PATTERN BASED FEATURE SELECTION AND CLASSIFICATION SCHEME FOR CANCER DIAGNOSIS DATA ANALYSIS

Mrs. M.karthika¹ and Mr. S.Muruganandam²

¹M.Phil Full time Research Scholar, PG and Research Department Of Computer Science
²Asst. Professor, PG and Research Department of Computer Science Vivekanandha College of Arts & Sciences for Women (Autonomous), Namakkal-637 205, Tamil Nadu, India

Abstract - Data classification methods are applied to discover the category of the transactional data items. The classification task is divided into two phases. Pattern extraction on the labeled transaction is carried out under the training or learning phase. All the classification operations are performed under the testing phase. Support Vector Machine (SVM), Artificial Neural Networks (ANN) and decision tree based techniques are adapted for the classification process. The correlation coefficient based feature selection mechanism is combined with the SVM and ANN techniques for the breast cancer diagnosis process. The breast tumor classification system is constructed with Biclustering and Support Vector Machine (SVM) techniques. The correlation coefficient based feature selection and feature score assignment models are combined to improve the feature discovery process. The pattern extraction operations are carried out using the Biclustering algorithm. The discovered patterns are used in the SVM based classification process. The mini batch gradient descent training algorithm is integrated with the SVM classification method. The self adaptive learning rate estimation mechanism is also used to improve the classification results.

Keywords - Breast Tumor Analysis, Support Vector Machine, Neural Networks, Feature Selection, Clusters and Classification

I. INTRODUCTION

Nowadays, computer science and medical area are nested in order to provide proper prognosis or diagnosis of the human diseases. Many computational techniques are used for the identification of the health problems. Data mining has turned into a crucial procedure for registering applications in the space region of medicine. In this study, it is aimed to identify the breast cancer with the help of data mining classification methods. The dataset named Wisconsin Diagnostic Breast Cancer Database (WDBC) is obtained from Wisconsin Madison University. The classification techniques used on WDBC are Decision Trees (DT), Support Vector Machines (SVM), Artificial Neural Network (ANN) and also the ensemble of them. SPSS Clementine software was used for the experimental studies. The models used in Clementine are support vector machine model, C5.0 model and neural net model. Furthermore, feature selection algorithm is used in order to reduce the dimensionality of the dataset. In order to measure the performance, 10-fold cross validation technique is used on WDBC dataset. That is, the data are partitioned by the ratio 90:10% for training and testing. This is done ten times by adifferent10% being tested each time.

Breast cancer is the most common cancer among females. It is the second most important cause of death among women, as it comes directly after lung cancer. The disease is characterized by malignant tumors when cells in the breast tissue divide and grow without normal controls on cell death and cell division. In fact, it is the most common form of cancer in females; that is affecting approximately 10% of all them at some period of their life. The breast cancer attacks the Arab countries in the last ten years
compared with other countries. The disease targets women in Arab countries of age of 30, while infecting women above 45 years in European countries. Although scientists do not know the exact causes of most types of breast cancer, they know some of the risk factors that increase the probability of woman infection, such attributes are: age, genetic risk and family history.

Medical scientists consider that mammography screening as the most dependable method of early detection of breast cancer. Nevertheless, the digital mammogram images are sometimes difficult to be read due to their low contrast and differences in the types of tissues. In such cases, fine needle aspiration cytology is adopted. The tissue has to be removed for examination using breast biopsy techniques. A false positive detection may cause an unnecessary biopsy. Statistics show that only 20-30 percentages of breast biopsy cases proved cancerous. A false negative detection, an actual tumor remains hidden that could lead to higher costs or even to the cost of a human life. However, the existing tumors are of different types, different shapes and some of them have the individuality of the normal tissue. So, it is necessary to develop better identification methods to recognize breast cancer.

Data mining methods can help to reduce the number of false positives and false negative decisions. The objective is to assign patients to either a ‘benign’ group that does not have breast cancer or a ‘malignant’ group that has strong evidence of having breast cancer based on cytological proven tumor data. Decision tree approaches are the most widely used data mining methods for several reasons. It has the ability to generate understandable rules and to handle both continuous and categorical variables. This paper investigates the effectiveness of four efficient decision tree models (C&R, CHAID, QUEST and C5.0) on the diagnosis of the Wisconsin Breast Cancer dataset. The results are compared to those obtained using radial basis function kernel support vector machine (RBF-SVM). SVM has been chosen as it is considered a good candidate because of its high generalization performance.

II. RELATED WORK

There are many studies [3] done for the diagnosis or the prognosis of the breast cancer using University of California Irvine (UCI) dataset called Wisconsin Diagnostic or Prognostic Breast Cancer Databases [7] (WDBC or WPBC). The frequently used method in these studies is a linear programming-based classification method which is called MSM-Tree (MSM-T) . While Olvi et al. and Wolberg et al. used only the MSM-Tree technique for the classification in their studies, Wolberg et al. in1995 also used the Logistic Regression technique. For the estimation of predictive accuracy, in all these studies 10-fold cross validation is chosen; Wolberg et al. used leave-one-out cross validation for the prognosis of the breast cancer. All these relevant studies get the same accuracy which is 97.5% with the help of MSM-T; whereas, Wolberg et al. gets the accuracy of 96.2% using the Logistic Regression technique.

Nowadays, it is still being challenged on data mining classification techniques applied for breast cancer diagnosis or prognosis. For instance, a framework for diagnosis and prognosis of tumor utilizing FP (frequent pattern mining) growth algorithm is introduced and Decision Tree algorithm to anticipate the likelihood of disease in setting to age. Also this study [1] shows that when contrasted with apriori algorithm, fp-growth is more effective as time required to execute is not exactly apriori and likewise memory use is less in fp-growth. So they find that fp-growth is more proficient to utilize. Furthermore, Sumbaly et al. (2014) have picked J48 Decision Tree algorithm [5] to build up the model of the study called Diagnosis of Breast Cancer using Decision Tree Data Mining Technique. In this study [2] 10-fold cross validation [4] is utilized to specify test and training information. The J48 algorithm is utilized on the dataset utilizing WEKA (Java Toolbox for different information mining method) [6] after data preprocessing (CSV group). As a result, they gain an accuracy of 94.5637% with the performance of J48 Decision Tree. The last but not the least, Gupta et al. (2011) gives an investigation of different review and technical papers on breast cancer prognosis and diagnosis issues and investigates that data mining
methods offer extraordinary guarantee to reveal examples covered up in the information that can help the clinicians in making of decision.

### III. BREAST CANCER DIAGNOSIS

Worldwide, breast cancer is the fifth most common cause of cancer death. In 2005, breast cancer caused 502,000 deaths worldwide. Among women worldwide, breast cancer is the most common cause of cancer death. In the United States, breast cancer is the third most common cause of cancer death. In 2007, breast cancer is expected to cause 40,910 deaths in the U.S. Among women in the U.S., breast cancer is the most common cancer and the second-most common cause of cancer death. Women in the U.S. have a 1 in 8 lifetime chance of developing invasive breast cancer and a 1 in 33 chance of breast cancer causing their death. In the U.S., both incidence and death rates for breast cancer have been declining in the last few years. Nevertheless, a U.S. study conducted in 2005 by the Society for Women's Health Research indicated that breast cancer remains the most feared disease, even though heart disease is a much more common cause of death among women.

Early breast cancer can in some cases present as breast pain or a painful lump. Since the advent of breast mammography, breast cancer is most frequently discovered as an asymptomatic nodule on a mammogram, before any symptoms are present. A lump under the arm or above the collarbone that does not go away may be present. When breast cancer has invaded the dermal lymphatics - small lymph vessels of the skin, its presentation can resemble skin inflammation and thus is known as inflammatory breast cancer. In inflammatory breast cancer, the breast cancer is blocking lymphatic vessels and this can cause pain, swelling, warmth and redness throughout the breast. Although there may have been no previous signs of breast cancer and the cancer might be missed in screening mammograms, Inflammatory Breast Cancer is at least locally advanced at presentation (LABC) and Stage IIIB. Immediate staging tests are required to rule out distant metastases which might already be present making it Stage IV.

Changes in the appearance or shape of the breast can raise suspicions of breast cancer. Another reported symptom complex of breast cancer is Paget's disease of the breast. This syndrome presents as eczematoid skin changes at the nipple and is a late manifestation of an underlying breast cancer. Most breast symptoms do not turn out to represent underlying breast cancer. Benign breast diseases such as fibrocystic mastopathy, mastitis, functional mastodynia and fibroadenoma of the breast are more common causes of breast symptoms. The appearance of a new breast symptom should be taken seriously by both patients and their doctors, because of the possibility of an underlying breast cancer at almost any age.

Breast cancer presents as metastatic disease, that is, cancer that has spread beyond the original organ. Metastatic breast cancer will cause symptoms that depend on the location of metastasis. More common sites of metastasis include bone, liver, lung and brain. Unexplained weight loss can occasionally herald an occult breast cancer, as can symptoms of fevers or chills. Bone or joint pains can sometimes be manifestations of metastatic breast cancer, as can jaundice or neurological symptoms. Pleural effusions are not uncommon with metastatic breast cancer. Obviously, these symptoms are "non-specific," meaning they can also be manifestations of many other illnesses. Breast cancer screening is an attempt to find unsuspected cancers. The most common screening methods are self and clinical breast exams, x-ray mammography, Breast Magnetic resonance imaging (MRI), ultrasound, Miraluma and genetic testing.

### IV. BREAST CANCER DIAGNOSIS USING ARTIFICIAL NEURAL NETWORK AND SUPPORT VECTOR MACHINES

Scientists and researchers have conducted several experiments and researches regarding breast cancer diagnosis due to the spread of breast cancer within a large number of women which led to the
loss of many lives. The development of medical techniques confirmed the possibility to survive and recover from the disease when detected in its early stages. The evolution of the field of medicine and technology has led to the availability of a huge amount of data stored and made available to researchers which contributed to the innovation and uses of several techniques to detect the symptoms of this disease.

After tumor discovery, the challenge radiologists and doctors face is how to distinguish between a benign and malignant tumor to take appropriate actions. The development of machine learning made it possible to build a model that can learn and diagnose the tumor based on past diagnosis collected from patients. Breast cancer analysis methods and techniques have been improved over the last ten years and many automated classification models have been developed over the past few years. Results vary from one technique to the other. Comparing different models help find a model that achieve high accuracy.

Many methods were proposed to solve this classification problem. Two of the widely used classification algorithms used to solve this problem is ANN and SVM. Other earlier work in this area has been done on the same dataset used in this paper. SVM with Radial Basis (RBF) and Polynomial functions as the kernel functions for the classifier and achieved 97.13%. The other two papers used ANN one used ANN back propagation with genetic algorithms and in the first cleaning process of data where it gave 100% accuracy meanwhile the second cleaning process of data gave 83.36% accuracy, the second paper used Feed forward neural network model and back propagation learning algorithm with momentum and variable learning rate and the study proved that multilayer neural network is quicker than one layer neural network.

Moreover, other studies for breast cancer have been done where they used SVM with RBF kernel function and F-Score feature selection and they achieved a high result of 99.51%. Decision tree modeling algorithms used and has been compared with results that were achieved by radial basis function kernel support vector machine (RBF- SVM) and SVM was the highest in accuracy for both Training 99.976% and Testing 96.635%. Furthermore, they used SVM with three types of different kernel functions, linear, polynomial and RBF, with genetic programming, and when increasing number of runs, the result for each class increases giving 97.67 for malignant and 99.78 for benign after 100 runs. In the experimented ANN by training for feed forward neural network using island model based differential evolution algorithm and it achieved 99.97%. Another study compared SVM and ANN for breast cancer diagnosis and ANN achieved higher result. Also another study experimented with SVM, ANN, Decision Trees and ensemble of those methods and the ensemble gave the best accuracy.

In this paper we propose to use SVM and ANN combined with features selection to build a classification model of the breast cancer diagnosis to determine whether the tumor is benign or malignant. The features were selected using the correlation coefficient between each feature and the target variable to form feature subsets to choose the combination that gives best results. The experiments have been carried out using Weka and tested using 10-fold cross validation combined with use of optimal parameters. Results from empirical work indicated that SVM showed better performance results on classifying the samples with 97.1388 % accuracy compared to ANN that achieved 96.7096 % accuracy.

V. ISSUES ON BREST CANCER DIAGNOSIS TECHNIQUES

Support Vector Machine (SVM), Artificial Neural Networks (ANN) and decision tree based techniques are used for the classification process. The Support Vector Machine (SVM) is a classification technique build with kernal methods. The kernal functions allow the kernal method to generate a non-linear decision using linear methods. Kernel methods are used to construct a hyper plane by dividing the attribute space. The kernal method classifies the data using the support vectors which are the instances that outline each hyper plane. The Artificial Neural Network (ANN) technique performs the data
classification with information and its weight relationships. Input layer, hidden layer and output layers are used in the ANN based classification process. The correlation coefficient based feature selection model is combined with the SVM and ANN for the breast cancer analysis.

- Feature selection process is not optimized
- Limited classification accuracy levels
- Computational complexity is increased in the feature selection process
- Complex learning rate assignment process

VI. PATTERN BASED FEATURE SELECTION AND CLASSIFICATION SCHEME

The breast tumor classification system is constructed with Biclustering and Support Vector Machine (SVM) techniques. The correlation coefficient based feature selection and feature score assignment models are combined to improve the pattern discovery process. The mini batch gradient descent training algorithm is integrated with the SVM classification method. The self adaptive learning rate estimation mechanism is also used to improve the classification results. Clustering and classification schemes are integrated to analyze the cancer patient diagnosis data values. The feature selection process is combined with pattern discovery process. The Support Vector Machine (SVM) classifier is improved with optimal features. The system is partitioned into five major modules. They are Data preprocess, Pattern discovery with Biclusters, Feature selection process, Optimal feature selection process and Data classification.

Data preprocess is performed to correct missing and noise transactions. The Biclustering algorithm is applied for the pattern discovery process. The feature selection is carried out with correlation coefficients. The mini batch gradient descent algorithm is applied for the optimal feature selection process. The Support Vector Machine (SVM) based classifier is used to detect the breast tumors.

The cancer diagnosis details are imported from textual data files. Noisy and missing data values are corrected in the data preprocess. Redundant data values are removed from the database. Missing elements are assigned using aggregation based substitution mechanism. Cleaned data values are referred as optimal data sets. The Biclustering technique is applied to fetch the transaction patterns. The Biclustering scheme simultaneously clusters the data matrix in both dimensions with row and column relationships. A Biclusters is a local coherent pattern made with subset of tumor cases and features. The biclustering process is divided into three phases. They are seed discovery, bicluster construction and merge and removal of redundant biclusters. The bicluster seeds are discovered using hierarchical clustering on the each columns. The biclusters are constructed by utilizing seeds with mean-square-residue (MSR) score. The biclusters are compared to handle the merge and removal of redundant biclusters.

The feature selection process is carried out on the breast tumor data sets. The feature scores are assigned with reference to the features and its associated class details. The features are selected using the correlation coefficient based feature selection mechanism. The features are used in the clustering and classification process. The correlation coefficient based feature selection process is combined with bicluster based pattern selection process. The batch gradient descent algorithm is required to analyze the whole data for single parameter update. The Mini batch gradient descent algorithm is used to support the optimal feature selection process. The region based data analysis is carried out for the optimal feature selection process. The breast tumor classification process is performed using the Support Vector Machine (SVM) classifier and Bicluster based SVM methods. The Support Vector Machine classifier uses the fixed learning rate for the classification process. The correlation coefficient based feature selection scheme is combined in the Bicluster based SVM scheme. The self adaptive learning rate estimation mechanism is used to improve the classification results.
VII. EXPERIMENTAL ANALYSIS

The breast cancer classification system is built to classify the cancer diagnosis data values. Feature selection process is carried out to identify the features from the breast cancer diagnosis data values. The class identification and cancer data analysis system is developed to estimate the cancer disease risk levels. The classification process is divided into two levels. The pattern feature selection process is applied to select transactional features from the cancer diagnosis data values. The breast cancer gene expression data is used in the system. The dataset is downloaded from the University of California, Irwin (UCI) Machine learning repository. The system is tested with Feature Selection based Support Vector Machine (FS-SVM) and Optimized Feature selection based Support Vector Machine (OFS-SVM) techniques. The Feature Selection based Support Vector Machine (FS-SVM) scheme is composed with Bicluster and Correlation coefficient based feature selection technique and Support Vector Machine technique. The Optimal Feature Selection based Support Vector Machine (OFS-SVM) scheme is built with Biclusters and minibatch gradient descent scheme and Support Vector Machine methods. The automatic learning rate estimation model is used to discover the training rate levels autonomously.

The feature selection based breast cancer classification process is tested with three performance measures. They are detection latency, false positive rate and false negative rate. The detection latency analysis compares the class identification period for the breast cancer diagnosis data. Figure 7.1. Shows the detection latency analysis between the Feature Selection based Support Vector Machine (FS-SVM) and Optimized Feature selection based Support Vector Machine (OFS-SVM) methods. The analysis result shows that the Optimized Feature selection based Support Vector Machine (OFS-SVM) scheme reduces the detection latency 20% than the Feature Selection based Support Vector Machine (FS-SVM) scheme.

The false positive rate and false negative rate measures are employed to estimate the decision making accuracy level of the system. The false positive rate analysis is estimated with the positive discriminatory results and the falsely assigned positive results. Figure 7.2.. Shows the False Positive Rate analysis between the Feature Selection based Support Vector Machine (FS-SVM) scheme and Optimized Feature selection based Support Vector Machine (OFS-SVM) schemes. The analysis result shows that the Optimized Feature selection based Support Vector Machine (OFS-SVM) scheme reduces the False Positive Rate 30% than the Feature Selection based Support Vector Machine (FS-SVM) scheme.

The false negative rate analysis is estimated with the negative discriminatory results and the falsely assigned negative results. Figure 7.3. Shows the False Negative Rate analysis between the Feature Selection based Support Vector Machine (FS-SVM) scheme and Optimized Feature selection based Support Vector Machine (OFS-SVM) schemes. The analysis result shows that the Optimized Feature selection based Support Vector Machine (OFS-SVM) scheme reduces the False Negative Rate 25% than the Feature Selection based Support Vector Machine (FS-SVM) scheme.

Figure No: 7.1. Detection Latency Analysis between FS-SVM and OFS-SVM on cancer Dataset
VIII. CONCLUSION

The medical decision support systems are built with the data mining based techniques. Support Vector Machine (SVM) and Artificial Neural Network (ANN) techniques are used for the cancer data classification process. Biclustering based pattern selection process is integrated with the correlation coefficient based feature selection process. The Mini batch gradient descent algorithm for optimal feature selection process is adapted to improve the SVM classification scheme. The system integrates the clustering and classification techniques for the breast cancer data analysis process. The pattern extraction operations are carried out with the Biclustering technique. The feature selection process is
improved with patterns and correlation coefficient models. The classification accuracy is improved with minimum computational complexity.

REFERENCES

[1] Majali, J., Niranjan, R., Phatak, V. and Tadakhe, O. Data Mining Techniques For Diagnosis And Prognosis of Breast Cancer. International Journal of Computer Science and Information Technologies (IJCSIT), Vol. 5 (5), 2014, 6487-6490.

[2] Sumbaly, R., N. Vishnusri and S. Jeyalatha Diagnosis of Breast Cancer using Decision Tree Data Mining Technique. International Journal of Computer Applications Volume 98 – No. 10, 2014.

[3] Kharya, S. Using Data Mining Techniques for Diagnosis And Prognosis of Cancer Disease, International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), 2012.

[4] Gokhan Zorluoglu and Mustafa Agaoglu, “Diagnosis of Breast Cancer Using Ensemble of Data Mining Classification Methods” International Journal of Bioinformatics and Biomedical Engineering, Vol. 1, No. 3, 2015, pp. 318-322.

[5] Bhargava, N., Sharma, G., Bhargava, R. and Mathuria, M. Decision Tree Analysis on J48 Algorithm for Data Mining. Proceedings of International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 6, 2013.

[6] Gupta, S., Kumar, D. and Sharma, A. Data Mining Classification Techniques Applied for Breast Cancer Diagnosis and Prognosis. Indian Journal of Computer Science and Engineering (IJCSE), 2011.

[7] Bache, K. and Lichman, M. UCI Machine Learning Repository. Irvine, CA: University of California, School of Information and Computer Science, 2013.