

# Diagnosing Breast Masses in Digital Mammography Using Feature Selection and Ensemble Methods

Shu-Ting Luo · Bor-Wen Cheng

Received: 18 January 2010 / Accepted: 23 April 2010  
© Springer Science+Business Media, LLC 2010

**Abstract** Methods that can accurately predict breast cancer are greatly needed and good prediction techniques can help to predict breast cancer more accurately. In this study, we used two feature selection methods, forward selection (FS) and backward selection (BS), to remove irrelevant features for improving the results of breast cancer prediction. The results show that feature reduction is useful for improving the predictive accuracy and *density* is irrelevant feature in the dataset where the data had been identified on full field digital mammograms collected at the Institute of Radiology of the University of Erlangen-Nuremberg between 2003 and 2006. In addition, decision tree (DT), support vector machine—sequential minimal optimization (SVM-SMO) and their ensembles were applied to solve the breast cancer diagnostic problem in an attempt to predict results with better performance. The results demonstrate that ensemble classifiers are more accurate than a single classifier.

**Keywords** Digital mammography · Feature selection · Breast cancer · Ensemble classifiers

## Introduction

Breast malignancy is one of the most common cancers among women. It is a major health problem and represents a significant worry for many women and their physicians [1]. During the last decade with development of more effective diagnostic techniques and improvements in treatment methodologies, breast cancer outcomes have improved. An important factor in this disease is the early detection and accurate diagnosis. The long-term survival rate for women with breast cancer is improved by detecting the disease in its early stage [2].

Mammography is the screening modality proven to detect breast cancer at early stage and diagnosis of breast cancer in women [3–5]. The efficacy of mammography is limited in extremely dense breasts where sensitivity to detect cancer maybe as low as 60–70% [6]. Both digital and film mammography use X-rays to produce an image of the breast. Pisano [7, 8] provided the overall diagnostic accuracy of digital and film mammography as a means of screening for breast cancer is similar. However, digital mammography performed better than film for pre- and perimenopausal women younger than 50 years with dense breasts [8]. In addition, digital mammography allows improvement in image storage and transmission because images can be stored and sent electronically. Diagnostic mammography is done for women with signs or symptoms of breast cancer. Any sign of cancer should be communicated to the radiologist with the referral for a diagnostic mammogram. Radiologists basically look for two types of patterns in mammography: micro-calcifications and masses [9].

Some researchers have developed a variety of statistical methods for mammographic diagnosis of breast cancer

---

S.-T. Luo (✉) · B.-W. Cheng  
Graduate School of Industry Engineering and Management,  
National Yunlin University of Science and Technology,  
123 University Road, Section 3,  
Douliou, Yunlin 64002, Taiwan  
e-mail: g9521806@yuntech.edu.tw

B.-W. Cheng  
e-mail: chengbw@yuntech.edu.tw

[10–14]. Rakowski and Clark utilized multiple logistic regression to select significant correlates of screening mammogram and used classification-tree (CHAID) to combine the significant correlates into exclusive and exhaustive subgroups [13]. In addition, Chhatwal et al. reported that logistic regression model can discriminate between benign and malignant in decision making for the early detection of breast cancer and identify the most important features associated with breast cancer [14]. Moreover, Heine and colleagues show how parametric statistical methods can be useful for in identifying normal tissue in mammograms [12].

Recently, many studies have been made on the problem of breast cancer diagnosing based on digital mammography [15, 16]. Some scholars applied data mining techniques to predict diagnosis for digital mammography [17, 18]. Data mining techniques offer precise, accurate, and fast algorithms for such classification using dimensionality reduction, feature extraction, and classification routines. Neural networks have improved accuracy rate for the classification of benign and malignant patterns in digitized mammography [19–21]. Recently, the fashionable technique support vector machine (SVM) [22, 23] have been applied for mammogram classification and have improved the prediction performance of breast cancer diagnosis.

Feature selection is also commonly used in machine learning. It has already seen application in statistics, pattern recognition, and data mining. The aim of feature selection is to filter out redundant or irrelevant features from the original data [24]. Feature selection, a pre-processing step in the data mining process, is the step to select and extract more valuable information in massive related materials. It can improve the model's performance as well as reduce the effort of training the model.

Ensemble classifier is now an active area of research in machine learning and pattern recognition [25]. Many studies have been published, both theoretical and empirical, which demonstrate the advantages of the combination paradigm over the individual classifier models [26]. Recently, two ensemble methods, Bagging and Boosting, have also obtained wide popularity [27]. Bagging uses the predictions of multiple base classifiers through majority voting. Boosting, a meta-classifier, combines weak classifiers and takes a weighted majority vote of their predictors.

This study aimed to calculate and compare the degree of importance for the features of the dataset. In addition, we tried to remove least important features to check whether it could help improve the results of breast cancer prediction. In this study, decision tree (DT) and support vector machine—sequential minimal optimization (SVM-SMO) and their ensembles were employed to solve the problem in an attempt to predict results with better performance. A large publicly

available mammography reference database was adopted to evaluate and compare our proposed approaches.

## Methods

### Decision tree

Decision tree (DT) provides a powerful technique for classification and prediction. Since DT includes both data exploration to identify relationships and modeling to provide decision rules that can be expressed in natural language, they are an insightful first step in the modeling process, even if another methodology is used for the final model. The DT is conceived in order to maximize the correct classification of all the training examples provided. The generated structure is thus subject to over fitting, in which the classifier specializes to the training examples, showing poor performance on new data. One benefit of using DT is that the extracted knowledge is organized in a structure that can be easily explainable by humans [28]. In this study, a predictive model was made by applying DT to the prepared data. DT was carried out using the J48 algorithm in WEKA (Waikato environment for knowledge analysis) [29]. WEKA is the software can either be applied directly to a dataset or called from other programs and it is a set of machine learning algorithms for data mining tasks [30]. We adopted WEKA for mining (applying the J48 algorithm) breast cancer register data. The application contains tools for data preparation, classification, clustering and visualization. In WEKA, the J48 algorithm is the equivalent of the C4.5 algorithm written by Quinlan [31].

### Support vector machine-sequential minimal optimization (SVM-SMO)

Support vector machine (SVM) represents a learning technique which follows principles of statistical learning theory [32]. It is a supervised machine-learning tool with wide application in classification studies. For example, it has been widely used for solving problems in pattern recognition, classification and regression. Generally, the main idea of SVM comes from binary classification, namely to find a hyperplane as a segmentation of the two classes to minimize the classification error. The Sequential Minimal Optimization (SMO) algorithm proposed by John Platt in 1998 [33], is a simple and fast method for training a SVM. The main idea is derived from solving dual quadratic optimization problem by optimizing the minimal subset including two elements at each iteration. The advantage of SMO is that it can be implemented simply and analytically. There are different kernel functions used in SVM. The selection of the appropriate kernel function

is very important, since the kernel defines the feature space in which the training set examples will be classified. Using of different kernel functions in SVM will lead to different performance results. In this study, the best result was obtained using radial kernel functions.

### Ensemble constructing techniques

Ensemble methods in machine learning aim to induce a collection of diverse predictors which are both accurate and complementary, so that, better prediction accuracy on previously unseen data is obtained when the decisions of the different learners are combined. The goal is to generate from a given training dataset a collection of diverse predictors whose errors are uncorrelated. Ensembles built in this manner often exhibit significant performance improvements over a single predictor in many regression and classification problems [34]. Ensembles can be built using different base classifiers: decision stumps [35] decision trees [36] support vector machines [37, 38], etc. In this work, we adopted the ensemble constructing techniques. All of the techniques combine DT and SVM classifiers to form different ensemble classifiers.

### Bagging

Bagging, which is also known as bootstrap aggregating, is a method that repeatedly samples from a dataset according to uniform probability distribution. It is a meta-algorithm to improve classification and regression models in terms of stability and classification accuracy. Although Bagging is usually applied to decision tree classifiers, it can be used with any type of model. Bagging improves generalization error by reducing the variance of the base classifiers [39]. If a base classifier is unstable, Bagging helps to reduce the errors associated with random fluctuations in the training data. If a base classifier is stable, then the error of the ensemble is primarily caused by bias in the base classifier.

### AdaBoost

Freund and Schapire [40] formulated AdaBoost, short for Adaptive Boosting. It is a well known, effective technique

for increasing the accuracy of learning algorithms. However, it has the potential to over-fit the training set because its objective is to minimize error on the training set. The sequence of base classifiers, produced by AdaBoost from the training set, is applied to the validation set, creating a modified set of weights. The training and validation sets are switched, and a second pass is performed. Re-weighting and re-sampling are two methods implemented in AdaBoost. The fixed training sample size and training examples are re-sampled according to a probability distribution used in each iteration. In term of re-weighting, all training examples with weights assigned to each example are used in each iteration to train the base classifier [41]. There are many variants on the idea of Boosting. We describe a widely used method called AdaBoostM1 that is designed specifically for classification.

### MultiBoosting

MultiBoosting is an extension to the highly successful AdaBoost technique for forming decision committees and can be viewed as combining AdaBoost with wagging. It is able to harness both AdaBoost's high bias and variance reduction with wagging's superior variance reduction [42].

### Model evaluation step

Empirical studies adopted k-fold cross validation that in order to obtain a reliable result with low mean square error (MSE) and bias [43]. The performance of each classifier was assessed with a stratified 10-fold cross validation method. Each run of cross validation is comprised of an independent training and testing database, where 90% of the data is put in the training set and the remaining 10% of the data is put into the test set. For each classification model, statistical results of 10 repetitions of 10-fold cross validation were averaged and calculated. In addition, the following statistics were calculated: sensitivity, specificity and accuracy.

ROC Curve has been widely accepted as the standard for describing and comparing the accuracy of diagnostic tests [44]. It can also be constructed from clinical prediction rules. The accuracy of the test depends on how well the test

**Table 1** Mammographic mass dataset: description of attributes

| Attribute                                    | Range and description   |
|--|---|
| BI-RADS assessment (ordinal, non-predictive) | The ranging from 1 (definitely benign) to 5 (highly suggestive of malignancy).                |
| Age (integer)                                | The patient's age has been widely distributes from 18 to 96 (in years).                       |
| Shape (nominal)                              | mass shape: round=1, oval=2, lobular = 3, irregular = 4                                       |
| Margin (nominal)                             | mass margin: circumscribed=1, microlobulated=2, obscured = 3, ill-defined = 4, spiculated = 5 |
| Density (ordinal)                            | mass density high=1, iso=2, low = 3, fat-containing = 4                                       |

**Table 2** The attributes obtained by feature selection methods

| Used METHODS | Attributes |     |       |        |         |
|--------------|------------|-----|-------|--------|---------|
|              | BIRADS     | Age | Shape | Margin | Density |
| FS           | 4          | 2   | 3     | 1      | /       |
| BS           | 4          | 3   | 2     | 1      | /       |

/: The attribute was omitted

separates the group being tested into those with and without the disease in question. An ROC curve is constructed by plotting the sensitivity versus 1-specificity, for varying cutoff values. The area under the curve (AUC) is the evaluation criteria for the classifier. It can be statistically interpreted as the probability of the classifier to correctly classify malignant cases and benign cases. In this work, the AUC is obtained by a nonparametric method based on the Wilcoxon statistic, using the trapezoidal rule, to approximate the area [45, 46]. AUC can be used for comparing two different ROC curves from the same samples of cases.

## Experiments and methodologies

### Database overview

In this study, a real-world breast cancer database from the UCI repository of machine learning databases [47] was chosen and the description of attributes of the dataset was illustrated in Table 1. The database was used to predict the severity (benign or malignant) of a mammography mass lesion from breast imaging reporting data system (BI-RADS) attributes and the patient's age. The BI-RADS, developed by the American College of Radiology, provides a standardized classification for mammographic studies. It was designed to reduce variability in mammography practice and used by a number of computerized mammography tracking systems. In addition, BI-RADS is a quality assurance tool, reduce confusion in breast imaging interpretations, and facilitate outcome monitoring [48]. The database contains a BI-RADS assessment, the patient's age and three BI-RADS attributes

(mass shape, mass margin, mass density) and is based on digital mammograms collected at the Institute of Radiology of the University of Erlangen-Nuremberg between 2003 and 2006 [49]. It consists of 961 records and each record in the database has one dependent and five independent variables. There are 516 benign and 445 malignant masses. The database does not reflect all variables that are collected by radiologists during mammography practice, hence, it is one of the limitations of this work.

### Feature selection

Feature selection is an important issue in building classification systems. It is advantageous to limit the number of input features in a classifier in order to have a good predictive and less computationally intensive model [50]. In this study we used two well-known feature selection techniques, which are forward selection (FS), and backward selection (BS) [51]. The statistical software SPSS was adopted for applying feature selection methods (FS and BS).

Forward selection and backward selection select the least important descriptors based on stepwise addition or elimination of features. In forward selection, variables are progressively incorporated into larger and larger subsets. In backward selection, the values of the selection criterion using all the features are calculated. Then starting from the initial set of features the algorithm temporarily deletes each feature, calculates the value of the selection criterion, and deletes the feature with the largest value of the selection criterion from the set.

### Experimental validation

An experiment was set up to compare DT and SVM with Bagging, AdaBoost, and MultiBoosting. In all ensemble methods, DT and SVM were used as the base classifiers. In implementing of the experiment, we used the WEKA software to gain access to different classifiers. DT construction method was the J48 algorithm and SVM construction method was SMO algorithm from the WEKA. We adopted the 10-fold cross-validation method to evaluate the prediction performance in this study.

**Table 3** Summary of performance using single DT and DT ensembles

| Classifiers      | 5-attributes |             |              | 4-attributes |             |              |
|------------------|--------------|-------------|--------------|--------------|-------------|--------------|
|                  | Specificity  | Sensitivity | Accuracy     | Specificity  | Sensitivity | Accuracy     |
| DT               | 0.829        | 0.836       | 0.831(0.005) | 0.832        | 0.835       | 0.831(0.006) |
| Bagging DT       | 0.836        | 0.833       | 0.833(0.004) | 0.835        | 0.835       | 0.834(0.004) |
| AdaboostM1 DT    | 0.809        | 0.811       | 0.808(0.007) | 0.814        | 0.816       | 0.815(0.009) |
| MultiBoosting DT | 0.822        | 0.820       | 0.821(0.006) | 0.823        | 0.821       | 0.822(0.005) |

The numbers in parentheses are the standard errors

**Table 4** Summary of performance using single SVM-SMO and SVM-SMO ensembles

| Classifiers           | 5-attributes |             |              | 4-attributes |             |              |
|-----------------------|--------------|-------------|--------------|--------------|-------------|--------------|
|                       | Specificity  | Sensitivity | Accuracy     | Specificity  | Sensitivity | Accuracy     |
| SVM-SMO               | 0.815        | 0.817       | 0.812(0.005) | 0.816        | 0.814       | 0.812(0.004) |
| Bagging SVM-SMO       | 0.823        | 0.814       | 0.818(0.009) | 0.824        | 0.813       | 0.820(0.013) |
| AdaboostM1 SVM-SMO    | 0.811        | 0.808       | 0.808(0.007) | 0.813        | 0.813       | 0.813(0.005) |
| MultiBoosting SVM-SMO | 0.815        | 0.811       | 0.811(0.007) | 0.816        | 0.811       | 0.812(0.005) |

The numbers in parentheses are the standard errors

**Results**

Selected features using feature selection methods

By setting some criteria for the two feature selection methods to take irrelevant features out, thus improving the results of breast cancer prediction, Table 2 shows the attributes that were extracted from the mammographic mass dataset using FS and BS. As seen in Table 2, FS and BS both reduced the attribute, *density*. Each feature is ranked respectively with two feature selection methods and the number in the column represents the degree of importance. The result shows *margin* is the most important attribute and *density* is the least important attribute (Table 2).

Diagnostic accuracy

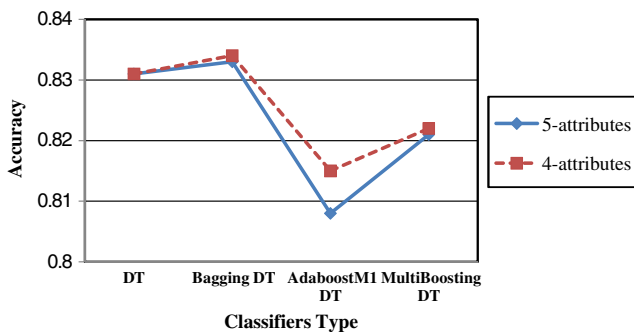
The performance using single SVM-SMO, DT and their ensembles for 5- and 4- attributes (omitted *density*) were shown in Tables 3 and 4. The items of the tables include total accuracy (i.e., the percentage of correctly classified patterns), sensitivity (i.e., the probability that a case identified as malignant is indeed malignant), and specificity (i.e., the probability that a case identified as benign is indeed benign). In addition, the accuracy comparisons of these methods were illustrated in Figs. 1 and 2. As shown in Fig. 1, the results demonstrate that ensembles are not

always better than a single DT classifier. However, the accuracy of AdaBoostM1 DT was slightly worse than those of ensembles DT and single DT. In addition, after applying FS and BS to remove least important feature (4-attributes), the accuracy rates were better than original features (5-attributes) except single DT. Feature selection is thus useful for improving the result of accuracy in our experiment. The Bagging DT adopting 4 attributes had the best performance (accuracy: 0.834) among all these approaches.

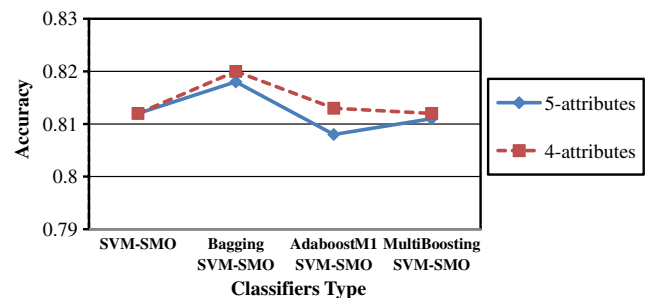
As shown in Fig. 2, the results indicate that SVM-SMO ensembles were also not always better than a single SVM-SMO classifier. However, the accuracy of AdaBoostM1 SVM-SMO was slightly worse than those of ensembles SVM-SMO and single SVM-SMO. Feature selection is also positive for improving the result of accuracy as using SVM-SMO. The accuracy rates of after removing the *density* attribute was better than those adopting 5-attributes when using Bagging SVM-SMO, AdaBoostM1 SVM-SMO and MultiBoosting SVM-SMO, except SVM-SMO. The Bagging SVM-SMO adopting 4-attributes had the highest accuracy (0.820) among all these approaches.

Model evaluation

The average of AUC ( $\overline{AUC}$ ), the corresponding standard error (S.E. derived from 30 AUC values), and 95%



**Fig. 1** The results of using single DT and DT ensembles



**Fig. 2** The results of using single SVM-SMO and SVM-SMO ensembles

**Table 5** Comparison of the AUC using the different classifiers (standard error)

| Classifiers           | Estimated $\overline{\text{AUC}}$ |               | 95% confidence intervals |              |
|-----------------------|-----------------------------------|---------------|--------------------------|--------------|
|                       | 5-attributes                      | 4-attributes  | 5-attributes             | 4-attributes |
| DT                    | 0.866 (0.004)                     | 0.866 (0.004) | 0.865–0.867              | 0.858–0.874  |
| Bagging DT            | 0.886 (0.004)                     | 0.885 (0.004) | 0.884–0.888              | 0.881–0.891  |
| AdaBoostM1 DT         | 0.875 (0.005)                     | 0.876 (0.004) | 0.873–0.877              | 0.871–0.879  |
| MultiBoosting DT      | 0.879 (0.009)                     | 0.880 (0.004) | 0.876–0.882              | 0.876–0.882  |
| SVM-SMO               | 0.813 (0.005)                     | 0.813 (0.004) | 0.811–0.815              | 0.812–0.814  |
| Bagging SVM-SMO       | 0.867 (0.005)                     | 0.869 (0.005) | 0.865–0.869              | 0.865–0.869  |
| AdaBoostM1 SVM-SMO    | 0.875 (0.004)                     | 0.876 (0.005) | 0.873–0.876              | 0.873–0.877  |
| MultiBoosting SVM-SMO | 0.874 (0.012)                     | 0.877 (0.004) | 0.870–0.878              | 0.872–0.876  |
| Elter et.al (2007)    |                                   |               |                          |              |
| DT                    | 0.838 (0.017)                     | –             | –                        | –            |
| CBR                   | 0.857 (0.016)                     | –             | –                        | –            |
| ANN                   | 0.847 (0.017)                     | –             | –                        | –            |

confidence interval (CI) using different classifiers (DT, SVM-SMO and their ensembles) are shown in Table 5. In addition, the performance measures using DT, CBR and ANN reported in [49] were compared with the methods we proposed in this study. Among the classifiers, the Bagging DT has the best performance of the 5-attributes and 4-attributes, respectively. The results of using DT method were not alike, 0.866 (0.004) vs. 0.838 (0.017), because the algorithms were different. The algorithm we adopted was the C4.5 algorithm, however, it was the ID3 algorithm in [49].

Tables 6, 7 and 8 show the comparisons and significance level of eight different methods. When the two classifiers are significant or close to significant,  $p$ -values are indicated in bold. ( $p$ -value from pairwise two-tailed  $z$ -test) in these three tables. Both under 5- and 4- attributes, the AUC of SVM-SMO is all significant different with the others classifiers ( $p$ -values<0.001), so, it is the worst method in our experiment. In addition, there are significant difference

between SVM-SMO and its three ensembles. The AUC value of SVM-SMO is lower than those of its ensembles, hence, the proposed ensemble are significantly helpful to improve the prediction performance. Under under 5-attributes, only Bagging DT has significant difference ( $p$ -values=0.004) as compared to DT. However, there under 4-attributes, the difference between DT and its three ensembles are significant ( $p$ -values=0.001, 0.077, 0.013). Hence, the effects of ensemble methods are more obvious when under 4-attributes.

## Discussion

Mammography is used to aid in the early detection and diagnosis of breast diseases in women. Diagnostic breast cancer is widely used to facilitate the diagnosis of breast cancers in women who present with symptoms or signs of

**Table 6** Significance level when two AUCs on the test set are compared under 5-attributes

|                       | DT | Bagging DT   | AdaboostM1 DT | Multiboosting DT | SVM-SMO        | Bagging SVM-SMO | AdaboostM1 SVM-SMO | Multiboosting SVM-SMO |
|-----------------------|----|--------------|---------------|------------------|----------------|-----------------|--------------------|-----------------------|
| DT                    | NA | <b>0.004</b> | 0.160         | 0.187            | < <b>0.001</b> | 0.876           | 0.112              | 0.527                 |
| Bagging DT            |    | NA           | 0.086         | 0.477            | < <b>0.001</b> | <b>0.003</b>    | <b>0.052</b>       | 0.343                 |
| AdaboostM1 DT         |    |              | NA            | 0.698            | < <b>0.001</b> | 0.258           | 1.000              | 0.939                 |
| Multiboosting DT      |    |              |               | NA               | < <b>0.001</b> | 0.244           | 0.685              | 0.739                 |
| SVM-SMO               |    |              |               |                  | NA             | < <b>0.001</b>  | < <b>0.001</b>     | < <b>0.001</b>        |
| Bagging SVM-SMO       |    |              |               |                  |                | NA              | 0.212              | 0.590                 |
| AdaboostM1 SVM-SMO    |    |              |               |                  |                |                 | NA                 | 0.937                 |
| Multiboosting SVM-SMO |    |              |               |                  |                |                 |                    | NA                    |



**Table 7** Significance level when two AUCs on the test set are compared under 4-attributes

|                       | DT | Bagging DT   | AdaboostM1 DT | Multiboosting DT | SVM-SMO        | Bagging SVM-SMO | AdaboostM1 SVM-SMO | Multiboosting SVM-SMO |
|-----------------------|----|--------------|---------------|------------------|----------------|-----------------|--------------------|-----------------------|
| DT                    | NA | <b>0.001</b> | <b>0.077</b>  | <b>0.013</b>     | < <b>0.001</b> | 0.639           | 0.118              | 0.086                 |
| Bagging DT            |    | NA           | 0.112         | 0.377            | < <b>0.001</b> | <b>0.013</b>    | 0.160              | 0.157                 |
| AdaboostM1 DT         |    |              | NA            | 0.480            | < <b>0.001</b> | 0.274           | 1.000              | 0.860                 |
| Multiboosting DT      |    |              |               | NA               | < <b>0.001</b> | 0.086           | 0.532              | 0.596                 |
| SVM-SMO               |    |              |               |                  | NA             | < <b>0.001</b>  | < <b>0.001</b>     | < <b>0.001</b>        |
| Bagging SVM-SMO       |    |              |               |                  |                | NA              | 0.322              | 0.212                 |
| AdaboostM1 SVM-SMO    |    |              |               |                  |                |                 | NA                 | 0.876                 |
| Multiboosting SVM-SMO |    |              |               |                  |                |                 |                    | NA                    |

the disease. A meaningful and reproducible evaluation of diagnostic methods was necessary and building an effective breast cancer diagnostic model has been an important task. In previous studies [1–5], the issue has attracted a lot of researchers’ interest. In the past, the breast cancer diagnostic problem has been one of the main application areas of classification problems [19–23]. Many modeling, like statistical methods [10–14] are becoming a very popular alternative in handling breast cancer diagnostic tasks. Over the last few years, many studies have shown that data mining techniques such as Artificial Neural Network [19–21] and Support Vector Machine [22, 23] achieved better performance than did statistical methods. Recently, classifier combination is now an active area of research in Machine Learning [24].

Many approaches cannot filter irrelevant or correlated information in the representation and quality of data will affect their performance. Feature selection methods applied on selecting more representative variables certainly improve the performance of prediction. In this work, we adopted two feature selection methods, FS and BS, to take

out irrelevant features thus improving the results of breast cancer prediction. FS and BS both reduced the *density* attribute and kept the others. The results shown in Figs. 1 and 2 indicated that the performances of 4-attributes are better than those of 5-attributes, except in terms of the single classifier (DT, SVM-SMO) was be equal to accuracy. In this study, the result demonstrates that the *margin* is the most important attribute and the *density* is the least important attribute in the mammographic mass dataset.

Feature selection in data mining has been well studied, where the main goal is to find a feature subset that produces higher classification accuracy. The main idea of feature selection is to choose a subset of input variables by eliminating features with little or no predictive information. It can significantly improve the comprehensibility of the resulting classifier models and often build a model that generalizes better to unseen points. Thus it may find a good subset of features that forms high quality of clusters for a given number of clusters. In this work, feature selection is helpful for improving the results of accuracy of the problem. Therefore, it can thus provide clinicians

**Table 8** Significance level when two AUCs on the test set are compared between 5-attributes and 4-attributes

| 4-attributes 5-attributes | DT             | Bagging DT     | AdaboostM1 DT  | Multiboosting DT | SVM-SMO        | Bagging SVM-SMO | AdaboostM1 SVM-SMO | Multiboosting SVM-SMO |
|---------------------------|----------------|----------------|----------------|------------------|----------------|-----------------|--------------------|-----------------------|
| DT                        | 1.000          | <b>0.001</b>   | <b>0.001</b>   | <b>0.013</b>     | < <b>0.001</b> | 0.693           | 0.118              | <b>0.052</b>          |
| Bagging DT                | < <b>0.001</b> | 0.860          | <b>0.077</b>   | 0.724            | < <b>0.001</b> | <b>0.008</b>    | 0.118              | 0.112                 |
| AdaboostM1 DT             | 0.160          | 0.118          | 0.876          | 0.377            | < <b>0.001</b> | 0.349           | 0.888              | 0.755                 |
| Multiboosting DT          | 0.187          | 0.542          | 0.761          | 0.919            | < <b>0.001</b> | 0.331           | 0.771              | 0.839                 |
| SVM-SMO                   | < <b>0.001</b> | < <b>0.001</b> | < <b>0.001</b> | < <b>0.001</b>   | 1.000          | < <b>0.001</b>  | < <b>0.001</b>     | < <b>0.001</b>        |
| Bagging SVM-SMO           | 0.876          | <b>0.005</b>   | 0.160          | 0.042            | < <b>0.001</b> | 0.777           | 0.203              | 0.118                 |
| AdaboostM1 SVM-SMO        | 0.112          | 0.077          | 0.860          | 0.377            | < <b>0.001</b> | 0.349           | 0.876              | 0.724                 |
| Multiboosting SVM-SMO     | 0.527          | 0.385          | 0.874          | 0.635            | < <b>0.001</b> | 0.701           | 0.878              | 0.813                 |

with insight into their databases and lead to further understanding of the disease manifestation.

Ensemble classification has received much attention in the machine learning community and has demonstrated promising capabilities in improving classification accuracy. The results revealed that the ensemble method also improves the performance of prediction. Generally, ensembles can be built using different base classifiers that are more accurate than a single classifier. In this study, the results of  $\overline{AUC}$  demonstrate ensemble classifiers are more accurate than a single classifier.

Moreover, from the results in Tables 3 and 4, we also found that single DT outperformed single SVM-SMO and DT ensembles outperformed SVM-SMO ensembles. Classifier performance depends greatly on the characteristics of the data to be classified. There is no single classifier that works best on all given problems. It is not easy to determine a suitable classifier for a given problem. The DT concept is much easier to comprehend than SVM-SMO. While SVM-SMO generally requires some kernels and their related parameters to be selected in order to get better performance. In contrast DT is an easy algorithm to use for classification and does not need excessive complex knowledge. In addition, ensemble classifiers are powerful tools for diagnostic in breast cancer performance measurement. Therefore it might be the fitting assistance to predict diagnostic for clinicians.

In this study, our results were compared with those of Elter and colleagues' work [49]. The comparisons are shown in Table 5. In most of cases, our results are better than those of [49] for the same dataset. The  $\overline{AUC}$  results of DT, SVM-SMO and their ensembles of this work outperformed than those of DT, CBR and ANN in [49], but the performance of SVM-SMO was slightly worse. Previous researches for diagnostic prediction of digital mammography have not employed feature selection to improve the performance. In our experiment, the most results of 4-attributes outperformed than those of 5-attributes. Therefore, feature selection and ensemble methods we proposed in this study can effectively improve the performance in this problem.

## Conclusions

In this work, we compared two well-known feature selection methods (FS and BS) used in the mammographic mass dataset. The results show that *density* is irrelevant feature and the accuracy rates are better than those of adopting all original features. Feature selection is useful for improving the performance in our experiment. In addition, this study employed DT, SVM-SMO and their ensembles to solve breast cancer diagnostic problem in an attempt to predict accuracy with better performance. The results reveal

ensemble classifiers outperformed than single classifiers, hence, ensembles are also beneficial for clinicians in predicting breast cancer diagnosis. The proposed approaches, feature selection and ensemble methods, by this study really improve the predictive performance for the mammograms dataset.

**Acknowledgement** The authors like to express our appreciations to Prof. Gordon Turner-Walker for his help in correcting earlier versions of this paper. We also would like to thank the anonymous reviewers for their valuable comments and insightful suggestions.

## References

- Razavi, A. R., Gill, H., Åhlfeldt, H., and Shahsavari, N., Predicting metastasis in breast cancer: comparing a decision tree with domain experts. *J. Med. Syst.* 31:263–273, 2007.
- Brenner, H., Long-term survival rates of cancer patients achieved by the end of the 20th century: a period analysis. *Lancet.* 360:1131–1135, 2002.
- Nystrom, L., Andersson, I., Bjurstam, N., Frisell, J., Nordenskjöld, B., and Rutqvist, L. E., Long-term effects of mammography screening: updated overview of the Swedish randomised trials. *Lancet.* 359(9310):909–919, 2002.
- Bjurstam, N., Bjorneld, L., Warwick, J., Sala, E., Duffy, S. W., Nyström, L., et al., The Gothenburg breast screening trial. *Cancer.* 97(10):2387–2396, 2003.
- Rijnsburger, A. J., van Oortmarssen, G. J., Boer, R., Draisma, G., Miler, A. B., et al., Mammography benefit in the Canadian National Breast Screening Study-2: a model evaluation. *Int. J. Cancer.* 110(5):756–762, 2004.
- Carney, P. A., Miglioretti, D. L., Yankaskas, B. C., Kerlikowske, K., Rosenberg, R., Rutter, C. M., et al., Individual and combined effects of age, breast density, and hormone replacement therapy use on the accuracy of screening mammography. *Ann. Intern. Med.* 138(3):168–175, 2003.
- Pisano, E. D., Gattonis, C., Hendrick, E., Yaffe, M., Baum, J. K., Acharyya, S., et al., Diagnostic performance of digital versus film mammography for breast-cancer screening. *N. Engl. J. Med.* 353:1773–1783, 2005.
- Pisano, E. D., Gattonis, C., Hendrick, E., Yaffe, M., Baum, J. K., Acharyya, S., et al., Diagnostic accuracy of digital versus film mammography: exploratory analysis of selected population subgroups in DMIST. *Radiology.* 246(3):376–383, 2008.
- Kulkarni, A. D., *Computer Vision and Fuzzy-Neural Systems*. Prentice-Hall, Englewood-Cliffs, 2001.
- Karssemeijer, N., Adaptive noise equalization and recognition of microcalcification clusters in mammograms. *Int. J. Pattern. Recog. Artificial. Intell.* 7(6):1357–1376, 1993.
- Priebe, C. E., Lorey, R. A., Marchette, D. J., Solka, J. L., and Rogers, G. W., Nonparametric spatio-temporal change point analysis for early detection in mammography. In: Gale, A. G., Astley, S. M., Dance, D. R., and Cairns, A. Y. (Eds.), *Digital mammography*. Elsevier, Amsterdam, pp. 111–120, 1994.
- Heine, J. J., Deans, S. R., Cullers, D. K., Stauduhar, R., and Clarke, L. P., Multiresolution statistical analysis of high-resolution digital mammograms. *IEEE. Trans. Med. Imaging.* 5(16):503–515, 1997.
- Rakowski, W., and Clark, M. A., Do groups of women aged 50–75 match the national average mammography rate? *Am. J. Prev. Med.* 15(3):187–197, 1998.



14. Chhatwal, J., Alagoz, O., Lindstrom, M. J., Kahn, C. E., Jr., Shaffer, K. A., and Burnside, E. S., A logistic regression model based on the national mammography database format to aid breast cancer diagnosis. *Am. J. Roentgenol.* 192(4):1117–1127, 2009.
15. Sameti, M., and Ward, R. K., A fuzzy segmentation algorithm for mammogram partitioning. In: Doi, K., Giger, M. L., Nishikawa, R. M., and Schmidt, R. A. (Eds.), *Third international workshop on digital mammography*. Elsevier, Amsterdam, pp. 471–474, 1996.
16. Qian, W., Sunden, P., Sjostrom, H., Fenger-Krog, H., and Brodin, U., Comparison of image quality for different digital mammogram image processing parameter settings versus analogue film. *Electromedica.* 71(1):2–6, 2003.
17. Verma, B., and Zakos, J. A., Computer-aided diagnosis system for digital mammograms based on fuzzy-neural and feature extraction techniques. *IEEE T. Inf. Technol. Biomed.* 5(1):46–54, 2001.
18. Acharya, U. R., Ng, E. Y. K., Chang, Y. H., Yang, J., and Kaw, G. J. L., Computer-based identification of breast cancer sing digitized mammograms. *J. Med. Syst.* 32(6):499–507, 2008.
19. Rafayah, M., Qutaishat, M., and Abdallah, M., Breast cancer diagnosis system based on wavelet analysis and fuzzy-neural. *Expert. Syst. Appl.* 28(4):713–723, 2005.
20. Verma, B., and Panchal, R., Neural networks for the classification of benign and malignant patterns in digital mammograms. In: Fulcher, J. (Ed.), *Advances in applied artificial intelligence*. Idea Group, USA, 2006.
21. Brijesh, B., Novel network architecture and learning algorithm for the classification of mass abnormalities in digitized mammo-grams. *Artif. Intell. Med.* 42(1):67–79, 2008.
22. Li, Y., and Jiang, J., Combination of SVM knowledge for microcalcification detection in digital mammograms. *Lect. Notes Comput. Sci.* 3177:359–365, 2004.
23. de Oliveira Martins, L., Junior, G. B., Correa Silva, A., de Paiva, A. C., and Gattass, M., Detection of masses in digital mammo-grams using K-means and support vector machine. *Electron. Lett. Comput. Vis. Image. Ana.* 8(2):39–50, 2009.
24. Yang, J., and Olafsson, S., Optimization-based feature selection with adaptive instance sampling. *Comput. Oper. Res.* 33 (11):3088–3106, 2006.
25. Rodriguez, J. J., Kuncheva, L. I., and Alonso, C. J., Rotation forest: a new classifier ensemble method. *IEEE Trans. Pattern Anal. Mach. Intell.* 28(10):1619–1630, 2006.
26. Kuncheva, L. I., *Combining pattern classifiers: methods and algorithms*. Wiley-IEEE Press, New York, 2004.
27. Schapire, R. E., The strength of weak learnability. *Mach. Learn.* 5 (2):197–227, 1990.
28. Mitchell, T., *Machine learning*. McGraw-Hill, New York, 1997.
29. Witten, I. H., and Frank, E., *Data mining: practical machine learning tools with java implementations*. Morgan Kaufmann, San Francisco, 2000.
30. Razavi, A.R., Gill, H., Åhlfeldt, H., and Shahsavari, N.: A data pre-processing method to increase efficiency and accuracy in data mining. In: Miksch, S., Hunter, J., Keravnou, E. (eds.) *10th Conference on Artificial Intelligence in Medicine*. Springer-Verlag GmbH, Aberdeen, pp. 434–443, 2005.
31. Quinlan, J. R., *C4.5: Programs for machine learning*. CA: Morgan Kaufmann, San Mateo, 1993.
32. Vapnik, V. N., *The nature of statistical learning theory*. Springer, Berlin, 1995.
33. Platt, J.C.: Sequential minimal optimization: a fast algorithm for training support vector machines. Technical Report MSR-TR-98-14, Microsoft Research, 1998.
34. Melville, P., and Monney, R. J., Creating diversity in ensembles using artificial data. *Inf. Fusion.* 6(1):99–111, 2005.
35. Schapire, R. E., Freund, Y., Bartlett, P. L., and Lee, W. S., Boosting the margin: a new explanation for the effectiveness of voting methods. *Ann. Statist.* 26(5):1651–1686, 1998.
36. Breiman, L., Random forests. *Mach. Learn.* 45(1):5–32, 2001.
37. Kim, H. C., Pang, S., Je, H. M., Kim, D., and Bang, S. Y., Constructing support vector machine ensemble. *Pattern. Recognit.* 36(12):2757–2767, 2003.
38. Valentini, G., and Dietterich, T. G., Low bias bagged support vector machines. In: Fawcett, T., and Mishra, N. (Eds.), *International conference on machine learning*. AAAI press, California, 2003.
39. Breiman, L., Bagging predictors. *Mach. Learn.* 24(2):123–140, 1996.
40. Freund, Y., and Schapire, R. E., A decision-theoretic generalization of on-line learning and an application to Boosting. *J. Comput. Syst. Sci.* 55(1):119–139, 1997.
41. Zhang, C. X., Zhang, J. S., and Zhang, G. Y., An efficient modified Boosting method for solving classification problems. *J. Comput. Appl. Math.* 214(2):381–392, 2008.
42. Webb, G. I., MultiBoosting: a technique for combining Boosting and wagging. *Mach. Learn.* 40(2):159–197, 2000.
43. Delen, D., Walker, G., and Kadam, A., Predicting breast cancer survivability: a comparison of three data mining methods. *Artif. Intell. Med.* 34:113–127, 2005.
44. Centor, R. M., Signal detectability: the use of ROC curves and their analyses. *Med. Decis. Mak.* 11:102–106, 1991.
45. Hanley, J. A., and McNeil, B., The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology.* 143(1):29–36, 1982.
46. DeLong, E. R., DeLong, D. M., and Clarke-Pearson, D. L., Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics.* 44:837–845, 1988.
47. Newmann, D.J., Hettich, S., Blake, C.L., Merz, C.J.: UCI repository of machine learning database. <http://archive.ics.uci.edu/ml/datasets/Mammographic+Mass>, Irvine, CA: University of California, Department of Information and Computer Science, (1998)
48. Kopans, D. B., D’Orsi, C. J., Adler, D. D., et al., *Breast Imaging Reporting and Data System (BIRADS)*. American College of Radiology, Reston, 1993.
49. Elter, M., Wendtland, R. S., and Wittenberg, T., The prediction of breast cancer biopsy outcomes using two CAD approaches that both emphasize an intelligible decision process. *Med. Phys.* 34 (11):4164–4172, 2007.
50. Zhang, G. P., Neural networks for classification: a survey. *IEEE Trans. Syst. Man Cybern. Part C Appl. Rev.* 30(4):451–462, 2000.
51. Zangwill, L. M., Chan, K., Bowd, C., Hao, J., Lee, T. W., Weinreb, R. N., et al., Heidelberg retina tomograph measurements of the optic disc and parapapillary retina for detecting glaucoma analyzed by machine learning classifiers. *Invest. Ophthalmol. Vis. Sci.* 45(3):3144–3151, 2004.