International Journal of Computational Intelligence Systems, Vol.2, No. 1 (March, 2009), 60-68

ACCURACY EVALUATION OF C4.5 AND NAÏVE BAYES CLASSIFIERS USING ATTRIBUTE RANKING METHOD

S.SIVAKUMARI
Department of Computer Science and Engineering,
Faculty of Engineering, Avinashilingam University for Women,
Coimbatore, Tamilnadu, India.
hod_cse_au@yahoo.co.in

R. PRAVEENA PRIYADARSINI
Department of Computer Science and Engineering,
Faculty of Engineering, Avinashilingam University for Women,
Coimbatore, Tamilnadu, India.

P. AMUDHA
Department of Computer Science and Engineering,
Faculty of Engineering, Avinashilingam University for Women,
Coimbatore, Tamilnadu, India.

Received: 15-05-2008
Revised: 21-01-2009

Abstract

This paper intends to classify the Ljubljana Breast Cancer dataset using C4.5 Decision Tree and Naïve Bayes classifiers. In this work, classification is carriedout using two methods. In the first method, dataset is analysed using all the attributes in the dataset. In the second method, attributes are ranked using information gain ranking technique and only the high ranked attributes are used to build the classification model. We are evaluating the results of C4.5 Decision Tree and Naïve Bayes classifiers in terms of classifier accuracy for various folds of cross validation. Our results show that both the classifiers achieve good accuracy on the dataset.

Keywords: Breast cancer, C4.5 Decision Tree, Naïve Bayes Classifiers, Information gain.

1. Introduction

Data mining is a step in knowledge discovery in data bases (KDD), which is the overall process of converting raw data into useful information. The data mining tasks are divided into two categories: predictive and descriptive. The predictive task is used to predict the value of target attribute based on the value of explanatory variable. One of the predictive modeling tasks is classification, which is a process of finding a model that describes and distinguishes data classes or concepts, for the purpose of being able to use the

model to predict the class of objects whose class label is unknown[1].

Breast cancer is a malignant tumor that has developed from the cells of the breast. It is the second leading cause of death for women and the second most dangerous cancer[2]. An analysis of the most recent data has shown that the survival rate is 88% after 5 years of diagnosis and 80% after 10 years of diagnosis[3]. Early and accurate detection of breast cancer is critical to the well-being of patients. Analysis of breast cancer data leads to cancer identification and classification which will facilitate proper treatment selection and drug development.

In our study, we have used the breast cancer dataset available in UCI Machine Learning Repository obtained from the University Medical Centre, Institute of Oncology, Ljubljana, Yugoslavia[4]. The main objective of this paper is to classify the dataset using C4.5 Decision Tree and Naïve Bayes classifiers. C4.5 generates decision tree for classification. It eliminates the problem of unavailable values, continuous attributes value ranges, pruning of decision trees and rule derivation. Naïve Bayes classifier estimates the class conditional probability by assuming that the attributes are conditionally independent within a class. When applied as a pre-processing step to machine learning, feature selection is valuable in dimensionality reduction, isolating the most important information and thereby improving classifier performance[5]. In this paper, we attempt to study the impact of information gain ranking method on the accuracy of the classifiers. We have used 10-fold cross validation method for evaluating the performance of the classifiers.

This paper is organized as follows: Section 2 discusses the literature survey. Section 3 explains the methodology carriedout for the classification analysis and section 4 presents the experimental results and discussions. Conclusion and future work are given in the last section.

2. Literature Survey

A literature survey showed that there have been several search studies on the prediction of survival rate using statistical and neural network approaches. Data mining approaches like decision trees have also been used in studies related to medical diagnosis and survival rate[6]. Burk *et al,* compared the 5 year, 10 year predictive accuracy of various statistical models with predictive accuracy of artificial neural network[7]. Lundin *et al.* used artificial neural network and logistic regression models to predict 5 -, 10 -, 15 - year breast cancer survival[8]. Pendharkar *et al.* used several data mining techniques for exploring patterns in breast cancer[9]. In the past usage of the Ljubljana dataset, an accuracy range of 66% to 78% has been achieved by several researchers. The results of the experiments performed by Michalski, *et al.,* regarding the prognosis of breast cancer recurrence-events yielded 66% classification accuracy[10]. Clark and Niblett performed a study to attempt to predict the recurrence-events of the breast cancer within 5 years[11,12]. Both the studies utilized 70/30 split of the data set comprising details of 286 patients. The resulting accuracy achieved was between 65% to 72% based on various algorithms used. Cestink et *al.,* achieved 78% accuracy[13]. Zhang and Su compared Naïve Bayes with decision tree learning algorithm c4.4 in terms of ranking, measured by Area Under the Curve (AUC)[14]. Jiang and Guo presented a lazy Naïve Bayesian classifier for ranking and compared it to Naïve Bayes and C4.4 measured by AUC[15] Huang et.al. compared the classification performance of Naïve Bayes, decision tree and Support Vector Machines (SVM) in terms of AUC and accuracy[16]

3. Methodology

In this paper we have analyzed the breast cancer dataset using C4.5 Decision Tree and Naïve Bayes algorithms. The classification task is performed using two methods. In the first method, all the attributes in the dataset are considered for building the classification model. In the proposed second method, the attributes are studied using information gain on the most representative attribute. The information thus obtained is used for studying the relationship of representative attribute with the remaining attributes. The accuracy of the classifiers is compared for various folds of cross validation.

3.1 C4.5 Decision Tree

Decision tree is a flow-chart like tree structure where each internal node denotes a test on an attribute and each branch represents the outcome of the test, and leaf nodes represent classes or class distribution. In order to classify an unknown sample, the attribute values of the sample are tested against the decision tree. Decision trees can be easily converted to classification rules. The Decision tree induction is based on greedy algorithm which constructs the trees in a top-down, recursive, divide and conquer method. C4.5 Decision Tree algorithm, which is the extension of the well known decision tree induction algorithm ID3, recursively visits each decision node selecting the optimal split. The process is continued until no further split is possible[1]. The algorithm uses the concept of information gain or entropy reduction to select the optimal split. Information gain is the increase in information produced by partitioning the training data according to the candidate split. The C4.5 algorithm chooses the split with highest information gain as the optimal split[17]. The information gain measure is used to select the best test attribute at each node in the tree. To avoid overfitting problem, C4.5 uses post-pruning method and thus increases the accuracy of the classification.

3.2. Naïve Bayes Classifier

Naïve Bayes classifier is one of the commonly used supervised learning methods. It deals with any number of attributes or classes and the results of this classifier are not affected by small quantity of noise in data. Naïve Bayes classifier predicts the class membership probabilities. It assumes class conditional independence and is based on Bayes theorem. The application of Bayes theorem in Naïve Bayesian classifier is explained in Eq.(1)

$$P(Ci \mid X) > P(Cj \mid X) \quad for\ 1 \le j \le m, j \neq i \qquad (1)$$

To maximize P (Ci | X), Bayes rule is applied as stated in Eq. (2)

$$P(C_i \mid X) = \frac{P(X \mid C_i)\, P(C_i)}{P(X)} \qquad (2)$$

P (X) is constant for all classes and P (C_i) is calculated as in Eq. (3),

$$P(C_i) = \frac{Number\ of\ training\ sample\ in\ a\ class}{Total\ number\ of\ training\ samples} \qquad (3)$$

To evaluate P(X | C_i), the naïve assumption of class conditional independence is used as in Eq. (4),

$$P(X|C_i) = \prod_{k=1}^{n} P(x_k \mid C_i) \qquad (4)$$

The given sample X is assigned to the class C_i for which P(X | C_i) P (C_i) is the maximum[1].

3.3 Performance Evaluation

3.3.1 10-fold Cross Validation

10-fold cross-validation is the standard way of measuring the accuracy of a learning scheme on a particular dataset. The data is divided randomly into 10 parts in which the class is represented in approximately the same properties as in full dataset. During each run, one of partitions is chosen for testing, while the remaining nine-tenths are used for training. Again, the procedure is repeated 10 times so that each partition is used for training exactly once.

4. Experimental Results and Discussions

This section presents the results of the experiment conducted to study the performance of the classifiers using WEKA tool kit[18].

4.1 Description of Dataset

In this study, the Ljubljana breast cancer dataset is taken for classifying breast cancer patients. The dataset includes 201 instances of no-recurrence-events class and 85 instances of recurrence-events class. The ten attributes are detailed in Table 1.

Table 1. Attribute Information

Attribute name	Value interval	Type
class	no-recurrence-events, recurrence-events	nominal
age	10-99	nominal
menopause	1t 40, ge40, premeno	nominal
tumor-size	0-59	nominal
inv-nodes	0-39	nominal
node-caps	yes, no	nominal
deg-malig	1,2,3	numeric
breast	left, right	nominal
breast-quad	left-up, left-low, right-up, right-low, central	nominal
irradiat	yes, no	nominal

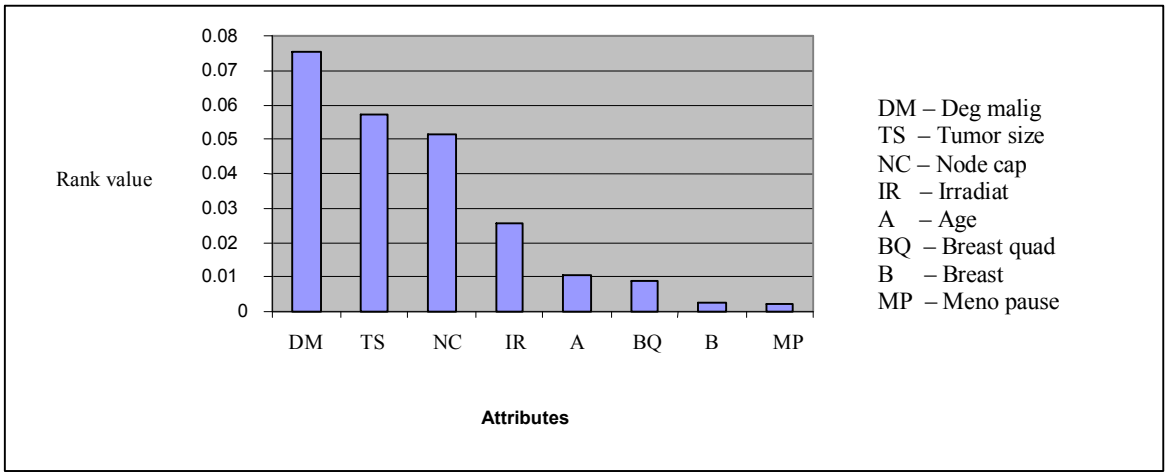
4.2 Classification Methods

4.2.1 *Method I*

Here, the classification is done using all the attributes of the dataset. The performance is evaluated for various folds of cross validation.

4.2.2 *Method II*

To study the performance of the classifiers, the attributes from the dataset are ranked using information gain ranking filter. The most informative attributes on the representative attribute is selected for the classification task. Ranking of the attributes are depicted in Table 2.



Graph 1: Attribute Ranking

The seven top ranked attributes are considered for building the classifier model.

4.3 Classification Results

To evaluate the performance of the algorithms, several measures have been employed in this study. They are listed below:

- Correctly Classified Instances and Incorrectly Classified Instances are the percentage of instances whose predicted value agrees with actual value and disagree with actual value respectively
- Mean Absolute Error averages the magnitude of individual error.
- Root Mean Squared Error calculates error in various computations.
- Relative Absolute Error is the total absolute error normalized by the error of predictor
- Root Relative Squared Error is the total squared error divided by the total squared error of a predictor

- True Positive Rate (TP Rate) = $\frac{TP}{TP+TN}$
- False Positive Rate(FP Rate) = $\frac{FP}{FP+TN}$
- Precision = $\frac{TP}{TP+FP}$
- Recall = $\frac{TP}{TP+FN}$
- F-Measure is the harmonic mean of precision and recall
- Area Under Curve (AUC) measures the ability of classifier to find the difference between two outcomes.

The True Positives (TP) and True Negatives (TN) are correct classifications. A False Positive (FP) occurs when the outcome is incorrectly predicted as positive when it is actually negative. A False Negative (FN) occurs when the outcome is incorrectly predicted as negative when it is actually positive[17].

Table 3 presents the results of evaluation of C4.5 Decision Tree with respect to different performance metrics using method I.

Table 3. Performance metrics of C4.5 using method I

Testing method		**5-fold (%)**	**6-fold (%)**	**7-fold (%)**	**8-fold (%)**	**9-fold (%)**	**10-fold (%)**
Correctly classified instances		237	233	237	236	237	236
Incorrectly classified instances		49	53	49	50	49	50
Mean absolute error		16.56	16.53	16.6	16.58	16.61	16.8
Root mean squared error		29.06	29.77	28.97	29.2	28.94	29.5
Relative absolute error		47.50	47.43	48.63	47.57	47.65	48.2
Root relative squared error		69.73	71.44	69.42	70.05	69.43	70.89
TP Rate	Premeno	78	78	78	78	78	78
	ge40	93	89.8	93	92.2	93	92.2
FP Rate	Premeno	7.4	11	7.4	8.1	7.4	8.1
	ge40	24.8	24.2	24.8	24.8	24.8	24.8
Precision	Premeno	92.1	88.6	92.1	91.4	92.1	91.4
	ge40	75.5	75.3	75.5	75.3	75.5	75.3
Recall	Premeno	78	78	78	78	78	78
	ge40	93	89.9	93	92.2	95	92.2
F-Measure	Premeno	84.5	83	84.5	84.2	84.5	84.2
	ge40	83.3	82	83.3	82.9	83.3	82.9
AUC	Premeno	89	88.3	89.1	88.7	88.7	87.6
	ge40	88.2	88.2	88.2	88.4	87.8	86.8
Accuracy	-	82.86	81.46	82.86	82.51	82.86	82.51
Error Rate	-	17.13	18.53	17.13	17.48	17.13	17.48

In this method highest accuracy is obtained in the seventh fold of cross validation and the accuracy does not vary much when the number of folds of cross validation increases. Also the TP rate, Precision, Recall and AUC metric are high in this fold.

The results of evaluation of C4.5 using method II for the same performance metrics is given in Table 4.

Table 4. Performance metrics of C4.5 using method II

Testing Method		5-Fold (%)	6-Fold (%)	7-Fold (%)	8-Fold (%)	9-Fold (%)	10-Fold (%)
Correctly Classified Instances		237	237	237	237	237	237
Incorrectly Classified Instances		49	49	49	49	49	49
Mean Absolute Error		16.56	16.61	16.6	16.56	16.61	16.55
Root Mean Squared Error		29.06	29	28.9	28.92	28.94	28.95
Relative Absolute Error		47.50	47.67	47.63	47.52	47.65	47.51
Root Relative Squared Error		69.73	69.58	69.52	69.38	69.43	69.47
TP Rate	Premeno	78	78	78	78	78	78
	ge40	93	93	93	93	93	93
FP Rate	Premeno	7.4	7.4	7.4	7.4	7.4	7.4
	ge40	24.8	24.8	24.8	24.8	24.8	24.8
Precision	Premeno	92.1	92.1	92.1	92.1	92.1	92.1
	ge40	75.5	75.5	75.5	75.5	75.5	75.5
Recall	Premeno	78	78	78	78	78	78
	ge40	93	93	93	93	93	93
F-Measure	Premeno	84.5	84.5	84.5	84.5	84.5	84.5
	ge40	83.3	83.3	83.3	83.3	83.3	83.3
AUC	Premeno	89	89.1	89.1	88.9	88.7	88.5
	ge40	88.2	88.1	88.2	88.1	87.8	87.8
Accuracy	-	82.86	82.86	82.86	82.86	82.86	82.86
Error Rate	-	17.13	17.13	17.13	17.13	17.13	17.13

In method II, the number of correctly classified instance remains the same irrespective of the number of folds and hence other metrics are also more or less the same. The results of evaluation of Naïve Bayes classifier using method I is given in Table 5.

Table 5. Performance metrics of Naïve Bayes using method I

Testing Method		5-Fold (%)	6-Fold (%)	7-Fold (%)	8-Fold (%)	9-Fold (%)	10-Fold (%)
Correctly Classified Instances		237	232	234	234	237	232
Incorrectly Classified Instances		59	54	52	52	49	54
Mean Absolute Error		18.49	18.13	18.07	17.9	17.98	17.84
Root Mean Squared Error		30.6	30.04	30.1	29.84	29.78	29.66
Relative Absolute Error		53.04	52.04	51.85	51.37	51.59	51.20
Root Relative Squared Error		73.42	72.08	72.22	71.61	71.46	71.1826
TP Rate	Premeno	80.7	81.3	80.7	82	83.3	80.7
	ge40	82.2	85.3	87.6	86	86.8	86
FP Rate	Premeno	17.6	14.7	12.5	14	13.2	14
	ge40	22.3	21.7	22.3	21	19.7	22.3
Precision	Premeno	83.4	85.9	87.7	86.6	87.4	86.4
	ge40	75.2	76.4	76.4	77.1	78.3	76
Recall	Premeno	80.7	81.3	80.7	82	83.3	80.7
	ge40	82.2	85.3	87.6	86	86.8	86
F-Measure	Premeno	82	83.6	84	84.2	85.3	83.4
	ge40	78.5	80.6	81.6	81.3	82.4	80.7
AUC	Premeno	89.3	90	89.9	90.1	90.3	90.3
	ge40	87.8	88.6	88.5	88.6	89	88.8
Accuracy	-	79.37	81.11	81.81	81.81	**82.86**	81.11
Error Rate	-	20.67	18.88	18.18	18.18	17.13	18.88

In this method highest accuracy is obtained in the nineth fold of cross validation and the accuracy does not vary much when the number of folds of cross validation increases.

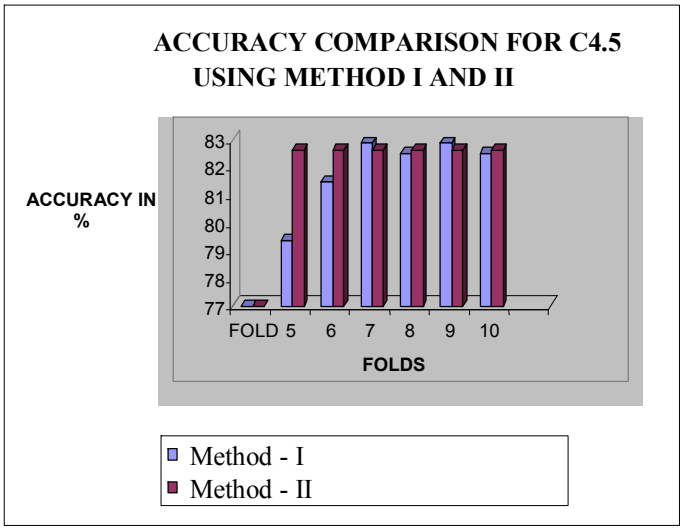
The results of evaluation of Naïve Bayes Classifier using method II is given in Table 6

Table 6. Performance metrics of Naïve Bayes using method II

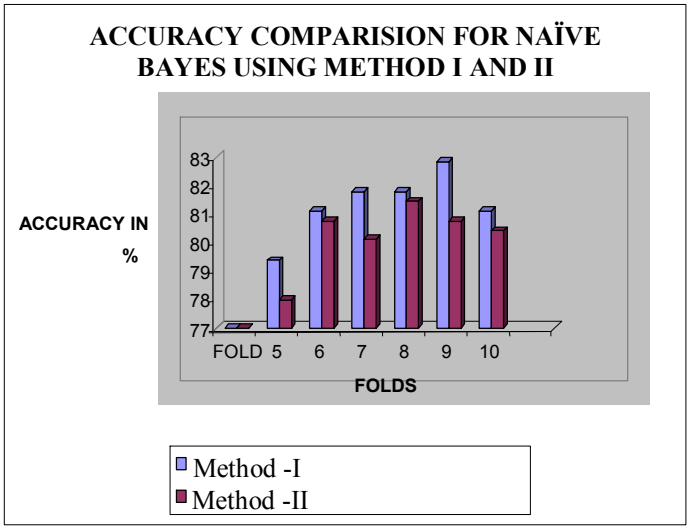
Testing method		**5-fold (%)**	**6- fold (%)**	**7-fold (%)**	**8-fold (%)**	**9-fold (%)**	**10-fold (%)**
Correctly Classified Instances		224	232	229	230	229	229
Incorrectly Classified Instances		62	56	57	56	57	57
Mean Absolute Error		18.43	18.27	18.01	17.98	18.17	17.82
Root Mean Squared Error		30.46	30.18	29.92	29.85	29.96	29.62
Relative Absolute Error		52.87	52.41	51.69	51.59	52.14	51.15
Root Relative Squared Error		73.10	72.43	71.80	71.61	71.84	71.06
TP Rate	Premeno	80	80.7	80	81.3	80.7	80
	ge40	79.8	85.3	85.3	86	85.3	85.3
FP Rate	Premeno	19.9	14.7	14.7	14	14.7	14.7
	ge40	22.9	22.3	22.9	21.7	22.3	22.9
Precision	Premeno	81.6	85.8	85.7	86.5	85.8	85.7
	ge40	74.1	75.9	75.3	76.6	75.9	75.3
Recall	Premeno	80	80.7	80	81.3	80.7	80
	ge40	79.8	85.3	85.3	86	85.3	85.3
F-measure	Premeno	80.8	83.2	82.8	83.8	83.2	82.8
	ge40	76.9	80.3	80	81	80.3	80
AUC	Premeno	89.4	89.6	90.1	90	90.1	90.3
	ge40	87.8	88.1	88.6	88.7	88.7	88.8
Accuracy	-	77.97	80.76	80.14	81.46	80.76	80.41
Error rate	-	22.02	19.23	19.58	18.53	19.23	19.58

In this method highest accuracy is obtained in the eighth fold of cross validation and minimum accuracy is obtained in the fifth fold.

The graphs 2 and 3 depict the accuracy of the classifiers using both the methods for various folds of cross validation.



Graph 2: Accuracy Comparison for C4.5 using method I and method II



Graph 3: Accuracy Comparison using method I and method II using Naive Bayes

Graphs 2 and 3 show that C4.5 decision tree classifier gets optimized in accuracy in the fifth fold of cross validation whereas Naïve Bayes classifier's accuracy varies for various folds of cross validation.

5. Conclusion and future work

In this paper, we have studied the data mining techniques to classify Ljubljana breast cancer dataset using C4.5 Decision Tree and Naïve Bayes algorithms. When all the attributes in the dataset are used C4.5 Decision Tree algorithm provides the accuracy in the range of 81.4% to 82.56% for various folds of cross-validation and the maximum accuracy is obtained in the fifth and seventh fold. The accuracy rate of Naïve Bayes algorithm varies in the range of 79.37% to 82.86% for various folds of cross validation and the maximum accuracy is obtained in the nineth fold.

When only the top ranked attributes are used for classification, C4.5 Decision Tree algorithm gets optimized by reaching constant accuracy rate irrespective of number of folds of cross validation. The accuracy rates of Naïve Bayes algorithm varies in the range of 77.97% to 81.46% and the maximum accuracy is obtained in the eighth fold.

The comparison between C4.5 and Naïve Bayes Classifiers reveals that C4.5 yields highly accurate results within few folds of cross validation considering the attribute with high information gain for classification while the Naïve Bayes Classifiers' performance does not show much significant improvement.

The accuracy of Naïve Bayes Classifers with attribute ranking has decreased in all folds of cross validation. Further extension of this work can be carriedout to improve the accuracy of Naïve Bayes Classifiers using boosting techniques with attribute ranking. Also, the algorithms can be evaluated in terms of runtime performance measure.

Acknowledgements

The authors would like to thank Dr. S.C. Sharma, Principal, R. V. College of Engineering, Bangalore, for his valuable guidance to carryout this work.

The authors would also like to thank M. Zwitter and M. Soklic, University Medical centre, Institute of oncology, Ljubljana, Yugoslavia for providing the breast cancer data set for running our experiments.

References

1. Jaiwei Han, Micheline Kamber, *Data mining Concepts and Techniques* (Morgan Kaufman, 2001).
2. Xianchun Xiong et al., *Analysis of Breast Cancer using Data Mining & Statistical Techniques,* IEEE Proceedings of 6th International Conference on Software Engineering (2005), pp. 82-87.
3. American Cancer Society, Breast Cancer Facts & Figures 2005-2006. Atlanta: American Cancer Society, Inc. (http://www.cancer.org/)
4. UCI Repository of Machine Learning Databases, University of California, Irvine, Dept of Information and Computer Science. http://www.ics.uci.edu/~mlearn/ML Repository.html
5. M. Prasad, *Online Feature Selection for Classifying Emphysema in HRCT Images,* International Journal of Computational Intelligence Systems, vol.1, no.2, (2008), pp.127-133.
6. Delen D, Walker G, Kadam A, *Predicting breast cancer survivability: A comparison of three data mining methods,* Artificial Intelligence in Medicine. 34(2) (2005), pp. 113-127.
7. Burk, H.B, et al., *Artificial Neural Networks Improve the Accuracy of Cancer Survival Prediction,* Cancer.79(4) *(*1997), pp. 857-862.
8. Lundin M, et al., *Artificial Neural Networks Applied to Survival Prediction in Breast Cancer,* Oncology. 57(4) (1999), pp. 281-286.
9. Pendharkar PC, et al., *Association, Statistical, Mathematical and Neural Approaches for Mining Breast Cancer Patterns,* Expert systems with Applications.17 (1999), pp. 223-232.
10. Michalski, et al., *The multi purpose incremental learning system aq15 and its testing application to three medical domain, in* Proceedings of the fifth National Conference on Artificial Intelligence. (Philadelphia, A: Morgan Kaufman.ss,1986), pp. 1041-1045.

11. Clark, P and Niblett, T, Induction in Noisy Domains, Progress in Machine Learning, in Proceedings of 2nd European working session on Learning, eds. I.Bratko & N.Lavac, , (Sigma Press 1987), pp.11-30.
12. Clark, and Niblett,T, *The CN2 induction algorithm,* Machine Learning Journal. 3(4) (1989), pp. 261-283.
13. Cestnik,G. et al., *Assistant-86: A Knowledge Elicitation Tool for Sophisticated Users* Progress in Machine Learning, eds. I.Bratko & N. Lavrac (Sigma Press 1987), pp. 31-45.
14. Harry Zhang and Jiang Su, *Naïve Bayesian Classifers for Ranking*, in Proceedings of 15th European Conference on Machine Learning (Springer 2004), pp. 501-512.
15. Liangxiao Jianq and Yuanyuan Guo, *Learning Lazy Naïve Bayesian Classifiers for Ranking*, in Proceedings of 17th International Conference on Tools with Artificial Intelligence (2005), pp. 412-416,.
16. Jin Huang et.al., *Comparing Naïve Bayes, Decision Trees and SVM with AUC and Accuracy*, in Proceedings of 3rd International Conference on Datamining (IEEE Computer Society Press, 2003), pp.553- 556.
17. Daniel T, Larase, Discovering Knowledge in Data. An introduction to Data mining, (John Wiley & Sons, Inc, 2005).
18. Ian H. Witten and Eibe Frank, Datamining: Practical Machine Learning Tools and Techniques, 2nd edn. (Elsevier, 2005.)