KKU Res. J. 2012; 17(1):1-13
http : // resjournal.kku.ac.th

Bagging Random Tree for Analyzing Breast Cancer Survival

Jaree Thongkam[1] * and Vatinee Sukmak[2]

[1,2]*Faculty of Informatics, Mahasarakham University, Mahasarakham, Thailand*
**Corresponding author: jaree.thongkam@gmail.com*

Received july 26,2011
Accepted january 26,2612

Abstract

Building the survivability prediction models is a challenging task because they provide an important approach to assessing risk and prognosis. In this paper, we investigated the performance of combining of the Bagging with several weak learners to build 5-accurate breast cancer survivability prediction models from the Srinagarind hospital database in Thailand. These models could assist medical professional in monitoring survival rates and up-to-date estimations of long-term survival rates. In order to evaluate the performance of models, Area Under the receiver operating characteristic Curve (AUC), sensitivity and specificity were employed. Moreover, the stratified 10-fold cross-validation was utilized to reduce the bias of experiments. Experimental results showed that combining the Bagging with Random Tree is superior to Bagging with weak learning including Decision Stump, REPTree and J48, and single weak learning alone.

Keywords : data mining, Bagging, breast cancer survivability

1. Introduction

Cancer is an abnormal cell which become a major cause of death and hardly prevented (1, 2). In 2003-2007, breast cancer has been the most common cause of cancer death in women. The death rates per 100,000 women are 23.4 in white women, 32.4 in black and 12.2 in Asia/Pacific Islander (1). Analyzing the survival rate is the main concern for estimating a particular patient suffering from a disease over a particular time period in the case of prognosis (3, 4). Kaplan-Meier and Cox-Proportional hazard are the traditional tools for analyzing the survival rate (4, 5). In order to develop a better service to patients, classification in data mining is an alternative tool which can achieve better results (5-8). Classification such as Decision Tree, Rule-Based and Neural Networks has become one of the most widely used techniques. They are managed to extract models describing important data classes and to predict future data classes (9).

In the field of data mining for medical prognosis, many research studies have made use of these classification techniques to build prediction models. For instance, Delen, Walker and Kadam (10) used decision tree (C5) to build the 5-year breast cancer survivability prediction model from a large data set. Their results showed that C5 is superior to Artificial Neural Networks (ANN) and logistic regression in terms of accuracy, sensitivity and specificity. Similarly, Jonsdottir, Hvannberg, Sigurdsson and Sigurdsson (11) presented that the decision tree classifier (C4.5) outperforms the Naïve

Bayesian classifier in a small data set. Nevertheless, few research studies have applied Bagging to build a prediction model, since it is the most intuitive and simplest ensemble models with a surprisingly good performance (12-14). Also, Bagging can be exploited to combine with several learners (classifiers) to reach the higher prediction outcomes (15).

In this paper we investigated the generalization performance of Bagging with weak learners and single weak learners based on decision tree techniques in order to enhance the prediction models for decision-making system in the prognosis of 5-year breast cancer survivability.

The paper is structured as follows: Section 2 reviews the basic concepts of Bagging, Random Tree, Decision stump, REPTree and J48. Section 3 presents the methodologies and experimental design applied in this paper. The experimental results are explained in Section 4. The discussions and conclusion are given in Sections 5 and 6, respectively.

2. Classification Techniques

This section reviews a brief concept of classification techniques including Bagging, Random Tree, Decision Stump, REPTree and J48.

2.1. Bagging

Bagging is an ensemble of classifiers collected from several classifiers which are combined to classify data in a data set (15). This technique creates classifiers and adjusts each classifier on a randomly drawn training set with the probability of drawing any given instance being equal. Besides, instances drawn with replacement resulting in some instances may be selected multiple times while others may not be selected at all. In this way, each classifier could return a higher test set error than a classifier using all of the data. In addition, when these classifiers are combined, the resulting ensemble produces lower test set error than a single classifier. Some researchers have successfully employed the Bagging technique to improve the prediction models. For instance, Blanco, Ricket and Martín-Merino (16) have achieved in combining Bagging technique with SVM classifiers and reducing the variance of the test error in email anti-spam filtering. On the other hand, Buciu, Kotropoulos and Pitas (17) demonstrated that Bagging unable to improve accuracy of the base model.

2.2.Random Tree

Random Tree is a weak machine learning model which constructs a tree considered K randomly chosen attributes at each node (18). Also it performs no pruning and has an option to allow estimation of class probabilities based on a hold-out set. Moreover, it is suited with binary classes, missing class values and nominal classes. However, it seems that on its own tends to be too weak in classification problems (19). Besides, a few researchers have employed this technique for classifying their data. For instance, Chen, Shou, Hu and Guo (19) demonstrated that accuracy of Random Tree is better than REPTree, PART, NaiveBayes, RBFNetwork. However, it is lower than C4.5 in identifying traffic in the broadband network.

2.3. Decision Stump

A Decision Stump is a weak learner which builds simple binary decision 'stumps' (1 level decision tress) for both numeric and nominal classification problems (18). It handles mission values by extending a third branch from the stump or treating the missing values as a separate attribute value. It is usually utilized with a boosting algorithm and regression for classification purposes. It is not commonly used on its own, since very few problems can be accurately classified using a single feature.

2.4.REPTree

REPTree is a fast decision tree learner used to build a decision or regression tree models (20). It prunes only sort values for numeric attributes once with back-fitting. Furthermore, it can deal with missing

values utilizing splitting the corresponding instances into pieces which is similar to C4.5. Moreover, it commonly combines with Bagging (18). Few research studies have demonstrated the performance of REPTree. This is because REPTree achieved lower accuracy than C4.5 for identifying the traffic via broadband network (19).

2.5. J48

J48 is a Java implementation of C4.5 which is a classic decision tree algorithm in machine learning (21). It is used to build a tree structure for classifying a data set related to a class attribute consisting of nodes and leaves (22-24). It employs Gain Ratio for selecting the best attribute from instances before applying the top-down greedy strategy to build a tree. In this way, models generated from C4.5 are easy to interpret from a tree structure and only needs a short computation time (24-26). Therefore, much research has utilized C4.5 to build the prediction models. For instance, Yao, Liu, Lei and Yin (25) successfully exploited C4.5 to build prediction models. However, it has limitation in overfitting and time consuming in computation (25, 26).

3. Methodologies

In order to build and interpret the breast cancer survivability models, background of breast cancer survivability is reviewed. The data preparation and pre-processing steps are presented in order to understand the source of data sets.

3.1.Breast cancer survivability

In the breast cancer context, "survival" is the length of time lived after the initial diagnosis of cancer (27). Similarly, Delen et al. (10) denoted "survival" as a patient remaining alive for a specified period of time after the diagnosis of cancer. Currently, many research studies have utilized five years period to analyze survivability of the patient. This may be due to the fact that the improvement of early detection and treatments, death as a result of breast cancer has been gradually decreased in the recent years (1).

In relation to the attributes utilized in the breast cancer survivors, many research studies found that age of the first diagnosis is a risk factor that increases the probability of a woman to develop the breast cancer (28). Unmarried patients with cancer have decrease overall survival (29). Basis of diagnosis is an attribute that related to treatments to the reduction in breast-cancer mortality (30). Therefore, the analysis of the relationship of each attribute in a field of medical prognosis can assist medical practitioners for patient management in predicting survival time for breast cancer patients.

3.2.Data preparation

In this paper, breast cancer data sets were obtained from Srinagarind Hospital. This hospital is the medical school hospital in the Northeastern Thailand established in 1972 as a part of the faculty of medicine at Khon Kaen University. The data set consist of 14 attributes from 1985-2002. The attributes are shown in Table 1.

Table 1 shows the attribute names in used and types of attributes in this paper. These attributes were chosen as the powerful prognostic factors identified in most studies. Topography attribute consists of nine values which point out the position of cancer in breast that related to the choice of treatments. Moreover, the extent of disease is aggregated attribute with morphology to see the patterns related to the breast cancer survival periods. The state of breast caner is a chronological factor based on spread to other areas. The class attribute is composed of two classes including 'Dead' and 'Alive'. The 'Dead' class refers to patients who died within five years following the diagnosis. On the other hand, the 'Alive' class refers to patients who have survived for five years or more after the diagnosis. The initial numbers of instances include 466 instances for the "Dead" class and 392 instances for the "Alive" class.

Table 1. Input attributes

Item No.	Attributes	Used Names	Types	Values	Value Names
1	Age of first diagnosis	Age	Number		
2	Marital status	Mars	Category(3)	1	Single
				2	Married
				3	Non
3	Basis of diagnosis	Basis	Category(6)	1	History & Physical exam.
				2	Endoscopy & Radiology
				3	Surgery & Autopsy (no histology)
				5	Cytology or Hematology
				6	Histology of Metastasis
				7	Histology of Primary
4	Topography	Top	Category(9)	500	C50.0 Nipple
				501	C50.1 Central portion of breast
				502	C50.2 Upper-inner quadrant of breast
				503	C50.3 Lower-inner quadrant of breast
				504	C50.4 Upper-outer quadrant of breast
				505	C50.5 Lower-outer quadrant of breast
				506	C50.6 Axillary tail of breast
				508	C50.8 Overl. lesion of breast
				509	C50.9 Breast, NOS
5	Morphology	Mor	Category(14)	8000	Neoplasm
				8001	Tumor cells
				8010	Epithelial tumor
				8041	Small cell carcinoma, NOS
				8070	Squamous cell carcinoma, NOS
				8140	Adenocarcinoma, NOS
				8480	Mucinous adenocarcinoma
				8500	Infiltrating duct carcinoma
				8501	Comedo carcinoma, NOS
				8510	Medullary carcinoma, NOS
				8520	Lobular carcinoma, NOS
				8530	Inflammatory carcinoma
				8541	P. dis. & infil. duct carc.,breast
				8800	Soft tissue tumor
6	Stage	Stage	Category(4)	1	Stage I
				2	Stage II
				3	Stage III
				4	Stage IV
7	Extent	Ext	Category(4)	2	Localized
				3	Direct extension
				4	Regional lymph nodes
				5	Distant metastases
8	Received surgery	Surg	Category(2)	1	Received treatment
				2	Do not Received treatment
9	Received radiation	Radi	Category(2)	1	Received treatment
				2	Do not Received treatment
10	Received chemothe-rapy	Chem	Category(2)	1	Received treatment
				2	Do not Received treatment
11	Received hormone	Horm	Category(2)	1	Received treatment
				2	Do not Received treatment
12	Received Supportive	SupT	Category(2)	1	Received treatment
				2	Do not Received treatment
13	Received Others	Other	Category(2)	1	Received treatment
				2	Do not Received treatment
14	Survivability	Classes	Category(2)	0	Dead
				1	Alive

3.3.Data Pre-processing

Pre-process is an important step in data mining. It is used to improve the data quality by eliminating outliers, balancing classes and selecting attributes (31). In this paper, we investigated the prediction models generated from Bagging with weak learners and single weak learners after exploiting three steps of pre-processing in Figure 1.

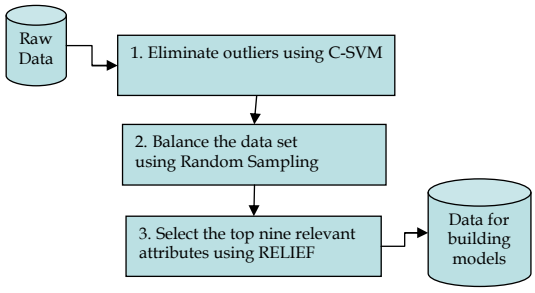

Figure 1. Pre-processing steps

Figure 1 displays the three pre-processing steps as follows:

1) Apply C-Support Vector Classification filtering (C-SVCF) to identify and eliminate outliers from both 'Dead' and 'Alive' classes. As a result, the numbers of instances corresponding 5-year breast cancer survivability data sets comprise instances being 368 instances for the "Dead" class and 250 instances for the "Alive" class.
2) Utilize random sampling to increase the size of the minority class to the same size of the majority class by using the ratio between the majority and minority classes. Consequently, the numbers of instances corresponding 5-year breast cancer survivability data sets contain instances being 368 instances for the "Dead" class and 367 instances for the "Alive" class.
3) Select the top nine relevant attributes arranged using RELIEF selecting the relevant attributes for the input data set based on the weighting scores. Much research study has utilized RELIEF to select appropriate attributes in their data. For example, Hall and Holmes (32) presented that C4.5 achieves a higher performance after applying RELIEF for selecting dependent attributes. Similarly, Vu, Ohn and Kim (33) demonstrated that RELIEF algorithm achieves better sensitivity and specificity than T-test in an ovarian 8-7-02 data set but lower than T-test in an ovarian 4-3-02 data set using radial SVM for classifying the cases.

Therefore, the final attributes in the data set comprise nine attributes including extent, stage, age, morphology, received radiation, topography, basis of diagnosis, received surgery and marital status. In comparison, Brenner et al. (6) selected the 11 attributes based on the relevant cancer literature including age, sex, year of diagnosis, month of diagnosis, year of end of follow-up, month of end of follow-up, vital status at the end of follow-up, length of follow-up, first calendar year of period of interest, last calendar year of period of interest, and dimension array as the attributes in their work. Delen et al. (10) utilized only the completed instances in the attributes. Their attributes included race, marital status, primary site code, histology, behavior, grade, extension of disease, lymph node involvement, radiation, stage of cancer, site specific surgery code, age, tumor size, number of positive nodes, number of nodes, number of primaries and classes.

3.4.Evaluation Methods

WEKA version 3.6.5 (20) was utilized as a data mining tool to evaluate the performance and effectiveness of the 5-breast cancer prediction models built from several techniques. This is because the WEKA program offers a well defined framework for experimenters and developers to build and evaluate their models. The parameters used in each model are utilized by each technique as follows:

1) Bagging uses four parameters including 100% size of each bag of the training set size, calculate out of bag sets to "False", 10 iterations and zero random seed.
2) Random Tree uses six parameters including 0 of the K value, no allow of un-classification field instances, 0 maximum depth, 1 minimum; 0 number of folds and one random seed.
3) Decision Stump does not need to set any parameter.
4) REPTree uses four parameters including no restriction for the maximum depth tree, two minimum total weights of the instances in a leaf, 0.0010 minimum proportion of the variance on all the data using three folds of regression trees and one random seed.
5) J48 uses four parameters including 0.25 confidence factor for pruning, two numbers of instances per leaf, three amounts of data used for reduced-error pruning and one random seed.

Moreover, stratified 10-fold cross-validation is employed to select data into training and test sets for minimizing bias and variance associated with the random sampling (34). In this way, both classes in the training and test sets have an approximately equal rating to the original data set. In this study, three evaluation methods including AUC, sensitivity and specificity are applied based on a confusion matrix in a matrix representation of the prediction results (see the Figure 2).

		Predicted Classes	
		'Dead'	'Alive'
Outcomes	'Dead'	*TP*	*FN*
	'Alive'	*FP*	*TN*

Figure 2. The confusion matrix

As represented in Figure 2 above, the confusion matrix is used to compute true positive (TP) which refers to the number of correct predictions in a positive class; false positive (FP) which refers to the number of incorrect predictions in a positive class; true negative (TN) which refers to the number of correct predictions in a negative class; and false negative (FN) which refers to the number of incorrect predictions in a negative class.

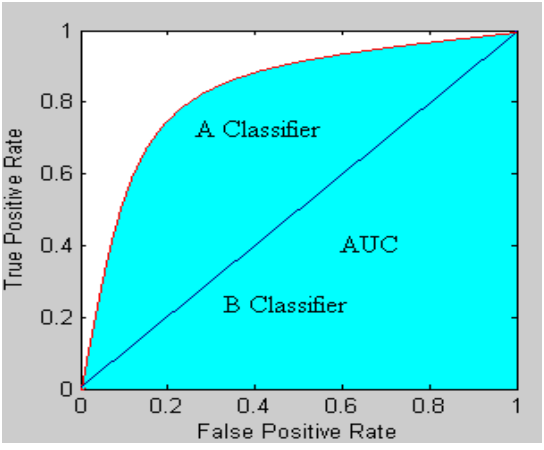

Figure 3. The area under the ROC curve

1) AUC

Area Under the receiver operating characteristic Curve (AUC) recently has been proposed as an alternative measurement criterion for evaluating the predictive ability of learning algorithms by randomly selecting the instance of one class which has a smaller estimated probability among other classes (35) (36). The AUC of A and B classifiers is exhibited in Figure 3 below.

Figure 3 shows that the AUC of the A classifier is larger than the B classifier, meaning that A classifier is better than the B classifier. Each line is a relative trade-offs between true positive and false positive. Besides, it can be interpreted into a numeric which has scores between 0 and 1. However, this study will present in form of percentage of the score for easier to point out results. Many research studies have utilized AUC scores for comparing classifiers' performance (37-39). In addition, Huang and Ling (38) found that AUC is a more accurate measurement method than the ROC curve.

2) Sensitivity

Sensitivity is the evaluation method in binary classification problems to indicate the performance of predicting the death cases. It refers to the true positive rate which has a formula as follows.

$$sensitivity = \frac{TP}{TP + FN} \quad (1)$$

Sensitivity of prediction models can be interpreted into percentage which has value between 1 and 100. One hundred percent refers to the best predicting model while 0 refers to the worst predicting model.

3) Specificity

Specificity is the evaluation method in binary classification problems to indicate the performance of predicting the living cases.

$$specificity = \frac{TN}{TN + FP} \quad (2)$$

Prediction model specificity can be interpreted into percentage which has value between 0 and 100. One hundred percent refers to the best predicting model while 0 refers to the worst predicting model.

4. Experimental Results

In this paper, the prediction models generated from Bagging with Random Tree (B+RT), Decision Stump (B+DS), REPTree (B+REPT) and J48 (B+J48), and weak learners (Random Tree (RT), Decision Stump (DS), REPTree (REPT) and J48) were evaluated based on AUC, sensitivity and specificity. The results were provided using 10 times of 10-fold cross validation for each model. Also the average results obtained from the 10 test sets for each fold.

4.1.AUC Comparison

Area Under the receiver operating characteristic Curve (AUC) is an effective evaluation method. In this section it is employed to evaluate the performance of 5-year breast cancer survivability prediction models. Each fold of AUC of prediction models generated from B+RT, B+DS, B+REPT, B+J48, RT, DS, REPT and J48 is displayed and the AUC average of the prediction models is showed in Figures 4 and 5, respectively.

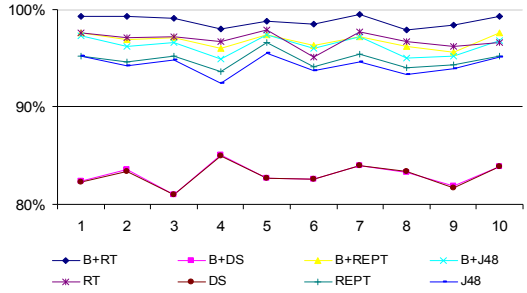

Figure 4. The AUC of the prediction models

Figure 4 demonstrates the results of the AUC score in details. The experimental results showed that Bagging with Random Tree can achieve the highest AUC score in every fold. On the other hand, RT seems to have a lower AUC score than B+RT. This is because Bagging can select suitable instances for Random Tree to build the breast cancer survivability prediction model. Besides, AUC scores of most weak learners had improved after making use of Bagging except Decision. This may be due to the fact that the data set was too small to extend a third branch. Nevertheless, B+RT is superior to B+DS, B+REPT, B+J48, RT, DS, REPT and J48.

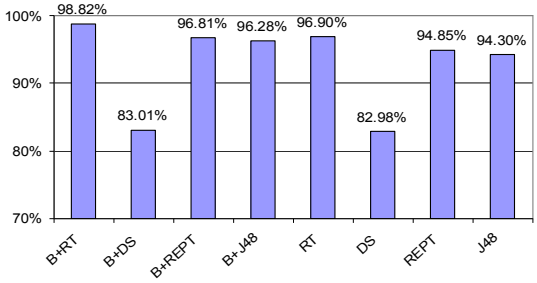

Figure 5. The average of AUC of the prediction models

Figure 5 illustrates the average of 5-year breast cancer survivability AUC score generated from B+RT, B+DS, B+REPT, B+J48, RT, DS, REPT and J48. The results exhibited that B+RT achieved the highest AUC score up to 98.82%. Following this B+REPT has the AUC score up to 96.81%, B+J48 has the AUC score up

to 96.28%, and B+DS has the AUC score up to 83.01%. Therefore, B+RT is better than B+DS, B+REPT, B+J48, RT, DS, REPT and J48 based on the average AUC score.

4.2.Sensitivity Comparison

Sensitivity is used to evaluate the performance of prediction models which predict the dead cases. Each fold of sensitivity of 5-year breast cancer survivability prediction models generated from Bagging with weak learners and single weak learners is displayed in Figure 5. Also average of sensitivity is displayed in Figures 6 and 7.

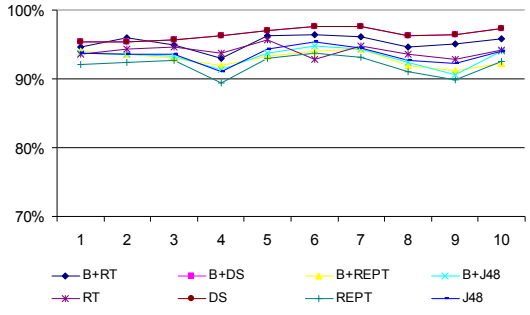

Figure 6. The sensitivity of the prediction models

Figure 6 shows 10 folds of sensitivity generated from B+RT, B+DS, B+REPT, B+J48, RT, DS, REPT and J48. The results indicated that most prediction models have a similar movement in each fold. Although, B+DS and DS are likely on the top of the graph, B+RT achieved the highest sensitivity in the second fold.

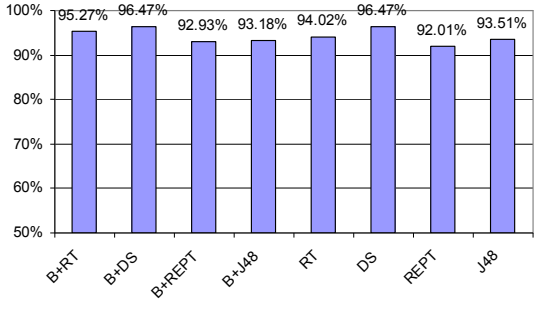

Figure 7. The average of sensitivity of the prediction models

Figure 6 displays the average of sensitivity of 5-year breast cancer survivability prediction models generated from Bagging with four weak learners and four single weak learners. The results exhibited that the average sensitivities of B+RT, B+DS, B+REPT, B+J48, RT, DS, REPT and J48 are up to 95.27%, 96.47%, 92.93%, 93.18%, 94.02%, 96.47%, 92.01% and 93.51%, respectively. These indicated that Bagging can improve some prediction models.

4.3. Specificity Comparison

Specificity is employed to evaluate the performance of prediction models which predict the patient still alive more than 5 years after the first diagnosis. Ten folds and average of specificity of 5-year breast cancer survivability prediction models generated from Bagging with weak learners and single weak learners are displayed in Figures 8 and 9.

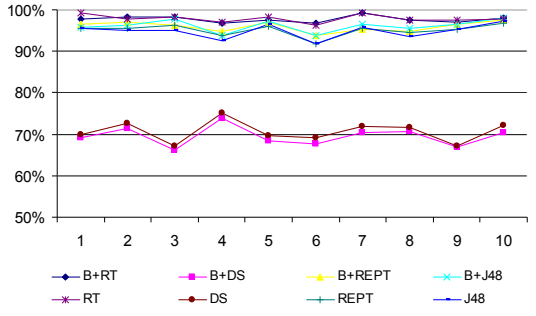

Figure 8. The specificity of the prediction models

Figure 8 presents 10 folds of the specificity of 5-year breast cancer survivability prediction models based on B+RT, B+DS, B+REPT, B+J48, RT, DS, REPT and J48. The experimental results pointed out that most prediction models have similar results. Even though B+RT and RT can manage to have the highest specificity, RT is mostly achieved the highest specificity.

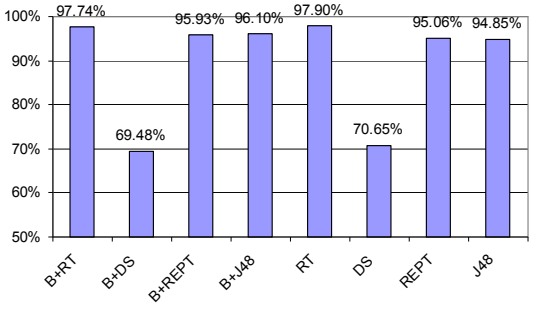

Figure 9. the average of specificity of the prediction models

Figure 9 illustrates the average results of prediction models specificity generated from B+RT, B+DS, B+REPT, B+J48, RT, DS, REPT and J48. B+RT, B+REPT, B+J48, RT, REPTree and J48 achieved significant results in specificity up to 97.74%, 95.93%, 96.10%, 97.90%, 95.06% and 94.85%, respectively. Besides, B+DS and DS have insignificant results in specificity. This may be due to the fact that DS only concentrates the positive class. These results pointed toward that RT outperforms B+RT, B+DS, B+REPT, B+J48, DS, REPT and J48.

4.4.Performance comparison in difference attributes

In order to evaluate the performance of models generated from the data sets which have 14 and 10 attributes, the average of the AUC score, sensitivity and specificity of each data set was employed. The average of the AUC score, sensitivity and specificity of 5-year breast cancer survivability prediction models generated from Bagging with weak learners and single weak learners is illustrated in Table 2.

Table 2. Performance comparisons in both data sets with 14 and 10 attributes

	Performance in 14 attributes			Performance in 10 attributes		
Classifiers	AUC	Sensitivity	Specificity	AUC	Sensitivity	Specificity
B+RT	98.80	95.22	97.47	98.82	95.27	97.74
B+DS	82.93	96.47	69.48	83.01	96.47	69.48
B+REPT	96.92	93.07	96.13	96.81	92.93	95.93
B+J48	96.26	93.32	95.67	96.28	93.18	96.10
RT	96.45	93.51	97.44	96.90	94.02	97.90
DS	82.98	96.47	69.48	82.98	96.47	70.65
REPT	94.84	92.01	95.04	94.85	92.01	95.06
J48	94.24	93.29	94.96	94.30	93.51	94.85
Average	92.93	94.17	89.46	*92.99*	*94.23*	*89.71*

Table 2 presents the results of the prediction models generated from B+RT, B+DS, B+REPT, B+J48, RT, DS, REPT and J48. The results showed that the average of the AUC, sensitivity and specificity of the prediction models of 14 attributes is 92.93%, 94.17% and 89.46%, respectively. Still, the average of the AUC, sensitivity and specificity of the prediction models of 10 attributes is 92.99%, 94.23% and 89.71%, respectively. This indicated that using the 10 suitable attributes is slightly better than using the 14 attributes.

4.5.Bagging Random Tree model

The Bagging Random Tree model is a decision tree model consisting of nodes and leaves. Nodes represent rules categorizing data according to attributes and leaves represent the condition in each rule. The model also provides the re-substitution error rate in each leaf. This error rate is the relationship between the number of the incorrect cases (E) and training cases covered by the leaf (N). The re-substitution error rate is shown in Equation 3.

$$\text{Re-substitution error rate} = E/N. \quad (3)$$

In this way, the B+RT decision tree model is easy to interpret from a tree structure. In this section, B+RT technique produces 10 decision trees. Each iteration comprises 188, 170, 188, 238, 153, 214, 277, 207, 223 and 196 leaves, respectively. In order to interpret the model generated from B+RT, the top 10 leaves without 0 error rate of the first tree are given in Figure 10.

Figure 10 demonstrates the top 10 leaves of the B+RT model. The main root of this decision tree is the extents of breast cancer including localized, direct extension, regional lymph nodes and distant metastases. The interpretation of this model is presented below.

```
ext = 2 : 1 (43/0)
ext = 3
|   mor = 8000
|   |   status = 1 : 1 (5/0)
|   |   status = 2
|   |   |   top = 504 : 0 (1/0)
|   |   |   top = 509
|   |   |   |   age < 41 : 1 (1/0)
|   |   |   |   age >= 41 : 0 (2/0)
|   mor = 8041 : 1 (1/0)
|   mor = 8070 : 0 (1/0)
|   mor = 8140 : 1 (3/0)
|   mor = 8500
|   |   age < 51.5 : 1 (158/0)
|   |   age >= 51.5
|   |   |   age < 67
|   |   |   |   age < 52.5
|   |   |   |   |   top = 504 : 0 (2/0)
|   |   |   |   |   top = 509 : 1 (2/0)
```

Figure 10. An example of the B+RT model

1) If a patient has localized extent ('ext' = '2') at the first diagnosis then this patient is predicted to live for five years or longer after the first diagnosis with 43/0 re-substitution error rates.
2) If a patient has direct extension ('ext' = '3'), has neoplasm ('mor'='8000') and is single then this patient is predicted to live for five years or more after the first diagnosis with 5/0 re-substitution error rates.
3) If a patient has direct extension ('ext' = '3'), has neoplasm ('mor'='8000'), is married and has upper-outer quadrant of breast ('top' = '504') then this patient is predicted to live for five years or more after the first diagnosis with 1/0 re-substitution error rates.

5. Discussions

In the field of medical prognosis, survival rate analysis commonly uses clinical data for predicting the survival of particular patients suffering from diseases over particular time periods (6, 40). In this paper, the performance of 5-year breast cancer prediction models utilizing Bagging with weak learners and single weak learners was investigated. AUC, sensitivity and specificity calculated from confusion matrix are exploited to evaluate the models. The several findings are presented below.

Firstly, we found that the Random Tree models have the highest generalization performance (AUC, sensitivity and specificity). However, when combining DS or REPT or J48 with Bagging, it seemed that they achieved a higher performance than a single weak learner when evaluate with AUC and sensitivity. In congruent with Kotsiantis, Tsekouras and Pintelas (41) combining Bagging with M5 learner can increase the average accuracy from 25 data sets.

Secondly, combining Bagging with Random Tree performed slightly better than Random Tree alone based on specificity. However, it is significant better than Random Tree based on AUC scores. On the other hand, Das and Sengur (42) demonstrated that combing Bagging with Multi Layer Perceptron (MLP) has the same results with MLP alone based on sensitivity and specificity in the heart disease database.

Lastly, we found that combining Bagging with Random Tree achieved AUC up to 98.82%, while Adaboost with Random Forest used in the same the data set achieved better AUC up to 99.09% (43). Similarly, Banfield, Hall, Bowyer and Kegelmeyer (44) found that using Adaboost outperforms Bagging. This may be the fact that Adaboost selects the instances based on reducing error rate, whereas Bagging uses random sampling with replacement to reduce the bay and variance (45) (46).

6. Conclusion

In this paper, we proposed a combination of Bagging with Random Tree to construct 5-year breast cancer survivability prediction models for assisting the medical professional to improve survival rate. However, this method is not aimed at replacing the medical professional and researchers, but rather to complement their invaluable efforts to save more patient lives. Therefore, the patterns found via these studies should be evaluated

by medical practitioners. The performance of the combining Bagging with Random Tree was illustrated using 10 times of stratified 10-fold cross-validation, AUC, sensitivity and specificity. The results showed that this method provided the AUC, sensitivity and specificity up to 98.82%, 95.27% and 97.74%, respectively. As for further work, we plan to employ this approach to analyze in the prostate data sets in order to evaluate the reliable of this approach.

Acknowledgements

Special thanks to IT and Cancer Department staffs at Srinagarind Hospital for kindly providing the data.

References

(1) National Cancer Institute. Surveillance, Epidemiology, and End Results (SEER) Program Public-Use Data (1973-2008). Cancer Statistics Branch; 2011.

(2) Delen D, Patil N. Knowledge extraction from prostate cancer data. The 39thAnnual Hawaii International Conference on System Sciences; 2006; 1-10.

(3) Ryu YU, Chandrasekaran R, Jacob VS. Breast cancer prediction using the isotonic separation technique. The Conference of European Operational Research. 2007;181(2):842-54.

(4) Borovkova S. Analysis of survival data. Available from: http://www.math.leidenuniv.nl/~naw/serie5/deel03/dec2002/pdf/borovkova.pdf

(5) Ohno-Machado L. A comparison of Cox proportional hazards and artificial neural network models for medical prognosis. An International Journal of Computers in Biology and Medicine. 1997;27(1):55-65.

(6) Brenner H, Gefeller O, Hakulinen T. A computer program for period analysis of cancer patient survival. European Journal of Cancer 2002;38(5):690-5.

(7) Burke HB, Rosen DB, Goodman PH. Comparing artificial neural networks to other statistical methods for medical outcome prediction. The IEEE International Conference on Neural Networks: 1994.

(8) Burke HB, Goodman PH, Rosen DB, Henson DE, Weinstein JN, Jr. FEH, et al. Artificial neural networks improve the accuracy of cancer survival prediction Cancer. 1997;79:857-62.

(9) Skevofilakas MT, Nikita KS, Templaleksis PH, Birbas KN, Kaklamanos IG, Bonatsos GN. A decision support system for breast cancer treatment based on data mining technologies and clinical practice guidelines. The 27th Annual International Conferences on Medicine and Biology Society. 2005:2429-32.

(10) Delen D, Walker G, Kadam A. Predicting breast cancer survivability: a comparison of three data mining methods. The Journal of Artificial Intelligence in Medicine. 2005;34(2):113-27.

(11) Jonsdottir T, Hvannberg ET, Sigurdsson H, Sigurdsson S. The feasibility of constructing a predictive outcome model for breast cancer using the tools of data mining. The Journal of Expert Systems with Applications. 2008;34(1):108-18.

(12) Henderson JC, Brill E. Bagging and boosting a treebank parser. The First Conference on North American chapter of the Association for Computational Linguistics; 2000.

(13) Buhlman P, Yu B. Analyzing Bagging. The Annuals of Statistics. 2002;30(4)(4):927-61.

(14) Xu W, Zuo M, Zhang M, He R. Constraint bagging for stock price prediction using neural networks. The International Conference on Modelling, Identification and Control 2010:606-10.

(15) Breiman L. Bagging predictors. Machine Learning. 1996;24:123-40.

(16) Blanco Á, Ricket AM, Martín-Merino M. Combining SVM classifiers for email antispam filtering. In: Sebastián S, editor. The 9th International Work-Conference on Artificial Neural Networks; 2007; 903-10.

(17) Buciu I, Kotropoulos C, Pitas I. Demonstrating the stability of support vector machines for classification. The Journal of Signal Processing. 2006;86(9):2364-80.

(18) Witten IH, Frank E, Trigg L, Hall M, Holmes G, Cunningham SJ. Weka: Practical machine learning tools and techniques with java implementations. The International Workshop: Emerging Knowledge Engineering and Connectionist-Based Information Systems; 1999; 192-6.

(19) Chen N, Shou G, Hu YH, Guo ZG. An experimental research of traffic identification algorithms in broadband network. The International Symposium on Computer Network and Multimedia Technology. 2009:1 - 4

(20) Witten IH, Frank E. Data mining: practical machine learning tools and techniques. 2 ed. San Francisco: Morgan Kaufmann; 2005.

(21) Quinlan JR. C4.5: programs for machine learning. San Mateo, California: Morgan Kaufmann. 1993.

(22) Han J, Kamber M. Data mining: concepts and techniques. 2nd. ed. San Francisco: Morgan Kaufmann, Elsevier Science; 2006.

(23) Ruggieri S. Efficient C4.5 [classification algorithm]. The IEEE Transactions on Knowledge and Data Engineering; 2002;438-44.

(24) Zhou Z-H, Jiang Y. Medical diagnosis with C4.5 rule preceded by artificial neural network ensemble. The IEEE Transactions on Information Technology in Biomedicine: 2003:37-42.

(25) Yao Z, Liu P, Lei L, Yin J. R-C4.5 decision tree model and its applications to health care dataset. The International Conference on Services Systems and Services Management. 2005; 1099-103.

(26) He P, Chen L, Xu X-H. Fast C4.5. The International Conference on Machine Learning and Cybernetics. 2007:2841-6.

(27) Clinical Best Practice. Breast Cancer in Australia: an overview. Available from: http://www.nbcc.org.au/bestpractice/statistics/

(28) Jerez-Aragones JM, Gomez-Ruiz JA, Ramos-Jimenez G, Munoz-Perez J, Alba-Conejo E. A combined neural network and decision trees model for prognosis of breast cancer relapse. The Journal of Artificial Intelligence in Medicine. 2003;27(1):45-63.

(29) James s.Goodwin, William C. Hunt, Charles R. Key, Samet JM. The effect of marital staus on stage, treatment and survival of cancers patients. JAMA. 1987;258(21):3125-30.

(30) Berry DA, Cronin KA, Plevritis SK, Fryback DG, Clarke L, Zelen M, et al. Effect of Screening and Adjuvant Therapy on Mortality from Breast Cancer. The Journal of Medicine for the Cancer Intervention and Surveillance Modeling Network. 2005;353:1784-92.

(31) Wongpun S, Srivihok A. Comparison of attribute selection techniques and algorithms in classifying bad behaviors of vocational education students. The IEEE International Conference on Digital Ecosystems and Technologies. 2008; 2008; 526-31.

(32) Hall MA, Holmes G. Benchmarking attribute selection techniques for discrete class data mining. The IEEE Transactions on Knowledge and Data Engineering; 2003.

(33) Vu T-N, Ohn S-Y, Kim C-W. RISC: A new filter approach for feature selection from proteomic data. In: Zhang D, ed. The Lecture Note in Computer Science: Springer-Verlag Berlin Heidelberg 2007;17-24.

(34) Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. The International Joint Conference on Artificial Intelligence; 1995; 1995; 1137-43.

(35) He X, Frey EC. Three-class ROC analysis-the equal error utility assumption and the optimality of three-class ROC surface using the ideal observer. The IEEE Transactions on Medical Imaging. 2006:979-86.

(36) Woods K, Bowyer KW. Generating ROC curves for artificial neural networks. The IEEE Transactions on Medical Imaging. 1997:329-37.

(37) Henzinger MR, King V, Warnow T. Constructing a tree from homeomorphic subtrees, with applications to computational evolutionary biology. The Seventh Annual ACM-SIAM Symposium on Discrete Algorithms. 1996; 1996; 333-40.

(38) Huang J, Ling CX. Using AUC and accuracy in evaluating learning algorithms. The IEEE Transactions on Knowledge and Data Engineering: IEEE 2005:299-310.

(39) Jiang Y. Uncertainty in the output of artificial neural networks. The International Joint Conference on Neural Networks. 2007; 2551-6.

(40) Ohno-Machado L. Modeling medical prognosis: survival analysis techniques. The Journal of Biomedical Informatics. 2002;34(6):428-39.

(41) Kotsiantis SB, Tsekouras GE, Pintelas PE. Bagging Model Trees for Classification Problems. Advances in Informatics: The Lecture Notes in Computer Science. 2005;328-37.

(42) Das R, Sengur A. Evaluation of ensemble methods for diagnosing of valvular heart disease. Expert Systems with Applications. 2010;37(7):5110-5.

(43) Thongkam J, Xu G, Zhang Y, Huang F. Toward breast cancer survivability prediction models through improving training space. Expert System with Application. 2009;36(10):12200-9.

(44) Banfield RE, Hall LO, Bowyer KW, Kegelmeyer WP. A comparison of decision tree ensemble creation techniques. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2007:173-80.

(45) Li X, Wang L, Sung E. AdaBoost with SVM-based component classifiers. Engineering Applications of Artificial Intelligence. 2008;21(5):785-95.

(46) Yu W, Cheng De L. Learning by Bagging and Adaboost based on Support Vector Machine. 5th IEEE International Conference on Industrial Informatics. 2007; 663-8.