the minority class. Nonetheless, the minority class is usually the primary interest class. Therefore, the breast cancer diagnosis problem should be classified from the perspective of class imbalanced.

Generally speaking, several techniques based on sampling techniques, algorithm solutions, and cost sensitivity [10,11] have been used in the literature for classifying the imbalanced datasets. Sampling techniques operate on the data level by either undersampling or oversampling strategies to provide a balanced distribution. The cost-sensitive solutions assign different cost of misclassification errors for various classes. Algorithm solutions modify the existing algorithms for handling the imbalanced problem [12]. The performance of these methods depends heavily on parameter setting, especially for sampling rate and misclassification cost of classes, which play a crucial role in building a prediction mode with high generalization performance.

The KNN is an effective method for classification, although it is simple and non-parametric. In addition, KNN has been confirmed to be effective for unbalanced data classification [13, 14]. In a study conducted by Zhang et al. [13], KNN was confirmed to be an effective technique for imbalance learning classification problem. These findings were in line with the obtained results of the study conducted by Yudong et al. [14], where KNN was found to be an effective method compared to other well-known approaches. However, the major drawbacks of KNN were low efficiency, low noise tolerance, and high dependence on the value of k parameter. Therefore, it is better to propose an improved Laplacian score (LS) which can provide effective feature selection prior to classification for feature selection. In doing so, KNN algorithm with automatic k neighbors is proposed for classification. More specifically, a grid search technique using 10-fold cross-validation is used to find the optimal parameter of k.

Accordingly, the current research aimed to propose a KNN-LS model that hybridizes KNN and improved LS to alleviate the problems of the class imbalance in classification. The reason for employing the improved LS algorithm was that it reduced dimensionality and avoided the iterative training on different subsets since feature selection could play an important role in classification and good feature selection method could lead to high classification [4, 12, 15]. The KNN was utilized as the base classifier to automatically produce a diagnostic system. However, the major drawback with respect to KNN was its dependency on the selection of a “good value” for k, which was tackled by a grid search technique using 10-fold cross-validation. The effectiveness and performance of LS-KNN were evaluated on WDBC and BCWO datasets. The experimental results showed that the current approach could work well with breast cancer datasets and it could be a good alternative to the well-known machine learning methods.

Materials and Methods

The LS algorithm was first proposed by He et al., and the algorithm [15] based on the observation of the local geometric structure and originally applied in face recognition. In the current study, the researchers employed the same algorithm for feature selection. The KNN algorithm with automatic k parameter was used to improve the imbalanced breast cancer dataset classification performance.

Data description

The performance of the proposed method in this study was evaluated using two data sets WBCD and BCWO taken from the UCI machine learning repository. The BCWO dataset had 699 samples with16 instances of missing values. The few missing values were discarded from the dataset; therefore, the remaining 683 samples were used in the current experiment (444 benign and 239 malignant). Table 1 shows the attribution information of BCWO.

The WDBC dataset is another representative dataset for breast cancer. This dataset had 569 samples (i.e., 357 benign and 212 malignant). This dataset contained 32 features in 10 categories for each cell nucleus. Table 2 tabulates the mean value, standard error, and maximum values for each category.

Table 1. Descriptive statistics of breast cancer Wisconsin (origional)

number	Attribute	Minimum	Maximum	Mean	Standard deviation
1	Clump thicknesss	1	10	4.442	2.821
2	Uniformity of cell size	1	10	3.151	3.065
3	Uniformity of cell shape	1	10	3.215	2.989
4	Marginal	1	10	2.830	2.865
5	Single eptithelial cell size	1	10	3.234	2.223
6	Bare nuclei	1	10	3.545	3.644
7	Bland chromatin	1	10	3.445	2.450
8	Normal nucleoli	1	10	2.870	3.053
9	Mitoses	1	10	1.603	1.733

Table 2. Descriptive statistics of Wisconsin diagnostic breast cancer

Attribute	Attribute	Mean	Standard error	Maximum
1	Radius	28.11-6.98	2.873-0.112	36.04-7.93
2	Texture	39.28-9.71	4.89-0.36	49.54-12.02
3	Perimeter	188.50-43.79	21.98-0.76	251.20-50.41
4	Area	2501.0-143.50	542.20-6.80	4254.0-185.20
5	Smoothness	0.163-0.053	0.031-0.002	0.223-0.071
6	Compactness	0.345-0.019	0.135-0.002	1.058-0.027
7	Concavity	0.427-0.00	0.396-0.00	1.252-0.00
8	Concave points	0.201-0.00	0.053-0.00	0.291-0.00
9	Symmetry	0.304-0.106	0.079-0.008	0.664-0.157
10	Fractal dimension	0.097-0.05	0.030-0.001	0.208-0.055

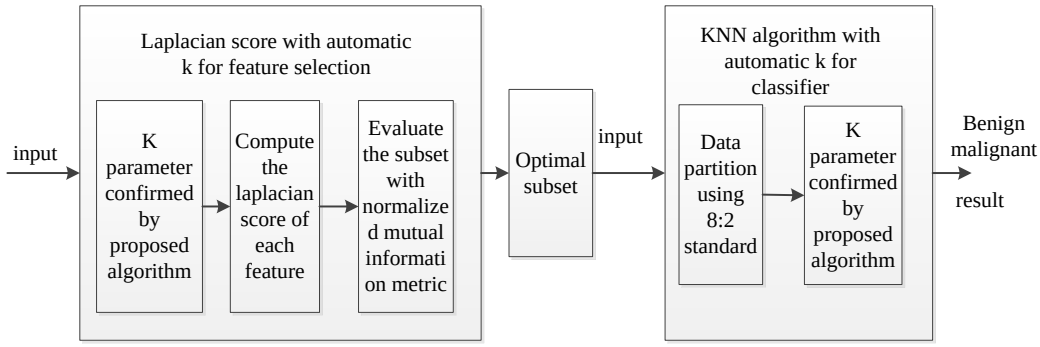

Figure 1. Block diagram of the proposed hybrid algorithm

LS-KNN algorithm

This section dealt with the presentation of the proposed KNN-based method combined with feature selection to overcome the class imbalance problem in breast cancer diagnosis. The feature selection method in this study was the improved LS algorithm, which could reduce dimensionality and avoid the iterative training on different subsets [4, 12, 15]. Figure 1 illustrates a block diagram of the proposed LS-KNN algorithm.

The block diagram included two major steps of preprocessing the feature selection and classification. In feature selection phase, the improved LS was used to extract the features of the breast tumor. In classification phase, KNN with automatic k was employed to classify the tumors. It is worth mentioning that the k in feature selection phase was the parameter to determine the neighbors when constructing a graph. However, in the classification phase, k was the parameter in KNN algorithm as a vote to classify the tumor. The LS-KNN hybrids improved LS and KNN algorithms to simultaneously determine an optimal subset of features and classify tumor.

Improved Laplacian score algorithm for feature selection

The LS is proposed by He [15] based on the theory of Laplacian Eigenmaps [16] and Locality Preserving Projection (LPP) [17]. This method is mainly used to show the ability of locality preserving power based on the assumption that two points are close if they are related to the same topic.

The feature selection process mainly consisted of four stages. First, the features were arranged based on LS in a descending order. Second, n feature subsets were constructed, the first feature subset was consisted of one feature which had the highest LS score, the $n-th$ feature subset was consisted of n features with top n LS scores. Third, the select feature subset were clustered by k-means algorithm. Fourth, the normalized mutual information ($\overline{MI}$) of each feature subset was computed by the labels of cluster algorithm and ground truth. The optimal subset was the feature subset with the highest $\overline{MI}$.

The improved LS was the nearest neighbor parameter in the first stage. This score could be obtained automatically when the nearest neighbor graph was constructed since in the original application it turned out to be a constant value of five. Due to the reasons that the constant value is not capable of delivering satisfactory performance for all the situation and the k value was so sensitive to the graph construct, there was a need for a test set algorithm which can automatically determine the k parameter. In the iteration process, KNN algorithm was utilized to train the processed dataset with different k, 10-fold cross-validation was employed to randomize sampling, and the k with the highest accuracy was considered as the optimal value. In this process, the value of k increased by two at a time, the k was an odd number within the range of 3-21.

In this algorithm, $\overline{MI}$ was used as a metric to evaluate the performance of feature subset [16-18] in the following formula. *Where*, C denotes the set of cluster obtained from the ground truth, C' signifies the set of cluster obtained from improved LS algorithm, $p(c)$ and $p(c')$ refer to the probabilities that data were selected from the clusters C and C', respectively. In addition, $p(c,c')$ is the joint probability that data were selected from both C and C' at the same time, $MI(C,C')$ denotes mutual information metric which is computed by equation 1, and $\overline{MI}(C,C')$ refers to normalized mutual information [18] which is computed by *Equation 2*.

$$MI(C,C') = \sum p(c,c') \log_2 \frac{p(c,c')}{p(c).p(c')} \quad (1)$$

$$\overline{MI}(C,C') = \frac{MI(C,C')}{\max(H(C),H(C'))} \quad (2)$$

Where, $H(C)$ and $H(C')$ are the entropies of clusters C and C', respectively. The value of $\overline{MI}(C,C')$ is within the range of 0-1. 1 means are identical, and 0 mean is totally different. Therefore, the higher the MI information, the better the feature subset.

KNN algorithm with optimal k for classification

In this paper, the KNN algorithm was used as the classifier, which was based on the theory of statistical learning and the principle of majority voting. Previous studies [13, 19] suggested KNN as one of the simplest classification methods, especially for distribution-unknow data. However, the major drawbacks with respect to KNN was the difficulty in determining k parameter since it was sensitive to the performance. Therefore, in this paper, the main challenge was to determine the optimal values of k. In doing so, a new algorithm was introduced to determine and confirm k.

A test set algorithm was proposed to determine the optimal k. The KNN algorithm with different k was employed to train the dataset, for which 10-fold cross-validation was used as sample method. The distance was computed by Euclidean distance. The performance of KNN was evaluated by Kappa, a commonly used statistics for evaluating model performance of unbalanced measures [20]. In this process, k was an odd number within the range of 3-21. The k with highest Kappa value was selected as the optimal parameter. Therefore, k parameter was adjustable to subsets with different sizes. It is worth mentioning that Kappa was more valuable to consider than accuracy in class imbalance learning.

Measures for performance evaluation

After the performance of different feature subsets were tested, the one with highest MI was selected as the optimal feature subset for classification. Next, the k parameter in KNN could be optimized by the algorithm. The performance of the proposed LS-KNN was compared with LS-SVM, LPP-SVM and (original) LS-KNN. The reason for this was that SVM was one of the most popular and widely implemented data mining algorithms in the domain of cancer diagnosis [21]. These results were reported in terms of accuracy, sensitivity, specificity, G-mean, and Matthews correlation coefficient (MCC). Large values of these criteria represented good classification performance.

As pointed out by Raeder [19], the choice of evaluation metric plays an important role in imbalanced learning. The G-mean is the geometric mean of the recalls of minority and majority classes. The aim of MCC is to measure the quantity of correlation between predictions and real target value [22]. Therefore, the consideration of MCC and G-mean are more important than accuracy since G-mean and MCC are the widely used overall performance measures in class imbalance learning [23, 24]. These measures are defined as follows:

$$accuracy = \frac{TP+TN}{TP+FP+FN+TN} \quad (3)$$

$$sensitivity = \frac{TP}{TP + FN} \quad (4)$$

$$specificity = \frac{TN}{FP + TN} \quad (5)$$

$$G - mean = \sqrt{\frac{TP}{TP + FN} \times \frac{TN}{TN + FP}} \quad (6)$$

$$MCC = \frac{TN \times TP - FN \times FP}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (7)$$

Where, *TP*, *TN*, *FP*, *FN* denotes the true positive, true negative, false positive, and false negative, respectively. Table 3 tabulates the values in a confusion matrix. In this study, non-cancer instances outnumbered the cancer instances. Therefore, non-cancer instances were considered "negative" and cancer instances were assigned "positive".

Table 3. Confusion matrix

	Predicted	Predicted
Actual	True positive	False negative
Actual	False positive	True negative

Results

In order to evaluate the effectiveness of the proposed LS-KNN algorithm for breast cancer diagnosis, two experiments were conducted on WDBC and BCWO datasets taken from UCI repository. The whole datasets were divided into two disjoint subsets with holdout method, namely 80% for training and 20% for testing in all the conducted experiments. The main purpose of the study was to investigate the ability of keeping the majority classification accuracy as well as the ability of improving the minority classification accuracy. In doing so, the experiment was performed on R platform with a Pentium CPU 2.19 GHz and 4 GB RAM, using R 3.3.3. The 'class' and 'caret' packages were used for KNN and 10-CV algorithms.

Experiment I

One of the data sources used in this experiment was BCWO dataset. The BCWO is the complete and representative dataset for testing breast cancer diagnosis model.

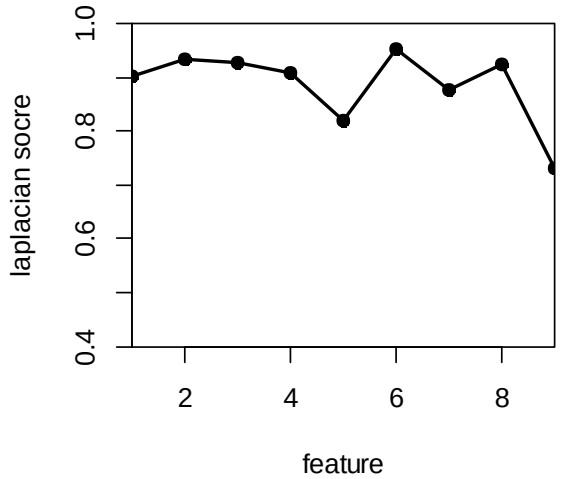

Figure 2. Laplacian score of the number

Figure 2 provides LS of different features using an improved LS algorithm. As can be observed from the results listed, the ordered features by Laplacian score were 6, 2, 3, 8, 4, 1, 7, 5, and 9. It should be pointed out that the horizontal axis indicates each feature (e.g., 2 in the horizontal axis indicates the second feature). Moreover, as Figure 3 illustrates MI information of different feature subsets need to be computed in order to justify the optimal feature subset of the dataset.

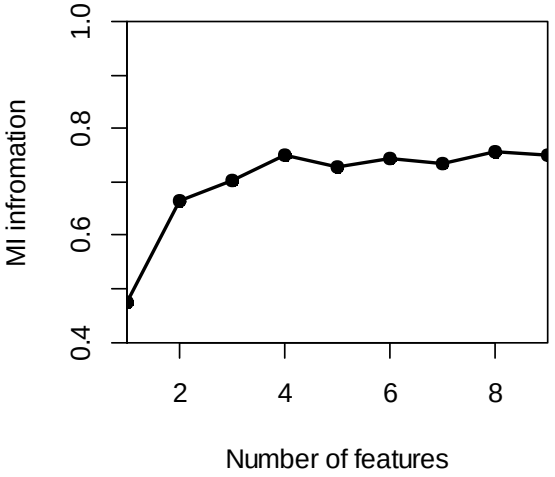

Figure 3. MI information using different subsets of features

In Figure 3, the horizontal axis shows the number of features, which have been arran1ged in a descending order according to their LS decrease (e.g., 2 in horizontal axis means the first two features). The vertical axis showed the MI information which has been explained in Section 2. The optimal subset was obtained when the number of feature in horizontal axis was equal to 4 because the vertical axis had reached the highest MI information at that time. The selected features were 6, 2, 3, and 8 for bare nuclei, cell size, cell shape, and normal nucleoli, respectively. Therefore, this subset could be used as the input of the KNN classifier. In KNN algorithm, the parameter of the k was calculated through a grid-search technique using 10-fold cross-validation sampling method. The optimal k parameter of KNN algorithm on BCWO dataset was k=5. The KNN algorithm with k=5 was utilized to classify the tumor. Table 4 presents the values of the confusion matrix by the proposed classification model.

The performance of LS-KNN on BCWO dataset was also compared with nine state-of-the-art methods form literature and three traditional methods. In order to evaluate the effectiveness of improved LS and KNN, the obtained results of LS-KNN were compared with Locality Preserving Projection-SVM (LPP-SVM), LS-SVM, and (original) LS-KNN. It should be pointed out that (original) LS-KNN refers to original LS for feature selection, and LS-KNN refers to the improved LS for feature selection. Table 5 reports the results of LS-KNN and different methods for BCWO dataset. The classification accuracy, sensitivity, G-mean, and MCC

are used as criteria for comparing the performances of these two methods. The symbol "~" in Table 5 indicates that the data are not derived from the literature. The performance of LS-KNN is italic.

Table 4. Confusion matrix for the K-nearest neighbor classifier on breast cancer Wisconsin (origional) dataset

Classifier	Predict result	Reference result	
		Benign	Malignant
KNN with k=5	Benign	102	0
	Malignant	1	35

According to Table 5, for BCWO dataset, 99.27% accuracy, 100% sensitivity, 99.02% specificity, 99.51% G-mean, and 98.13% MCC with 4 features was obtained by LS-KNN. The proposed LS-KNN outperformed the results reported by a number of studies [6, 8-9, 25-27] except the one conducted by El-Baz [5]. The comparison of the current algorithm with the algorithms in the studies conducted by Akay [3] and the one performed by Chen et al. [3, 4] revealed that the results were similar in term of accuracy; however, the proposed method had almost perfect sensitivity, specificity, and G-mean. Thus, the proposed LS-KNN method and the literature [3, 4] method performed similarly in breast cancer dataset. In addition, there was no significant difference between the performance level of the proposed LS-KNN and the one introduced by El-Baz [5].

As can be seen in Table 5, LS-KNN outperformed LPP-SVM and LS-SVM in terms of all measures. It should be noted that in this experiment, the LS-KNN was the same as the (original) LS-KNN since the optimal k in constructing the nearest neighbor graph turned out to be 5, which was similar to the default parameter.

Experiment II

The second experiment on WDBC dataset, another representative breast cancer dataset, was conducted with the purpose of demonstrating the robustness of LS-KNN algorithm in breast cancer prediction. Figure 4 illustrates the LS for each feature in WDBC dataset.

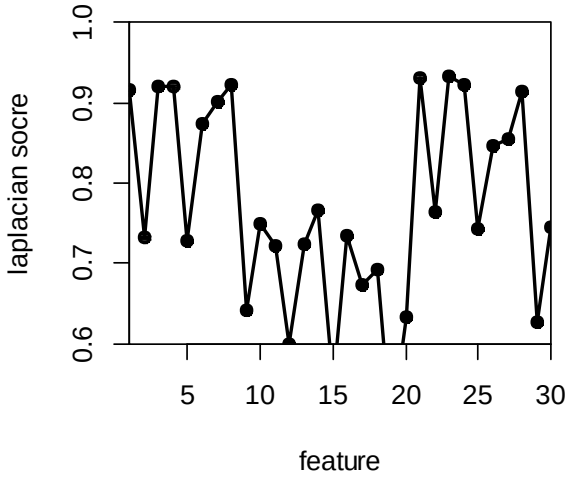

Figure 4. Laplacian score of the feature

As can be seen in Figure 4, the ordered features were 23, 21, 8, 24, 3, 4, 1, 28, 7, 6, 27, 26, 14, 22, 10, 30, 25, 16, 2, 5, 13, 11, 18, 17, 9, 20, 29, 12, 15, and 19. The horizontal axis indicated each feature (e.g., 5 in horizontal axis indicates the fifth feature). Figure 5 shows MI information of different feature subsets are computed in order to justify the optimal feature subset of dataset, and the result is shown in.

In Figure 5, the horizontal axis shows the number of features which are arranged based on LS score in a descending order (e.g., 10 in horizontal axis means the first ten features, which have been ordered based on LS). Vertical axis shows the MI information which has been explained in Section 2. The optimal feature subset was obtained when the number of feature in horizontal axis was equal to 14 since the vertical axis has reached the highest MI information. The optimal feature subset was 23, 21, 8, 24, 3, 4, 1, 28, 7, 6, 27, 26, 14, and 22.

Table 5. Comparison of metrics on breast cancer Wisconsin (origional) dataset

Algorithm	Accuracy	Sensitivity	Specificity	G-mean	MCC
F-score+SVM [3]	99.51%	~	~	~	~
PSO-RBF Kernel[4]	99.3%	~	~	~	~
PSO-SVM [25]	93.55%	~	~	~	~
AR1+AR2+NN [26]	98.4%	~	~	~	~
GA-MOO-ANN[27]	98.10%	~	~	~	~
K-Means, SVM [6]	97.38%	~	~	~	~
RS-KNN [5]	99.41%	100%	99.23%	99.61%	~
PSO-KDE [9]	98.53%	~	~	~	~
Aisl [8]	98.3%	94.3	99.6	~	~
(original)LS-KNN	99.27%	100%	99.02%	99.51%	98.13%
LPP-SVM	95.59%	97.27%	92.47%	90.01%	90.31%
LS-SVM	96.21%	96.87%	94.98%	91.8%	91.44%
LS-KNN	99.27%	100%	99.02%	99.51%	98.13%

SVM: support vector machine
PSO: particle swarm optimization
RBF: radial basis function kernel
AR: association rules
GA-MOO-ANN: genetic algorithm-based multi-objective optimization of an artificial neural network
RS: rough set
NN: neural network
KNN: nearest neighbor-K
KDE: kernel density estimation
LS: laplacian score
LPP:locality preserving projections

Table 6. Confusion matrix for K-nearest neighbor classifier on Wisconsin diagnostic breast cancer dataset

classifier	Predict result	Reference result	
		Benign	Malignant
KNN with k=5	Benign	86	0
	Malignant	1	26

Table 7. Comparison of metrics on Wisconsin diagnostic breast cancer dataset

Algorithm	Predicting accuracy	Sensitivity	Specificity	G-mean	MCC
QKCLDA [28]	97.4%	~	~	~	~
Filtered+Logistic regression [29]	96.62%	~	~	~	~
K-SVM [6]	97.4%	~	~	~	~
PSO(4-2) [30]	93.38%	~	~	~	~
PSO+Boosted C5.0[9]	96.38%	97.7%	94.28%	95.97%	~
Aisl [8]	98.0%	95.9%	98.7%	97.29%	~
BBHA-RF [22]	97.38%	95.79%	98.57%	97.17%	~
FSMLP [31]	100%	100%	100%	100%	~
(original) LS-KNN	98.23%	96.15%	98.85%	97.49%	95%
LPP-SVM	92.48%	90.46%	93.6%	84.93%	84.17%
LS-SVM	94.27%	91.83%	95.73%	87.64%	87.46%
LS-KNN	99.12%	100%	98.85%	99.42%	97.56%

QKCLDA: quasi-conformal kernel common locality discriminant analysis
PSO: particle swarm optimization
FSMLP: feature selection using multilayer perceptron
LPP:locality preserving projections
SVM: support vector machine
BBHA-RF: binary black hole algorithm-random forest
LS: laplacian score
nearest neighbor-K :KNN

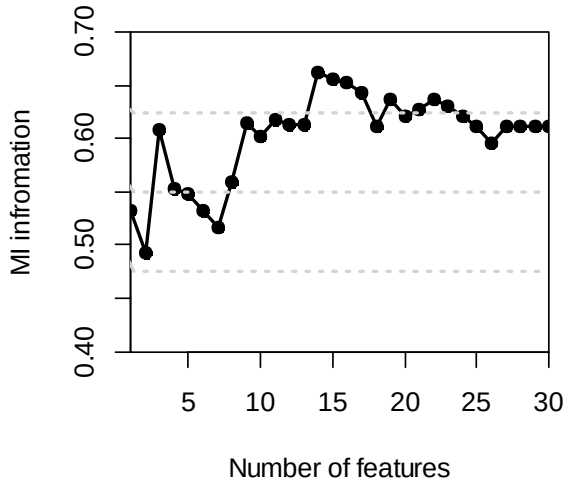

Figure 5. MI information using different subsets of features

The optimal subset was used as the input of the KNN classifier. In KNN algorithm, the parameter of k was calculated through a grid search technique using 10-fold cross-validation. On WDBC dataset when k=15, 17, 19, and 21, the Kappa values were all the same due to using 10-cv method for the sample which have averaged the results of multiple splitting. Therefore, k=19 was randomly selected in the current experiment. Table 6 presents the values of the confusion matrix by the proposed classification model. Table 7 shows the comparison of the performance level of the proposed algorithm in comparison with other predicting methods. The obtained results supports the fact that the proposed algorithm increases the prediction performance on WDBC dataset. The design of traditional comparison algorithm was the same as the one on BCWO dataset. The symbol "~" in Table 7 indicates that data is not derived from the literature. The performance of LS-KNN is italic.As can be seen in Table 7, LS-KNN obtains an accuracy, sensitivity, specificity, G-mean, and MCC of 99.12%, 100%, 98.85%, 99.42%, and 97.56% with 14 features. The LS-KNN method provides better performance than other approaches reported in the literature [6, 8, 9, 22, 28-30]. Moreover, LS-SVM outperformed LPP-SVM, LS-SVM and (original) LS-KNN. Regarding WDBC dataset, 100% classification accuracy for 80-20 scheme was obtained by feature selection using multilayer perceptron (FSMLP) method [31].

Discussion

Although investigation of the prediction methods for breast cancer is not a new endeavour, there is a scarcity of research in exploring the class imbalance nature of breast cancer dataset. In this regard, a hybrid method based on LS and KNN algorithm was proposed to reduce this negative effect. Compared to the methods mentioned in the literature, LS-KNN could find the optimal feature subset in a sensible computational cost, and provide better classification performance. Based on the experimental analysis, it can be concluded that firstly, the proposed LS-KNN method was a better hybrid classifier for the imbalanced dataset due to the obtained results of G-mean and MCC.

Secondly, in term of accuracy, the proposed LS-KNN could maintain a good classification accuracy of overall class data except for FSMLP [31] in WDBC dataset, as well as F-score+SVM [3], RSO-RBF [4] and RS-EKNN [5] in BCWO dataset, respectively. The obtained results in WDBC dataset were mainly based on the feature selection method. In practice, FSMLP algorithm was computationally inefficient because of exhaustive search. The FSMLP and the proposed LS-KNN with the difference in performance in BCWO

dataset was less than 1%, and their performance levels were not significantly different. The difference in classification accuracy is mainly due to sampling selection method. In addition, El-Baz [5] reported the use of a majority voting technique to combine the results of the individual classifiers in classification phase, which could significantly increase the complexity of the algorithm**.** However, this type of analysis was not conducted in the current research.

Thirdly, the performance of the proposed LS-KNN was stable compared to other methods in the literature in term of sensitivity of specificity indices. Specifically, the achieved MCC by LS-KNN was remarkable, which led to the conclusion that LS-KNN could significantly improve the classification accuracy for minority class, while keeping the classification of the majority high. Although the value of MCC was almost perfect, it was impossible to compare the MCC value with that of other methods since there was no access to the data of studies in the literature.

Finally, the comparison of the current method with LPP-SVM, LS-KNN, (original) LS-KNN, and the mentioned methods in a study by Pashaei and Aydin [22] revealed the superior performance of improved LS and KNN with automatic k. From the experimental results on datasets, it can be said that LS-KNN obtained good classification accuracy and selected fewer features.

Conclusion

Predicting breast cancer has been widely studied in the literature; however, few studies focus on the optimal feature subset and class-imbalance problem existing in breast cancer prediction. This paper aimed to propose a hybrid of LS and K-nearest neighbor algorithm to handle these problems. In this regard, the improved LS algorithm was utilized to obtain the optimal feature subset, and the KNN algorithm with automatic k was employed to classify the class-imbalance data. The use of LS-KNN was advantageous since feature subset can be obtained automatically and it rendered high performance for imbalanced dataset. This method was evaluated on two famous breast cancer datasets of UCI which led to satisfactory results regarding different performance measurements. More specifically, it had a good performance in sensitivity meaning that it had a good ability to handle the minority class. All these results lend support to the effectiveness of our algorithm.

Due to the importance of breast cancer diagnosis, a need is felt to further the study by focusing on the sample selection problem in class-imbalance learning. Moreover, it is suggested to solve the class-imbalance problem through finding some special samples which are not homogenous and have a big difference. In other words, the tumor can be classified by small number of samples. Furthermore, researchers can adopt some techniques to utilize the small number of labeled samples.

Acknowledgment

This project was supported by National Key Technology Science and Technique Support Program of China (Grant no. 2013BAH49F02) and Shaanxi Technology Committee Industrial Public Relation Project (no. 2018GY-146).

References

1. Mohammadpoor M, Shoeibi A, Shojaee H. A Hierarchical Classification Method for Breast Tumor Detection. Iranian Journal of Medical Physics. 2016;13(4):261-8.
2. Sahan S, Polat K, Kodaz H. A new hybrid method based on fuzzy-artificial immune system and k-nn algorithm for breast cancer diagnosis. Computers in Biology and Medicine. 2007;37(3): 415-23.
3. Mehmet Fatih Akay. Support vector machines combined with feature selection for breast cancer diagnosis. Expert Systems with Applications. 2009;36(2):3240-7.
4. Chen Hui-Ling, Yang Bo, Liu Jie, Liu Da-You. A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis. Expert Systems with Applications. 2011;38(7):9014-22.
5. El-Baz A H. Hybrid intelligent system-based rough set and ensemble classifier for breast cancer diagnosis. Neural Computing and Applications. 2015;26(2):437-46.
6. Bichen Zheng, Sang Won Yoon, Sarah S Lam. Breast cancer diagnosis based on feature extraction using a hybrid of k-means and support vector machine algorithms. Expert Systems with Applications. 2014;41(1):1476-82.
7. Pashaei E, Ozen M, Aydin N, editors. Improving medical diagnosis reliability using Boosted C5.0 decision tree empowered by Particle Swarm Optimization. Engineering in Medicine and Biology Society. 37th Annual International Conference of the IEEE. 2015.
8. Peng L, Chen W, Zhou W, Li F, Yang J, Zhang J. An immune-inspired semi-supervised algorithm for breast cancer diagnosis. Computer Methods and Programs in Biomedicine. 2016;134(C):259-65.
9. Sheikhpour R, Sarram M A, Sheikhpour R. Particle swarm optimization for bandwidth determination and feature selection of kernel density estimation based classifiers in diagnosis of breast cancer. Applied Soft Computing. 2016;40:113-31.
10. Bayan C, Fisher R. Classifying imbalanced data sets using similarity based hierarchical decomposition. Pattern Recognit. 2015;48:1653‑72.
11. Galar M, Fernandez A, Barrenechea E, Bustince H, Herrera F. A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. IEEE Trans. Syst. Man Cybern. ‑ Part C. 2012;42(4):463‑84.
12. Shirkavand A, Mohammadreza H. Detection of Melanoma Skin Cancer by Elastic Scattering Spectra: A Proposed Classification Method. Iranian Journal of Medical Physics. 2017;14(3):162-6.
13. Zhang J, Mani I, editors. kNN Approach to Unbalanced Data Distributions: A Case Study involving Information Extraction. workshop on

Learning from imbalanced Datasets. in Proceedings of the International Conference on Machine Learning. 2003: AAAI Press; 2003: 42-8.
14. Zhang Y, Lu S, Zhou X, Yang M, Wu L, Liu B. Comparison of machine learning methods for stationary wavelet entropy-based multiple sclerosis detection: decision tree, k-nearest neighbors, and support vector machine. Imulation. 2016;92(9):861-71.
15. He X, Cai D, Niyogi P. Laplacian score for feature selection. Advances in neural information processing systems. Neural Inofrmation Processing Systems. 2006.
16. Belkin M, Niyogi P. Laplacian eigenmaps and spectral techniques for embedding and clustering. Advances in Neural Information Processing Systems. 2009;14(6):585-91.
17. Han J. Data mining: concepts and techniques. San Francisco: Morgan Kaufmann Publishers Inc. 2005.
18. Kohavi R, John G H. Wrappers for feature subset selection. Artificial Intelligence. 1997;91(1-2):273-324.
19. Raeder T, Forman G, Chawla N V. Learning from Imbalanced Data: Evaluation matters. In: Dawn E. Holmes, Lakhmi C. Jain. Data Mining: Foundations and Intelligent Paradigms. Berlin Heidelberg: Springer Berlin Heidelberg. 2012:315-31.
20. Dehghani-Bidgoli Z, Baygi MHM, Kabir E, Malekfar R. Common Raman Spectral Markers among Different Tissues for Cancer Detection. Iranian Journal of Medical Physics. 2014;11(4):308-15.
21. Dastjerdi MV, Zadeh ZD, Mousavi SJ, Askari HR, Soltanolkotabi M. Hair analysis by means of laser induced breakdown spectroscopy technique and support vector machine model for diagnosing addiction. Iranian Journal of Physics Research. 2018;17(5):661-7.
22. Pashaei E, Aydin N. Binary black hole algorithm for feature selection and classification on biological data. Applied Soft Computing. 2017; 56: 94-106.
23. Maarten van Someren, Gerhard Widmer. Learning When Negative Examples Abound: 1997: 9th European Conference on Machine Learning Prague; 1998 April 23-25; Berlin, Germany. 1997:1224:146-53.
24. Kubat M, Matwin S. Addressing the Curse of Imbalanced Training Sets: One-Sided Selection. In Proceedings of the Fourteenth International Conference on Machine Learning. 1997; 179-86.
25. Das SR, Panigrahi PK, Das K, Mishra D. Improving RBF Kernel Function of Support Vector Machine using Particle Swarm Optimization. International Journal of Advanced Computer Research. 2012;2(7):130-5.
26. Palaniappan S, Pushparaj T. A novel prediction on breast cancer from the basis of association rules and neural network. International Journal of Computer Science and Mobile Computing. 2013;2(4):269-77.
27. Ahmad F, Isa NA, Hussain Z, Sulaiman SN. A genetic algorithm-based multi-objective optimization of an artificial neural network classifier for breast cancer diagnosis. Neural Computing and Applications. 2013;23(5):1427-35.
28. Li J-B, Peng Y, Liu D. Quasiconformal kernel common locality discriminant analysis with application to breast cancer diagnosis. Information Sciences. 2013;232(2):256-69.
29. Bamakan S M H, Gholami P. A Novel Feature Selection Method based on an Integrated Data Envelopment Analysis and Entropy Model. Procedia Computer Science. 2014;31:632-8.
30. Xue B, Zhang M, Browne W N. Particle swarm optimisation for feature selection in classification: Novel initialisation and updating mechanisms. Applied Soft Computing. 2014;18(4):261-76.
31. Sridevi T, Murugan A. A Novel Feature Selection Method for Effective Breast Cancer Diagnosis and Prognosis. International Journal of Computer Applications. 2014;88(11):28-33.