J Med Syst (2015) 39: 40
DOI 10.1007/s10916-015-0210-x

NON-INVASIVE DIAGNOSTIC SYSTEMS

Applying Under-Sampling Techniques and Cost-Sensitive Learning Methods on Risk Assessment of Breast Cancer

Jia-Lien Hsu · Ping-Cheng Hung · Hung-Yen Lin · Chung-Ho Hsieh

Received: 21 November 2014 / Accepted: 10 December 2014 / Published online: 25 February 2015

Abstract Breast cancer is one of the most common cause of cancer mortality. Early detection through mammography screening could significantly reduce mortality from breast cancer. However, most of screening methods may consume large amount of resources. We propose a computational model, which is solely based on personal health information, for breast cancer risk assessment. Our model can be served as a pre-screening program in the low-cost setting. In our study, the data set, consisting of 3976 records, is collected from Taipei City Hospital starting from 2008.1.1 to 2008.12.31. Based on the dataset, we first apply the sampling techniques and dimension reduction method to preprocess the testing data. Then, we construct various kinds of classifiers (including *basic classifiers*, *ensemble methods*, and *cost-sensitive methods*) to predict the risk. The cost-sensitive method with random forest classifier is able to achieve recall (or sensitivity) as 100 %. At the recall of 100 %, the precision (positive predictive value, PPV), and specificity of cost-sensitive method with random forest classifier was 2.9 % and 14.87 %, respectively. In our study, we build a breast cancer risk assessment model by using the data mining techniques. Our model has the potential to be served as an assisting tool in the breast cancer screening.

Keywords Breast cancer · Cost-sensitive learning · Sampling

This article is part of the Topical Collection on *Systems-Level Quality Improvement*

J.-L. Hsu (✉) · P.-C. Hung · H.-Y. Lin
Department of Computer Science and Information Engineering, Fu Jen Catholic University, New Tapei City, Taiwan, Republic of China
e-mail: alien@csie.fju.edu.tw

C.-H. Hsieh (✉)
Department of General Surgery, Shin Kong Wu Ho-Su Memorial Hospital, Taipei, Taiwan, Republic of China
e-mail: M012363@ms.skh.org.tw

Introduction

According to the report of World Health Statistics in 2013, cancer is a leading cause of death [1]. In United States, the breast cancer is the first woman cancer incidence and the second caner mortality [1]. In Taiwan, according to the "Statistics Report of Bureau of Health Promotion, Department of Health 2012", there were more than nine thousands women suffering from breast cancer, and one thousand and eight hundreds women of breast cancer death. Also in Taiwan, the breast cancer is the first woman cancer incidence and the forth caner mortality. The cancer statistics reveals that the breast cancer is one of the most serious threat to women's health.

Referring to Table 1, there are three most common screening methods of breath cancer prediction, including mammography, ultrasound, and MRI. These screening methods may reduce breast cancer mortality and increase breast cancer survival rate. In Taiwan, the BHP (Bureau of Health Promotion) provides a bi-annual mammography in women aged 45-69. In addition, there is an evidence that early detection through mammography screening and adequate follow-up of women could significantly reduce mortality from breast cancer [2, 3]. However, these screening methods demand a considerable cost. The mammography screening may not be cost-effective. In the meanwhile, the over-diagnosis of screening mammography to detect breast cancer has been reported [4]. Bleyer and Welch estimated

Table 1 Screening methods for breast cancer [5–7]

Method	SPC [5]	SEN [5]	SEN [6]	SPC [7]	SEN [7]
Mammography	95.5 %	50.0 %	25 – 59 %		20 – 50 %
Ultrasound	91.8 %	50.0 %			
Mammography plus Ultrasound	89.4 %	77.5 %	49 – 67 %		
MRI			93 – 100 %	37 – 97 %	71 – 100 %

Note: Specificity (SPC); Sensitivity (SEN)

breast cancer was overdiagnosed in 1.3 million U.S. women in the past 30 years. In 2008, the researches estimated breast cancer was overdiagnosed in more than 70,000 women, and this accounted for 31 % of breast cancers diagnosed.

In this paper, we would like to propose a computational model to evaluate the risk of breast cancer which is only based on patient questionnaire information. In prior to mammography screening, our computational method can be served as a pre-diagnosis program in the low-cost setting.

We make use of Weka (Waikato Environment for Knowledge Analysis, a collection of machine learning algorithms for data mining tasks) to build a computational predict model for breast cancer risk assessment. There are various kinds of classification methods implemented, in which we conclude these methods into three categories: "*basic classifier*", "*ensemble method*", and "*cost-sensitive method*".

In the first category of "*basic classifier*", we choose the J48 (Trees), LMT (Trees), NaïveBayes (Bayes), LibSVM (Functions), IBk (Lazy), RBFNetwork (Functions) described as follows.

- J48 classifier: The J48 classifier is using the C4.5 algorithm to generate a decision tree for prediction. Based on the concept of information entropy, a tree-based model is constructed in which the easily-interpreted model may reach a reasonable precision.
- LMT (logistic model tree) classifier: The LMT classifier is a classification model, which combines decision tree and logistic regression learning.
- Naïve Bayes (NB) classifier: The Naïve Bayes classifier is based on Bayes' theorem of probabilistic statistical classifier. Usually, the Naïve Bayes classifier is robust to isolated noise points and irrelevant attributes which following the statistical principle for combining prior knowledge of the classes gathered from data.
- LibSVM (support vector machines, SVM) classifier: The LibSVM classifier constructs a hyperplane to separate the different classes of data. The LibSVM classifier maximize the margin around the separate hyperplane.
- IBk classifier: The IBk classifier is the k nearest-neighbor algorithm. The k nearest-neighbor algorithm is a type of lazy learning and instance-based learning. The IBk classifier has an advantage of constructing arbitrary-surface boundaries. The IBk classifier is also applicable for data in high variance distribution.
- RBFNetwork (RBF) classifier: The RBFNetwork classifier is an instance-based learning method which implements a normalized Gaussian radial basis function network to predict.

Table 2 The cost matrix for cost-sensitive methods

		Predicted Class	
		Class = +	Class = −
Actual Class	Class = +	*CostTP*	*CostFN*
	Class = −	*CostFP*	*CostTN*

In the second category of "*ensemble method*", we choose VOTE (Meta), AdaBoostM1 (Meta), Bagging (Meta), Stacking (Meta), RandomForest (Trees) described as follows. The "*ensemble method*" make use of multiple "basic classifiers" to obtain better predict performance than that could be obtained from any of the constituent classifiers. In other words, an ensemble is a technique for combining many *weak* classifiers in an attempt to produce a *strong* classifiers.

- VOTE classifiers: The VOTE classifier is a common theoretical framework for combining classifiers which use distinct pattern representations to accomplish a compound classification where all the pattern representations are used jointly to make a decision [8].
- AdaBoostM1 classifier: In the beginning, the AdaBoostM1 classifier assigns weight to each training instances. Then, the AdaBoostM1 works by repeatedly running a given weak learning algorithm on various distributions over the training data, and then combining the classifiers produced by the weak learner into a single composite classifier. The output of the

Table 3 The attributes of BIRADS data set

i-th	Attributes	High risk	Low risk	Overall
	N (%)	76 (2.50 %)	2959 (97.50 %)	3035 (100 %)
NO.1	age	46.78 (9.63)	41.83 (7.28)	41.96 (7.39)
NO.2	body height (cm)	158.63 (5.00)	158.78 (5.08)	158.79 (5.08)
NO.3	body weight (kg)	56.90 (8.40)	56.76 (8.71)	56.76 (8.70)
NO.4	1st degree relatives cancer	0.08 (0.27)	0.03 (0.17)	0.03 (0.17)
NO.5	2nd degree relatives cancer	0.04 (0.20)	0.03 (0.19)	0.03 (0.19)
NO.6	3rd degree relatives cancer	0.07 (0.38)	0.04 (0.20)	0.04 (0.20)
NO.7	menstration period (days)	30.42 (5.74)	27.65(6.03)	27.72 (6.04)
NO.8	is mens regular			
	True	58 (76.32 %)	2160 (73.00 %)	2218 (73.08 %)
	False	18 (23.68 %)	799 (27.00 %)	817 (26.92 %)
NO.9	contraceptive year	0.34 (1.50)	0.24 (1.31)	0.25 (1.31)
NO.10	hormone year	0.36 (1.42)	0.08 (0.76)	0.08 (0.79)
NO.11	parturition times	1.68 (1.09)	1.75 (1.59)	1.75 (1.58)
NO.12	breast feeding times	0.59 (1.00)	0.82 (0.96)	0.82 (0.96)

Data are presented as means (SD) or numbers (%)

weak learners is combined into a weighted sum that represents the final output of the boosted classifier.

It reduces the bias of the weak learner by forcing the weak learner to concentrate on different parts of the instance space, and it also reduces the variance of the weak learner by averaging several hypotheses that were generated from different subsamples of the training set.

- Bagging classifier: The Bagging classifier is a special case of the model averaging approach. The Bagging classifier randomly create subsets of original data, then aggregate each subsets predictions to determine a final prediction.
- Stacking classifier: The Stacking classifier is a different way of combining models, in which the Stacking

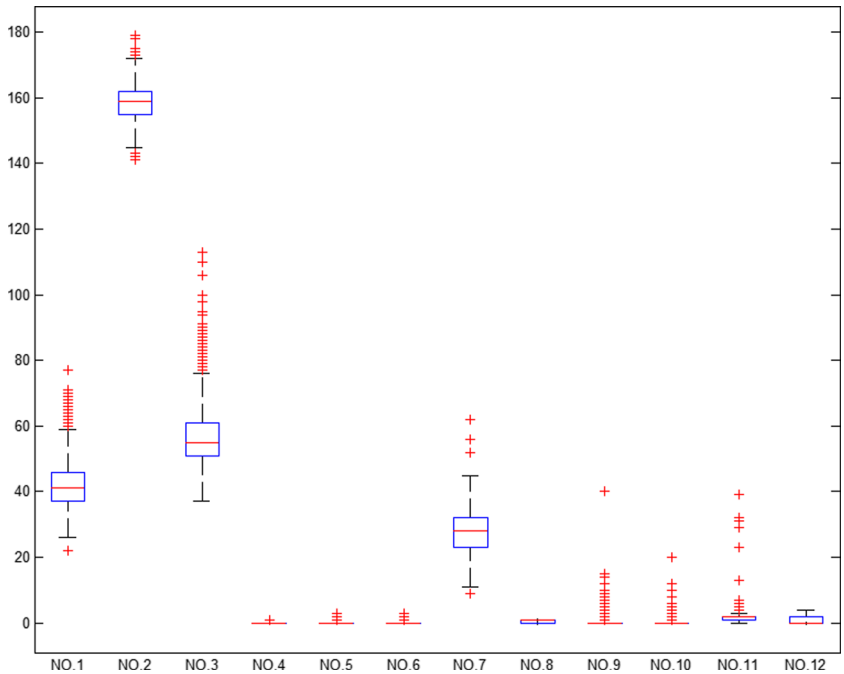

Fig. 1 The boxplot of numeric attributes in BIRADS data set (The 'NO.*i*' denotes the *i*-th attribute in Table 3)

classifier works by deducing the biases of the generalizers with respect to a provided learning set [9].

- Random Forest (RF) classifier: The random forest classifier combines a multiple of decision tree. Each decision tree are independent predictions, the largest number votes for the final result class.

Regarding the third category of "cost-sensitive methods" [10, 11], the cost-sensitive classification aims to reach a minimal cost class results on a class imbalance dataset. When applying the cost-sensitive method, we have to preset the cost matrix shown in Table 2. The cost-sensitive classifier is attempting to find prediction the class with

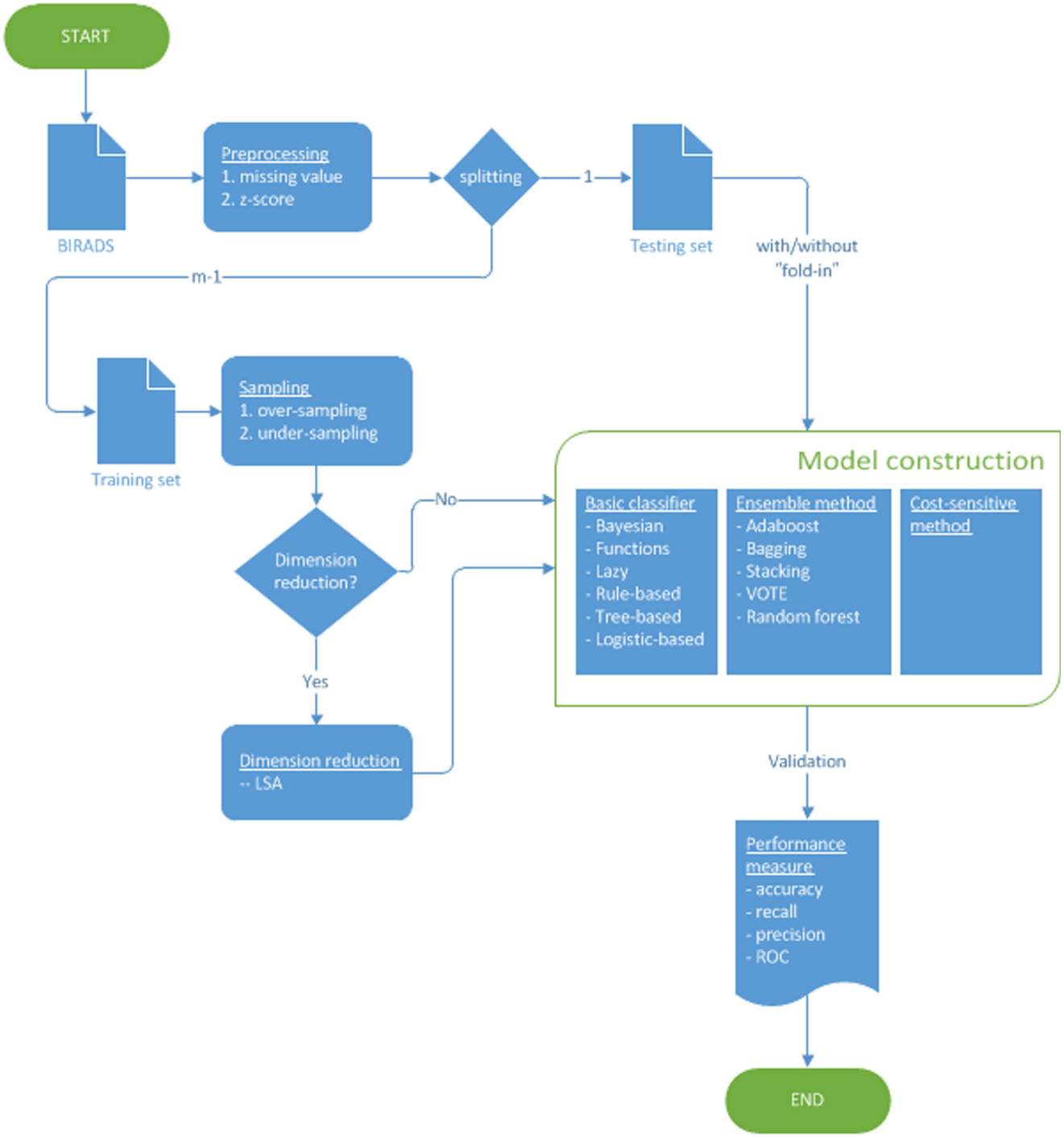

Fig. 2 Process flow of performing experiments

minimum misclassification cost, by re-weighting training data instances according cost matrix assigned to each class. In our study, we will set a higher cost of *false negative* (*FN*) case, since the *FN* misjudgment might be serious and could result in delay seeking medical treatment for possible patients.

Because the class imbalance nature of data source, in this project, we first apply sampling approaches to obtain the entire set of data of interest and to improve the detection of rare cases.

In addition, since unequal penalty of making decision (including *TP*, *FP*, *FN*, and *TN*), we would like to build a computation model to predict patient of high risk with almost 100 % recall and reasonable high precision. An alternative metric, which will be detailed in section "Materials and methods", is also introduced to describe the performance of classification models.

Materials and methods

In this section, we will first introduce our computing platform and describe our dataset. Then, we present the approach, as well as the performance measure.

The BIRADS data is collected from Taipei City Hospital, starting from 2008.01.01 to 2008.12.31. We enrolled women who received breast cancer screening program with sonography in Taipei City. Data were collected by filling a questionnaire before examination. The assessment category of sonography are used to determine the target attribute as high risk or low risk in the "Breast Imaging-Reporting and Data System". There are 3,976 records in our BIRADS dataset in which only 94 records are *true* (High risk). Since some missing values in this data set, we exclude those incomplete records, and there are 3,035 records left [12]. In reference to Table 3, there are thirteen attributes and the "High risk" attribute is the target to be predicted. The characteristic of BIRADS data set is illustrated in Fig. 1.

To estimate the performance accuracy of computational model, we use m-fold cross-validation in which the original dataset is randomly divided into m equal size partitions. Of the m partitions, a single partition is considered as the validation data for testing the model, and the remaining $(m-1)$ partitions are used as the training data. In the experiments, we repeated the cross-validation process m times, and report the average results.

In reference to Fig. 2, we introduce the approach of our study and illustrate the process flow of performing experiments.

Given the BIRADS data set, in the preprocessing step, those records containing missing values are excluded first.

Table 4 The confusion matrix

		Predicted Class	
		Class = +	Class = −
Actual Class	Class = +	TP	FN
	Class = −	FP	TN

Then, all twelve attributes are normalized and replaced by *z-scores*.

In the next step, we apply the *stratified splitting procedure* in which 67 % records are the *training set* of containing both "high risk" and "low risk" records. Similarly, in the *testing set*, there are 33 % records of containing both "high risk" and "low risk" records.

Since the data set is of imbalanced classes, we only have limited "high risk" records (76 among 2959). We apply sampling techniques in priori to the model construction. In our study, we apply the *under-sampling technique* and the *over-sampling technique* for training data, respectively.

In the under-sampling technique, the data set of majority class will be shrunk. In the over-sampling technique, the data set of minor class will be expanded. After the sampling technique, the size of positive and negative classes are comparable, and the class boundary could be more clear.

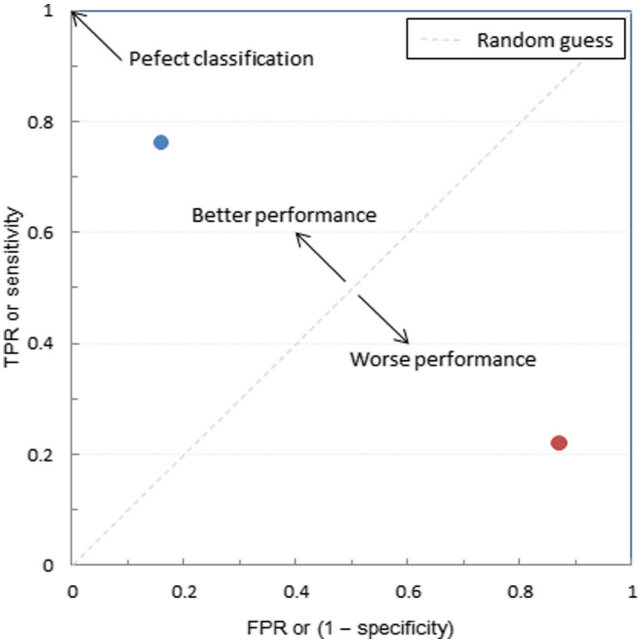

Fig. 3 An illustration of ROC space [15]

Table 5 Performance vs. sampling technique (under-sampling; training data size: five percent of input data set)

Evaluation	NB	SVM	IBk	LMT	J48	RBF	RF
Accuracy	69.26 %	45.77 %	54.07 %	57.46 %	56.08 %	75.62 %	49.39 %
Recall/TPR/Sensitivity	0.395	0.658	0.632	0.579	0.553	0.329	0.658
Specificity/1−FPR	0.700	0.453	0.538	0.575	0.561	0.767	0.490
Precision	0.033	0.030	0.034	0.034	0.031	0.035	0.032
AUC	0.597	0.557	0.585	0.576	0.574	0.554	0.597

Table 6 Performance vs. sampling technique (over-sampling; training data size: ten percent of input data set)

Evaluation	NB	SVM	IBk	LMT	J48	RBF	RF
Accuracy	68.96 %	87.91 %	58.29 %	58.39 %	54.33 %	75.55 %	54.07 %
Recall/TPR/Sensitivity	0.408	0.066	0.553	0.632	0.671	0.342	0.605
Specificity/1−FPR	0.697	0.900	0.584	0.583	0.540	0.766	0.539
Precision	0.033	0.017	0.033	0.037	0.036	0.036	0.033
AUC	0.599	0.483	0.572	0.616	0.622	0.553	0.608

Table 7 Performance vs. ensemble method (AdaBoost with various '*base*' classifiers; training data size: five percent of input data set)

Evaluation	NB	SVM	IBk	LMT	J48	RBF	RF
Accuracy	59.11 %	42.47 %	54.07 %	56.67 %	55.85 %	61.55 %	53.64 %
Recall/TPR/Sensitivity	0.487	0.658	0.632	0.513	0.632	0.513	0.697
Specificity/1−FPR	0.594	0.419	0.538	0.568	0.557	0.618	0.532
Precision	0.030	0.028	0.034	0.030	0.035	0.033	0.037
AUC	0.571	0.543	0.585	0.583	0.613	0.569	0.655

Table 8 Performance vs. ensemble method (Bagging with various '*base*' classifiers; training data size: five percent of input data set)

Evaluation	NB	SVM	IBk	LMT	J48	RBF	RF
Accuracy	65.63 %	47.94 %	54.60 %	50.51 %	52.59 %	69.29 %	55.72 %
Recall/TPR/Sensitivity	0.487	0.618	0.618	0.671	0.592	0.421	0.592
Specificity/1−FPR	0.661	0.476	0.544	0.501	0.524	0.700	0.565
Precision	0.036	0.029	0.034	0.033	0.031	0.035	0.033
AUC	0.597	0.515	0.615	0.629	0.601	0.572	0.599

Table 9 Performance vs. ensemble method (Stacking with various '*base*' classifiers; training data size: five percent of input data set)

Base classifiers	Recall/ TPR/ Sensitivity	Specificity/ 1−FPR	Precision	AUC
Three classifiers				
J48 + NB + IBk	0.857	0.295	0.031	0.614
IBk + LMT + SVM	0.597	0.404	0.025	0.501
LMT + RBF + SVM	0.545	0.550	0.031	0.563
Four classifiers				
LMT + SVM + NB + J48	0.714	0.431	0.032	0.611
LMT + SVM + IBk + RBF	0.519	0.573	0.031	0.551
IBk + RBF + NB + J48	0.935	0.238	0.031	0.602
Five classifiers				
NB + SVM + IBk + LMT + J48	0.649	0.428	0.029	0.568
SVM + IBk + LMT + J48 + RF	0.513	0.438	0.023	0.612
IBk + LMT + J48 + RF + RBF	0.461	0.536	0.025	0.478

Table 10 Performance vs. ensemble method (VOTE with various '*base*' classifiers; training data size: five percent of input data set)

Base classifiers	Recall/ TPR/ Sensitivity	Specificity/ 1−FPR	Precision	AUC
Three classifiers				
J48 + NB + IBk	0.671	0.539	0.036	0.635
IBk + LMT + SVM	0.697	0.543	0.038	0.623
LMT + RBF + SVM	0.684	0.447	0.031	0.595
Four classifiers				
LMT + SVM + NB + J48	0.553	0.562	0.031	0.622
LMT + SVM + IBk + RBF	0.632	0.562	0.036	0.619
IBk + RBF + NB + J48	0.618	0.564	0.035	0.641
Five classifiers				
NB + SVM + IBk + LMT + J48	0.632	0.572	0.037	0.639
SVM + IBk + LMT + J48 + RF	0.697	0.542	0.038	0.636
IBk + LMT + J48 + RF + RBF	0.618	0.560	0.035	0.634

In our study, we also make use of the dimension reduction technique to further reduce the data size, but the data characteristic still maintains. We consider the LSA for the dimension reduction [13, 14].

The LSA (*Latent Semantic Analysis*) is a document processing technique originated in the field of information retrieval, in which we apply a series of operations from linear algebra, known as *matrix decomposition*, to construct a *low-rank* approximation to the *term-document matrix*. Some typical applications of low-rank approximation is to index and retrieve documents, as well as to cluster documents.

Given an m by n term-document matrix $\mathbf{A}$, a SVD (*singular value decompostion*) of $\mathbf{A}$ can be written as

$$\mathbf{A} = \sum_{i=1}^{rank(\mathbf{A})} \sigma_i \mathbf{u}_i \mathbf{v}_i^T = \mathbf{U}\Sigma\mathbf{V}^T \quad (1)$$

where σ_i is the i^{th} singular value of $\mathbf{A}$.

Usually, we choose the first k singular values to obtain the rank k approximation, as follows.

$$\mathbf{A}'_k = \mathbf{U}'_k \Sigma'_k \mathbf{V}'^T_k \quad (2)$$

The low-rank approximation matrix $\mathbf{A}'$ yields a new representation for each document in the collection, which is expected to combine and merge the dimensions associated with terms that have similar meanings.

As a result, the original records represented as vectors in the twelve dimensional space can be reduced as the corresponding vectors in the k dimensional space. The k dimension axes are also considered as the k "*concepts*".

Note that if we apply LSA on the training set, we have to apply the *folding-in* process on the testing set to cast the records into low-rank representation for the further processing, such that the dimensionality of testing set match that of training set.

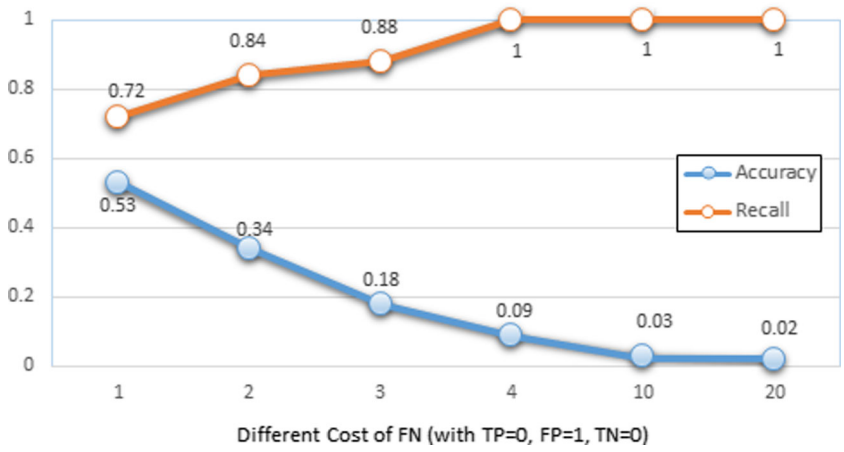

Fig. 4 Cost-sensitive method (RF classifier) use different cost setting of false negative case

Table 11 Performance vs. cost-sensitive methods (various '*base*' classifiers; training data size: five percent of input data set)

Evaluation	NB	SVM	IBk	LMT	J48	RBF	RF
Accuracy	9.98 %	11.70 %	28.63 %	16.14 %	14.66 %	7.28 %	17.00 %
Recall/TPR/Sensitivity	1	1	0.921	1	0.987	1	1
Specificity/1−FPR	0.077	0.094	0.270	0.140	0.125	0.049	0.149
Precision	0.027	0.028	0.031	0.029	0.028	0.026	0.029
AUC	0.620	0.547	0.606	0.606	0.579	0.581	0.593
Cost (TP,FN)	−1,17	0,1	0,9	0,4	−1,9	0,4	−2,16
Cost (FP,TN)	1,0	1,0	1,0	1,0	1,0	1,0	1,−1

For each record $\vec{q}$ in the testing set, the folding-in process is

$$\vec{q}_k = \Sigma_k^{-1} \mathbf{U}_k^T \vec{q} \tag{3}$$

Regarding the model construction for classification, in our study, we apply "*basic classifier*", "*ensemble method*" and "*cost-sensitive method*" in search of the best practice of risk assessment. The three kinds of approaches have been described in section "Introduction".

In the last step of validation, the classification performance is measured by the common metrics presented as follows.

Data with imbalanced class distribution are common in some of real and medical applications. Only the *accuracy* measure is not suitable for evaluating classification model derived from imbalanced data set. In this study, the *baseline* of classification could be as high as 97.64 %. Without careful consideration, trivial applying of existing classification algorithms may not effectively detect instances of the rare class. That is, all instances are predicted to be *low risk*. As a result, those derived models become useless even though the accuracy is high enough.

In this subsection, we introduce some performance metrics, including *accuracy*, *recall*, *precision*, and *ROC*. In reference to the confusion matrix shown in Table 4, the metrics are convenient ways of comparing classifiers and defined as follows.

Accuracy A correct classifier mean predicts the same class as the original class of the test data. The accuracy of prediction system is the degree of closeness between the predicted class and actual class.

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \tag{4}$$

Recall and precision Recall is the fraction of relevant instances that are retrieved, while precision the fraction of retrieved instances that are relevant. Precision can be thought of as a measure of exactness, whereas recall is a measure of completeness.

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

ROC The ROC (Receiver Operating Characteristics) of a classifier shows its performance as a relative trade-off

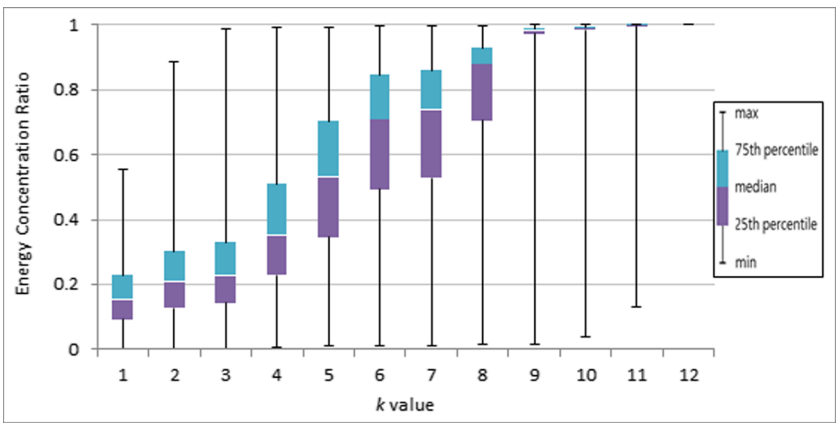

Fig. 5 The effectiveness of applying LSA on BIRADS for various k values

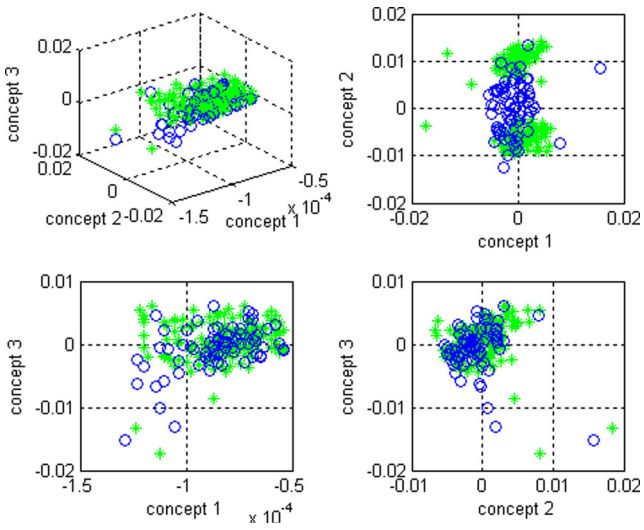

Fig. 6 The visualization of BIRADS data by applying LSA, we only show the first three concepts (i.e., y_1, y_2, y_3); $k = 7$; The 'circle' denotes "high risk"; the 'plus' denoted "low risk")

between sensitivity (true positive rate) and specificity (one minus the false positive rate).

$$TPR = \frac{TP}{TP + FN} \tag{7}$$

$$FPR = \frac{FP}{FP + TN} \tag{8}$$

The *TPR* (true positive rate), which can be interpreted as *benefits*, defines how many correct positive results occur among all positive samples available during the test. On the other hand, *FPR* (false positive rate), which can be interpreted as *cost*, defines how many incorrect positive results occur among all negative samples available during the test.

As shown in Fig. 3, a ROC space is defined by FPR and TPR as the x axis and the y axis, respectively. A single point (i.e., the pair of values illustrated in the ROC space) indicates a prediction result of a classifier.

The best possible prediction method would yield a point in the upper left corner or coordinate (0, 1) of the ROC space, representing 100 % sensitivity (no false negatives) and 100 % specificity (no false positives). The (0, 1) point is also called a perfect classification. A classifier of completely random guess would give a point which lies along a diagonal line from the left bottom (0, 0) to the top right (1, 1) corners.

In addition, the area under ROC curve, denoted by AUC, is a single index for measuring the performance of classifier. The larger the AUC, the better is overall performance of classifier.

Results

Some key findings are summarized as follows.

Sampling technique (over-sampling vs. under-sampling) − In reference to Tables 5 and 6, the average recall of classifiers by using under-sampling is 54.3 %, while the average recall by using over-sampling is 46.8 %. In general, we conclude that the performance of under-sampling is better than that of over-sampling. Note that, if not applying sampling techniques, prediction would tend to majority instances, and the high-accuracy classifiers still become meaningless.

Ensemble method − We conclude that, in general, the performance of ensemble method is better than that of basic and single classifiers. For some ensemble classifiers, such as Stacking, the high recall meets our goal of this study (in reference to Table 9). The performance is summarized in Tables 7, 8, 9 and 10.

Cost-sensitive classifier − The higher cost (high value) of FN implies the high penalty of making wrong decisions. The performance of cost-sensitive methods for various assignments is illustrated in Fig. 4. With respect to various base classifiers, the performance is summarized in Table 11.

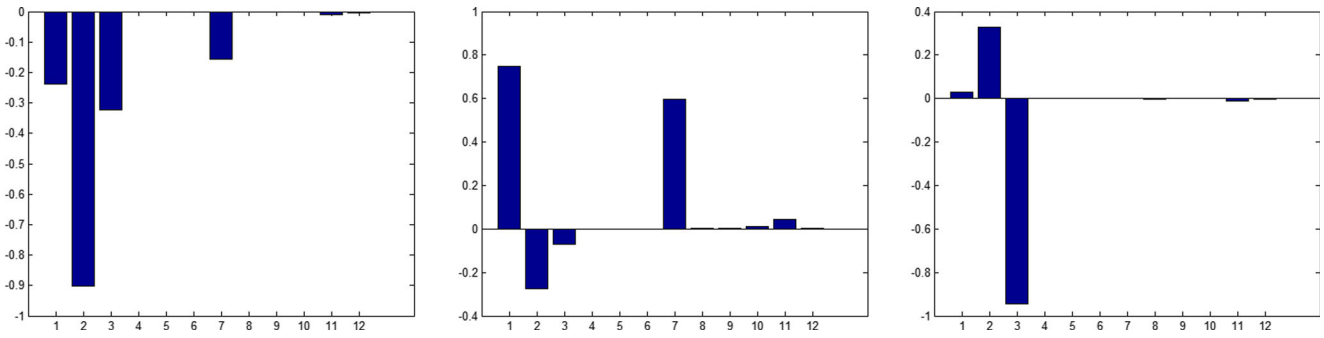

Fig. 7 An illustration of three "concepts" derived by LSA (denoted by y_1, y_2 and y_3, left to right)

As applying the cost-sensitive classifier based on various classifiers, the recall almost reach 100 %. In our study, the cost-sensitive classifier with RF is the best setting.

Dimension reduction (LSA) − With respect to the BIRADS data set consisting of twelve attributes, we apply LSA to reduce the dimension of data set.

In our study shown in Fig. 5, the best assignment of k-value of SVD is seven, which indicates that the original data set in twelve dimensional space can be reduced in the seven dimensional space. Meanwhile, the performance, in terms of accuracy, recall/precision, and AUC, almost remain the same (Fig. 6).

The *energy concentration ratio* is a measure of data set information concentration range. The *energy* $E(\vec{x})$ of a vector $\vec{x}$ in n-dimensional space is defined as the sum of energies at every point of the vector:

$$E(\vec{x}) = \sum_{i=1}^{n} |x_i|^2 \tag{9}$$

where $\vec{x} = (x_1, x_2, ...x_n)$.

Given an original vector $\vec{x} = (x_1, x_2,..., x_{12})$, and the transformed vector $\vec{y} = (y_1, y_2, ..., y_{12})$, the *energy concentration ratio of k* (denoted, *ratio-of-k*) is a measurement of *top-k strongest coefficient* of the transformed vector.

$$\textit{ratio-of-k} = \frac{\sum_{i=1}^{k} y_i^2}{\sum_{i=1}^{12} y_i^2},\ 1 \le k \le 12. \tag{10}$$

The ratio is between 0 and 1. The higher *ratio-of-k* indicates that choosing the first-k coefficient is better enough to represent the whole vector and having less *squared error* (sum of squares of omitted coefficient) between the *reduced vector* and the *original vector*.

By applying LSA, the seven ($k = 7$)"concepts" (denoted by y_1, y_2, y_3, y_4, y_5, y_6 and y_7) are described as follows.

For better illustration, these derived seven "concepts" (i.e. (11) ∼ (17)) are visualized in Figs. 7 and 8.

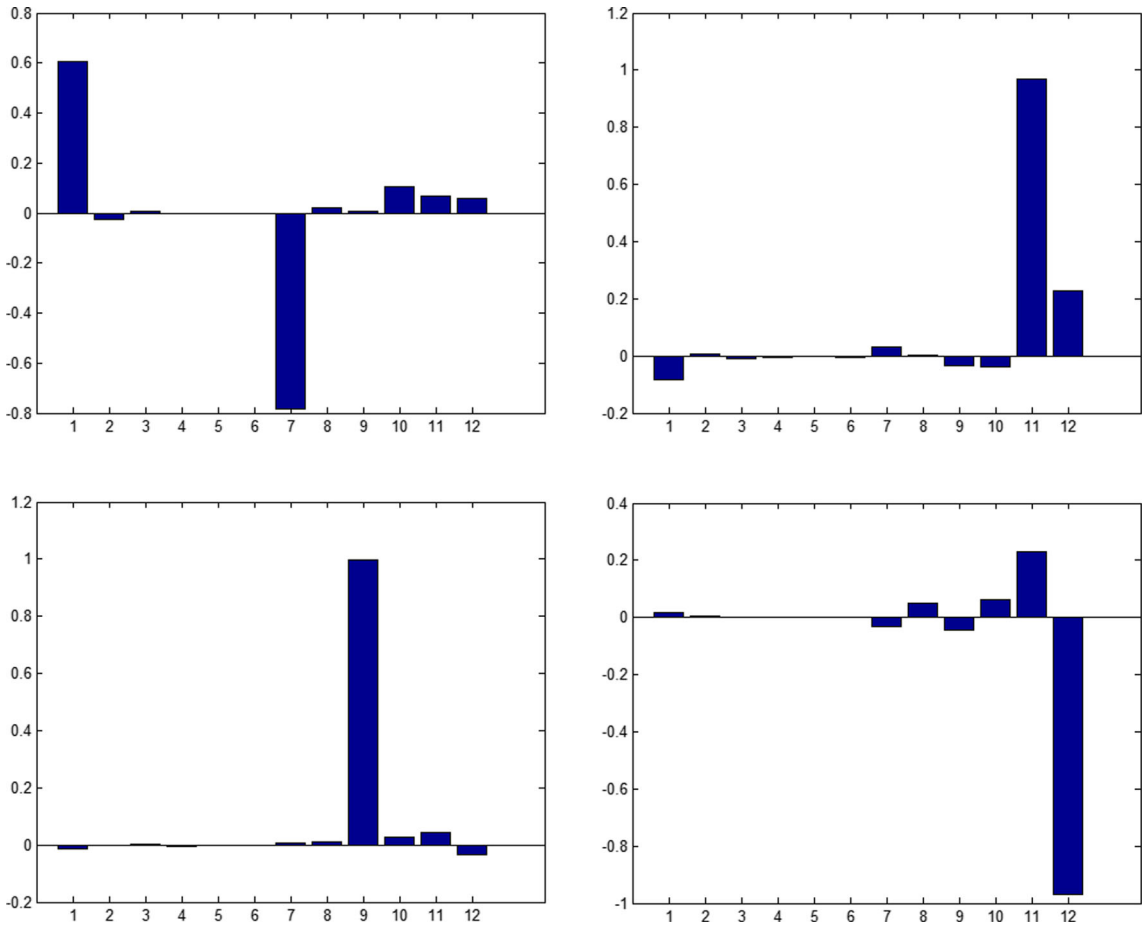

Fig. 8 An illustration of four "concepts" derived by LSA (denoted by y_4, y_5, y_6 and y_7, left to right, up to down)

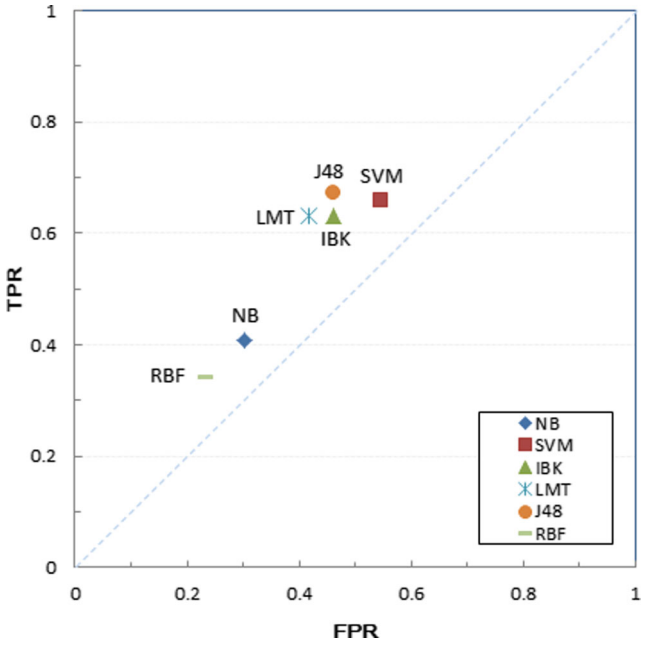

Fig. 9 An illustration of overall performance in the ROC space (under-sampling; basic classifiers)

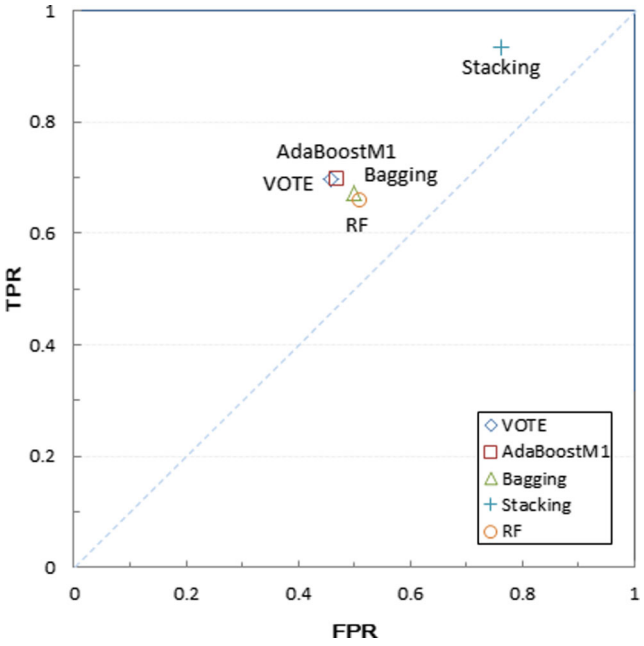

Fig. 10 An illustration of overall performance in the ROC space (under-sampling; ensemble classifiers)

For example, considering the '1-st concept' (in reference to (11) and Fig. 7), the concept is predominated by the first, second, third, and seventh attributes.

$$y_1 = (-0.2385)x_1 + (-0.9019)x_2 + (-0.3234)x_3 + (-0.0002)x_4 + (-0.0002)x_5 + (-0.0002)x_6 + (-0.1577)x_7 + (-0.0015)x_8 + (-0.0014)x_9 + (-0.0005)x_{10} + (-0.0099)x_{11} + (-0.0046)x_{12} \quad (11)$$

$$y_2 = (0.7487)x_1 + (-0.2770)x_2 + (-0.0718)x_3 + (-0.0004)x_4 + (0.0009)x_5 + (0.0010)x_6 + (0.5961)x_7 + (0.0036)x_8 + (0.0028)x_9 + (0.0105)x_{10} + (0.0449)x_{11} + (0.0015)x_{12} \quad (12)$$

$$y_3 = (0.0303)x_1 + (0.3300)x_2 + (-0.9434)x_3 + (0.0000)x_4 + (-0.0006)x_5 + (0.0002)x_6 + (0.0026)x_7 + (-0.0046)x_8 + (0.0015)x_9 + (-0.0002)x_{10} + (-0.0104)x_{11} + (-0.0030)x_{12} \quad (13)$$

$$y_4 = (0.6068)x_1 + (-0.0274)x_2 + (0.0067)x_3 + (-0.0009)x_4 + (-0.0005)x_5 + (-0.0005)x_6 + (-0.7814)x_7 + (0.0192)x_8 + (0.0082)x_9 + (0.1076)x_{10} + (0.0688)x_{11} + (0.0603)x_{12} \quad (14)$$

$$y_5 = (-0.0823)x_1 + (0.0082)x_2 + (-0.0111)x_3 + (-0.0042)x_4 + (-0.0022)x_5 + (-0.0038)x_6 + (0.0330)x_7 + (0.0041)x_8 + (-0.0353)x_9 + (-0.0376)x_{10} + (0.9684)x_{11} + (0.2269)x_{12} \quad (15)$$

$$y_6 = (-0.0119)x_1 + (-0.0001)x_2 + (0.0008)x_3 + (-0.0037)x_4 + (-0.0010)x_5 + (-0.0005)x_6 + (0.0064)x_7 + (0.0128)x_8 + (0.9980)x_9 + (0.0257)x_{10} + (0.0439)x_{11} + (-0.0337)x_{12} \quad (16)$$

$$y_7 = (0.0153)x_1 + (0.0035)x_2 + (0.0019)x_3 + (-0.0006)x_4 + (0.0025)x_5 + (-0.0009)x_6 + (-0.0334)x_7 + (0.0499)x_8 + (-0.0447)x_9 + (0.0625)x_{10} + (0.2299)x_{11} + (-0.9682)x_{12} \quad (17)$$

In summary, we illustrate the overall performance in the ROC space and P-R diagram (*precision-recall* diagram) as follows.

1. Performance in the ROC space: see Figs. 9, 10, and 11.

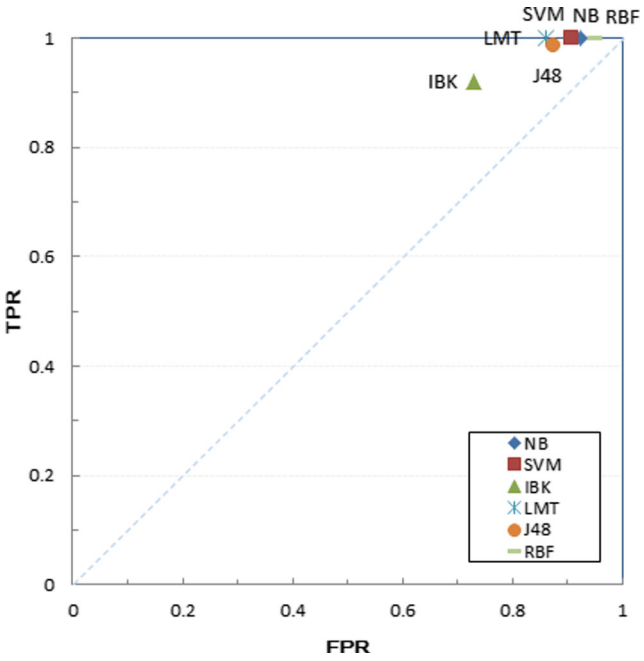

Fig. 11 An illustration of overall performance in the ROC space (under-sampling; cost-sensitive classifiers)

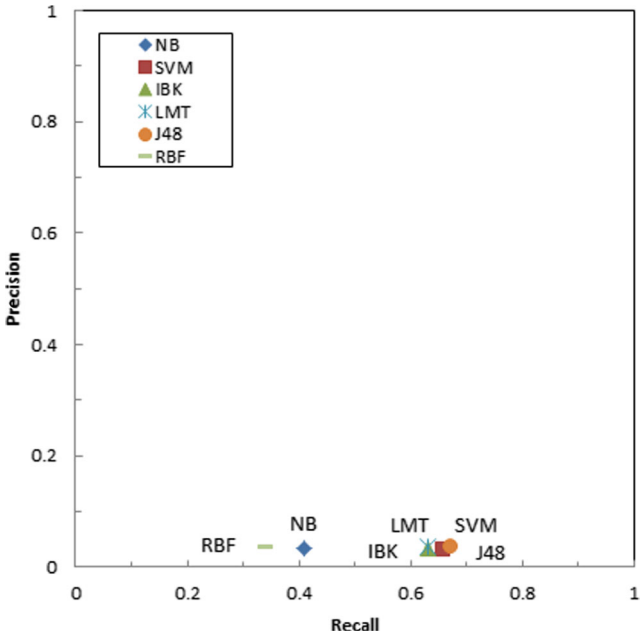

Fig. 12 An illustration of overall performance in the P-R diagram (under-sampling; basic classifiers)

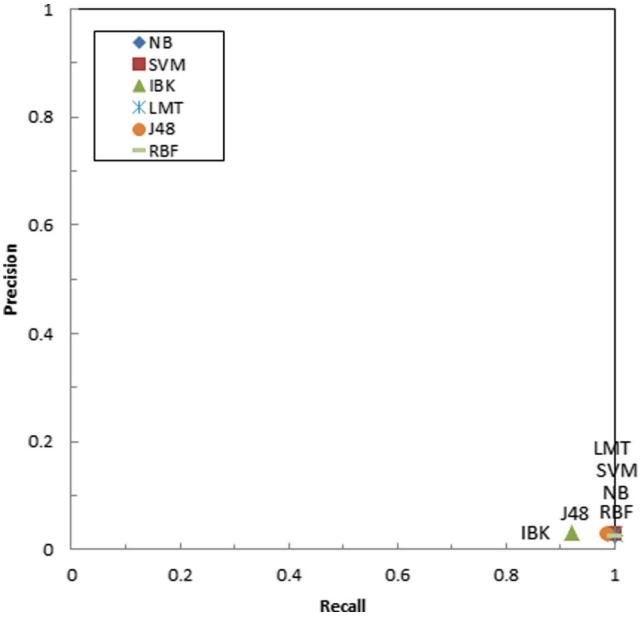

Fig. 14 An illustration of overall performance in the P-R diagram (under-sampling; cost-sensitive classifiers)

2. Performance in the P-R diagram: see Figs. 12, 13, and 14.

Discussion

In this section, we present the discussion based on the experiment results.

1. In our study, we apply the cost-sensitive method to construct a computation model which meets our goal of high recall and reasonable precision.
2. The higher cost setting of FN case (indicating the penalty of misclassification), the better we are able to approach our goal of "no false dismissals".
3. As shown in Fig. 15, that is a trade-off between *recall* and *specificity*. When the recall is 100 %, the specificity

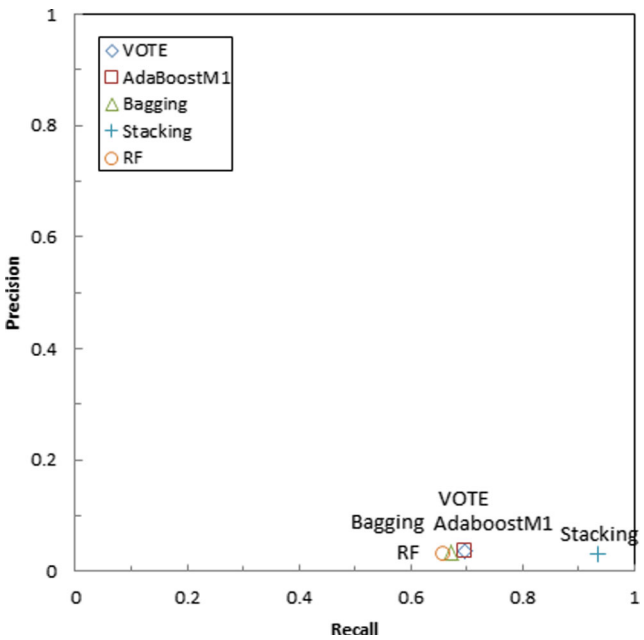

Fig. 13 An illustration of overall performance in the P-R diagram (under-sampling; ensemble classifiers)

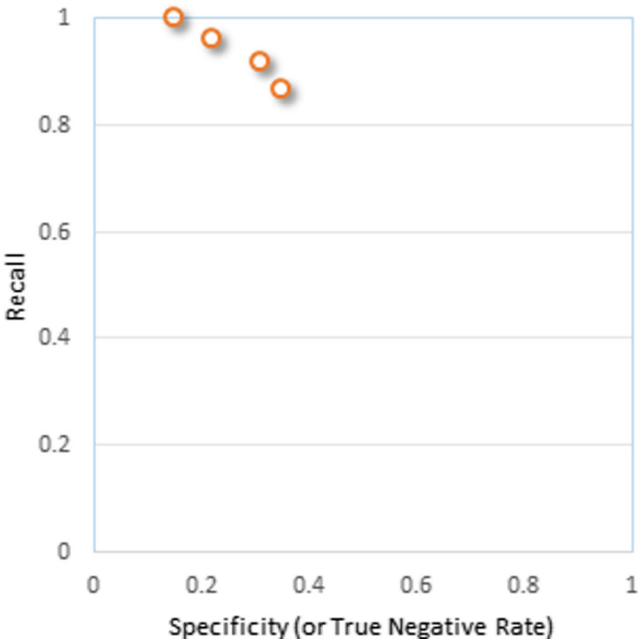

Fig. 15 Cost-sensitive methods with random forest classifiers (recall vs. specificity)

is 14.87 %. If we slightly decrease the FN cost, the recall will be down to 86 % and the specificity will be up to 34.84 %.

4. When building classification model of imbalance data, the sampling technique is crucial to reinforce the class boundary.
5. In our study, we also apply the dimension reduction technique (LSA) to reduce the size of data set for model construction. To be more specific, the reduced data set is about 58 % of the original data set. Meanwhile, the performance almost remains the same.

Conclusion

In our paper, we make use of patient health information to build a computational model for predicting the risk of breast cancer. Our goal is to construct a low-cost pre-diagnosis program which guarantees "no false dismissals" (i.e., a 100 % recall/sensitivity). The system architecture consists of four major components, including the pre-processing module, the sampling modules, the dimension reduction module, and classifiers. Based on our performance evaluation, we conclude that: (1) apply the under-sampling technology; (2) apply LSA dimensional reduction; (3) choose cost-sensitive method in which the random forest is the base classifier. Our approach is able to achieve the recall/sensitivity as 100 %. The precision and specificity is 2.9 % and 14.87 %. As a result, before mammography screening and early diagnosis, our model could be successfully applied to predict the risk of breast cancer in the clinical application.

Acknowledgments Financial support for this study was provided in part by a grant from the National Science Council, Taiwan, under Contract No. NSC-102-2218-E-030-002 . The funding agreement ensured the authors' independence in designing the study, interpreting the data, writing, and publishing the report.

References

1. Siegel, R., Naishadham, D., Jemal, A., Cancer statistics, 2013. *CA: Cancer J. Clin.* 63(1):11–30, 2013. Available from: doi:10.3322/caac.21166.
2. Kim, J., and Shin, H., Breast cancer survivability prediction using labeled, unlabeled, and pseudo-labeled patient data. *J. Am. Med. Inform. Assoc.* 20(4):613–618, 2013.
3. Uhry, Z., Hédelin, G., Colonna, M., Asselain, B., Arveux, P., Rogel, A., et al., Multi-state Markov models in cancer screening evaluation: a brief review and case study. *Stat. Methods Med. Res.* 19(5):463–486, 2010.
4. Bleyer, A., and Welch, H.G., Effect of three decades of screening mammography on breast-cancer incidence. *N. Engl. J. Med.* 367(21):1998–2005, 2012.
5. Blume, J.D., Cormack, J.B., Mendelson, E.B., Lehrer, D., Pisano, E.D., Jong, R.A., et al., Combined screening with ultrasound and mammography vs mammography alone in women at elevated risk of breast cancer. *J. Am. Med. Assoc.* 299(18):2151–2163, 2008.
6. Lord, S.J., Lei, W., Craft, P., Cawson, J.N., Morris, I., Walleser, S., et al., A systematic review of the effectiveness of magnetic resonance imaging (MRI) as an addition to mammography and ultrasound in screening young women at high risk of breast cancer. *Eur. J. Cancer* 43(13):1905–1917, 2007. Available from: http://www.sciencedirect.com/science/article/pii/S0959804907004844.
7. Breast Cancer Screening (PDQ), Breast Cancer Screening Modalities Beyond Mammography (Health Professional Version) [homepage on the Internet]. National Cancer Institute; c2014 [updated 2014 Oct. 3; cited 2014 Oct. 6]. Available from: http://www.cancer.gov/cancertopics/pdq/screening/breast/healthprofessional/page9.
8. Kittler, J., Hatef, M., Duin, R.P.W., Matas, J., On combining classifiers. *IEEE Trans. Pattern Anal. Mach. Intell.* 20(3):226–239, 1998. Available from: doi:10.1109/34.667881.
9. Wolpert, D.H., Stacked generalization. *Neural Netw.* 5:241–259, 1992.
10. Elkan, C., The Foundations of cost-sensitive learning. In: Proceedings of the 17th International Joint Conference on Artificial Intelligence - Volume 2. IJCAI'01, pp. 973–978. San Francisco, CA: Morgan Kaufmann Publishers Inc., 2001. Available from: http://dl.acm.org/citation.cfm?id=1642194.1642224.
11. Seiffert, C., Khoshgoftaar, T.M., van Hulse, J., Napolitano A., A Comparative Study of Data Sampling and Cost Sensitive Learning. In: Proceedings of the 2008 IEEE International Conference on Data Mining Workshops, pp. 46–52, 2008.
12. Garca-Laencina, P., Sancho-Gmez, J.L., Figueiras-Vidal, A., Pattern classification with missing data: a review. *Neural Comput. Applic.* 19(2):263–282, 2010. Available from: doi:10.1007/s00521-009-0295-6.
13. Evangelopoulos, N.E., Latent semantic analysis. *Wiley Interdiscip. Rev. Cogn. Sci.* 4(6):683–692, 2013. doi:10.1002/wcs.1254.
14. Deerwester, S.C., Dumais, S.T., Landauer, T.K., Furnas, G.W., Harshman, R.A., Indexing by latent semantic analysis. *J. Am. Soc. Inf. Sci.* 41(6):391–407, 1990.
15. Fawcett, T., An introduction to, R O C analysis. *Pattern Recognit. Lett.* 27(8):861–874, 2006.