Mining Big Data: Breast Cancer Prediction using DT -SVM Hybrid Model

Eman Elkady

Mining Big Data: Breast Cancer Prediction using DT - SVM Hybrid Model

K.Sivakami, *Assistant Professor, Department of Computer Application*

Nadar Saraswathi College of Arts & Science, Theni.

Abstract - Breast Cancer is becoming a leading cause of death among women in the whole world; meanwhile, it is confirmed that the early detection and accurate diagnosis of this disease can ensure a long survival of the patients. This paper work presents a disease status prediction employing a hybrid methodology to forecast the changes and its consequence that is crucial for lethal infections. To alarm the severity of the diseases, our strategy consists of two main parts: 1. Information Treatment and Option Extraction, and 2. Decision Tree-Support Vector Machine (DT-SVM) Hybrid Model for predictions. We analyse the breast Cancer data available from the Wisconsin dataset from UCI machine learning with the aim of developing accurate prediction models for breast cancer using data mining techniques. In this experiment, we compare three classifications techniques in Weka software and comparison results show that DT-SVM has higher prediction accuracy than Instance-based learning (IBL), Sequential Minimal Optimization (SMO) and Naïve based classifiers.

Index Terms - **b**reast cancer; classification; Decision Tree- Support Vector Machine, Naïve Bayes, Instance-based learning, Sequential Minimal Optimization, and weka;

1 INTRODUCTION

Data mining, also known as knowledge discovery in databases is defined as “the extraction of implicit, previously unknown, and potentially useful information from data”. It encompasses a set of processes performed automatically, whose task is to discover and extract hidden features (such as: various patterns, regularities and anomalies) from large datasets.

Big data is a broad term for datasets so large or complex that traditional data processing applications are inadequate. Analysis of data sets can find new correlations, to "spot business trends, prevent diseases, and combat crime and so on”. Scientists, practitioners of media and advertising and governments alike regularly meet difficulties with large data sets in areas including Internet search, finance and business informatics [1].

Building accurate and efficient classifiers for large databases is one of the essential tasks of data mining and machine learning research [10]. Usually, classification is a preliminary data analysis step for examining a set of cases to see if they can be grouped based on similarity" to each other. The ultimate reason for doing classification is to increase understanding of the domain or to improve predictions compared to unclassified data. Many different types of classification techniques have been proposed in literature that includes DT-SVM, SMO, IBL, etc.

Clustering is a Data mining technique which segments a heterogeneous data into a number of homogeneous subgroups or clusters. In clustering, the records are grouped together based on self-similarity without any predefined classes or examples. The quality of a cluster is measured by cluster diameter which is the distance between any two objects in a cluster. In this study k-means clustering is chosen because of its popularity and it's proved effectiveness.

In data mining breast cancer research has been one of the important research topics in medical science during the recent years The classification of Breast Cancer data can be useful to predict the result of some diseases or discover the genetic behavior of tumors. There are many techniques to predict and classification breast cancer pattern. This work empirically compares performance of different classification rules that are suitable for direct interpretability of their results.

Breast cancer is one of the most common cancers among women. Breast cancer is one of the major causes of death in women when compared to all other cancers. Cancer is a type of diseases that causes the cells of the body to change its characteristics and cause abnormal growth of cells. Most types of cancer cells eventually become a mass called tumor. The occurrence of breast cancer is increasing globally. It is a major health problem and represents a significant worry for many women [1].

Early detection of breast cancer is essential in reducing life losses. However earlier treatment requires the ability to detect breast cancer in early stages. Early diagnosis requires an accurate and reliable diagnosis procedure that allows physicians to distinguish benign breast tumors from malignant ones. The automatic diagnosis of breast cancer is an important, real-world medical problem. Thus, finding an accurate and effective diagnosis method is very important. In recent years machine learning methods have been widely used in prediction, especially in medical diagnosis [2]. Medical diagnosis is one of major problem in medical application. The classification of Breast Cancer data can be useful to predict the outcome of some diseases or discover the genetic behavior of tumors [3]. A major class of problems in medical science involves the diagnosis of disease, based upon various tests performed upon the patient. For this reason the use of classifier systems in medical diagnosis is gradually increasing.

The aim of this dissertation is to develop the various phase. They are:

- Predictive Framework for Breast Cancer Disease
- Hybrid model
- Generation of rule to predict Breast Cancer
- Qualitative measures used for Comparison (accuracy, error rate, sensitivity, and specificity).
- GUI Development

1.2 OVERVIEW

Breast cancer is becoming a leading cause of death among women in the whole world, meanwhile, it is confirmed that the early detection and accurate diagnosis of this disease can ensure a long survival of the patients. In this research work, a decision intelligence technique based support vector machine classifier (DT-SVM) is proposed for breast cancer diagnosis. In the proposed DT-SVM, the issue of model selection and feature selection in SVM is simultaneously solved using Net beans and WEKA analytical tool. A weighted function is adopted to design the objective function of DT-SVM, which takes into account the average accuracy rates of SVM, the number of support vectors and the selected features simultaneously.

2 RELATED WORKS

Delen et al in their work preprocessed the SEER data for to remove redundancies and missing information. They have compared the predictive accuracy of the SEER data on three prediction models indicated that the decision tree (C5) is the best predictor with 93.6% accuracy on the holdout sample.

Endo et al. 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

Kotsiantis et al. did a work on Bagging, Boosting and Combination of Bagging and Boosting as a single ensemble using different base learners such as C4.5, Naïve Bayes, OneR and Decision Stump. These were experimented on several benchmark datasets of UCI Machine Learning Repository.

Bittern et al. used artificial neural network to predict the survivability for breast cancer patients. They tested their approach on a limited data set, but their results show a good agreement with actual survival.

Vikas Chaurasia et al. used RepTree, RBF Network and Simple Logistic to predict the survivability for breast cancer patients.

Djebbari et al. consider the effect of ensemble of machine learning techniques to predict the survival time in breast cancer. Their technique shows better accuracy on their breast cancer data set comparing to previous results. Liu Ya-Qin‘s [5] experimented on breast cancer data using C5 algorithm with bagging to predict breast cancer survivability.

Bellaachi et al. used naive bayes, decision tree and back-propagation neural network to predict the survivability in breast cancer patients. Although they reached good results (about 90% accuracy), their results were not significant due to the fact that they divided the data set to two groups; one for the patients who survived more than 5 years and the other for those patients who died before 5 years.

Tsirogiannis‘s et al. applied bagging algorithm on medical databases using the classifiers neural networks, SVM‘S and decision trees. Results exhibits improved accuracy of bagging than without bagging.

3 RELATED METHODS

The overall design procedure is shown is shown in Figure 3.1.

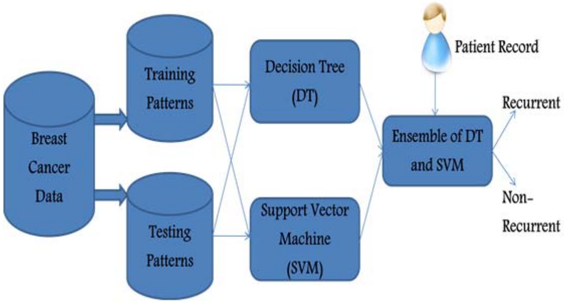
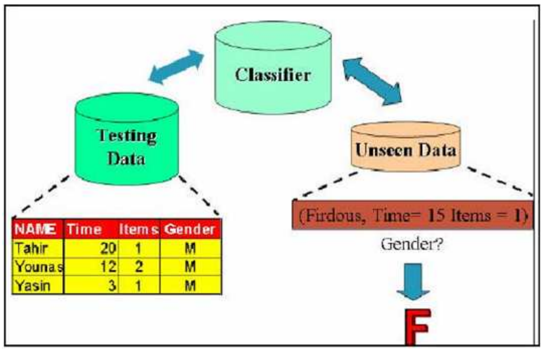


Figure.3.1: Ensemble Model for Breast cancer identification

The design is the structure of any scientific work. It gives direction and systematizes the research. Most scientists are interested in getting reliable observations that can help the understanding of a phenomenon. DT – SVM method has been adopted in this research was so carefully designed.

3.1 EXPERIMENTAL DESIGN

Classification predicts categorical class labels (discrete or nominal) and also classifies data (constructs a model) based on the training set and the values (class labels) in a classifying attribute and uses it in classifying new data. The figure 3.2 shoes classification of new data.



NAME	Time	Items	Gender
Tahir	20	1	M
Younas	12	2	M
Yasin	3	1	M

Figure 3.2: Data classification

3.1.1 SUPPORT VECTOR MACHINE

Classification in SVM is an example of Supervised Learning. Known labels help indicate whether the system is performing in a right way or not. This information points to a desired response, validating the accuracy of the system, or be used to help the system learn to act correctly. A step in SVM classification involves identification as which are intimately connected to the known classes is called feature selection. Feature selection and SVM classification together have been used even, when prediction of unknown samples is not necessary. They can be used to identify key sets which are involved in whatever processes distinguish the classes. The flow diagram for the SVM method is shown is given in figure 3.3.

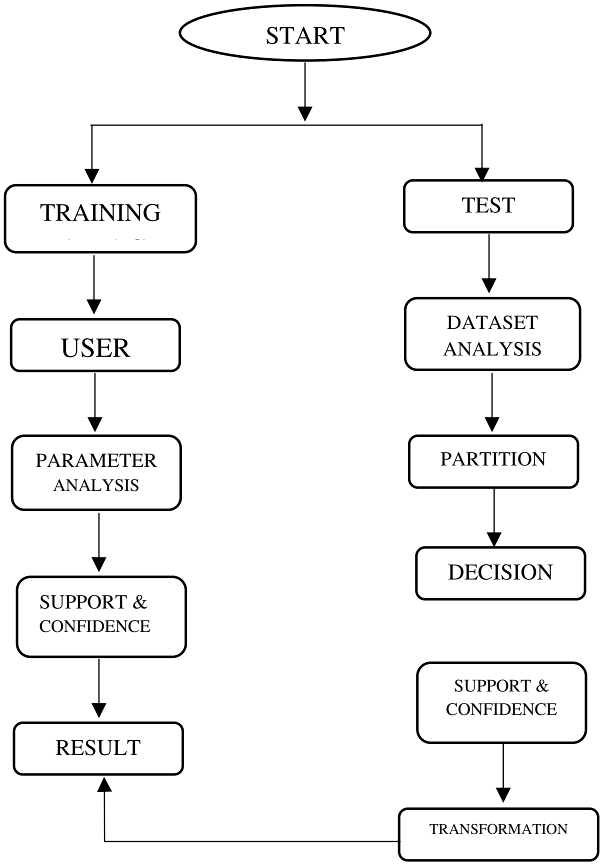


Figure 3.3: SVM Data Flow

3.1.2 Decision Tree

DT is a most popular and powerful classification technique where each internal node (non-leaf node) denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (or terminal node) holds a class label. DT with four decision nodes and five leaf nodes are shown in figure 3.5. The top most nodes in a tree are the root nodes. DT is so popular because construction of DT classifiers does not require any domain knowledge or parameter setting and, therefore, is appropriate for exploratory knowledge discovery. Various decision tree based techniques are widely accepted and applied on health care diagnosis process. On the other hand some statistical technique like

SVM is also used as classifier for health care diagnosis. Interactive Dichotomiser 3 (ID3) and C4.5 are the two very popular DT algorithms proposed by Quinlan [20]. ID3 uses Entropy and Information Gain to construct a decision tree.

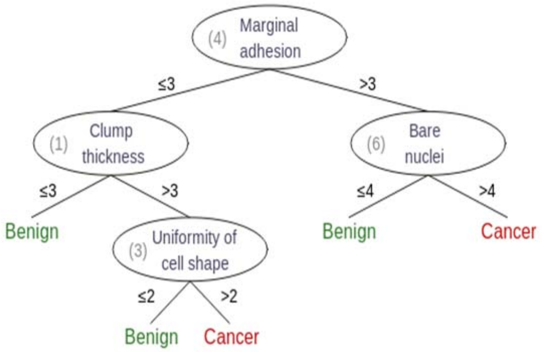


Figure 3.4: Decision Tree with 4 decision nodes and 5 leaf nodes

Entropy: A decision tree is built top-down from a root node and involves partitioning the data into subsets that contain instances with similar values (homogenous). ID3 algorithm uses entropy to calculate the homogeneity of a sample. If the sample is completely homogeneous the entropy is zero and if the sample is an equally divided it has entropy of one.

To build a decision tree, we need to calculate two types of entropy using frequency tables as follows:

a) Entropy using the frequency table of one attribute:

$$E(S) = \sum_{i=1}^{c} - P_i \, log_2 \, P_i$$

b) Entropy using the frequency table of two attributes:

$$E(T, X) = \sum_{c \in X} P(c)E(c)$$

Information Gain: The information gain is based on the decrease in entropy after a dataset is split on an attribute. Constructing a decision tree is all about finding attribute that returns the highest information gain (i.e., the most homogeneous branches).

3.1.3 INSTANCE BASED LEARNING

IBL is a basic instance-based learner which finds the training instance closest in Euclidean distance to the given test instance and predicts the same class as this training distance. If several instances qualify as the closest, the first one found is used. IBL algorithms do not construct extensional CDs. These functions are two of the three components in the following framework that describes all IBL algorithms:

- Similarity Function: This calculates the similarity between training instances i and the instances in the concept depiction. Similarities are numeric-valued.
- Classification Function: This obtains the similarity functions results and the classification performance records of the instances in the concept description. It returns a classification for i.
- CD Updater: This retains records on classification performance and decides which instances to include in the concept description. Inputs include i, the classification results, the similarity results, and a current concept description. It returns the modified concept description.

3.1.4 SEQUENTIAL MINIMAL OPTIMIZATION

SMO is an algorithm for solving the quadratic programming (QP) problem that arises during the training of SVM. SMO is widely used for training support vector machines and is implemented by the popular LIBSVM tool.

SMO is an iterative algorithm for solving the optimization problem described above. SMO breaks this problem into a series of smallest possible sub-problems, which are then solved analytically.

3.2 DATA SOURCES

In this study, we have performed our conduction on the Wisconsin Breast Cancer Dataset (WBCD) taken from UCI machine learning repository (UCI Repository of Machine Learning Databases). The dataset contains 699 instances taken from needle aspirates from patients' breasts, of which 458 cases belong to benign class and the remaining 241 cases belong to malignant class. It should be noted that there are16 instances which have missing values, in this study all the missing values are replaced by the mean of the attributes. Each record in the database has nine attributes. These nine attributes were found to differ significantly between benign and malignant samples.

3.3 DATA ANALYSIS

1. Number of Instances: 699

2. Number of Attributes: 10 + class attribute

S.NO	ATTRIBUTE	DOMAIN
1.	Sample code number	id number
2.	Clump Thickness	1-10
3.	Uniformity of Cell Size	1-10
4.	Uniformity of Cell Shape	1-10
5.	Marginal Adhesion	1-10
6.	Single Epithelial Cell Size	1-10
7.	Bare Nuclei	1-10
8.	Bland Chromatin	1-10
9.	Normal Nucleoli	1-10
10.	Mitoses	1-10
11.	Class	2 - Benign, 4 for malignant

Table 3.1 Attribute Information: (class attribute has been moved to last column)

3. Missing attribute values: 16

- There are 16 instances in Groups 1 to 6 that contain a single missing (i.e., unavailable) attribute value, now denoted by "?".

4. Class distribution:

- Benign: 458 (65.5%)
- Malignant: 241 (34.5%)

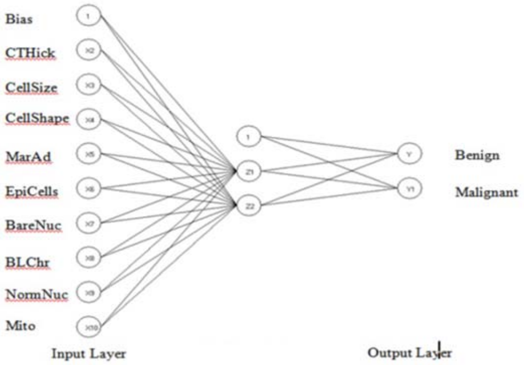


Figure 3.5: Diagrammatic Representation of Breast Cancer Attributes

4 RESULTS & DISCUSSION

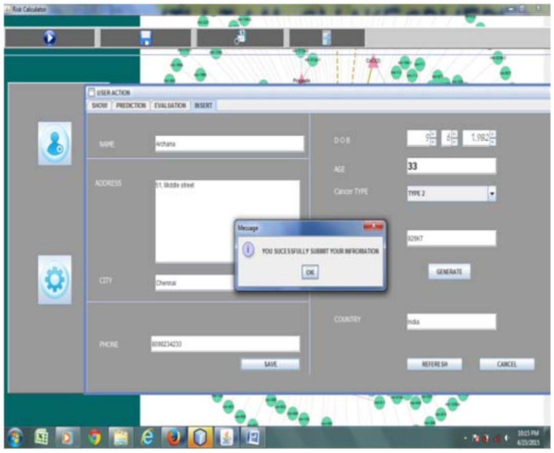
The proposed DT-SVM diagnostic system is implemented using Weka tool with Java platform. For SVM, LIBSVM implementation was utilized. The empirical experiment was conducted on Intel Quad-Core Xeon 5130 CPU (2.0 GHz) with 4GB of RAM.

In order to guarantee the valid results, the k-fold Crossover Validation (CV) was used to evaluate the classification accuracy [45].

Data was divided into ten subsets. Each time, one of the ten subsets is used as the test set and the other nine subsets are put together to form a training set. Then the average error across all ten trials is computed. The advantage of DT-SVM method is that all of the test sets are independent and the reliability of the results could be improved. We attempted to design our experiment using two loops. The inner loop is used to determine the optimal parameters and best feature subset. The outer loop is used for estimating the performance of the SVM classifier.

We selected Breast cancer data set and selected different class of patient groups from the collection for our experiments. The group data is pre-processed and made as simple text files thereby removing all the headers, replies and other non- related information. The preprocessed data from these selected classes are divided as 60% training data and 40% testing data. But we selected only 10 parameters for training and around 400 patients' documents for testing. The feature selection was simple and did not involve any special techniques, to test the classifier with such features. In our implementation of the classifier we used below interface to train the SVM. The classes which we selected for our experiments are different age groups, breast size and menopause. The classifier was trained with these classes and then tested with the test data.

The figure 4.1 is act as an interface to feed individual patients data to predict the response which is diagnosed against the classified ones using DT-SVM. Show method used to view all available data set.

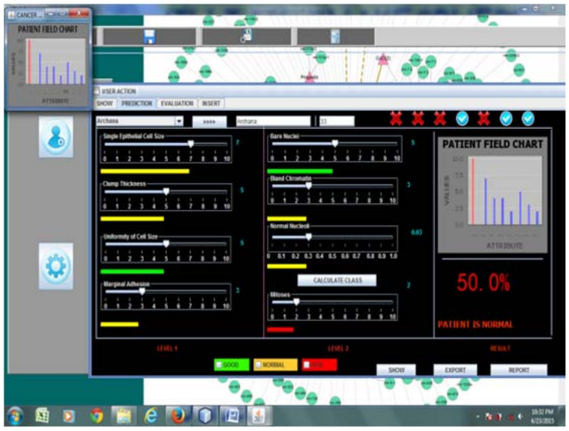


NAME Archana
ADDRESS 51, Middle street
CITY Chennai
PHONE 9090234233
D O B 9 6 1,982
AGE 33
Cancer TYPE TYPE 2
829KT
COUNTRY India

YOU SUCESSFULLY SUBMIT YOUR INFROMATION

Figure 4.1 Interface to feed individual patient data

To forecast the severity of breast cancer with help of slider by feeding the field values are given in figure 4.2. The attribute values are also displayed in chart.



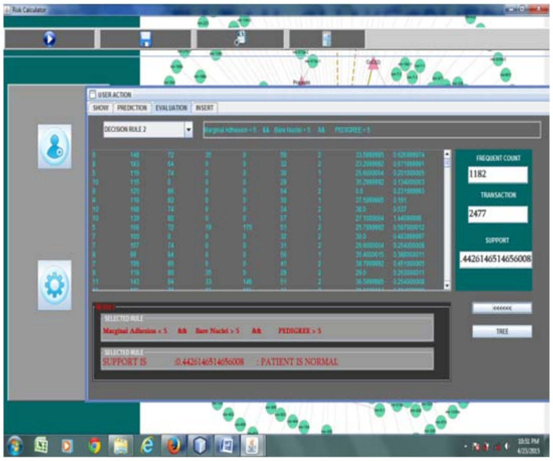
Archana
Archana
33
Single Epithelial Cell Size 7
Clamp Thickness 5
Uniformity of Cell Size 5
Marginal Adhesion 3
Bare Nuclei 5
Bland Chromatin 3
Normal Nucleoli 0.03
Mitoses 2

50. 0%

PATIENT IS NORMAL

Figure 4.2: Predictive framework to forecast the severity of cancer.

The figure 4.3 is used to evaluate the dataset based on the class distribution benign and malignant with predefined decision rule by calculating the support & confidence values.



Marginal Adhesion < 5 && Bare Nuclei > 5 && PEDIGREE > 5

FREQUENT COUNT: 1182

TRANSACTION: 2477

SUPPORT: .4426146514656008

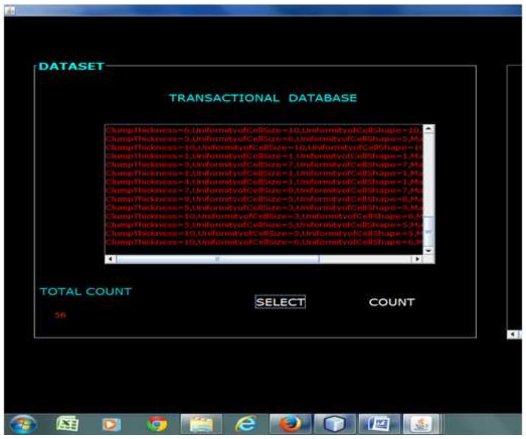
SELECTED RULE: Marginal Adhesion < 5 && Bare Nuclei > 5 && PEDIGREE > 5

SELECTED RULE: SUPPORT IS :0.4426146514656008 : PATIENT IS NORMAL

Figure 4.3: Rule to evaluate dataset for class malignant

To find the rule to predict breast cancer using DT, the transaction database will be imported are given in figure 4.4. The number of transactions will be calculated with respect to the attribute values. It also shows the Class distribution: Benign, Malignant.

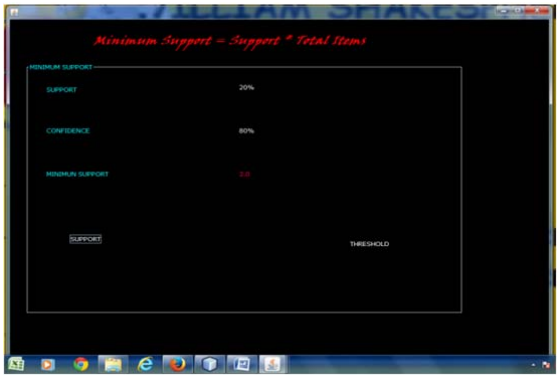
In order to find the rule based on transactional database. We must truncate the discarded data by calculating the minimum support threshold through interesting measures such as support and confidence values are shown in figure 4.5.



ClumpThickness=6,UniformityofCellSize=10,UniformityofCellShape=10,
ClumpThickness=5,UniformityofCellSize=6,UniformityofCellShape=5,M
ClumpThickness=10,UniformityofCellSize=10,UniformityofCellShape=10
ClumpThickness=1,UniformityofCellSize=1,UniformityofCellShape=1,M
ClumpThickness=3,UniformityofCellSize=7,UniformityofCellShape=7,M
ClumpThickness=1,UniformityofCellSize=1,UniformityofCellShape=1,M
ClumpThickness=4,UniformityofCellSize=1,UniformityofCellShape=1,M
ClumpThickness=7,UniformityofCellSize=8,UniformityofCellShape=7,M
ClumpThickness=9,UniformityofCellSize=5,UniformityofCellShape=8,M
ClumpThickness=5,UniformityofCellSize=3,UniformityofCellShape=3,M
ClumpThickness=10,UniformityofCellSize=3,UniformityofCellShape=6,M
ClumpThickness=5,UniformityofCellSize=5,UniformityofCellShape=5,M
ClumpThickness=10,UniformityofCellSize=5,UniformityofCellShape=5,M
ClumpThickness=10,UniformityofCellSize=6,UniformityofCellShape=6,M

TOTAL COUNT: 56

hFigure 4.4: Imported transaction database to find the rule



Minimum Support = Support * Total Items

SUPPORT 20%

CONFIDENCE 80%

MINIMUN SUPPORT 2.0

Z

Figure 4.5: Calculation of Minimum Support Threshold values

The truncated database will be imported to a DT (ID3) algorithm to find the predictive rule for breast cancer. In figure 4.6 present the complete DT with 25 decision nodes with three leaf nodes with class of malignant.

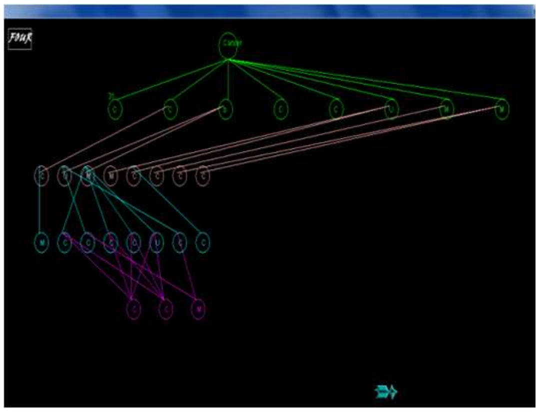


Figure 4.6: Generation of Final Decision Tree

We have designed a Prototype of Breast cancer disease prediction model. This model will predict the breast cancer's disease class based on the rules created by ID3 algorithms.Figure.4.1 shows the interface for input, which takes Medical profiles of a patient such as age, sex, blood pressure and blood sugar etc as input and it can predict about presence or absence of Breast cancer disease. The path from root to leaf node shows if the patient has these combinations of cell values is high, and then the patient is affected by cancer. This is the rule we generate through DT.

It has been observed that the classes which are predicted at the top level of the tree have good accuracy and the ones predicted at the lower levels have poor accuracy. In our experiment which was classified at the root node has 95% accuracy.

DESCRIPTION	ACCURACY	ERROR RATE	CORRECTLY CLASSIFIED INSTANCE	INCORRECTLY CLASSIFIED INSTANCE
DT+SVM	91%	2.58	459	240
IBL	85.23%	12.63	184	515
SMO	72.56%	5.96	325	374
NAÏVE	89.48%	9.89	291	408

The performance of a chosen (IBL, SMO and Naïve based) classifiers are performed through weka tool and it's validated based on error rate and accuracy. The classification accuracy is predicted in terms of Sensitivity and Specificity. The evaluation parameters are the specificity, sensitivity, and overall accuracy. Hence DT+SVM perform well in classifying the breast cancer data, compared to all other algorithms.

From the above figures and table we find that highest accuracy of Classification model is DT - SVM (91%), low error rate (2.58%), correctly classified instance (459) and incorrectly classified instance (240) in breast cancer data as shown in Figure 4.11, 4.12, 4.13 and 4.14.

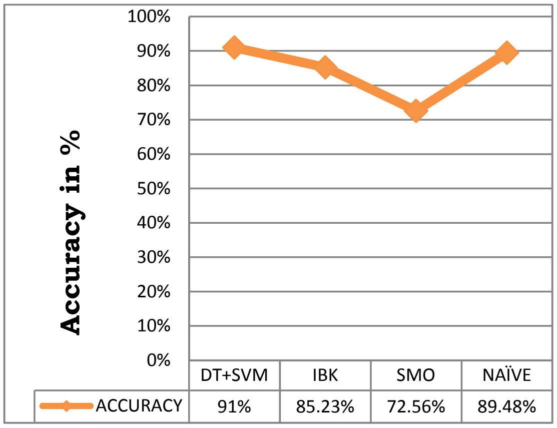


Figure 4.11: Accuracy of Classification Methods

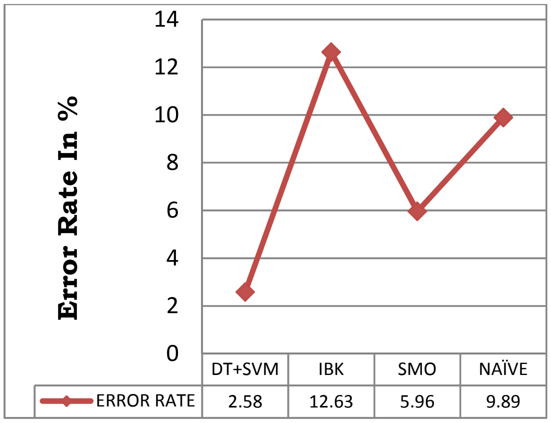


Figure 4.12: Error Rate of Classification Methods

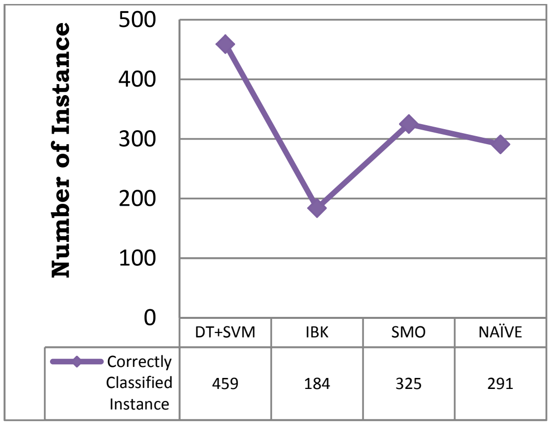


Figure 4.13: Correctly Classified Instance of Classification Methods

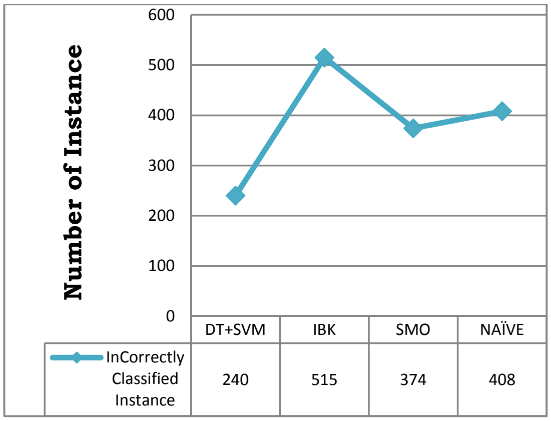


Figure 4.13: InCorrectly Classified Instance of Classification Methods

5 CONCLUSION

This work proposed a hybrid classification algorithm for breast cancer patients which integrates DT and SVM algorithms. The proposed algorithm was composed of two main phases. The first phase is Information treatment and option extraction followed by DT-SVM hybrid model predictions. Classification had two main phases, Training and Testing phases. The input parameters for SVM were optimized using DT algorithm. The SVM algorithm was used to classify breast cancer patients into one of two classes (Benign/Malignant).

In comparison of data mining techniques, this research used accuracy indicator to evaluate classification efficiency of different algorithms. The proposed algorithm was compared with different classifier algorithms using Weka tool. Specifically, we used popular data mining methods: DT-SVM, SMO, IBL, and Naïve. An important challenge in data mining and machine learning areas is to build precise and computationally efficient classifiers for Medical applications.

The experimental results given in the Figure 4.10 and Table 4.1 showed the effectiveness of the proposed algorithm. Overall, DT-SVM classification accuracy is better than other classifier algorithm. However, from a relatively low error rate, the results show that the DT-SVM will be the best prognosis in clinical practice.

The optimum breast cancer disease predictive model obtained in this study adopts DT-SVM classification algorithm, this research may provide references for future research on selecting the optimal predictive models to lower the incidence of breast cancer.

REFERENCES

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

[2] G. Ravi Kumar, Dr. G. A. Ramachandra, K.Nagamani, " An Efficient Prediction of Breast Cancer Data using Data Mining Techniques", International Journal of Innovations in Engineering and Technology (IJIET), Vol. 2 Issue 4 August 2013.

[3] Miss Jahanvi Joshi, Mr. RinalDoshi , Dr. Jigar Patel, "Diagnosis and Prognosis Breast Cancer Using Classification Rules", International Journal of Engineering Research and General Science Volume 2, Issue 6, October-November, 2014,

[4] A.Bellachia and E.Guvan,"Predicting breast cancer survivability using data mining techniques", Scientific Data Mining Workshop, in conjunction with the 2006 SIAM Conference on Data Mining, 2006.

[5] 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.

[6] Breast Cancer Wisconsin Data [online]. Available: http://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancerwisconsin/breast-cancer-wisconsin.data.

[7] Brenner, H., Long-term survival rates of cancer patients achieved by the end of the 20th century: a period analysis. Lancet. 360:1131–1135, 2002.

[8] D. Delen, G. Walker and A. Kadam (2005), Predicting breast cancer survivability: a comparison of three data mining methods, Artificial Intelligence in Medicine, vol.34, pp.113-127.

[9] Ian H. Witten and Eibe Frank. Data Mining: Practical machine learning tools and techniques, 2nd Edition. San Fransisco:Morgan Kaufmann; 2005.

[10] J. Han and M. Kamber, Data Mining—Concepts and Technique (The Morgan Kaufmann Series in Data Management Systems), 2nd ed. San Mateo, CA: Morgan Kaufmann, 2006.

[11] J. R. Quinlan, C4.5: Programs for Machine Learning. San Mateo, CA:Morgan Kaufmann; 1993.

[12] Mitchell, T. M., Machine Learning, McGraw-Hill Science/Engineering/Math, 1997

[13] P.-N. Tan, M. Steinbach, and V. Kumar, Introduction to Data Mining. Reading, MA: Addison-Wesley, 2005.

[14] Razavi, A. R., Gill, H., Ahlfeldt, H., and Shahsavar, N., Predicting metastasis in breast cancer: comparing a decision tree with domain experts. J. Med. Syst. 31:263–273, 2007.

[15] S.B.Kotsiantis and P.E.Pintelas,"Combining Bagging and Boosting", International Journal of Information and Mathematical Sciences, 1:4 2005.

[16] Vapnik, V. N., The nature of statistical learning theory. Springer, Berlin, 1995.

[17] Weka: Data Mining Software in Java, http://www.cs.waikato.ac.nz/ml/weka/

[18] Witten H.I., Frank E., Data Mining: Practical Machine Learning Tools and Techniques, Second edition, Morgan Kaufmann Publishers, 2005.

[19] Y Rejani- "Early detection of breast cancer using SVM". 2009 –arxiv

[20] Ilias Maglogiannis, E Zafiropoulos "An intelligent system for automated breast cancer diagnosis and prognosis using SVM based classifiers" Applied Intelligence, 2009 – Springer.

[21] Zhang Qinli; Wang Shitong; Guo Qi; "A Novel SVM and Its Application to Breast Cancer Diagnosis" http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=4272649.

[22] Quinlan.J.R. (1993).C4.5:Programs for machine learning (1st edition), San Francisco, Morgan Kaufmann Publishers,1993.

[23] Pendharkar PC, Rodger JA, Yaverbaum GJ, Herman N, Benner M (1999) Association, statistical, mathematical and neural approaches for mining breast cancer patterns. Expert Systems with Applications 17: 223-232.

[24] Zhou ZH, Jiang Y (2003) Medical diagnosis with C4.5 Rule preceded by artificial neural network ensemble. IEEE Trans Inf Technol Biomed 7: 37-42.

[25] Lundin M, Lundin J, Burke HB, Toikkanen S, Pylkkanen L, et al. (1999) Artificial neural networks applied to survival prediction in breast cancer. Oncology 57: 281-286.

[26] Shweta Kharya, "Using data mining techniques for diagnosis and prognosis of Cancer Disease", International Journal of Computer Science, Engineering and Information

Technology (IJCSEIT), Vol.2, No.2, April 2012, pp. 55-66.

[27] P.Ramachandran, N.Girija and T.Bhuvaneswari, “Health care Service Sector: Classifying and finding Cancer spread pattern in Southern india using data mining techniques”, International Journal on Computer Science and Engineering (IJCSE), Vol. 4 No. 05 May 2012, pp. 682-687.

[28] S.B.Kotsiantis and P.E.Pintelas,”Combining Bagging and Boosting”, International Journal of Information and Mathematical Sciences, 1:4 2005.

[29] K. P. Bennett and O. L. Mangasarian: "Robust Linear Programming Discrimination of Two Linearly Inseparable Sets", Optimization Methods and Software 1, 1992, 23-34.