KKU Engineering Journal

http://www.en.kku.ac.th/enjournal/th/

Improving quality of breast cancer data through pre-processing

Vatinee Sukmak and Jaree Thongkam*

Faculty of Nursing, Mahasarakham University, Mahasarakham, Thailand, 44150.

Faculty of Informatics, Mahasarakham University, Mahasarakham, Thailand, 44150.

Received June 2013

Accepted September 2013

Abstract

Using data mining for medical prognosis becomes a promising approach recently. In the mining process, the raw data are commonly suffering from outlier and imbalanced problems which affect the performance of the model in predicting the unseen data. Thus, choosing appropriate data mining algorithms has a straight forward impact on the prediction model. The objective of this study is to investigate the use of three kinds of data pre-processing techniques including outlier filtering, Synthetic Minority Over-sampling TEchnique (SMOTE) and attribute selections for improving the quality of breast cancer data at Srinagarind Hospital in Thailand. Three types of decision rule building techniques, i.e. Decision Table with Naïve Bays (DTNB), Repeated Incremental Pruning to Produce Error Reduction (RIPPER) and PART Decision List were employed. The performance of proposed approaches was evaluated through the Area Under the receiver operating characteristics Curve (AUC) of the decision rules. Experimental results have shown that applying the suitable data pre-processing, especially the outlier filtering method, can lead to the significant improvement of the prediction performance of decision rule models.

Keywords: Breast cancer data, Pre-processing, Data mining, Decision rules

*Corresponding author. Tel.: 0-4375-4359 Ext 5352; fax: 0-4375-4359

Email address: jaree.thongkam@gmail.com

1. Introduction

Breast cancer is a common cancer and the second largest cause of cancer death among women in all countries[1]. Death from breast cancer accounts for 1.6% of female deaths every year worldwide[2]. The mortality rates of breast cancer vary among different countries. The overall survival rate show that breast cancer from 18 SEER geographic is at 89.2% [3] but 42.9% for Thai women [4]. Survival rates describe the percentage of people who survive a certain type of cancer for a specific amount of time. An overall survival rate includes people of all ages and health conditions diagnosed with their cancer, including those diagnosed very early and those diagnosed very late [5].

The traditional tools for predicting the survival rate include Cox-Proportional hazard and Kaplan-Meier which reapplied to estimate the survival rate of a particular patient suffering from a disease over a particular time period[6]. Currently, data mining in the field of medical prognosis has shown to be more accurate than those traditional tools in predicting the new sample of breast cancer [7]. It provides many techniques such as decision tree, Neural Network, and rule-based method. Firstly, decision tree's outcomes are easy to understand but the outcomes can be complicated when the tree becomes large [8, 9]. Secondly, Neural Network provides complicated outcome structure and hard to understand. Lastly, rule-based method is commonly exploited to form the decision rules which are more easily understood and can be combined with previous knowledge for medical practitioners to make appropriate decisions in their prognoses [10].

A decision rule refers to a set of 'If- Then' conditions which exhibit the information and knowledge in data [11, 12]. This decision rule becomes necessary when a decision tree is too large to be interpreted. Many researchers have utilized rule-based techniques to generate decision rules. For example, Zhou and Jiang [13] applied C4.5rule to build their decision rules. Their results indicated that C4.5rule produces rules with high generalization ability. In spite of this, Alshammari and Zincir-Heywood [14] employed RIPPER to classify motor traffic log files and compared its performance and effectiveness with basic AdaBoost. Their results showed that RIPPER outperforms basic AdaBoost in terms of accuracy. On the other hand, a PART decision list was developed based on both C4.5rule and RIPPER methods to generate decision rules without global optimization to provide more accurate decision rules than C4.5 and RIPPER [15].

However, survival data from the databases have some problems including missing data, outliers and imbalanced data frequently occur after applying data mining[16-18]. These problems directly affect the performance of the prediction models generated from data mining techniques [17-19]. Therefore, without data pre-processing to improve data quality, these techniques hardly produce a significantly, accurate model [20, 21]. Many research studies have employed several techniques in pre-processing to improve the quality of data including filing the missing values, eliminating the outliers and balancing the data. For instances, Mussa and Tshilidzi [22] used Neural Networks and genetic algorithms to filing missing data. Their results indicated that the number of missing data effects the accuracy of the prediction model. Besides, Gamberger, Šmuc and Mariĉ [23] presented algorithms to eliminate the outlier in medical domain. Their results showed that the

accuracy prediction models is significantly increased. Furthermore, Thong- kam, Xu, Zhang and Huang [24] introduced Support Vector Machine technique to detect outliers in cancers survivability prediction dataset. Their results demonstrated that this technique is superior to AdaBoost and Bagging techniques. To date, few research studies have investigated how much each technique affects the performance of prediction models especially in medical survival data.

Therefore, this study investigated several data pre-processing techniques including outlier filtering, a Synthetic Minority Over-sampling TEchnique (SMOTE) [25] and a combination of outlier filtering, SMOTE and Relief [26], and carried out the implementation of rule-based techniques for generating the decision rules.

The remainders of this paper include data pre-processing, decision rule, methodologies and experimental design, experimental results, discussions and conclusions.

2. Data pre-processing

Data pre-processing is an important step in data mining. It can be used to filter, transform, increase and decrease instances in data set mainly to improve data quality ensuring the model's performance. In this section, Outlier Filtering, SMOTE and Relief are briefly reviewed.

2.1 Outlier filtering

Outlier Filtering (OF) is commonly utilized to identify outliers using distance measures to detect outsider instances using *k*-Nearest Neighbors (*k*-NN) in order to improve the performance of classifiers [12, 27]. In this study the C-Support Vector Classification Filter (C-SVCF) [24] was employed to identify and eliminate outliers. This is because this C-SVCF algorithm is effective in classifying two classes in classification problems [28], and is a new generation learning algorithm based on recent advances in statistics, machine learning and pattern recognition [24, 29].

2.2 SMOTE

SMOTE is commonly applied to resize the imbalanced data. It utilizes a synthetic minority over-sampling technique to match the majority class by taking minority class instances and introducing synthetic instances [25]. Several research studies have exploited SMOTE to balance their data sets. For example, He, Han and Wang [30] applied SMOTE to ensemble classifiers trained on data sets. Their results indicated that SMOTE can enhance the C4.5 classifier performance. Similarly, Pelayo and Dick [31] employed SMOTE to resize the minority class to match the majority class. Their results demonstrated that SMOTE improves the accuracy of four NASA benchmark models.

2.3 Relief

Relief [26] is an instance-based attribute ranking algorithm utilizing the random sampling to locate the nearest neighbour from the same and opposite classes. This method is suited to selecting relevant attributes to improve the effectiveness of prediction models [32]. Relief for attribute selection defined from the weight w_j on input feature j can be computed in Equation 1.

$$w_j = P(x_j \neq x_d) - P(x_j \neq x_s) \quad (1)$$

Where x_j refers to the randomly selected training sample, and x_d and x_s are the two nearest

training instances to x_j in the same class and a different class, respectively. Moreover, $P(x_j \neq x_d)$ is the probability of nearest training instances to x_j in the same class, and $P(x_j \neq x_s)$ is the probability of nearest training instances to x_j in the different class.

3. Decision rules

In this section, the background of decision rules including DTNB, RIPPER and a PART Decision List is outlined.

3.1 Decision table with naïve bays

DTNB [33] is the combination model proceeded as the traditional stand-alone Decision Table (DT). However, it evaluates the merit associated with splitting the attributes into two disjoint subsets in order to searching at each point: one for the decision table and the other for Naïve Bays (NB). Therefore the selected attributes are modeled by NB at each step and the remainder by the decision table.

3.2 RIPPER

RIPPER [34] is a well-known decision rules induction method in which accuracy is increased by revising individual rules and combining cross-validation and minimum-description length techniques to prevent the over fitting problem [35]. Therefore, the accuracy of decision rules produced from RIPPER is increased [34, 35].

3.3 PART decision list

PART decision list [15] is a new method for inducting decision rules by combining both C4.5 and RIPPER to: a) find and generate a set of rules by exploiting a rule induction approach, b) adopt the divide and conquer strategies to build a sub-tree, and c) construct rules from the sub-tree. Although this method is similar to C4.5 rule, it avoids constructing a complete decision tree and leads to improving the training time. Unlike RIPPER, a PART decision list produces each decision rule corresponding to the leaf with the largest coverage in the partial decision tree [36]. In this way, the PART decision list generates and prunes a partial decision tree, handles missing data, numerical and discrete attributes, and provides accurate rule sets [15, 36, 37].

4. Methodologies

In order to build 5-year breast cancer survivability decision models, the background of breast cancer survivability is explained. The data preparation, data pre-processing framework and experimental design are presented.

4.1 Data preparation

In this paper, the breast cancer data sets were obtained from Srinagarind Hospital. This hospital is the only medical school associated hospital in northeastern Thailand established in 1972 as part of the faculty of medicine at Khon Kaen University. The total data of breast cancer became 4,462 records as of January 2007. However, some attributes are missing and unknown values comprise of more than 30%. This may be due to the fact that some patients were diagnosed in this hospital but received treatments in other hospitals. In order to compute the 5-year breast cancer survivability data sets, the selected attributes are displayed in Table 1.

Table 1 Input attributes

No.	Attributes Name	Initial	Values
1	Age	(Age)	Number
2	Marital status	(Mars)	Category(3)
3	Occupations	(Occ)	Category(19)
4	Basis of diagnosis	(Dx)	Category(6)
5	Topography	(Top)	Category(8)
6	Morphology	(Mor)	Category(12)
7	Stage	(Stage)	Category(4)
8	Extent	(Ext)	Category(4)
9	Received surgery	(Surg)	Category(2)
10	Received radiation	(Radi)	Category(2)
11	Received chemotherapy	(Chem)	Category(2)
12	Received hormone	(Horm)	Category(2)
13	Received Immune	(Immu)	Category(2)
14	Survivability (Class attribute)	(Class)	Category(2)

Table 1 shows the attributes and the attribute names used in this study. These attributes were chosen as the powerful prognostic factors identified in most studies [38] For example, Topography consists of nine values which point out the position of cancer in breast that related to the choice of treatments in the clinical trial. Moreover, the extent of disease is aggregated attribute with Morphology to see the patterns related to the breast cancer survival periods. 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 more than five years after the diagnosis. Moreover, the instances with unknown of values in each attribute are excluded. Therefore, the initial numbers of instances are displayed in Table 2.

Table 2 The initial numbers of instances

Data Sets	Years	'Dead'	'Alive'	Total
5-year Surviability	1985-2002	210	314	524

4.2 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 [39]. Padmaja, Dhulipalla, Bapi and Krishna[40] utilized a three step framework to improve data quality: 1) employ *k*-Nearest Neighbours (*k*-NN) to eliminate outliers in a minority class, 2) apply over-sampling to increase the size of the minority class, and 3) exploit under-sampling to reduce the size of the majority class. Nevertheless, in this paper, investigation has been conducted not only the combination of outlier elimination and over-sampling but also the integration of attribute selection approaches as follows.

1) Apply C-Support Vector Classification filtering (C-SVCF) to identify and eliminate outliers from both 'Dead' and 'Alive' classes.

2) Utilize SMOTE to increase the size of the minority class to the same size of majority class by using the different ratio between majority and minority classes.

3) Select the top nine relevant attributes arranged by Relief.

As a result, the number of instances corresponding 5-year breast cancer survivability data sets is tabulated in Table 3.

Table 3 The number of instances after applying pre-processing

Approaches	Number of instances			Diff. Ratio	Number of attributes
	Classes		Total		
	'Dead'	'Alive'			
Raw	210	314	524	49.52	14
SMOTE	313	314	627	0.31	14
OF	136	232	368	70.58	14
OF +SMOTE	233	232	465	0.42	14
OF +SMOTE+Relief	233	232	465	0.42	10

Note: Different ratio = ((majority class– minority class)/ majority class)*100

Table 3 illustrates the number of instances in both classes ('Dead' and 'Alive') of 5-year breast cancer survivability data sets after data pre-processing. It seems that using C-SVCF to filter outliers leads to increase imbalanced problem to the data sets.

4.3 Experimental design

In order to evaluate the performance of the preprocessing, stratified 10-fold cross-validation [12]and AUC were applied as the procedure to validate the decision rules in order to minimize the variance and bias associated with the results [41, 42]. This is because they are popular methods in data mining widely used in medical research.They are also able to increase the performance of models which results in a greater reliability. There are six main processes of stratified 10-fold cross-validation which are displayed as follow.

1) Divide the data set into a set of 'dead' and 'alive' subclasses.
2) Assign a new sequence number to each set of subclasses.
3) Divide each subclass into 10 subsets called fold.
4) Combine each fold of each subclass into a single fold.
5) Combine nine folds as a training set and remaining fold as a test set.
6) Repeat step five using different fold of test set each time nine times. The average of AUC score would be computed to demonstrate the performance of prediction models.

Ten iterations of stratified 10-fold cross-validation were used. This means that 100 prediction models of each technique were built in order to minimize the uncertainty of these experimental results. Besides, AUC is an alternative measure for evaluating the predictive ability of learning algorithms based on the true positive rate against the false positive rate [43-45]. Moreover, the average AUC is chosen as a performance selection criterion of filtering methods in the classification tasks. It usually has scores between 0 which is the lowest performance and 1 which is the highest performance so that the results are easy to be understood [46].

5. Experimental results

In these experiments, WEKA version 3.7 [36] was selected as a data mining tool. This is because it provides several data pre-processing and learning algorithms in data mining and machine learning. In order to evaluate the effectiveness of data pre-processing, the generalization performance of 5-year breast cancer survivability, decision rules generated from DTNB, RIPPER [34] and a PART Decision list [15] is employed. Furthermore, random seed one of stratified 10-fold cross-validation is exploited to divide 5-year breast cancer survivability data sets into training and test sets to reduce variance and bias of prediction results. Then, this

test was run for 10 iterations to increase the reliability of the results.

5.1 Area under ROC curve

Area Under the receiver operating characteristic Curve (AUC) is used to evaluate the performance of classifiers in both balanced and imbalanced classification problems such as that presented in medical data sets [47-49]. In this experiment, the AUC score of 5-year breast cancer survivability classifiers generated from DTNB, RIPPER and PART after data pre-processing is presented in Table 4.

Table 4 The average AUC score of 5-year decision models

Pre-processing	The averageAUC of Classifiers (%)		
	DTNB	RIPPER	PART
Raw	71.29±6.76	68.82±5.65	65.89±7.64
SMOTE	71.13±6.31	68.15±5.88	65.75±6.48
OF	92.99±5.92	89.36±5.74	90.88±5.86
OF +SMOTE	92.99±4.38	88.79±4.95	91.20±4.97
OF +SMOTE+Relief	92.80±4.77	89.72±4.72	92.13±4.46
Average	84.24	80.96	81.17

Table 4 demonstrates the overall AUC score of 5-year breast cancer survivability decision rules generated from DTNB, RIPPER and the PART decision list after applying SMOTE, OF, C-SVCF +SMOTE and OF+SMOTE+Relief. The experimental results showed that the average AUC scores of DTNB (92.80%) are superior to RIPPER (89.72%)and a PART Decision List (92.13%) in 5-breast cancer survivability data sets after employing the OF +SMOTE+Relief.

5.2 Decision rules for 5-year survivability

In this section, 5-year breast cancer survivability decision rules are built using the PART Decision List which provides visual rule base. Unlike DTNB, PART Decision List are unable to present the visual rules. Therefore, PART Decision List was employed to generate the rules from a whole data set which involves 465 instances after employingOF +SMOTE. In order to interpret decision rules, the example of the rules computed from the PART Decision List technique is illustrated in Table 5.

Table 5 Rule for decision-making

Rule No.	Conditions
Rule 1:	ext = 3 AND Dx = 7: 1 (158.0/2.0)
Rule 2:	ext = 5: 0 (104.0/8.0)
Rule 3:	ext = 4 AND age <= 68 AND occ = 1 AND top = 509 AND mor = 8500: 0 (78.0/3.0)
Rule 4:	ext = 2: 1 (19.0)
Rule 5:	mor = 8000 AND age > 37 AND age <= 67: 0 (10.0)

Table 5 presents the 5-year breast cancer survivability rules generated from the PART decision list. Each rule provides the number of the coverage (n_{coverage}) and uncorrected cases ($n_{incorrect}$). Therefore the accuracy of each rule can be computed as Equation 2.

$$accuracy = \frac{n_{\text{coverage}} - n_{incorrect}}{n_{\text{coverage}}} \quad (2)$$

Results of the first five rules, for example are interpreted as follows:

Rule 1: If a patient with direct extension (‘ext’ = ’3’) and histology of primary (‘Dx’ = ‘7’) at the first diagnosis, the model correctly predicted 98.73% ((158-2)/158) for this patient to live more than 5 years after the first diagnosis.

Rule 2: If a patient with distant metastases (‘ext’ = ‘5’) at the first diagnosis, the model correctly predicted 92.30% for this patient to live less than 5 years after the first diagnosis.

Rule 3: If a patient with regional lymph nodes (‘ext’ = ‘4’), age less than 68, labor occupation (‘occ’=’1’), breast (NOS) (‘top’ = ‘509’) and infiltrating duct carcinoma (‘mor’ = ‘8500’), the model correctly predicted 96.15% for this patient to live within 5 years after the first diagnosis.

Rule 4: If a patient with localized (ext = ‘2’) at the first diagnosis, the model correctly predicted 100% for this patient to live 5 or more years after the first diagnosis.

Rule 5:If a patient with neoplasm (‘mor’ = ‘8000’) and age between 37 to 67 years old, the model correctly predicted 100% for this patient to live within 5 years after the first diagnosis.

6. Discussions

Pre-processing including Outlier Filtering (OF), SMOTE and a combination of three methods (OF, SMOTE and Relief) are investigated in this paper in order to enhance the generalization performance of 5-year breast cancer survivability decision rules. These rules were generated by DTNB, RIPPER and PART decision list.

Firstly, we found that after applying outlier filtering and SMOTE, the accuracy of 5-year breast cancer survivability decision rules using DTNB was significantly increased up to 21.32%. These findings were consistent with those of Padmaja et al.[40] who found that the different performance improvements of classifiers are obtained after eliminating outliers rather than using over-sampling to resize the minority class to match the majority class. This may be because using the outlier filtering method reducing the insignificance resulting in increasing the performance of the model.

Secondly, the experimental results in this study presented that exploiting the right attribute selection method leads to the performance improvement of the decision rule models. This observation is also evidenced by Yi and Fuyong [50] who claimed that the most practical machine learning algorithms are less concerned with selecting the irrelevant attributes that may damage the accuracy of the model. This may due to the fact that some attribute selection method scan be able to identify the significance attributes in the data set.

Lastly, another important finding is that the average AUC of the model generated from DTNB is superior to the average AUC the model generated from RIPPER and PART in 5-year breast cancer survivability data sets. Correspondingly, Ozbakir and Delice [51] found that the DTNB is better than PART in heart cancer and world breast cancer data sets. This may due to the fact that DTNB is performed well in the data set with less outlier.

7. Conclusions

In this paper, we have investigated the use of data pre-processing and decision rule building for the study of Thai breast cancer survivability prediction. The data pre-processing including outlier filtering SMOTE and combination of outlier filtering with SMOTE, outlier filtering with SMOTE

and Reliefisutilized to improve the data quality. And the 5-year cancer survivability decision rule models are generated by DTNB, RIPPER and a PART decision list to demonstrate the effectiveness of the proposed approaches. The Experimental results have showed that applying outlier filtering, outlier filtering+SMOTE and SMOTE are able to generate the better decision rule models. Therefore, we believe that employing the suitable method for data pre-processing can lead to the enhancement of the generalization performance. Moreover the decision rules built from this study are expected to provide meaningful and interpretable findings to medical practitioners in real applications.

8. Acknowledgements

The authors expresses special thank to IT and Cancer Department staffs at Srinagarind Hospital Thailand for kindly providing the data.

9. References

[1] The American Cancer Society. Cancer facts and figures2006. http://www.cancer.org/ downloads/ STT/CAFF2006PWSecured.pdf, Accessed 2006.

[2] Peter B, Bernard L. World cancer report 2008 [pdfsonline].www.iarc.fr/en/publications/pdfs_online/wrc/2008/wcr_2008.pdf, Access 2008.

[3] Nation Cancer Institutes. www.seer.cancer.gov /statfacts/html/breast.html, Accessed 10 Jun 2013.

[4] Poum A, Kamsa-ard S, Promthet S. Survival rates of breast cancer: a hospital-based study from northeast of Thailand. Asian Pacific Journal of Cancer Prevention. 2012. 13(3): 791-4.

[5] Zeng T, Liu J. Mixture classification model based on clinical markers for breast cancer prognosis. Artificial Intelligence in Medicine. 2010; 48(2-3): 129-137.

[6] Borovkova S. Analysis of survival data. Vakantiecursus: 2002;5(3): 302-7.

[7] Ohno-Machado L. Modeling medical prognosis: survival analysis techniques. Biomedical Informatics. 2002;34(6): 428-439.

[8] Mitchell TM. Machine learning. McGraw-Hill. 1997.

[9] Flores BA, Gonzalez JA. Data mining with decision trees and neural networks for calcification detection in mammograms. Advances in Artificial Intelligence. 2004; p. 232–241.

[10] Olukunle A, Ehikioya S. A fast algorithm for mining association rules in medical image data. Canadian Conference on Electrical and Computer Engineering: 2002; p. 1181-1187.

[11] Tan P-N, Steinbach M, Kumar V. Introduction to data mining. Pearson Addison Wesley: 2006.

[12] Han J, Kamber M. Data mining:concepts and techniques. Morgan Kaufmann, Elsevier Science: 2006.

[13] Zhou Z-H, Jiang Y. Medical diagnosis with C4.5 rule preceded by artificial neural network ensemble. IEEE Transactions on Information Technology in Biomedicine. 2003; p. 37-42.

[14] Alshammari R, Zincir-Heywood AN. A flow based approach for SSH traffic detection. Proccessing of the IEEE International Conference on Systems, Man and Cybernetics. 2007; p. 296-301.

[15] Frank E, Witten IH. Generating accurate rule sets without global optimization. 15th International Conference on Machine Learning. 1998; p. 144-151.

[16] Corrigan D, Harte N, Kokaram A. Pathological motion detection for robust missing data treatment in degraded archived media. Proccessing of the IEEE International Conference on Image. 2006; p. 621-4.

[17] Lin J-H, Haug PJ. Exploiting missing clinical data in Bayesian network modeling for predicting medical problems. Biomedical Informatics. 2007; 41(1): 1-14.

[18] Tsumoto S. Problems with mining medical data. Twenty-Fourth Annual International Conference on Computer Software and Applications: 2000; p. 467-468.

[19] Thongkam J. Problems of Data Mining in Medical Health Systems. MSU J. 2011; 30(3): 360-365.

[20] 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. Expert Systems with Applications. 2008; 34(1): 108-118.

[21] 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.

[22] Mussa A, Tshilidzi M. The use of genetic algorithms and neural networks to approximate missing data in database. IEEE Third International Conference on Computational Cybernetics: 2005; p. 207-212.

[23] Gamberger D, Šmuc T, MariĆ I. Noise detection and elimination in data preprocessing experiments in medical domains. Applied Artificial Intelligence: 2000; 14(2): 205-223.

[24] Thongkam J, Xu G, Zhang Y, Huang F. Support vector machines for outlier detection in cancers survivability prediction. International Workshop on Health Data Management: 2008; p. 99-109.

[25] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. Artificial Intelligence and Research. 2002; 16: 321-357.

[26] Kira K, Rendell LA. A practical approach to feature selection. Proccessing of the 9th International Conference on Machine Learning. 1992; p. 249-256.

[27] Brodley CE, Friedl MA. Identifying and eliminating mislabeled training instances. Artificial Intelligence Research. 1996; 1: 799-805.

[28] Vapnik V. Statistical learning theory, Wiley. 1998.

[29] Yin Z, Yin P, Sun F, Wu H. A writer recognition approach based on SVM. Multi Conference on Computational Engineering in Systems Applications. 2006; p. 581-586.

[30] He G, Han H, Wang W. An over-sampling expert system for learing from imbalanced data sets. International Conference on Neural Networks and Brain. 2005; p. 537-541.

[31] Pelayo L, Dick S. Applying novel resampling strategies to software defect prediction. Annual Meeting of the North American Fuzzy Information Processing Society. 2007; p. 69-72.

[32] Dietterich TG. An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization. Machine Learning. 1999; 40(2): 1-22.

[33] Hall M, Frank E. Combining Naive Bayes and Decision Tables. Association for the Advancement of Artificial Intelligence. 2008; p. 1-2.

[34] Cohen WW. Fast effective rule induction Proccessing of. Twelfth International Conference on Machine Learning; 1995. p. 115-123.

[35] Xin J, Rongfang B. Improving Software Quality Classification with Random Projection. Proccessing of the 5th IEEE International Conference on Cognitive Informatics. 2006; p. 149-154.

[36] Witten IH, Frank E. Data mining: practical machine learning tools and techniques. Morgan Kaufmann. 2005.

[37] Thabtah F, Cowling P. Mining the data from a hyperheuristic approach using associative classification. Expert Systems with Applications. 2008; 34(2): 1093–1101.

[38] Xiong X, Kim Y, Baek Y, Rhee DW, Kim. S-H. Analysis of breast cancer using data mining & statistical techniques. The 6th International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing and First ACIS International Workshop on Self-Assembling Wireless Networks. 2005; p. 82-87.

[39] Wongpun S, Srivihok A. Comparison of attribute selection techniques and algorithms in classifying bad behaviors of vocational education students. Proccessing of the 2nd IEEE International Conference on Digital Ecosystems and Technologies. 2008; p. 526-531.

[40] Padmaja TM, Dhulipalla N, Bapi RS, Krishna PR. Unbalanced data classification using extreme outlier elimination and sampling techniques for fraud detection. International Conference on Machine Learning and Cybernetics on Advanced Computing and Communications. 2007; p. 511-516.

[41] Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. Proccessing of International Joint Conference on Artificial Intelligence. 1995; p. 1137-1143.

[42] Thongkam J, Xu G, Zhang Y. An analysis of data selection methods on classifier accuracy measures. KKU Eng J. 2008; 35: 1-10.

[43] Hand D, Mannila H, Smyth P. Principles of data mining. The MIT Press. 2001.

[44] 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; p. 979-986.

[45] Woods K, Bowyer KW. Generating ROC curves for artificial neural networks. The IEEE Transactions on Medical Imaging. 1997; p. 329-337.

[46] Huang J, Ling CX. Using AUC and accuracy in evaluating learning algorithms. The IEEE Transactions on Knowledge and Data Engineering; 2005. p. 299-310.

[47] Xie J, Qiu Z. The effect of imbalanced data sets on LDA: A theoretical and empirical analysis. Pattern Recognition Society. 2007; 40(2): 557-562.

[48] Alejo R, Garcia V, Sotoca JM, Mollineda RA, Sánchez JS. Improving the classification accuracy of RBF and MLP neural networks trained with imbalanced samples. The Intelligent Data Engineering and Automated Learning. 2006; p. 464-471.

[49] Estabrooks A, Jo T, Japkowicz N. A multiple resampling method for learning from imbalanced data sets. Computational Intelligence. 2004; 20(1): 18-36.

[50] Yi W, Fuyong W. Breast cancer diagnosis via support vector machines. Chinese Control Conference. 2006; p. 1853-56.

[51] Özbakır L, Delice Y. Exploring comprehensible classification rules from trained neural networks integrated with a time-varying binary particle swarm optimizer Engineering Applications of Artificial Intelligence. 2011; 24(3): 491-500.