An Efficient Prediction of Breast Cancer Data using Data Mining Techniques

G. Ravi Kumar
Research Scholar
Department of Computer Science & Technology
S K University, Anantapur, AP, India

Dr. G. A. Ramachandra
Associate Professor
Department of Computer Science & Technology
S K University, Anantapur, AP, India

K.Nagamani
Lecturer
Department of Computer Science
Rayalaseema University, Kurnool, AP, India

Abstract - Breast cancer is one of the major causes of death in women when compared to all other cancers. Breast cancer has become the most hazardous types of cancer among women in the world. Early detection of breast cancer is essential in reducing life losses. This paper presents a comparison among the different Data mining classifiers on the database of breast cancer Wisconsin Breast Cancer (WBC), by using classification accuracy. This paper aims to establish an accurate classification model for Breast cancer prediction, in order to make full use of the invaluable information in clinical data, especially which is usually ignored by most of the existing methods when they aim for high prediction accuracies. We have done experiments on WBC data. The dataset is divided into training set with 499 and test set with 200 patients. In this experiment, we compare six classification techniques in Weka software and comparison results show that Support Vector Machine (SVM) has higher prediction accuracy than those methods. Different methods for breast cancer detection are explored and their accuracies are compared. With these results, we infer that the SVM are more suitable in handling the classification problem of breast cancer prediction, and we recommend the use of these approaches in similar classification problems.

Keywords—breast cancer; classification; Decision tree, Naïve Bayes, MLP, Logistic Regression SVM, KNN and weka;

I. 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" [8] [16]. 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. Classification is one of the most studied problems in machine learning and data mining [8]. Predicting the outcome of a disease is one of the most interesting and challenging tasks in which to develop data mining applications.

The goal of the classification algorithms is to construct a model from a set of training data whose target class labels are known and then this model is used to classify unseen instances. The classification of Breast Cancer data can be useful to predict the outcome of some diseases or discover the genetic behavior of tumors.
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. 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. 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 structure of this paper is as follows: Section 2, we present related works. In section 3, classification techniques are discussed in details. In section 4 we present the experimental results and evaluation of the classification techniques and final results. The conclusion would be given in section 5.

II. RELATED WORKS

Bellachia et al [2] uses the SEER data to compare three prediction models for detecting breast cancer. They have reported that C4.5 algorithm gave the best performance of 86.7% accuracy.
Delen et al [6] 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 [3] 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. [13] 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.

III. CLASSIFICATION TECHNIQUES

Building accurate and efficient classifiers for large databases is one of the essential tasks of data mining and machine learning research. Building effective classification systems is one of the central tasks of data mining. Many different types of classification techniques have been proposed in literature that includes Decision Trees, Naive-Bayesian methods, Neural Networks, Logistic Regression , SVM and KNN etc.

1. DECISION TREE (J48)

Decision tree models are commonly used in data mining to examine data and induce the tree and its rules that will be used to make predictions[10].The prediction could be to predict categorical values (classification trees) when instances are to be placed in categories or classes.
Decision tree is a classifier in the form of a tree structure where each node is either a leaf node, indicating the value of the target attribute or class of the examples, or a decision node, specifying some test to be carried out on a single attribute-value, with one branch and sub-tree for each possible outcome of the test. A decision tree can be used to classify an example by starting at the root of the tree and moving through it until a leaf node is reached, which provides the classification of the instance.

2. NEURAL NETWORKS

Neural networks are capable of modeling extremely complex, typically non-linear functions[10]. It is made up of a structure or a network of numerous interconnected units (artificial neurons). Each of these units consists of input/output characteristics that implement a local computation or function. The function could be a computation of weighted sums of inputs which produces an output if it exceeds a given threshold. The output (whatever the result), could serve as an input to other neurons in the network. This process iterates until a final output is produced.

3. NAIVE BAYES (NB)

The Naive Bayes is a quick method for creation of statistical predictive models [16]. NB is based on the Bayesian theorem. This classification technique analyses the relationship between each attribute and the class for each instance to derive a conditional probability for the relationships between the attribute values and the class. During

training, the probability of each class is computed by counting how many times it occurs in the training dataset. This is called the "prior probability" P(C=c). In addition to the prior probability, the algorithm also computes the probability for the instance x given c with the assumption that the attributes are independent. This probability becomes the product of the probabilities of each single attribute. The probabilities can then be estimated from the frequencies of the instances in the training set.

4. LOGISTIC REGRESSION (LR)

LR is considered as the standard statistical approach to modeling binary data [16]. It is a better alternative for a linear regression which assigns a linear model to each of the class and predicts unseen instances basing on majority vote of the models. During prediction, instead of predicting the point estimate of the event itself, it builds a model to predict the odds of its occurrence. In two class problem for example, when the odds are greater than 50%, then the case is assigned to the class designated as "1" for YES and "0" for "YES" and "NO" instead.

5. SUPPORT VECTOR MACHINE (SVM)

SVMs are a set of related supervised learning methods that analyze data and recognize patterns, used for classification and regression analysis. SVM is an algorithm that attempts to find a linear separator (hyper-plane) between the data points of two classes in multidimensional space. SVM represents a learning technique which follows principles of statistical learning theory [14]. Generally, the main idea of SVM comes from binary classification, namely to find a hyperplane as a segmentation of the two classes to minimize the classification error. The SVM finds the hyperplane using support vectors (training tuples) and margins(support vectors). The Sequential Minimal Optimization (SMO) algorithm is a simple and fast method for training a SVM.

6. K-NEAREST NEIGHBOR (KNN)

K-Nearest Neighbor (KNN) classification [8] classifies instances based on their similarity. An object is classified by a majority of its neighbors. K is always a positive integer. The neighbors are selected from a set of objects for which the correct classification is known. The training samples are described by n dimensional numeric attributes. Each sample represents a point in an n-dimensional space. In this way, all of the training samples are stored in an n-dimensional pattern space. When given an unknown sample, a k-nearest neighbor classifier searches the pattern space for the k training samples that are closest to the unknown sample. "Closeness" is defined in terms of Euclidean distance. The unknown sample is assigned the most common class among its k nearest neighbors. When k=1, the unknown sample is assigned the class of the training sample that is closest to it in pattern space. In WEKA this classifier is called IBK

IV. RESULT AND DISCUSSION

The data set consist of 699 patients' record. Among them, 241 or 34.5% are reported to have breast cancers while the remaining 458 or 65.5% are not. In order to validate the prediction results of the comparison of the six popular data mining techniques and the 10-fold crossover validation is used. The k-fold crossover validation is usually used to reduce the error resulted from random sampling in the comparison of the accuracies of a number of prediction models. The entire set of data is randomly divided into k folds with the same number of cases in each fold. The training and testing are performed for k times and one fold is selected for further testing while the rest are selected for further training. The present study divided the data into 10 folds where 1 fold was for testing and 9 folds were for training for the 10-fold crossover validation. These diagnostic results of each patient's record in above dataset consist of ten variables that are summarized in Table 1. One of the 10 variables is the response variable representing the diagnostic status of the patient with or without breast cancers (i.e. malignant or benign). The training data are selected from the whole dataset randomly and directly fed into the proposed mining approach.

Table 1 provides the attribute information.

S. No	Attribute	Domain
1	Clump thickness	1-10
2	Uniformity of cell size	1-10
3	Uniformity of cell shape	1-10
4	Marginal adhesion	1-10
5	Single epithelial cell size	1-10

6	Bare nuclei	1-10
7	Bland chromatin	1-10
8	Normal nucleoli	1-10
9	Mitosis	1-10
	Class	2 for benign, 4 for malignant

1. Evaluation Methods

We have used the Weka toolkit to experiment with these three data mining algorithms [14]. The Weka is an ensemble of tools for data classification, regression, clustering, association rules, and visualization. WEKA version 3.6.9 was utilized as a data mining tool to evaluate the performance and effectiveness of the 6-breast cancer prediction models built from several techniques. This is because the WEKA program offers a well defined framework for experimenters and developers to build and evaluate their models.
The performance of a chosen classifier is validated based on error rate and computation time. The classification accuracy is predicted in terms of Sensitivity and Specificity. The computation time is noted for each classifier is taken in to account. The evaluation parameters are the specificity, sensitivity, and overall accuracy.

The sensitivity or the true positive rate (TPR) is defined by TP / (TP + FN); while the specificity or the true negative rate (TNR) is defined by TN / (TN + FP); and the accuracy is defined by (TP + TN) / (TP + FP + TN + FN)

- True positive (TP) = number of positive samples correctly predicted.
- False negative (FN) = number of positive samples wrongly predicted.
- False positive (FP) = number of negative samples wrongly predicted as positive.
- True negative (TN) = number of negative samples correctly predicted.

These values are often displayed in a confusion matrix as be presented in Table 2. Classification Matrix displays the frequency of correct and incorrect predictions. It compares the actual values in the test dataset with the predicted values in the trained model.

Table 2: Confusion Matrix. Predicted

	Predicted	
Actual	**Positive**	**Negative**
Positive	TP	FN
Negative	FP	TN

2. Results. The confusion matrix of each Classification method is presented in Table 3; the values to measure the performance of the methods (i.e. accuracy, sensitivity, specificity, error rate and time) are derived from the confusion matrix and showed in Table 4.

Table 3: Confusion Matrix of Training and Testing data

Algorithm	**Training Data (499)**			**Testing Data (200)**		
	Desired Result	**Output Result**		**Desired Result**	**Output Result**	
		Benign	Malignant		Benign	Malignant
J48	Benign	308	13	Benign	129	8
	Malignant	9	169	Malignant	8	55
Naive bayes	Benign	310	11	Benign	128	9
	Malignant	5	173	Malignant	2	61
MLP	Benign	307	14	Benign	130	7
	Malignant	12	166	Malignant	10	53
Logistic	Benign	314	7	Benign	131	6
	Malignant	9	169	Malignant	9	54
SVM(SMO)	Benign	315	6	Benign	131	6
	Malignant	6	172	Malignant	5	58
KNN(IBK)	Benign	314	7	Benign	130	7
	Malignant	17	161	Malignant	5	58

Table 4: Performance of Training and Testing data

Algorithm	Training Data (499)					Testing Data (200)				
	Acc	Senst	Spec	Err	Time	Acc	Senst	Spec	Err	Time
J48	95.59	0.96	0.949	4.41	0.09	92	0.942	0.873	8	0.08
Naive Bayes	96.79	0.966	**0.972**	3.21	0.05	**94.5**	0.934	**0.968**	**5.5**	0.05
MLP	94.78	0.956	0.933	5.22	4.03	91.5	0.949	0.841	8.5	1.94
Logistic	96.79	0.978	0.949	3.21	0.23	92.5	**0.956**	0.857	7.5	0.23
SVM(SMO)	**97.59**	**0.981**	0.966	**2.41**	0.73	**94.5**	**0.956**	0.921	**5.5**	0.69
KNN(IBK)	95.19	0.978	0.904	4.81	**0**	94	0.949	0.921	6	**0**

Acc - Accuracy, Senst - Sensitivity, Spec - Specificity, Err - Error Rate

From the above table we find that highest accuracy of Classification model is SVM (97.59%) and low error rate (2.41%) in both training and testing data as shown in figure 1and figure 2.

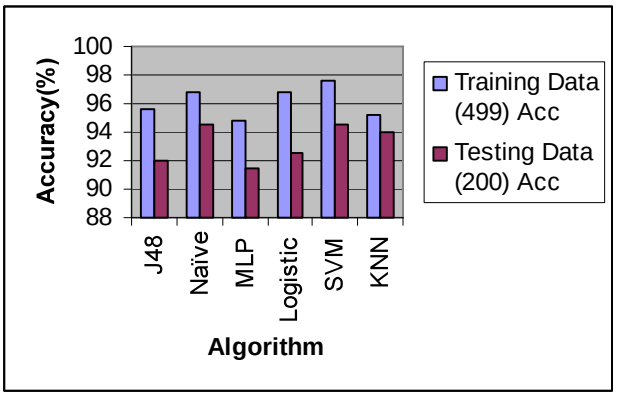

Figure 1: Accuracy of Classification methods

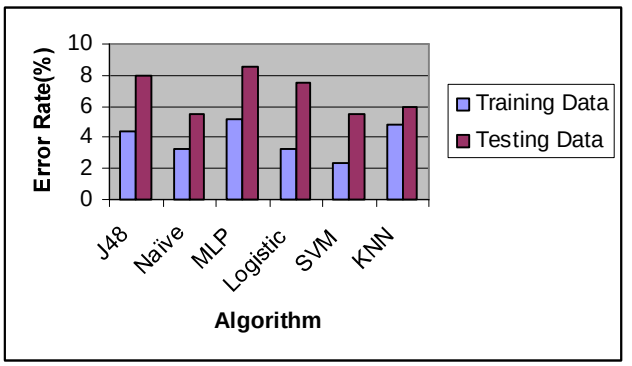

Figure 2: Error rate of Classification methods

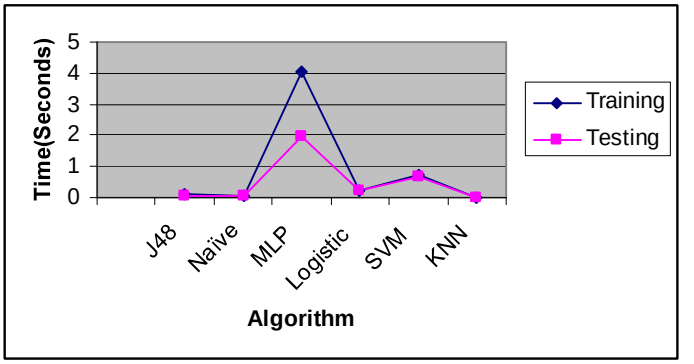

Figure 3: Execution Time of Classification methods

V. CONCLUSIONS

In this paper, the accuracy of classification techniques is evaluated based on the selected classifier algorithm .An important challenge in data mining and machine learning areas is to build precise and computationally efficient classifiers for Medical applications. The performance of SVM shows the high level compare with other classifiers. Hence SVM shows the concrete results with Breast Cancer disease of patient records. Therefore SVM classifier is suggested for diagnosis of Breast Cancer disease based classification to get better results with accuracy, low error rate and performance.

REFERENCES

[1] American Cancer Society. Breast Cancer Facts & Figures 2005-2006. Atlanta: American Cancer Society, Inc. (http://www.cancer.org/).
[2] 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.
[3] 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.
[4] Breast Cancer Wisconsin Data [online]. Available: http://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/breast-cancer-wisconsin.data.
[5] 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.
[6] 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.
[7] Ian H. Witten and Eibe Frank. Data Mining: Practical machine learning tools and techniques, 2nd Edition. San Fransisco:Morgan Kaufmann; 2005.
[8] 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.
[9] J. R. Quinlan, C4.5: Programs for Machine Learning. San Mateo, CA:Morgan Kaufmann; 1993.
[10] Mitchell, T. M., Machine Learning, McGraw-Hill Science/Engineering/Math, 1997
[11] P.-N. Tan, M. Steinbach, and V. Kumar, Introduction to Data Mining. Reading, MA: Addison-Wesley, 2005.
[12] 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.
[13] S.B.Kotsiantis and P.E.Pintelas,"Combining Bagging and Boosting", International Journal of Information and Mathematical Sciences, 1:4 2005.
[14] Vapnik, V. N., The nature of statistical learning theory. Springer, Berlin, 1995.
[15] Weka: Data Mining Software in Java, http://www.cs.waikato.ac.nz/ml/weka/
[16] Witten H.I., Frank E., Data Mining: Practical Machine Learning Tools and Techniques, Second edition, Morgan Kaufmann Publishers, 2005.