Iran J Public Health, Vol. 46, No.11, Nov 2017, pp.1563-1571

Original Article

Prognosis and Early Diagnosis of Ductal and Lobular Type in Breast Cancer Patient

Houriyeh EHTEMAM [1], ***Mitra MONTAZERI*** [2,3], ****Reza KHAJOUEI*** [4], ****Raziyeh HOSSEINI*** [5], ***Ali NEMATI*** [6], ***Vahid MAAZED*** [6]

1. *Health Services Management Research Center, Institute for Futures Studies in Health, Kerman University of Medical Sciences, Kerman, Iran*
2. *Modeling in Health Research Center, Institute for Futures Studies in Health, Kerman University of Medical Sciences, Kerman, Iran*
3. *Computer Engineering Department, Faculty of Engineering, Shahid Bahonar University, Kerman, Iran*
4. *Medical Informatics Research Center, Institute for Futures Studies in Health, Kerman University of Medical Sciences, Kerman, Iran*
5. *Health Information Sciences Dept., Faculty of Management and Medical Information Sciences, Kerman University of Medical Sciences, Kerman, Iran*
6. *Hematology and Oncology, Faculty of Medicine, Kerman University of Medical Sciences, Kerman, Iran*

***Corresponding Authors:** Email: r.khajouei@yahoo.com & rhosseini70011@gmail.com

(Received 07 Nov 2016; accepted 16 Mar 2017)

Abstract

Background: Breast cancer is one of the most common cancers with a high mortality rate among women. Prognosis and early diagnosis of breast cancer among women society reduce considerable rate of their mortality. Nowadays, due to this illness, try to be setting up intelligent systems, which can predict and early diagnose this cancer, and reduce mortality of women society.

Methods: Overall, 208 samples were collected from 2014 to 2015 from two oncologist offices and Javadalaemeh Clinic in Kerman, southeastern Iran. Data source was medical records of patients, then 64 data mining models in MATLAB and WEKA software were used, eventually these measured precision and accuracy of data mining models.

Results: Among 64 data mining models, Bayes-Net model had 95.67% of accuracy and 95.70% of precision; therefore, was introduced as the best model for prognosis and diagnosis of breast cancer.

Conclusion: Intelligent and reliable data mining models are proposed. Hence, these models are recommended as a useful tool for breast cancer prediction as well as medical decision-making.

Keywords: Diagnosis, Breast cancer, Ductal and lobular, Data mining models

Introduction

Cancer leads to physical and emotional stress (1) among all kinds of cancers is the most common cancer (2). Moreover, it has ascending growth in deprived areas (3). Surprisingly, this illness is rare among men. However, it is the most common cause of death in women (2). Breast cancer has various morphologies, which are used in classifying of this disease (4). Some researchers consider Ductal and Lobular to classify types of this cancer. These two morphologies (Ductal and Lobular) have different characteristics, but Ductal is the most common type, and approximately it has allocated 75% to 85% of breast cancers to own (5).

Identifying risk factors of breast cancer has become an important issue among physicians and pathologists (6). However, by medical technologies improvements, useful risk factors are measuring and recording (7). Early diagnosis of breast cancer is very effective in re-cover of patients, and it has positive impact on longevity of them. In spite of this cancer is so common, it will be the most curable when detect soon (8). Early diagnosis of breast cancer is very effective in recovery of the disease, and it has positive impact on longevity of patients, although this cancer is the most common types of cancer among women, it will be the most curable when detected early (9). In order to diagnosis of breast cancer, intelligent models are useful to increase the precision and accuracy of diagnosis (10). By advance-

ment in computerized software and hardware, the massive volume of data is recorded automatically, after that efficient analysis methods help to analyze the data efficiently (7).

Data mining is one of the technology improvements that serve to manage data. Widespread use of information systems lead to merge data mining with traditional methods (11).

Utilization of data mining techniques with the approach of extracting knowledge from information have many advantages, such as how to recognize diseases, reducing health care costs, reducing medical errors, and last but not least improve the performance of healthcare organizations (12).

Additionally, data-mining models can be a way to reduce errors in decision making by physicians. In medical levels, data mining effort is used to extract relationships and patterns from a large number of data to predict diseases (13). The result of these analyses should be comprehensible for everyone (14). Totally, data mining is defined as a process of selecting, exploring and modeling large volume of data used in order to discover new and usable patterns from data analyzing (15). According to Fig. 1, steps of extracting knowledge from database by using data mining were depicted in five stages.

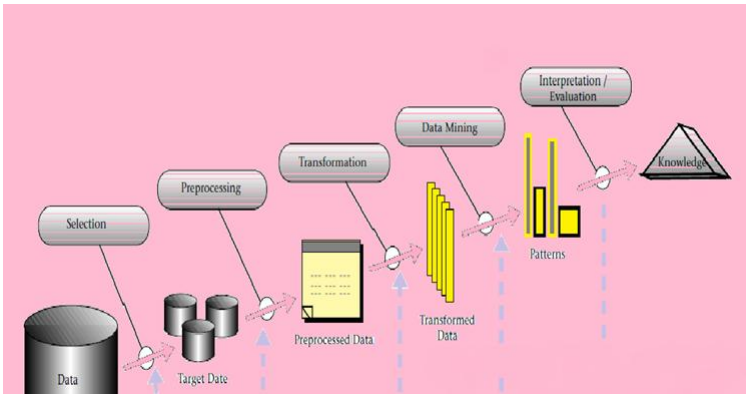

Fig. 1: Steps of knowledge discovery in databases with data mining process (16)

In the first stage, special data was selected among large volume of data. In the second stage, preprocessing methods was performed on data, for instance controlling a missing data. In third step, data were ready to transform based on hypothesis. Then, data-mining algorithms were selected, they decide about which patterns are more appropriate. In fifth stage, interpretation/evaluation was done. All previous steps will be evaluated again. Consequently, it prepared us an image from extracted patterns and models. Knowledge was the final product of this process. Eventually, we could present this knowledge without combined to other systems, or report it to other enthusiastic people (16).

Hence, we can use this intelligent method as accurate and reliable system to early diagnosis of benign or malignant of breast cancer (17). This method could lead to save many people from threat of death due to breast cancer, or enhance their longevity and quality of their life.

In this study, we aimed to present the most effective data mining models to identify breast cancer sooner.

Materials and Methods

Data collection

A list of breast cancer risk factors was taken from a previous study (18), and then they were confirmed by an oncologist. Samples based on these risk factors were gathered from records of breast cancer patients, and whole of their identity information kept secret. Medical records of 208 patients collected from two oncologist offices, and Javadalaemeh Clinic, from 2014 to 2015. In order to control missing data, the most frequent repeat was replaced for discrete data, and for

continuous missing one, the average of data in corresponding column is replaced (19).

These risk factors are as follow: age, sex, BMI, Marital Status (MS), Age Starting of First Menstruation (ASFM), the Number of Parturition (NP), the Number of Abortion (NA), Age Starting of Menopause (ASM), History of Breast Cancer, Uterine and Ovarian Cancer in First-Degree Relatives(HBCUOCFDR), History of Breast Cancer, Uterine and Ovarian Cancer in Second-Degree Relatives (HBCUOCSDR), History of Other Cancers in First-Degree Relatives (HOCFDR), History of Other Cancers in Second-Degree Relatives (HOCSDR), ER, PR, Existence of Tumor (ET), Size of Tumor (ST), Type of Cancer (TC).

Risk factors

Overall, 17 risk factors for breast cancer were used. The risk factors were divided into two groups (nominal and real). These risk factors are as follow: age (yr), sex (Male/Female), BMI (kg/m^2), Marital Status (Single/Married), Age Starting of First Menstruation (yr), the Number of Parturition (Number), the Number of Abortion (Number), Age Starting of Menopause (yr), History of Breast Cancer, Uterine and Ovarian Cancer in First-Degree Relatives(Yes/No), History of Breast Cancer, Uterine and Ovarian Cancer in Second-Degree Relatives (Yes/No), History of Other Cancers in First-Degree Relatives (Yes/No), History of Other Cancers in Second-Degree Relatives (Yes/No), ER (Positive/Negative), PR (Positive/Negative), Existence of Tumor (Yes/No), Size of Tumor (Cm), Type of Cancer (Ductal/Lobular).

Classification

The data were analyzed by WEKA and MATLAB software, and 64 data mining models classified them. Of all 17 risk factors, 16 of them were defined as independent risk factors, and one of them that was a specified type of cancer divided into Ductal and Lobular allocated class (dependent risk factor) tag to own. The stages of our method are shown in Fig.2. Initially, the collected breast cancer data were considered as input. Secondly, the data divided into train and test kind. In third stage, train data were learned based on a special technique and produce data mining models. After that, the model changed to learned model. In fourth step, the performance of the learned model became valid by test data. Finally, the final model was presented as output.

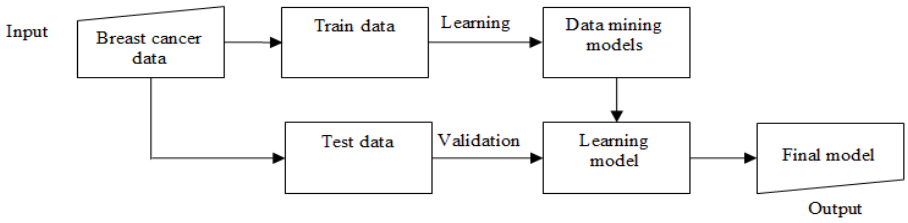

Fig. 2: Flow chart of proposed method

Experimental findings

Configuration of the proposed models

Samples that were belong to positive and negative class, were denoted as *P* and *N*, respectively. In each classification, four definitions can be explained as follow:

- positive group and anticipate correctly called True Positive (*TP*).
- positive group and anticipate incorrectly called False Positive (*FP*).
- negative group and anticipate correctly called True Negative (*TN*).
- negative group and anticipate incorrectly called False Negative (*FN*).

Therefore, the equations for precision and accuracy can be defined as follow:

$$\text{Precision} = \frac{TP}{(TP+FP)} \quad (1) \qquad \text{Accuracy} = \frac{(TP+TN)}{(P+N)} \quad (2)$$

Results

After choosing effective risk factors, two morphologies of this cancer were considered (Ductal and Lobular). Another phase of this paper was data mining. In this phase the data became valid by a special method explained in section entitle "method" and the valid data after some other process produced final model. In order to evaluation, *K*-Fold cross validation method was used. *K* was equal to 10 (*K*=10).

The results of the Binomial Test are shown in Table 1. First phase of our work was presented in Table 1 that was designed by SPSS software (Chicago, IL, USA). Error was reported as 0.05. In addition, the value of *P*-value for Ductal and Lobular had been achieved 0; thus, our final method had high accuracy.

Table 1: *P*-value which compares *P*-value between two type of breast cancer (Ductal and Lobular)

Binomial Test					
Type of cancer	***Category***	***n***	***Observed Prop.***	***Test Prop.***	***P-Value***
Group 1	Ductal	198	.95	.50	.000
Group 2	Lobular	10	.05		
Total		208	1.00		

Table 2 presents nominal risk factors (ER, PR, tumor size, parity, marital status and age) which were grouped based on frequency, percent, valid percent, and cumulative percent. In Table 3, 64 data mining models are shown. There are percentages of accuracy and precision, too.

Table 2: Grouping of nominal risk factors of breast cancer

Risk factor	***Group***	***Frequency***	***Percent[1]***	***Valid Percent[2]***	***Cumulative Percent[3]***
	Positive	137	65.2	65.2	66.2
ER	Negative	71	33.8	33.8	100.0
	Total	208	100.0	100.0	-
	Positive	148	70.5	70.5	71.4
PR	Negative	60	28.6	28.6	100.0
	Total	208	100.0	100.0	-
	Married	196	93.3	94.2	94.2
MS	Single	12	5.7	5.8	100.0
	Total	208	99.0	100.0	-
	Yes	179	86.1	86.1	86.1
ET	No	29	13.9	13.9	100.0
	Total	208	100.0	100.0	-
	Yes	15	7.2	7.2	7.2
HBCUOCFDR	No	193	92.8	92.8	100.0
	Total	208	100.0	100.0	-
	Yes	17	8.2	8.2	8.2
HBCUOCSDR	No	191	91.8	91.8	100.0
	Total	208	100.0	100.0	-
	Yes	19	9.1	9.1	9.1
HOCFDR	No	189	90.9	90.9	100.0
	Total	208	100.0	100.0	-
	Yes	18	8.7	8.7	8.7
HOCSDR	No	190	91.3	91.3	100.0
	Total	208	100.0	100.0	-
	0-5	179	85.2	86.1	86.1
Parity	6-11	27	12.9	13.0	99.0
	12-17	2	1.0	1.0	100.0
	Total	208	99.0	100.0	-
	Ductal	198	95.2	95.2	95.2
TC	Lobular	10	4.8	4.8	100.0
	Total	208	100.0	100.0	-

[1] Represents the percentages of all data, including the missing data, established by each category.

[2] Valid percent presents only the non-missing cases.
[3] Cumulative percent brings an easier way to compare different sets of data.

Table 3: Amount of precision and accuracy of the each model

NO.	*Machine learning model*	*Classification accuracy (%)*	*Precision (%)*
1.	Bayes-Net (20)	95.67	95.70
2.	Naïve-Bayes	91.83	95.00
3.	Naïve-Bayes-Updateable	91.83	95.00
4.	Logistic	90.86	95.00
5.	Multilayer-Perceptron	91.83	95.50
6.	RBF-Network	94.23	95.10
7.	Simple-Logistic	95.19	95.20
8.	Sequential-Minimal Optimization (21)	95.19	95.20
9.	Voted-Perceptron	95.19	95.20
10.	Instance-Based-Learning-algorithms	90.86	95.00
11.	IBK	90.38	95.00
12.	K-Star	91.82	95.00
13.	Locally-Weighted-Learning	94.71	95.20
14.	AdaBoost-ML	95.19	95.20
15.	Attribute-Selected-Classifier (22)	95.19	95.20
16.	Bagging	95.19	95.20
17.	Classification-Via-Clustering	69.23	94.70
18.	Classification-Via-Regression	94.71	95.20
19.	Cross-Validation-Parameter-Selection (23)	95.19	95.20
20.	Dagging	95.19	95.20
21.	Decorate (24)	95.19	95.20
22.	Ensembles of Nested Dichotomies (25)	95.19	95.20
23.	Ensemble-Selection (26)	95.19	95.20
24.	Filtered-Classifier (22)	95.19	95.20
25.	Grading	95.19	95.20
26.	Logit-Boost	95.19	95.20
27.	Multi-Boost-AB (27)	95.19	95.20
28.	Multi-Class-Classifier	90.86	95.00
29.	Multi-Scheme	95.19	95.20
30.	Ordinal-Class-Classifier (28)	95.19	95.20
31.	Raced-Incremental-Logit-Boost (29)	95.19	95.20
32.	Random-Committee	94.23	95.10
33.	Random-Sub-Space (30)	95.19	95.20
34.	Rotation-Forest	95.19	95.20
35.	Stacking	95.19	95.20
36.	Stacking-C	95.19	95.20
37.	Threshold-Selector	94.23	95.10
38.	Vote	95.19	95.20
39.	Hyper-Pipes (31)	95.19	95.20
40.	classification by Voting Feature Intervals	74.52	95.60
41.	Conjunctive-Rule	95.19	95.20
42.	Decision-Table	95.19	95.20
43.	Decision-Table-Naïve-Bayes (32)	95.19	95.20
44.	J-Repeated-incremental-pruning (33)	95.19	95.20
45.	Non-Nested-generalized-exemplars	92.79	95.10
46.	One-R (34)	95.19	95.20
47.	PART	94.23	95.10
48.	Ridor(35)	95.19	95.20
49.	Zero-R	95.19	95.20
50.	Alternating-Decision Tree (36)	95.19	95.20
51.	Best-FirstTree	95.19	95.20
52.	Decision-Stump	95.19	95.20
53.	Functional trees	94.71	95.20
54.	J48 (37)	95.19	95.20
55.	J48-graft (38)	95.19	95.20
56.	LAD-Tree	91.35	95.20
57.	NB-Tree (39)	95.19	95.60
58.	Random-Forest	93.75	95.10
59.	Random-Tree	90.38	95.40
60.	REP-Tree (40)	95.19	95.20
61.	Simple-Cart	95.24	95.20
62.	Class-Balanced-Nested-Dichotomies (41)	95.19	95.20
63.	(Data-Near-Balanced-ND (41)	95.19	95.20

64.	Nested-Dichotomies	95.19	95.20

As it is obvious in Table 3, VFI were the weakest model in prognosis and diagnosis of breast cancer, and Bayes-Net was identified as the best. Fig. 3 demonstrates the ROC curve of the four best models among 64 models (BN, MP, NB-Tree, and RT). This figure shows the performance of these models in WEKA software. The MP model has the highest ROC area value among the other four models.

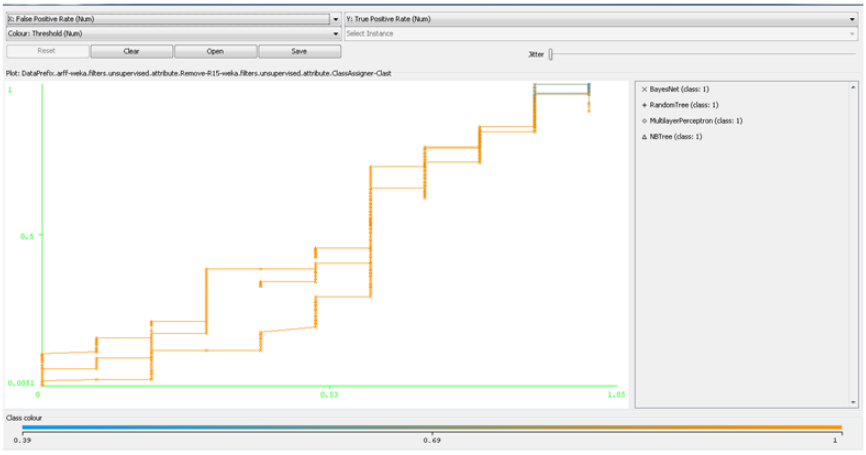

Fig. 3: ROC curve of four best models in WEKA software (BN, MP, NB-Tree, and RT)

Discussion

Breast cancer is one of the most common cancers in women. Early detection of breast cancer leads to declining mortality. Technology improvements can help early diagnosis of breast cancer. Data mining method is an intelligent model that diagnosis this cancer with more precision and accuracy. We aimed to help physicians by computerized models to prognosis of this cancer sooner, without expensive price, and have a less side effect on patients. Therefore, the data collections enter to validation process by *k*-fold method, other process were done, and finally the last model was generated. The data were collected from two physician offices and Javadalaemeh Clinic. Eventually, 208 patients were examined. Evaluation of 64 data mining models was done in Weka and MATLAB software. The evaluation was based on accuracy and precision.

In our study, Bayes-Net with accuracy of 95.67%, precision of 95.70% and sensitivity of 100% was found the best model for prediction and diagnosis of breast cancer. In addition, spread of Ductal is more than Lobular in Kerman.

Advantageous of BN model:

- BN had a high ability for prognosis (42).
- There was absence of access to valuable data sources BN still has a good performance (42).
- It had a high ability in controlling missing data.
- BN had a good ability to deal with unrelated data.

Comparison between ABML and BN models:

- The base of classification used in ABML was random classification (13) but BN was a model that incorporates two kinds of theory (presumption and graphical) to display a relationship between data (43).
- In both of them, percentages of sensitivity were the same, and percentage of accuracy and precision in BN is higher than RBFN.

- ABML had higher sensitivity to data noises, and BN has a good performance to make probability relationships.

Comparison between RF and BN:

- RF was made of some CART (Classification and Regression Trees). These CARTs used some random sample data among the main sample data (12), BN was made of algorithms that can predict with high precision and accuracy.
- RF was user-friendly model because it has just two parameters: The first parameter was number of random trees in forest, and the second parameter was number of predictor variables, which are set into subsets (12). BN had a perfect ability to predict values even in limitation of access to comprehensive data (42).
- In this study, BN model had higher percentages of accuracy, precision, and sensitivity than RF.

Comparison between Bagging and BN:

- Bagging was a model used to produce different models of a predictor (44). BN have algorithms that have many uses such as prognosis.
- Bagging had a considerable accuracy despite turmoil in learning collection it can modify accuracy (44). BN is a great way to represent real conclusions, and it is able to organize real conclusions (43).

Conclusion

To early predict and undergo prognosis of breast cancer utilization of data mining models is necessary. By a reliable data mining model, we can help physician to early diagnosis of breast cancer. Therefore, the cost of treatments dramatically decreases, and disease progression is prevented.

Ethical considerations

Ethical issues (Including plagiarism, informed consent, misconduct, data fabrication and/or falsification, double publication and/or submission, redundancy, etc.) have been completely observed by the authors.

Acknowledgements

Financial source was based on Institute for Futures Studies in Health, Kerman University of Medical Sciences.

Conflict of Interests

The authors declare that there is no conflict of interests.

References

1. Sephton SE, Sapolsky RM, Kraemer HC, Spiegel D (2000). Diurnal cortisol rhythm as a predictor of breast cancer survival. *J Natl Cancer Inst,* 92:994-1000.
2. Cook MB, Guénel P, Gapstur SM et al (2015). Tobacco and alcohol in relation to male breast cancer: an analysis of the male breast cancer pooling project consortium. *Cancer Epidemiol Biomarkers Prev,* 24:520-531.
3. Anderson AS, Macleod M, Mutrie N et al (2014). Breast cancer risk reduction-is it feasible to initiate a randomised controlled trial of a lifestyle intervention programme (ActWell) within a national breast screening programme? *Int J Behav Nutr Phys Act,* 11:156.
4. Borst MJ, Ingold JA (1993). Metastatic patterns of invasive lobular versus invasive ductal carcinoma of the breast. *Surgery,* 114:637-41.
5. Kurtz JM, Jacquemier J, Torhorst J et al (1989). Conservation therapy for breast cancers other than infiltrating ductal carcinoma. *Cancer,* 63:1630-5.
6. Kumar Y, Sahoo G (2013). Prediction of different types of liver diseases using rule based classification model. *Technol Health Care,* 21:417-432.
7. Delen D, Walker G, Kadam A (2005). Predicting breast cancer survivability: a comparison of three data mining methods. *Artif Intell Med,* 34:113-127.
8. Fiuzy M, Haddadnia J, Mollania N et al (2012). Introduction of a new diagnostic method for breast cancer based on Fine Needle

Aspiration (FNA) test data and combining intelligent system. *Iran J Cancer Prev,* 5:169-177.
9. Luo Z, Wu X, Guo S, Ye B (2008). Diagnosis of breast cancer tumor based on manifold learning and support vector machine. Information and Automation, 2008. ICIA 2008. International Conference on, IEEE, pp. 703-707.
10. Mangasarian OL, Street WN, Wolberg WH (2009). Breast Cancer Diagnosis and Prognosis via Linear Programming. July,
11. Khajouei R, Salehi S, Ahmadian L (2013). Methods Used for Evaluation of Health Information Systems in Iran. *J Health Adm,* 16:7-21.
12. Maroco J, Silva D, Rodrigues A et al (2011). Data mining methods in the prediction of Dementia: A real-data comparison of the accuracy, sensitivity and specificity of linear discriminant analysis, logistic regression, neural networks, support vector machines, classification trees and random forests. *BMC Res Notes,* 4:299.
13. Broomhead DS, Lowe D (1988). Radial basis functions, multi-variable functional interpolation and adaptive networks. Royal Signals and Radar Establishment Malvern (United Kingdom).
14. Lavrač N (1999). Selected techniques for data mining in medicine. *Artif Intell Med,* 16:3-23.
15. Bellazzi R, Zupan B (2008). Predictive data mining in clinical medicine: current issues and guidelines. *Int J Med Inform,* 77:81-97.
16. Fayyad U, Piatetsky-Shapiro G, Smyth P (1996). From data mining to knowledge discovery in databases. *AI Magazine,* 17:37-54.
17. Listgarten J, Damaraju S, Poulin B et al (2004). Predictive models for breast cancer susceptibility from multiple single nucleotide polymorphisms. *Clin Cancer Res,* 10:2725-37.
18. Hosseini A SF (2013). Determination of minimum data set to design a decision making system in order to prognosis breast cancer [BSc Thesis]. School of Management and Medical Information Science, Kerman University of Medical Science, Iran.
19. Montazeri M, Baghshah MS, Enhesari A (2015). Hyper-Heuristic algorithm for finding efficient features in diagnose of lung cancer disease. *arXiv preprint arXiv:1512.04652*.
20. Murphy K (2001). The bayes net toolbox for matlab. *Computing Science and Statistics,* http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.25.1216&rep=rep1&type=pdf
21. Platt J (1998). Fast training of support vector machines using sequential minimal optimization. *Advances in kernel methods—support vector learning,* 3.
22. Rojek I (2009). Classifier Models in Intelligent CAPP Systems. In: *Man-Machine Interactions.* Ed(s): Springer, pp. 311-319.
23. Kohavi R (1995). Wrappers for performance enhancement and oblivious decision graphs. Citeseer.
24. Melville P, Mooney RJ (2003). Constructing diverse classifier ensembles using artificial training examples. IJCAI'03 Proceedings of the 18th international joint conference on Artificial intelligence. pp. 505-510.
25. Dong L, Frank E, Kramer S (2005). Ensembles of balanced nested dichotomies for multi-class problems. Lecture Notes in Computer Science, vol 3721. Springer, pp. 84-95.
26. Caruana R, Niculescu-Mizil A, Crew G, Ksikes A (2004). Ensemble selection from libraries of models. Proceedings of the twenty-first international conference on Machine learning, ACM, pp. 18.
27. Webb GI (2000). Multiboosting: A technique for combining boosting and wagging. *Machine Learning,* 40:159-196.
28. Frank E, Hall M (2001). *A simple approach to ordinal classification*. ed. Springer.
29. Peng C, Liu L, Niu B et al (2011). Prediction of RNA-binding proteins by voting systems. *J Biomed Biotechnol,* 2011:506205.
30. Ho TK (1998). The random subspace method for constructing decision forests. *IEEE Trans Pattern Anal Mach Intell,* 20:832-844.
31. Khoshgoftaar TM, Seliya N (2004). The necessity of assuring quality in software measurement data. Software Metrics, 2004. Proceedings. 10th International Symposium on, IEEE, pp. 119-130.
32. Hall M, Frank E (2008). Combining Naive Bayes and Decision Tables. FLAIRS Conference, pp. 318-319.
33. Cohen WW (1995). Fast effective rule induction. Proceedings of the twelfth international

conference on machine learning, pp. 115-123.
34. Holte RC (1993). Very simple classification rules perform well on most commonly used datasets. *Machine Learning,* 11:63-90.
35. Devasena CL, Sumathi T, Gomathi V, Hemalatha M (2011). Effectiveness evaluation of rule based classifiers for the classification of iris data set. *Bonfring International Journal of Man Machine Interface,* 1:05-09.
36. Freund Y, Mason L (1999). The alternating decision tree learning algorithm. ICML, pp. 124-133.
37. Quinlan JR (1993). C4. 5: Programming for machine learning. *Morgan Kauffmann.*
38. Webb GI (1999). Decision tree grafting from the all-tests-but-one partition. IJCAI, Citeseer, pp. 702-707.
39. Kohavi R (1996). Scaling Up the Accuracy of Naive-Bayes Classifiers: A Decision-Tree Hybrid. KDD, Citeseer, pp. 202-207.
40. Park J, Tyan H-R, Kuo CJ (2006). Internet traffic classification for scalable qos provision. . Multimedia and Expo, 2006 IEEE International Conference on, IEEE, pp. 1221-1224.
41. Dong L, Frank E, Kramer S (2005). Ensembles of balanced nested dichotomies for multi-class problems. In: *Knowledge Discovery in Databases: PKDD 2005*. Ed(s): Springer, pp. 84-95.
42. Bockhorst J, Craven M, Page D, Shavlik J, Glasner J (2003). A Bayesian network approach to operon prediction. *Bioinformatics,* 19:1227-1235.
43. West P, Rutstein D, Mislevy RJ et al (2009). A Bayes net approach to modeling learning progressions and task performances. Learning Progressions in Science (LeaPS) Conference, Iowa City, IA, pp 257-292.
44. Breiman L (1996). Bagging predictors. *Machine Learning,* 24:123-140.