Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 3, Issue. 11, November 2014, pg.223 – 232

RESEARCH ARTICLE

A PROPORTIONAL LEARNING OF CLASSIFIERS USING BREAST CANCER DATASETS

M.Sadhana* A.Sankareswari, M.C.A., M.Phil.**
*M.Phil(Computer Science), Research Scholar,
Vivekanandha College for Women, Unjanai, Tiruchengode, India
******Assistant Professor in Computer Science
Vivekanandha College for Women, Unjanai, Tiruchengode, India

***ABSTRACT:** The aim of this research is to find out the best classifier with respect to accuracy on three different databases of breast cancer (Wisconsin Breast Cancer (WBC), Wisconsin Diagnosis Breast Cancer (WDBC) and Wisconsin Prognosis Breast Cancer (WPBC)) by using classification accuracy method. In the performance criterion of supervised learning classifiers such as Decision trees and SVM are compared, to find the best classifier in breast cancer datasets (WBC). Support Vector Machines (SVMs, also support vector networks) are supervised learning models with associated learning algorithms that analyze data and recognize patterns, used for classification and regression analysis.*

SVM proves to be the most accurate classifier with accuracy of 96.99%. In the performance of decision tree classifier with or without feature selection in breast cancer datasets, WBC. The selected attributes in the dataset are: Uniformity of Cell Size, Mitoses, Clump thickness, Bare Nuclei, Single Epithelial cell size, Marginal adhesion, Bland Chromatin and Class. In this paper compare the performance of SVM classifier and Decision Tree Classifier. Finally evaluate the performance based on the resource utilizations of classification and also based on the time efficiency. This paper mainly focuses on comparing the classifier and produces the classifier results and evaluates the performance of individual classifiers.

***Keywords:** SVM, Decision Tree, Data Mining, Wisconsin Breast Cancer, Wisconsin Diagnosis Breast Cancer, Wisconsin Prognosis Breast Cancer.*

1. INTRODUCTION

Data Mining is defined as extracting the data from large datasets. The main goal of data mining is extracting the information from data sets and transforms it to an understandable structure for future works. The actual data mining task is the automatic analysis of large data sets. It has been classified into cluster analysis (group of records), unusual records (anomaly detection), and dependencies (association rules mining).

Data mining (the analysis step of the "Knowledge Discovery in Databases" process, or KDD), an interdisciplinary subfield of computer science, is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems.

Aside from the raw analysis step, it involves database and data management aspects, data pre-processing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating.

Data mining uses information from past data to analyze the outcome of a particular problem or situation that may arise. Data mining works to analyze data stored in data warehouses that are used to store that data is being analyzed. That particular data may come from all parts of business, from the production to the management. Managers also use data mining to decide upon marketing strategies for their product. They can use data to compare and contrast among competitors.

2. OVERVIEW OF BREAST CANCER

Breast cancer has become the leading cause of death in women in developed countries. The most effective way to reduce breast cancer deaths is detect it earlier. Early diagnosis requires an accurate and reliable diagnosis procedure that allows physicians to distinguish benign breast tumors from malignant ones without going for surgical biopsy. The objective of these predictions is to assign patients to either a “benign” group that is noncancerous or a “malignant” group that is cancerous. The prognosis problem is the long-term outlook for the disease for patients whose cancer has been surgically removed. In this problem a patient is classified as a ‘recur’ if the disease is observed at some subsequent time to tumor excision and a patient for whom cancer has not recurred and may never recur.

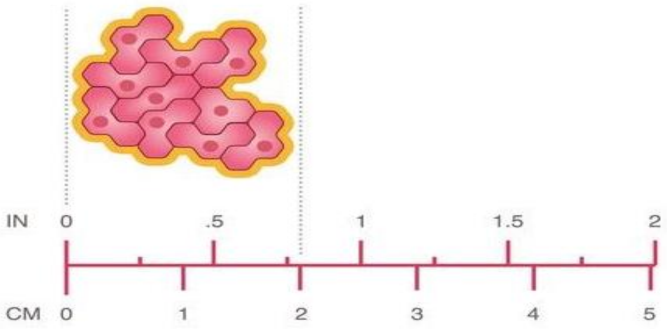

Fig 1 Breast Cancer Cell

Thus, breast cancer diagnostic and prognostic problems are mainly in the scope of the widely discussed classification problems. These problems have attracted many researchers in computational intelligence, data mining, and statistics fields.

Cancer research is generally clinical and biological in nature, data driven statistical research has become a common complement. Predicting the outcome of a disease is one of the most interesting and challenging tasks where to develop data mining applications. As the use of computers powered with automated tools, large volumes of medical data are being collected and made available to the medical research groups.

As a result, Knowledge Discovery in Databases (KDD), which includes data mining techniques, has become a popular research tool for medical researchers to identify and exploit patterns and relationships among large number of variables, and made them able to predict the outcome of a disease using the historical cases stored within datasets.

The objective of this study is to summaries various review and technical articles on diagnosis and prognosis of breast cancer. It gives an overview of the current research being carried out on various breast cancer datasets using the data mining techniques to enhance the breast cancer diagnosis and prognosis**.**

3. RELATED WORKS

Ross Quinlan, Classification, a data mining task is an effective method to classify the data in the process of Knowledge Data Discovery. A Classification method, Decision tree algorithms are widely used in medical field to classify the medical data for diagnosis. Feature

Selection increases the accuracy of the Classifier because it eliminates irrelevant attributes. This paper analyzes the performance of 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. The results show that a particular feature selection using CART has enhanced the classification accuracy of a particular dataset.

Chen, Y., Abraham, A., Yang, B ,Data mining is the process of analyzing large quantities of data and summarizing it into useful information. In medical diagnoses the role of data mining approaches increasing rapidly. Particularly Classification algorithms are very helpful in classifying the data, which is important in decision making process for medical practitioners. Further to enhance the classifier accuracy various preprocessing techniques and ensemble techniques were developed. A hybrid approach, CART classifier with feature selection and bagging technique has been considered to evaluate the performance in terms of accuracy and time for classification of various breast cancer datasets. For medical diagnosis various data mining techniques are available. For classification of medical data employed decision tree algorithm because it produce human readable classification rules which are easy to interpret

4. METHODOLOGY

4.1 Support Vector Machine

Support vector machine is very efficient and widely used. SVM classifier is mainly used for finding the decision values between the two datasets. It tests the various datasets. For a given kernel space input data and a linear model are projected and built. It produces several regression models and classifications. A regression model tries to find a continuous function.

The SVM model is a supervised machine learning technique, which is based on the statistical learning theory. It was firstly proposed by Cortes and Vapnik from his original work on structural risk minimization in and modified by Vapnik in. The algorithm of SVM is able to create a complex decision boundary between two classes with good classification ability. When the data are not linearly separable, the algorithm works by mapping the input space to higher dimensional feature space. SVM introduces the concept of ‘margin’ on either side of a hyper plane that separates the two classes. Maximizing the margins and thus creating the largest possible distance between the separating hyper plane and the samples on either side, is proven to reduce an upper bound on the expected generalization error. SVM may be considered a linear classifier in the feature space. On the other side it becomes a nonlinear classifier as a result of the nonlinear mapping from the input space to the feature one.

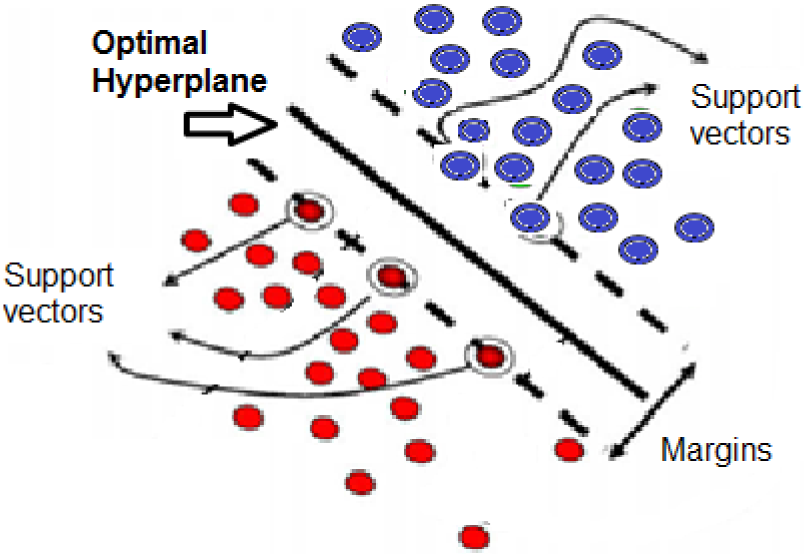

Fig 2 Optimal hyper plane separating the two classes and support vectors.

SVM Algorithm

Algorithm: Generate SVM

Input: Training Data, Testing Data

Output: Decision Value

Method:

```
    Step 1: Load Dataset
    Step 2: Classify Features (Attributes) based on class labels
    Step 3: Estimate Candidate Support Value
        While (instances! =null)
        Do
    Step 4: Support Value=Similarity between each instance in the attribute
```

```
        Find Total Error Value
Step 5: If any instance < 0
        Estimate
        Decision value = Support Value\Total Error
        Repeat for all points until it will empty
End If
```

4.2 Decision Tree

A decision tree is a flowchart-like structure in which internal node represents a "test" on an attribute (e.g. whether a coin flip comes up heads or tails), each branch represents the outcome of the test and each leaf node represents a class label (decision taken after computing all attributes). The paths from root to leaf represent classification rules.

In decision analysis a decision tree and the closely related influence diagram are used as a visual and analytical decision support tool, where the expected values (or expected utility) of competing alternatives are calculated.

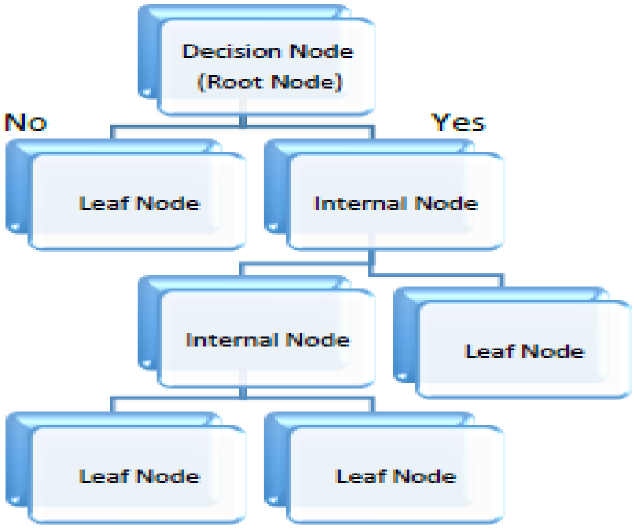

Fig 3 Decision Tree Model

Decision tree (DT) provides powerful techniques for classification and prediction. There are several algorithms to build DT model. As the name implies, this model recursively separates data samples into branches to construct a tree structure for the purpose of improving the prediction accuracy. Each tree node is either a leaf node or decision node.

Decision Tree Algorithm

Algorithm: Generate a decision tree from the training tuples of data partition D.

Input:

- Data partition D, which is a set of training tuples and their associated class labels;
- Attribute list, the set of candidate attributes;
- Attribute selection method, a procedure to determine the splitting criterion that "best" partitions the data tuples into individual classes. These criterions consist of a splitting attribute and, possibly, either a split point or splitting subset.

Output: A decision tree

Method:

```
(1) Create a node N;
(2) If tuples in D are all of the same class, C then
(3) Return N as a leaf node labeled with the class C
(4) If attribute list is empty then
(5) Return N as a leaf node labeled with the majority class in D
(6) Apply Attribute selection method (D, attribute list) to find the "best" splitting criterion
(7) Labelnode N with splitting criterion
(8) If splitting attribute is discrete-valued and multiway splits allowed then
(9) Attribute list $\leftarrow$ attribute list $-$ splitting attribute
(10) For each outcome j of splitting criterion
 (11) Let Dj be the set of data tuples in D satisfying outcome j
(12) If Dj is empty then
(13) Attach a leaf labeled with the majority class in D to node N
(14) Else attach the node returned by Generate decision tree (Dj, attribute list) to node N
        End for
(15) Return N
```

5. EXPERIMENTAL RESULTS

Dataset Description

Data Set	No. of Attributes	No. of Instances	No. of Classes
Wisconsin Breast Cancer (WBC)	11	699	2
Wisconsin Diagnosis Breast Cancer (WDBC)	11	699	2
Wisconsin Prognosis Breast Cancer (WPBC)	11	699	2

Table 1. Description of the Breast Cancer Data Sets

The Wisconsin Breast Cancer datasets from the Machine Learning Repository is used, to distinguish malignant (cancerous) from benign (non-cancerous) samples. A brief description of these datasets is presented in the table. Each dataset consists of some classification patterns or instances with a set of numerical features or attributes.

Experimental Evaluation for SVM and Decision Tree

Classifiers	SVM	Decision Tree
Accuracy	92.27%	94.54%

Table 2. For WBC

The comparison of accuracies for the two classifiers (SVM, Decision Tree) based on 10-fold cross validation as a test method. The accuracy of SVM(92.27%) is the best classifier and the accuracy obtained by decision tree is better than that produced by SVM.

Classifiers	SVM	Decision Tree
Accuracy	84.34%	85.1%

Table 3. For WDBC

The comparison of accuracies for the two classifiers (SVM, Decision Tree) for WDBC based on 10-fold cross validation as a test method. The accuracy of SVM(84.34%) is the best classifier and the accuracy obtained by decision tree is better than that produced by SVM.

Classifiers	SVM	Decision Tree
Accuracy	79%	82%

Table 4. For WPBC

The comparison of accuracies for the two classifiers (SVM, Decision Tree) for WPBC based on 10-fold cross validation as a test method. The accuracy of SVM(79%) is the best classifier and the accuracy obtained by decision tree is better than that produced by SVM.

From the Experimental Results, we identified that Decision tree is better than the SVM classification. Decision tree gives more accuracy when compared with the Existing System.

6. CONCLUSION AND FUTURE WORK

The classification of breast cancer is a medical application that poses a great challenge for researchers and scientists. The use of learning machine and artificial intelligence techniques has revolutionized the process of diagnosis and prognosis of the breast cancer. The aim of our study is to propose an approach for breast cancer distinguishing between different classes of breast cancer. This approach is based on the Wisconsin Diagnostic and Prognostic Breast Cancer datasets for feature selection, and the classification of different types of breast cancer . In this research, proposed the comparative study of classifiers by using the breast cancers. In this, evaluated the performance of SVM and Decision Tree. We predict the efficiency, accuracy and resource utilizations of classifiers and evaluate and compare the performance levels of classifiers. It evaluates the performance of different classifiers on breast cancer (Wisconsin Breast Cancer (WBC), Wisconsin Diagnosis Breast Cancer (WDBC) and Wisconsin Prognosis Breast Cancer (WPBC)). In future work Linear discriminant analysis (LDA) and the related Fisher's linear discriminant are methods used in statistics, pattern recognition and machine learning to find a linear combination of features which characterizes or separates two or more classes of objects or events. The resulting combination may be used as a linear classifier, or more commonly, for dimensionality reduction before later classification. Consider a set of observations for each sample of an object or event with known class *y*.

This set of samples is called the training set. The classification problem is then to find a good predictor for the class *y* of any sample of the same distribution given only an observation

REFERENCES

[1] U.S. Cancer Statistics Working Group. United States Cancer Statistics: 1999–2008 Incidence and Mortality Web-based Report. Atlanta (GA): Department of Health and Human Services, Centers for Disease Control

[2] S. Aruna, Dr S.P. Rajagopalan and L.V. Nandakishore (2011). Knowledge based analysis of various statistical tools in detecting breast cancer.

[3] AngelineChristobel. Y, Dr. Sivaprakasam (2011). An Empirical Comparison of Data Mining Classification Methods. International Journal of Computer information Systems,Vol. 3, No. 2, 2011.

[4] D.Lavanya, Dr.K.Usha Rani,..," Analysis of feature selection with classification: Breast cancer datasets",Indian Journal of Computer Science and Engineering (IJCSE),October 2011.

[5] E.Osuna, R.Freund, and F. Girosi, "Training support vector machines: Application to face detection". Proceedings of computer vision and pattern recognition, Puerto Rico pp. 130–136.1997.

[6] Vaibhav Narayan Chunekar, Hemant P. Ambulgekar (2009). Approach of Neural Network to Diagnose Breast Cancer on three different Data Set. 2009 International Conference on Advances in Recent Technologies in Communication and Computing.

[7] D. Lavanya, "Ensemble Decision Tree Classifier for Breast Cancer Data," international Journal of Information Technology Convergence and Services, vol. 2, no. 1, pp. 17-24, Feb. 2012.