# A Comparison of Data Mining Classification Algorithms using Breast Cancer Microarray Dataset: A study

**N.Poomani[1] Dr.R.Porkodi[2]**
[1]Research Scholar [2]Assistant Professor
[1,2]Department of Computer Science
[1,2]Bharathiar University, Coimbatore, Tamilnadu, India

*Abstract*— Classification is one of the most familiar data mining Technique and model finding process that is used for transmission the data into different classes according to particular condition. Further the classification is used to forecast group relationship for precise data instance. It is generally construct models that are used to predict potential statistics trends. Medical science industry has huge amount of data, but unfortunately most of this data is not mined to find out hidden information in data, advanced data mining techniques can be used to discover hidden pattern in data. This Present study and analysis of various classification algorithms in data mining. This paper also presents the comparative study on four classification algorithms such as naïve bayes, CART, J48Graft, JRip algorithms using breast cancer dataset. The comparative results show that the J48Graft algorithm gives best classification accuracy than the rest of the algorithms. The algorithms also compared based on the execution time and error rate.

***Key words:*** Document classification, document clustering, entropy, accuracy, classifiers, clustering algorithms

## I. INTRODUCTION

Data mining is one of the many applications of machine learning. It is the task of discovering exciting patterns from bulky amount of data where the data can be stored in database [1]. Data mining techniques are used in healthcare management for, Diagnosis and Treatment, Healthcare Resource Management, Customer Relationship Management and Fraud and Anomaly Detection. Data mining can facilitate Physicians discover effective treatments and best practices, and Patients take delivery of in good health and more reasonable healthcare services. A particular active area of research in bioinformatics is the application and growth of data mining techniques to solve biological problems. Analyzing large biological data sets requires making sense of the data by inferring formation or generalizations from the data Applications of data mining to bioinformatics consist of gene decision, protein function domain detection, function prototype detection, protein function inference, disease diagnosis, disease prognosis, disease treatment optimization, protein and gene interaction network reconstruction, data cleansing, and protein sub-cellular location prediction [2]. Data mining techniques have been extensively applied for breast cancer diagnosis. Diagnosis is used to predict the presence of cancer.

In this paper, we have attempted to classify breast cancer data using classification algorithm. Breast cancer is the second most general cause of deaths from cancer along with women in the United States. In 2006, it is expected that about 212000 new cases of invasive breast cancer will be diagnosed, along with 58000 new cases of non-invasive breast cancer and 40000 women are expected to die from this disease [27].

Recurrence of therapeutically resistant disseminated disease is the major problem of breast cancer. By the time the primary tumor is diagnosed clinically or microscopically in many patients, evident metastases would have occurred already. Chemotherapy or hormonal therapy reduces the risk of distant metastases by one third. About 70% patients receiving treatment would have survived without the therapy. Predicting outcomes of the disease more accurately helps physicians make informed decisions regarding the need of adjuvant treatment. This leads to the development of individually modified treatments to maximize the efficiency of the treatment. This ultimately contributes to decrease in overall breast cancer and overall health care cost. Also, helps in an improvement in patients' quality of life [23]. Cancer is a disease characterized by uncontrolled growth and spread of the abnormal cells. Cancer also has the capability to invade other tissues that can be caused by both external factors like chemicals, radiation, tobacco etc., and internal factors like hormones, inherited mutations, immune conditions, etc. It is said that there are more than hundred different types of cancers [24].

The paper organized as follows: section 1 describes the introduction on data mining with breast cancer, section 2 describes the literature review, section 3 describes the various classification and algorithms, section 4 gives the summary of breast cancer dataset, and section 5 discusses the experimental results and finally the paper is concluded in section 6.

## II. RELATED WORK

Xin Yao [3] had attempted to implement neural network for breast cancer diagnosis. Negative correlation training algorithm was used to decompose a problem automatically and solve them. In this article the author has discussed two approaches such as evolutionary approach and ensemble approach, in which evolutionary approach can be used to design compact neural network automatically. The ensemble approach was aimed to tackle large problems but it was in progress.

Ramadevi Yellasiri, C.R.Rao [4] had proposed a new classification model called Rough Set Classifier for classifying the voluminous protein data based on structural and functional properties of protein. This model is fast and accurate and it can be used as an efficient classification tool than the others. This Classifier provides 97.7% accuracy. It is a hybridized tool comprising Sequence Arithmetic, Concept Lattice and Rough Set Theory. It can reduce the domain search space to 9% without losing the potentiality for the classification of proteins. The information about the family is identified using special arithmetic and utilizes it for reducing the domain search space is proposed. The rules that are generated are stored in Sequence Arithmetic database.

Huilin Xiong And Xue –Wen Chen says the new approach called kernel function, which improves the performance of the classifier in genetic data. The efficiency of a kernel approach has been probed in which it is depends upon on optimizing a data -dependent kernel model. The K-nearest-neighbor (KNN) and support vector machine (SVM) could be used as a classifier for performance analysis. Data set utilized here, ALL-AML Leukemia Data, Breast-ER, Breast-LN, Colon Tumor Data, Lung Cancer Data and Prostate Cancer from micro array data. Kernel optimization schemes have been discovered to classify gene expression data. The performance is evaluated when applying the optimized kernel in classifying gene expression data. Compared with KNN, SVM as “ok svm”, with optimized kernel provides better accuracy [5].

K. Srinivas**,** B. Kavitha Rani and Dr. A. Goverdhann haddeveloped the classification based data mining techniques such as Rule Based, Decision tree, Naïve Bayes and Artificial Neural Network to the massive volume of healthcare data. Using an age, sex, blood pressure and blood sugar medical profiles it can predict the likelihood of patients getting a heart disease [6].

David B.fogel [7] et al. had presented the evolving neural networks for detecting breast cancer and the related works used for breast cancer diagnosis using back propagation method with multilayer perceptron. In contrast to back propagation found that evolution computational method and algorithms were used often, perform more classic optimization techniques.

Dr.S.Santhosh baboo and S.Sasikala [8] had done a survey on data mining techniques for gene selection classification. This article dealt with most used data mining techniques for gene selection and cancer classification; particularly they have focused on four main emerging fields. They are neural network based algorithms, machine learning algorithms, genetic algorithm and cluster based algorithms and they have specified future improvement in this field

Aruna et.al. [9] Had proposed comparison of classification algorithms on the Wisconsin Breast Cancer dataset. They have analyzed the classification results of only five classification algorithms namely Naive Bayes, Support Vector Machines (SVM), Radial Basis Neural Networks (RB-NN), Decision trees J48 and simple CART.

D. Lavanya [10] had considered Decision tree classifier-CART with and without feature selection in terms of accuracy, time to build a model and size of the tree on various Breast Cancer Datasets are observed. From the results it is clear that, though we considered only breast cancer datasets, a specific feature selection may not lead to the best accuracy for all Breast Cancer Datasets. The best feature selection method for a particular dataset depends on the number of attributes, attribute type and instances.

Dalen et.al had Compared ANN, decision tree and logistic regression techniques for breast cancer survival analysis. They used the SEER data’s twenty variables in the prediction models. The decision tree with 93.6% accuracy and ANN with 91.2% were found more superior to logistic regression with 89.2% accuracy [11].

V. Krishnaiah et al [12] had developed a prototype lung cancer disease prediction system using data mining classification techniques. The most effective model to predict patients with Lung cancer disease appears to be Naïve Bayes followed by IF-THEN rule, Decision Trees and Neural Network. For Diagnosis of Lung Cancer Disease Naïve Bayes observes better results and fared better than Decision Trees.

R. Geetha Ramani had considered the data mining application on medical research for a predicting and discovering pattern base on detected symptom on health condition for process take a mammography, dermatology, orthopedic thyroids for data pre-processing execute classification for clinical test data lode test data for verification for classifier Malady classification. They support decision tree generate by the quinlan’s algorithm is smaller than the decision tree by the random tree classification technique [13].

J. Padmavati had presented [14] the comparative study on WBC dataset for breast cancer prediction using RBF and MLP along with logistic regression. Logistic regression was performed using logistic regression in SPSS package and MLP and RBF were constructed using MATLAB. It was observed that neural networks took slightly higher time than logistic regression but the sensitivity and specificity of both neural network models had a better predictive power over logistic regression.

Endo et al [15] had implemented common machine learning algorithms to predict survival rate of breast cancer patient. This study is based upon data of the SEER program with high rate of positive examples (18.5 %). Logistic regression had the highest accuracy; artificial neural network showed the highest specificity and J48 decision trees model had the best sensitivity.

K. Rajiv Gandhi, Marcus Karnan had proposed their paper on constructed classification rules using the Particle Swarm Optimization Algorithm for breast cancer datasets. In this study to cope with heavy computational efforts, the problem of feature subset selection as a pre-processing step was used which learns fuzzy rules bases using GA implementing the Pittsburgh approach. It was used to produce a smaller fuzzy rule bases system with higher accuracy. The resulted datasets after feature selection were used for classification using particle swarm optimization algorithm. The rules developed were with rate of accuracy defining then underlying attributes effectively [16].

## III. CLASSIFICATION TECHNIQUES USED IN HEALTHCARE

Classification is one of the most extensively used methods of data mining in healthcare. The classification algorithms on Breast Cancer data can be useful to predict the outcome of some diseases or discover the genetic performance of tumor. Classification model is build relating a predefined set of classes or ideas. The model is constructed by analyzing database tuples described by attributes. The classification is used to predict categorical class labels and classify data based on the training set [17]

Classification techniques in data mining are capable of processing a huge quantity of data. It can predict categorical class labels and classifies data based on training set and class labels and hence can be used for classifying newly available data. Thus it can be out lined as a predictable part of data mining and is gaining more popularity. Classification contains two phase such as training phase and testing phase, in training phase every sample in the training set is assumed to belong to a predefined class. The testing phase is unknown test samples are measured to classify using the model build using the training set.

This paper gives the detailed description of four algorithms such as naïve bayes, classification and regression tree, JRip, and J48graft, and the algorithms are compared using breast cancer dataset.

### A. *Naïve Bayes Classifier :*

The Naïve Bayes is a simple probabilistic classifier. Naïve Bayes is based on the assumption of mutual independency of attributes. This classifier is highly scalable, requiring a no of parameters. The algorithm works on the assumption, that variables provided to the classifier are independent [25].

Naive Bayes is one of the commonly used supervised technique and most effective statistical and probabilistic classification Algorithms in data mining. it is used to healthcare sector in this survey this technique is applied on breast cancer dataset for result provide an improved accuracy with low computational effort and very high speed [21] The goal of naïve bayes is to predict the breast cancer class and instances as accurately as possible. bayes theorem is defined as following formula

$$P(h_i|x_i) = \frac{P(x_i|h_i)P(h_i)}{P(x_i|h_i)+P(x_i|h_2)P(h_2)} \qquad (1.1)$$

### B. *CART (Classification and Regression Tree):*

CART algorithm is based on the decision tree technique. It was introduced by breiman in 1984.Classification and regression tree analysis is used to refer both of the classification tree and regression tree procedures. Trees used for regression and trees used for classification have some similarities but also some difference, such as the procedure used to determine where to split. This method is used to estimate the relationship between dependent and independent function. Machine learning gives the parameters are include, heuristic=true, MinNumObj =2.0; Num Folds Pruning =5; seed=1, size=1. Use pruning=true.

$$Gain(D,S) = \frac{Gain(D,S)}{H(D_1,D_s)} \qquad (1, 2)$$

### C. *J48 Graft Classifier:*

J48 algorithm is open source algorithm in weka data mining tool. J48 classifier is a simple C4.5 decision tree for classification. It creates a binary tree. The decision tree approach is most useful in classification problem. With this technique, a tree is constructed to model the classification process. Once the tree is built, it is applied to each tuple in the database and results in classification for that tuple [26].

J48graft classifier is the most popular tree classifier till today. Weka classifier package has its own version of C4.5 classifier known as j48 or j48graft it is used in weka platform.J48graft is an optimized implementation of C4.5 this classifier is experimented is this study with the parameters:

ConfidenceFactor=0.25;minNumObj=2;numFolds=3;reduced error pruning=false, seed=1.

### D. *JRip (JRipper):*

JRip is a fast algorithm for learning "IF-THEN" rules. The Jrip algorithm was proposed by William Cohen (1995) like decision trees rule learning algorithms are popular because the knowledge representation is very easy to interpret repeated Incremental Pruning to Produce Error Reduction (RIPPER) [22] is one of the basic and most popular algorithms. Classes are examined in increasing size and an initial set of rules for the class is generated using incremental reduced-error pruning. In this study, we evaluated RIPPER through JRip, an implementation of RIPPER in WEKA with the parameters: Check Error Rate=True; true debug=false; folds = 3; MinNo = 2; optimizations = 2; seed = 1; use Pruning = true.

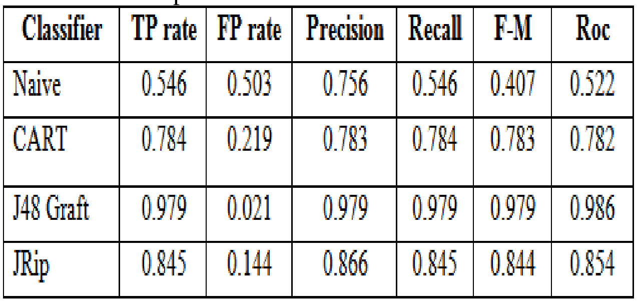
## IV. Dataset Description

To compare these classification algorithms using the breast cancer dataset is taken from UCI repository it contains the information about breast cancer [27].the dataset contains the information about 97 instance and 24482 attributes are used to classifier algorithm.

## V. Experimental Results

In this survey the accuracy of four data mining technique is compared by using the breast cancer dataset perform the weka tool using the four algorithms are compared based on the classification accuracy like TP rate. In this comparison the TP rate of j48 graft is 0.979 and the other algorithms are TP rate of naïve bayes is 0.546 and CART classifier accuracy algorithm is 0.784 and JRip TP rate is 0.845the classification accuracy, the accuracy of j48 graft algorithm is considered to be best when it compared to other classification algorithms.

In this study all data is considered as instances and features in the data are known as attributes. The simulation results are partitioned into several sub items for easier analysis and evaluation. Different performance matrix such as TPrate, FP rate and Precision, Recall, F-measure and ROC area are reported in TABLE I.



| Classifier | TP rate | FP rate | Precision | Recall | F-M | Roc |
|---|---|---|---|---|---|---|
| Naive | 0.546 | 0.503 | 0.756 | 0.546 | 0.407 | 0.522 |
| CART | 0.784 | 0.219 | 0.783 | 0.784 | 0.783 | 0.782 |
| J48 Graft | 0.979 | 0.021 | 0.979 | 0.979 | 0.979 | 0.986 |
| JRip | 0.845 | 0.144 | 0.866 | 0.845 | 0.844 | 0.854 |

Table 1: Performance Measures of Classification Algorithms

Fig 1 shows the comparison accuracy of four classifiers such as naïve bayes, CART, J48graft and JRip, based on 10-fold cross validation as a test method, the accuracy obtained by J48 is the best Classifier better than that produced by naive, JRip and CART. From the analysis of accuracy measures of j48 classifier performs well when compared to all Accuracy measures, TP rate, FP rate, Precision, Recall, Measure, and Roc Area, As a result is J48 classifier is better than others.

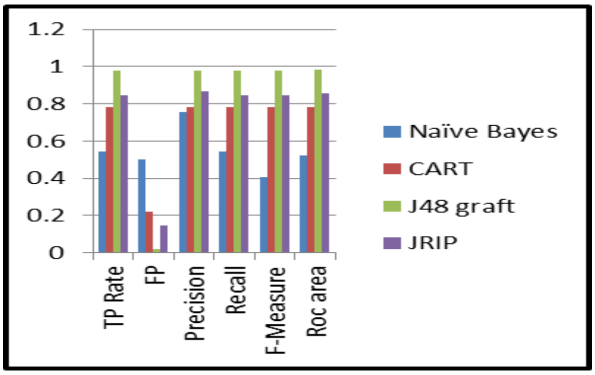


Fig. 1: Performance Measures

| Classifier | Classified Instance | In Classified Instance |
|---|---|---|
| Naïve Bayes | 53% | 44% |
| CART | 76% | 21% |
| J48 Graft | 95% | 2% |
| JRip | 82% | 15% |

Table 2: Performance of the Classifier Accuracy

This result section carried out some experiments in order to evaluate the performance of different algorithms. The breast cancer survivals in order to build a model, The TABLE II shows the result of correctly classified instance and incorrectly classified instance in breast cancer dataset. It describes the classification accuracy of Naïve Bayes, CART, and J48graft and JRip algorithms implemented in machine learning tool. The Fig 2 represent that J48Graft is more correctly classified instance, than the other classification algorithm.

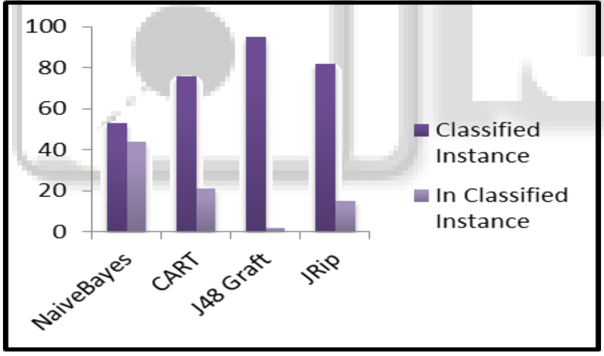


Fig. 2: Comparison of Correctly and Incorrectly Instances

| Classifier | Time Sec |
|---|---|
| Naïve Bayes | 0.66 |
| CART | 15.54 |
| J48 Graft | 3.96 |
| JRip | 3.44 |

Table 3: Time Comparison

From the TABLE III, it is observed that the shortest time which is around 0.66 seconds compared to other classify technique. CART algorithm requires the longest model building time which is around 15.54seconds and third one J48Graft classifier takes the time for build the model 3.96 seconds finally the JRip classifier requires the time for 3.44 seconds. The Fig 3 represented the naïve bayes classifier gives the shortest time result and the CART classifier produce the longest time to build the model.

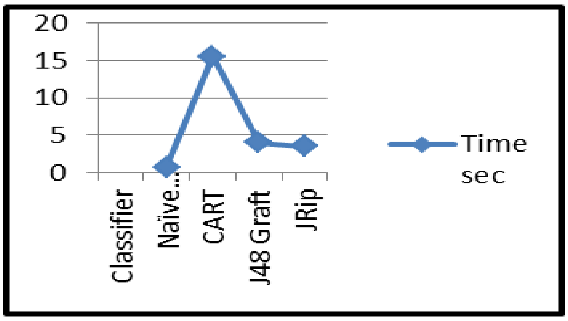


Fig. 3: Time comparison of classifier

| Classifier | Kappa Statistic | Mean Absolute Error | Mean Squared Error | Relative Absolute Error | Relative Squared Error |
|---|---|---|---|---|---|
| Naïve Bayes | 0.0456 | 0.4536 | 0.6735 | 90.9584 | 134.8801 |
| CART | 0.5658 | 0.3392 | 0.4118 | 68.0156 | 82.4738 |
| J48 Graft | 0.9587 | 0.0395 | 0.1405 | 7.9122 | 28.1295 |
| JRip | 0.6935 | 0.2459 | 0.3507 | 49.3151 | 70.2266 |

Table 4: Error Measurement

The TABLE IV shows the result of kappa statistic error, mean absolute error, mean squared error, relative absolute error and relative squared error in percentage for references and evaluation. The figure 4 represents the graphical analysis for training and simulation of error.

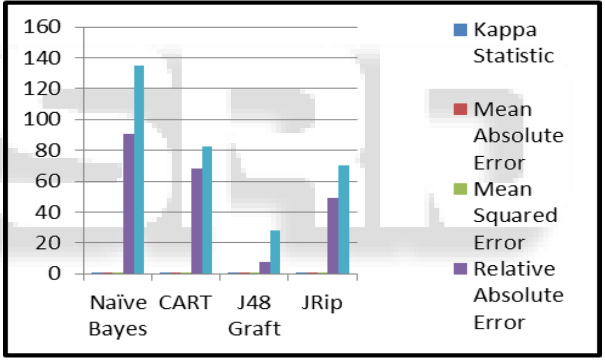


Fig. 4: Error Comparison

## VI. CONCLUSION

This paper have studied and compared on various supervised learning algorithms to predict the best classifier. The experimental result shows that the highest accuracy is found in J48graft classifier gives 0.979 with the lowest error rate 0.9587 among various classification algorithms. The next highest accuracy is found in JRip, accuracy rate is 0.845 with the lowest error rate 0.693.the CART algorithm has the next highest accuracy rate 0.784 with error rate is 0.565.the naïve bayes classifier proves the least accuracy rate 0.546 with error rate 0.045.based on the experimental result, it proves that the probabilistic model is not much suitable for classify breast cancer dataset. And decision tree based algorithms such as CART; JRip and J48 graft obtained better accuracy for breast cancer dataset.

## REFERENCES

[1] Fayyad, Piatetsky-Shapiro, Smyth, and Uthurusamy: "Advances in Knowledge Discovery and Data Mining", AAAI/MIT Press 1996.

[2] Jiawei Han and Micheline Kamber,"Data mining: concepts and techniques", San Francisco: Morgan Kaufmann Publishers, 2001.

[3] Xin Yao, Yong Liu "Neural Networks for Breast Cancer Diagnosis" 01999 IEEE.

[4] Ramadevi Yellasiri, C.R.Rao, "Rough Set Protein Classifier", Journal of Theoretical and Applied Information Technology, (2009).

[5] Huilin Xiong And Xue-Wen Chen,"Optimized KernelMachines for Cancer Classification Using gene Expression Data", Proceedings Of The 2005 IEEE Symposium On Computational Intelligence in Bioinformatics and Computational Biology, Pp.1-7, 2005.

[6] K. Srinivas , B. Kavitha Rani and Dr. A. Govrdhan, ―Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks‖ International Journal on Computer Science and Engineering (2010).

[7] David B.Fogel, Eugene C, Wasson, Edward M.Boughton "Evolving neural networks for detecting breast cancer". 1995 Elsevier Science Ireland Ltd.

[8] Dr.Santhosh baboo, S.Sasikala "A Survey on data mining techniques in gene selection and cancer classification"-April 2010 International journal of Computer science and information technology.

[9] Aruna, Dr S.P. Rajagopalan and L.V. Nandakishore, 2011 Knowledge Based Analysis Of Various Statistical Tools In Detecting Breast Cancer.

[10] D.Lavanya and Dr.K.Usha Rani," Analysis of feature selection with classification Breast cancer datasets", Vol.2-No.5, oct-nov: 2011, Pg.no:756-763.

[11] Delen Dursun, Walker Glenn and Kadam Amit, "Predicting breast cancer survivability: a comparison of three data mining methods," Artificial Intelligence in Medicine, vol. 34, pgno. 113-127, June 2005.

[12] Krishnaiah "Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques" International Journal of Computer Science and Information Technologies, Vol. 4 (1) 2013, 39 – 45 www.ijcsit.Com ISSN: 0975-9646.

[13] ShomonaGracia Jacob, R. GeethaRamani ―Mining of Classification patterns in clinical data through data mining algo‖ access from IEEE.

[14] Padmavati J., "A Comparative study on Breast Cancer Prediction Using RBF and MLP," International Journal of Scientific & Engineering Research, vol. 2, Jan. 2011

[15] A. Endo, T. Shibata and H. Tanaka (2008), Comparison of seven algorithms to predict breast cancer survival, Biomedical Soft Computing and Human Sciences, vol.13, pp.11-[16].

[16] Gandhi Rajiv K., Karnan Marcus and Kannan S., "Classification rule construction using particle swarm optimization algorithm for breast cancer datasets,"

[17] Survey of classification techniques in data mining in Proceedings of the International Multi Conference of Engineers and Computer Scientists 2009 Vol IIMECS 2009, March 18 -20, 2009, Hong Kong-classification

[18] Dr. K. Usha Rani, Parallel Approach for Diagnosis of Breast Cancer using Neural Network Technique.

[19] CL.Chang and C.H.Chen,"Applying decision tree and neural network to increase quality of dermatologic diagnosis", Expert Systems with Applications,Elsevier, vol. 36,(2009),pp. 35-4041.

[20] Abdelghani Bellaachia,Erhan guven, Predicting Breast cancer survivability using Data Mining Techniques

[21] I. H. Witten and E. Frank, "Data Mining: Practical Ma- chine Learning Tools and Techniques," 2nd Edition, Morgan Kaufmann, San Francisco, 2005.

[22] Breast cancer, from http: // www. cancer. gov/ cancertopics/types/breast access on [02-09-2014]

[23] American Cancer Society. Breast Cancer Facts & Figures 2005-2006. Atlanta: American Cancer Society, Inc.(http://www.cancer.org/)

[24] J. C. Bailar, T.A. Louis, P.W. Lavori, and M. Polansky, "A Classification for Biomedical Research Reports," *N Engl J Med*,, Vol. 311, No. 23, pp. 1482-1487, 1984

[25] Nong Y., *the Handbook of Data Mining* (Lawrence Earlbaum Associates, 2003)

[26] Aman Kumar Sharma, Suruchi Sahni, "A Comparative Study of Classification Algorithms for Spam Email Data Analysis", IJCSE, Vol. 3, No. 5, 2011, pp.189001895

[27] UCI Machine Learning Repository. http://www.ics.uci.edu/mlearn/MLRepository.html

[28] Breast cancer, from http://www.cancer.gov/ cancer topics/types/breast access on [02-09-2014]