2009 Third International Conference on Network and System Security

Early Breast Cancer Identification: Which Way to Go? Microarray or Image Based Computer Aided Diagnosis!

Jesmin Nahar[1], Kevin S. Tickle[1], A B M Shawkat Ali[1], Yi-Ping Phoebe Chen[2]
[1]School of Computing Sciences, Central Queensland University, Australia
E-mail :{ j.nahar, k.tickle, s.ali@cqu.edu.au}
[2]School of Information Technology, Deakin University, Australia
E-mail: phoebe@deakin.edu.au

Abstract—
The goal of this research is to develop a computer aided diagnostic (CAD) system that can detect breast cancer in the early stage by using microarray and image data. We verified the performance of six well known classification algorithms with various performance matrices. Although we do not suggest a unique classifier algorithm for a CAD system, we do identify a number of algorithms whose performance is very promising. The algorithms performance was validated by 3 images dataset; two have been used for the first time in this experiment. Multidimensional image filtering is adopted for the final data extraction. The image data classification performance is compared with microarray data. Results suggest the most effective means of breast cancer identification in the early stage is a hybrid approach.

Keywords- Breast cancer, image, microarray, computer aided diagnostic.

1.0 Introduction

Throughout the world cancer is a serious threat to human health. Cancer is a group of diseases with the common feature of uncontrolled growth of cells and has the ability to infiltrate and destroy normal body tissue. Cancer cells can spread all the way through the blood and lymph systems to different parts of the body. There are more than 100 different types of cancer. For the average woman the chance of developing invasive breast cancer at some stage of her is about 1 in 8, and the chance of dying from of breast cancer is 1 in 35 (1). In our research we investigate the efficiency of the machine learning algorithms on breast cancer-related image and microarray data. We consider two new breast image data in our experiment.

Image-based identification

In cancer research, image-based identification is going to become more popular day by day. Group of scientists have worked on image-based data for different cancer identification.
An image is an optical picture of an object. Medical image data mining is used to collect effective models, relations, rules, changes, irregularities and common laws from a mass amount of data and it has become a separate important discipline called Medical Imaging. The image-based machine-learning technique can accelerate the processing speed and accuracy of the diagnostic decisions made by doctors. In recent times (2) the rapid development of digital medical devices and medical information databases has included not only the structured information of patients, but also non-structured medical image information. Antonie, et al. suggests mammography as the most reliable method in the early detection of breast cancer. Due to the high volume of mammograms to be read by physicians, the accuracy rate tends to decrease, and then the automatic reading of digital mammograms becomes highly desirable. Additionally, as they suggest, double reading of mammograms (consecutive reading by two physicians or radiologists) increases the accuracy, but this is problematic due to its high cost (3). Wang, et al. used a decision tree algorithm for mammography classification and constructed a medical image classifier. They found that the system performed quite accurately (98% accuracy) and therefore revealed the potential of data mining in medical (2) treatment assistance. Rangayyan et al. developed the digital imaging and image analysis systems to detect mammographic features, classify them, and present visual prompts to the radiologist for breast cancer detection (4). Also, Walker et al. (5), Sheshadri and Kandaswamy (6), Rodrigues et al. (7), Twellmann et al. (8), Nattkemper et al. (9) worked on image data for early cancer detection.

Microarray-based identification

Currently, microarray tumor gene expression profiles are used for early cancer diagnosis. By allowing the monitoring of expressions level for thousands of genes simultaneously, such methods will guide researchers to a more complete understanding of the molecular differences among tumors, and thus to a finer and more reliable classification.
Lu and Han et al. suggests that single classifier is not superior to all the others in the aspect of classification accuracy. Statistically based classifiers shown better accuracy but they are not sufficiently good classifiers in the case of cancer classification (10). Berrar et al. proposed a

978-0-7695-3838-9/09 $26.00 © 2009 IEEE
DOI 10.1109/NSS.2009.75

probabilistic neural network (PNN) for muticlass cancer classification. The main disadvantage of PNNs is that all training data must be stored in the input layer, requiring a large amount of memory (11). Sharma Paliwal suggest the Gradient LDA (linear discriminant analysis), technique which avoids the singularity problem associated with the within-class scatter matrix and shows its usefulness for cancer classification. The proposed method achieves lower misclassification error as compared to several other previous techniques (12). In recent times several researches have been done on microarray data for early cancer detection (13-20).

It is established that breast cancer is the most common cancer and responsible for the highest death rate for females. It is also noticed that breast cancer is rare in males but men can get breast cancer, though rates are appreciably lower than for women (21). Therefore this paper addresses an important research issue. We provide emphasis of the importance of early detection of breast cancer. A CAD system can help anyone in early breast cancer diagnosis. The benefit of early diagnosis is that it can reduce the mortality rate of breast cancer patients. Currently, microarray and image based cancer classification is most popular in the IT community. Many scientists work in these two sectors for early cancer detection.

2.0 Background and Current Research

Modern computer technology is helping the researcher in breast cancer identification. As discussed earlier, researchers use two types of methods for breast cancer identification: image and microarray based methods. In most cases, patients in the early stages of breast cancer have no symptoms, so it is difficult to identify the cancer stage and condition of the patient. By the time symptoms are identified, the cancer cells have already spread to other parts of the body. The methodologies of current breast cancer treatment have many limitations and in most cases detection is not 100% accurate. That is the reason that a treatment program/plan utilizing a combination of tests, such as Clinical breast exam, Mammogram, Ultrasound and Magnetic Resonance Imaging (MRI), is considered to be the optimal means of detection at any age. On the other hand, to undergo a clinical breast exam, mammogram, ultrasound and MRI to determine the exact condition of the cancer are expensive for the patient. Modern microarray and image detection technologies are helping in this regard. Michaelson et al. proposes a new method to estimate tumor growth rate from information taken from the numbers and sizes of breast cancers found at screening. It appeared that the median doubling time for invasive breast cancer is approximately 130 days. (22). Rangayyan and Nguyen propose mammogram-based methods for the analysis and classification of breast masses. They used receiver operating curve methods to measure the accuracy and did not compare their performance with the fractal dimension based measure (23). Another study identifies a set of gene marks and mediates breast cancer metastasis to the lungs. A number of these genes serve double functions, providing growth advantages both in the primary tumor and in the lung microenvironment (24). Delen et al. argues that a decision tree (C5) is the best predictor with 93.6% accuracy on the holdout sample (this prediction accuracy is better than any reported in the literature). Artificial neural networks came second with 91.2% accuracy and the logistic regression models were the worst of the three with 89.2% accuracy (25). Cheng et al. found that the individual samples of most models had high rates of concordance in their outcome predictions (26).

In our research we investigate the efficiency of the machine learning algorithms on breast cancer-related image and microarray data. We consider two new breast image data in our experiment.

3.0 Data Preparation

We consider three images and two microarray breast cancer datasets in our experiment. The BcancerImage data we generated from cancer affected and normal breast images (27). The bcw_noise is noise data of breast-cancer-wisconsin (28).We consider this noise dataset to evaluate the machine learning algorithms' performance. The well known breast-cancer-wisconsin data is available in the UCI data repository (39). The beastmit is microarray data, which is available from the Massachusetts Institute of Technology (MIT) (29). The breastcancernsu is also microarray data, which is available from Singapore National University (NSU) (30). A basic description of these datasets has been summarized in Table 1.

Table 1. Breast image and microarray dataset description.

Name of Data	No of Attributes	No of Instances	Classes	Types
BcancerImage	153	361	2	Image
bcw_noise	18	682	2	Image
breast-cancer-wisconsin	10	698	2	Image
breastmit	1213	146	4	microarray
breastcancernsu	24481	1066	2	microarray

We extracted image data from the images. We considered 4 pairs of breast images. First we read the image data by using the Matlab function. After that we used multidimensional image filtering to remove the noise data from extracted data. The image filtering function is also implemented in the Matlab package.

4.0 Experimental Design

Six of the most popular machine learning algorithms, Naive Bayes, Sequential Minimizing Optimisation (SMO), IBk, AdaBoostM1, J48 and PART, were selected for this experiment. All the algorithms' description and Java based implementations are available in the data mining book (31). To evaluate the prediction accuracy of the above mentioned algorithms for the data, percentage split test options were used. Prediction metrics considered in this study are given below with their mathematical expression (32, 33).

Percent Correct (PC) p_i are the total number of predicted instances N is the total number of instances	$PC=\frac{1}{N}(p_i)\text{X}100\%$
Root Mean Squared Error (RMSE) a_i are the total number of actual instances	$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(p_i - a_i)^2}{N}}$
Mean Absolute Error (MAE)	$MAE = \frac{\sum_{i=1}^{N} \mid p_i - a_i \mid}{N}$
Root Relative Squared Error (RRSE) a - mean of a_i	$RRSE = \sqrt{\frac{\sum_{i=1}^{N}(p_i - a_i)^2}{\sum_{i=1}^{N}(a - a_i)^2}}$
Relative Absolute Error (RAE)	$RAE = \frac{\sum_{i=1}^{N} \mid p_i - a_i \mid}{\sum_{i=1}^{N} \mid a - a_i \mid}$
True Negative Rate p is the number of correct predictions that an instance is negative, q is the number of incorrect predictions that an instance is positive,	$TNR = \frac{p}{p+q} \times 100\ \%$
False Negative Rate r is the number of incorrect of predictions that an instance negative, and t is the number of correct predictions that an instance is positive.	$FNR = \frac{r}{r+t} \times 100\ \%$
F Measure R is the recall, have a weight of $\beta \in (0;+1)$ and P is the precision, have a weight of 1	$F = \frac{(\beta+1)RP}{(R+\beta P)}$
Kappa Statistic A is the number of times the appraisers agreed, E is the number of times appraisers would have agreed by chance	$K = \frac{A - E}{N - E}$

5.0 Experimental Results

We considered several parameters to evaluate the machine learning algorithms. It is constantly better to justify any algorithm's performance from different points of view. Here we reflected on two approaches to measure the performance. First, we observed the correctly classified performance for various algorithms and then we calculated the computational performance as well. Since the majority of the datasets hold less than 1000 instances, we preferred to use a 10 fold cross validation procedure for our experiment. Moreover, we ran each dataset 10 times. Therefore all the below results start with (100). We used Percent_correct, i.e. accuracy, Mean_absolute_error, Root_mean_squared_error, Relative_absolute_error, Root_relative_squared_error, True_negative_rate, False_negative_rate, F_measure, Kappa_statistic measures to evaluate the algorithms' performances. The mentioned above measures descriptions are available in any basic statistics books. Based on these measures we found that SMO is the best suited algorithm to produce CAD for breast cancer study.

Analysing: **Percent_correct**

Dataset	NaiveBayes	SMO	IBk	AdaBoostM1	J48	PART
BcancerImage	(100) 75.29	93.50	98.86	92.70	95.03	95.37
bcw_noise	(100) 96.43	96.88	96.03	95.27	94.57	95.01
breast-cancer-wisconsin	(100) 96.07	96.80	95.67	95.01	94.59	94.81
breastcancernsu	(100) 49.88	71.55	61.45	59.71	54.76	53.79
breastmit	(100) 51.26	90.90	86.04	65.04	73.49	73.67

Analysing: **Mean_absolute_error**

Dataset	NaiveBayes	SMO	IBk	AdaBoostM1	J48	PART
BcancerImage	(100) 0.24	0.07	0.01	0.13	0.06	0.05
bcw_noise	(100) 0.04	0.03	0.04	0.06	0.06	0.05
breast-cancer-wisconsin	(100) 0.04	0.03	0.04	0.06	0.07	0.06
breastcancernsu	(100) 0.50	0.28	0.39	0.40	0.46	0.46
breastmit	(100) 0.24	0.26	0.08	0.26	0.14	0.14

Analysing: **Root_mean_squared_error**

Dataset	NaiveBayes	SMO	IBk	AdaBoostM1	J48	PART
BcancerImage	(100) 0.48	0.24	0.06	0.24	0.20	0.18
bcw_noise	(100) 0.18	0.16	0.19	0.19	0.22	0.21
breast-cancer-wisconsin	(100) 0.18	0.16	0.20	0.19	0.22	0.20
breastcancernsu	(100) 0.71	0.50	0.59	0.56	0.65	0.66
breastmit	(100) 0.48	0.33	0.24	0.36	0.35	0.35

Analysing: **Relative_absolute_error**

Dataset	NaiveBayes	SMO	IBk	AdaBoostM1	J48	PART
BcancerImage	(100)	48.87	13.01	2.88	26.36	11.12	10.25
bcw_noise	(100)	8.00	6.85	9.04	12.55	14.02	11.71
breast-cancer-wisconsin	(100)	8.70	7.09	9.89	13.08	15.20	12.85
breastcancernsu	(100)	100.12	56.83	77.74	79.93	91.30	92.47
breastmit	(100)	70.03	75.45	22.58	74.37	40.69	39.38

Analysing: **Root_relative_squared_error**

Dataset	NaiveBayes	SMO	IBk	AdaBoostM1	J48	PART
BcancerImage	(100)	95.78	47.02	12.30	47.14	39.39	36.78
bcw_noise	(100)	37.21	34.30	39.75	39.53	45.38	43.75
breast-cancer-wisconsin	(100)	38.84	34.59	41.53	40.07	46.08	42.66
breastcancernsu	(100)	140.80	99.15	117.25	112.47	130.79	131.80
breastmit	(100)	115.81	79.17	57.91	85.49	83.99	83.93

Analysing: **True_negative_rate**

Dataset	NaiveBayes	SMO	IBk	AdaBoostM1	J48	PART
BcancerImage	(100)	0.70	0.98	0.99	0.93	0.96	0.96
bcw_noise	(100)	0.98	0.96	0.91	0.93	0.92	0.93
breast-cancer-wisconsin	(100)	0.95	0.97	0.97	0.97	0.95	0.96
breastcancernsu	(100)	0.97	0.70	0.72	0.59	0.54	0.52
breastmit	(100)	0.97	0.98	0.94	1.00	0.88	0.89

Analysing: **False_negative_rate**

Dataset	NaiveBayes	SMO	IBk	AdaBoostM1	J48	PART
BcancerImage	(100)	0.19	0.11	0.02	0.08	0.06	0.06
bcw_noise	(100)	0.04	0.02	0.01	0.03	0.04	0.04
breast-cancer-wisconsin	(100)	0.03	0.04	0.07	0.08	0.07	0.08
breastcancernsu	(100)	0.96	0.27	0.48	0.40	0.44	0.44
breastmit	(100)	1.00	0.11	0.11	1.00	0.48	0.48

Analysing: **F_measure**

Dataset	NaiveBayes	SMO	IBk	AdaBoostM1	J48	PART
BcancerImage	(100)	0.76	0.93	0.99	0.92	0.95	0.95
bcw_noise	(100)	0.97	0.98	0.97	0.96	0.96	0.96
breast-cancer-wisconsin	(100)	0.95	0.95	0.94	0.93	0.92	0.92
breastcancernsu	(100)	0.06	0.71	0.54	0.57	0.53	0.53
breastmit	(100)	0.00	0.88	0.81	0.00	0.46	0.44

Analysing: **Kappa_statistic**

Dataset	NaiveBayes	SMO	IBk	AdaBoostM1	J48	PART
BcancerImage	(100)	0.51	0.87	0.98	0.85	0.90	0.91
bcw_noise	(100)	0.92	0.93	0.91	0.90	0.88	0.89
breast-cancer-wisconsin	(100)	0.91	0.93	0.90	0.89	0.88	0.88
breastcancernsu	(100)	0.01	0.43	0.23	0.19	0.09	0.07
breastmit	(100)	0.33	0.87	0.80	0.42	0.62	0.62

Analysing: **Time_training**

Dataset	NaiveBayes	SMO	IBk	AdaBoostM1	J48	PART
BcancerImage	(100)	0.01	0.15	0.00	0.12	0.04	0.05
bcw_noise	(100)	0.01	0.15	0.00	0.09	0.02	0.03
breast-cancer-wisconsin	(100)	0.00	0.13	0.00	0.04	0.01	0.02
breastcancernsu	(100)	1.00	2.02	0.06	16.14	3.98	4.64
breastmit	(100)	0.11	0.99	0.01	0.42	0.86	1.28

Analysing: **Time_testing**

Dataset	NaiveBayes	SMO	IBk	AdaBoostM1	J48	PART
BcancerImage	(100)	0.00	0.00	0.06	0.00	0.00	0.00
bcw_noise	(100)	0.00	0.00	0.08	0.00	0.00	0.00
breast-cancer-wisconsin	(100)	0.00	0.00	0.04	0.00	0.00	0.00
breastcancernsu	(100)	0.24	0.02	1.04	0.00	0.00	0.00
breastmit	(100)	0.05	0.00	0.23	0.00	0.00	0.00

The computational time was measured in seconds. We considered the computational performance in the two different phases: model buildup time called Time training and model evolution time called Time testing. Overall SMO is a bit more expensive than some other algorithms. But the computational time differences are not a significant barrier to designing a CAD for breast cancer study using SMO.

6.0 Discussion

This research addressed the problem of early breast cancer identification that one of the world's most serious health issues. It would help the breast cancer CAD designer to choose an appropriate machine learning algorithm. We tested the algorithms' performances from different angles by using the most useful statistical measures. These measures verified the algorithms' classification performance. Moreover, we tested the computational performance for all the algorithms. Finally, we found that SMO is the best suited machine learning algorithm for breast cancer study. It performed better in the scenarios, image and microarray breast data analysis. Comparatively it was better at handling image data. Since medical practitioners are struggling to recognize the breast cancer problem in the early stage, it seems evident that microarray analysis could be a more useful technique for a breast cancer CAD system. We suggest both microarray and an image-based approach could be jointly adopted in the CAD system. This hybrid approach could verify more accurately if any breast cancer is present. We plan to extend this research within the larger breast cancer domain to discover the reason why the image data is more meaningful for breast cancer study.

REFERENCES

(1)Breast cancer incidence and mortality, (2008). http://search.live.com/images/results.aspx?q=picture+of++men+breast+cancer+&mkt=enau#focal=f542963b7f211062e24de410636ce8ae&furl=http%3A%2F%2Fwww.revolutio

nhealth.com%2Ffabric%2Fimages%2Fimage%2F1876.png, Accessed: 17th October, 2008.
(2) S. Wang, M. Zhou, and G. Geng, "Application of Fuzzy Cluster Analysis for Medical Image Data Mining," 2005, Proc of the IEEE International Conference on Mechatronics & Automation, pp 631-636.
(3) M-L. Antonie, O. R.. Zaiane, and A. Coman, "Application of Data Mining Techniques for Medical Image Classification," 2001, Proc. of the second International Workshop on Multimedia Data and Mining, pp. 94-101.
(4) R.M. Rangayyan, F.J. Ayres, and J.E.L. Desautels, "A review of computer-aided diagnosis of breast cancer: Toward the detection of subtle signs," 2007, Journal of the Franklin Institute, pp. 344 (3-4), 312-348.
(5) W. H. Jr. Land, T.Masters, Lo. J.Y, and D.W. McKee. "Application of Evolutionary Computation and Neural Network Hybrids for Breast Cancer Classification Using Mammogram and History Data," 2001 2, Digital Object Identifier 10.1109/CEC.2001.934320 pp. 1147 – 1154.
(6) H. S. Sheshadri, and A. Kandaswamy, "Experimental investigation on breast tissue classification based on statistical feature extraction of mammograms," 2007, Computerized Medical Imaging and Graphics, pp. 31(1), 46-48.
(7) P.S. Rodrigues, C. Ruey-Feng, and J.S. Suri, "Non-Extensive Entropy for CAD Systems of Breast Cancer Images," 2006,Computer Graphics and Image Processing, SIBGRAPI '06. 19th Brazilian Symposium, pp. 121 – 128.
(8) T. Twellmann, A. Meyer-Baese, O. Lange, S. Foo, and T.W. Nattkemper, "Model-free visualization of suspicious lesions in breast MRI based on supervised and unsupervised learning," 2008, Engineering Applications of Artificial Intelligence, pp. 21(2), 129–140.
(9) T. W. Nattkemper, B. Arnrich,, O. Lichte, W. Timm, A. Degenhard, L. Pointon, C. Hayes, M.O. Leach, and The UK MARIBS Breast Screening Study, "Evaluation of radiological features for breast tumour classification in clinical screening with machine learning methods," 2005, Artificial Intelligence in Medicine, pp.34, 129-139.
(10) Y. Lu, and J. Han, "Cancer classification using gene expression data,"2003, Information Systems, pp.28, 243-268.
(11) D. P. Berrar, C. S. Downes, and W. Dubitzky, "Multiclass Cancer Classification Using Gene Expression Profiling and Probabilistic Neural Networks,'2003, Proc. of the Pacific Symposium on Biocomputing, pp. 5-16.
(12) A. Sharma, and K. K. Paliwal, "Cancer classification by gradient LDA technique using microarray gene expression data,"2008, Data & Knowledge Engineering, pp.66(2), 338-347.
(13) A. Berns, "Cancer: gene expression in diagnosis,"2000 Nature, pp. 403, 491-492.
(14) C. Campbell, Y. Li, and M. Tipping, "An efficient feature selection algorithm for classification of gene expression data," 2001, NIPS 2001 Workshop on Machine Learning Techniques for Bioinformatics, Vancouver, Canada.
(15) L. J. V. Veer, H. Dai, M. J. V. D. Vijver, Y. D. He, A. A. M. Hart, M. Mao, L. Hans, H. L. Peterse, K. V. D. Kooy, M.J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhoven, C. Roberts, P.S. Linsley, R Bernards, and S. H. Friend, "Gene expression profiling predicts clinical outcome of breast cancer,"2002, Nature, pp.415, 530-536.
(16) S. L. Pomeroy, P. Tamayo, M, Gaasenbeek, L.M. Sturla, M. Angelo, M. E. Margaret E McLaughlin, J. Y. K. Kim, L. C. Goumnerova, P.M. Black, C. Lau, J. C. Allen, D. Zagzag, J.M. Olson, T. Curran, C. Wetmore, J.A. Biegel, T. Poggio, S. Mukherjee, R., Rifkin, A. Califano, G., Stolovitzky, D.N. Louis, J. P. Mesirov, E.S. Lander, and T. R. Golub, "Prediction of central nervous embryonal tumour outcome based on gene expression,"2002, Nature, pp. 415, 436-442.
(17) J. Nahar, S. Ali, YPP, Chen, "Microarray Data Classification Using Automatic SVM Kernel Selection,"2007, DNA and Cell Biology, pp. 26(10),707-712.
(18) H-S. Wong, and H-Q. Wang, "Constructing the gene regulation-level representation of microarray data for cancer classification,"2008, Journal of Biomedical Informatics, pp. 41(1): 95-10.
(19) Y. Lee, and C. K. Lee, "Classification of multiple cancer types by mulitcategory support vector machines using gene expression data,"2003, Bioinformatics, pp. 19, 1132-1139.
(20) L. E. Peterson, and M. A. Coleman, "Machine learning-based receiver operating characteristic (ROC) curves for crisp and fuzzy classification of DNA microarrays in cancer research," 2008, International Journal of Approximate Reasoning, pp. 47(1), 17-36.
(21)Statistics and Research (2008). http://www.breasthealth.com.au/statisticsresearch/index.html. Accessed, 2nd November, 2008.
(22) J. Michaelson, S Satija, R. Moore, G. Weber, E. Halpern, A. Garland, D. Kopans, and K. Hughes, "Estimates of the Sizes at Which Breast Cancers Become Detectable on Mammographic and Clinical Grounds,"2003, Journal of Women's Imaging, pp. 5(1),3-10.
(23) R.M. Rangayyan and T.M. Nguyen, "Pattern classification of breast masses via fractal analysis of their contours," 2005, International congress series, pp. 1281, 1041-1046.
(24) AJ. Minn, GP. Gupta, PM. Siegel, PD, Bos, W. Shu, DD. Girl, Agnes,V, AB. Olshen, WL. Gerald and J. Massague, "Genes that mediate breast cancer metastasis to lung," 2005, Nature, pp. 436(28).
(25) D. Delen, G. Walker, and A. Kadam, "Predicting breast cancer survivability: a comparison of three data mining methods," 2005, Artificial Intelligence in Medicine, pp. 34(2),113–127.
(26) F. Cheng, S.O. Daniel, W. Lodewyk, W. Britta, S.A. Dimitry, A.B.N. Nuyten, J.V.V. Laura, M.P. Charles,

"Concordance among Gene-Expression-Based Predictors for Breast Cancer," 2006, N Engl J Med. pp. 10(355),560-569.
(27) Breast image (2008). http://www.ibchelp.org/pictures.html, Accessed 2nd Nov 2008.
(28) D. J. Newman, S. Hettich, C.L. Blake, and C.J. Merz, "Repository of machine learning databases," 1998, (http://www.ics.uci.edu/~mlearn/MLRepository.html).
Irvine, CA: University of California, Department of Information and Computer Science.
(29) Cancer datasets (2008). http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi. Accessed 10th Nov 2008.
(30) J. Li, and H. Liu, "Kent Ridge Biomedical Data Set Repository,"2006, Singapore.
http://sdmc.i2r.a-star.edu.sg/rp/. Accessed 4th September, 2006.
(31) I. H. Witten, E. Frank, "Data Mining: Practical machine learning tools and techniques," 2005, 2nd Edition". Morgan Kaufmann, San Francisco.
(32)A. B. M. S. Ali and S. A. Wasimi, "Data Mining: Methods and Techniques," 2007, Victoria, Australia, Thomson.
(33) Nahar J., Chen YPP, Ali S., "Kernel Based Naive Bayes Classifier for Breast Cancer Prediction", Journal of Biological Systems, 15(1): 17-25, 2007.