Comparison Analysis of Different Classifier Algorithms and Selection of Highly Ranked Attributes on Breast Cancer Datasets

Rashmi Amardeep[1], Shalini R[2], Megha S Kencha Reddy[3]
[1]Faculty, Dept. of ISE, Sir MVIT, rashmi_is@sirmvit.edu,
[2]VI semester, Dept. of ISE, Sir M.VIT,
shalinirgowda666@gmail.com,[3]megha.skr@gmail.com

Abstract: **Breast cancer is the second most common cancer worldwide. Globally, breast cancer represents one in four of all cancers in women.Valuable knowledge i.e. behavioural patterns and frequent/rare item trends in data can be identified from the application of data mining techniques on breast cancer healthcare data. Breast Cancer datasets of the Wisconsin dataset considered from UCI machine learning repository has been used for analysis. The study is been implemented on different classification algorithms (Naïve Bayes, SMO, Attribute Selected Classifier, Decision Strump, J48) using Weka 3.8.3 to predict the best model. Feature selection is applied by Wrapper-Subset-Evaluation in all algorithms.**

***Keywords*: Association Classifiers, Wrapper-Subset-Evaluation**

I. INTRODUCTION

Breast cancer is cancer that develops from breast tissue. Signs of breast cancer may include a lump in the breast, a change in breast shape, dimpling of the skin, and fluid coming from the nipple, a newly inverted nipple, or a red or scaly patch of skin. Risk factors[4] for developing breast cancer include being female, obesity, lack of physical exercise, drinking alcohol, hormone replacement therapy during menopause, ionizing radiation, early age at first menstruation, having children late or not at all, older age, prior history of breast cancer, and family history. The most common types are ductal carcinoma in situ, invasive ductal carcinoma, and invasive lobular carcinoma. In situ breast cancers does not spread. Invasive or infiltrating cancers spread (invaded) into the surrounding breast tissue.Tests and procedures used to diagnose breast cancer include: Breast exam, Mammogram, Breast ultrasound, Removing a sample of breast cells for testing (biopsy), Breast magnetic resonance imaging (MRI).
Breast cancer is treated in several ways: Surgery, Chemotherapy, Hormonal therapy, Biological therapy, Radiation therapy. Breast cancer recurrence: The goal of treating early and locally advanced breast cancer is to remove the cancer and keep it from coming back (breast cancer recurrence). Local recurrence is usually found on a mammogram, during a physical exam by a health care provider or when you notice a change after a lumpectomy, mastectomy.

Data Mining also known as Knowledge Discovery in Databases, refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data stored in databases. KDD is an iterative process where evaluation measures can be enhanced, mining can be refined, new data can be integrated and transformed in order to get different and more appropriate results[3]. Data learning uses two types of techniques: supervised learning, which trains a model on known input and output data so that it can predict future outputs, and unsupervised learning, which finds hidden patterns or intrinsic structures in input data.

Supervised methods [9] are methods that attempt to discover the relationship between input attributes (referred to as independent variables) and a target attribute (referred to as dependent variables).It is useful to distinguish between two main supervised models: classification models (classifiers) and regression models. Regression models map the input space into a real-value domain. On the other hand, classifiers map the input space into pre-defined classes.There are many alternatives for representing classifiers, for

example, support vector machines, decision trees, probabilistic summaries, algebraic function etc.

Classification deals with assigning observations into discrete categories, rather than estimating continuous quantities. Classification is the problem of identifying to which of a set of categories (sub populations), a new observation belongs to, on the basis of a training set of data containing observations and whose categories membership is known [5].

It is a two step process such as :

1. Learning Step (Training Phase): Construction of Classification Model Different Algorithms are used to build a classifier by making the model learn using the training set available. Model has to be trained for prediction of accurate results.
2. Classification Step: Model used to predict class labels and testing the constructed model on test data and hence estimate the accuracy of the classification rules.

Feature Subset: Given an inducer I, and a dataset D with features XI, X2, . . . , X,,, from a distribution D over the labeled instance space, an optimal feature subset, Xopt(optimal), is a subset of the features such that the accuracy of the induced classifier C = Z(D) is maximal[10].

II. LITERATURE SURVEY

K.Sutha et al. [1] applied effective unified framework for feature selection which involves in the Selection of best feature subset without any redundant and noisy data. Proved that wrapper subset evaluation shows better performance having no drawbacks.

Miss Jahanvi Joshi et al. [2] made Comparative study of classification method with different dataset Wisconsin Brest Cancer. Experimental findings used 10 fold cross validation method. Result analysis of BayesNet , Logistic, Multilayer Perceptron ,SGD, SimpleLogistic, SMO, AdaBoostM1,AttributeSelected, Classification Via Regression, FilteredClassifier, MultiClass Classifier Classifier, J48, LMT classifier gives more accurate result.

Shelly Gupta et al. [3] studied best classification technique over a dataset a set of rules that can be generated for the particular dataset.SVM showed the most promising results for PIMA Indian Diabetes dataset (Tanagra) and StatLog Heart Disease (Tanagra) dataset with 96.74% and 99.25% accuracy rate respectively and C4.5 decision tree for BUPA Liver-disorders (Tanagra) dataset with an accuracy rate of 79.71% whereas for Wisconsin Breast Cancer (clementine) dataset Bayes Net, SVM, kNN and RBF-NN all shown the almost similar results with high accuracy rate and the highest accuracy rate achieved is 97.28%.

Vikas Chaurasia et al. [4] implemented WEKA version 3.6.9 as a data mining tool to evaluate the performance and effectiveness of the 3-breast cancer prediction models built from several techniques. The study concluded that Sequential Minimal Optimization (SMO) is more accurate classifier in comparison to BFTree and Chi-square test, Info Gain test and Gain Ratio test analysis is done to determine the importance of each variable individually. Different algorithms provide very different results, i.e. each of them accounts the relevance of variables in a different way.

S. Syed Shajahaan et al. [5] studied supervised learning algorithm: Naïve Bayes, ID3, k-nearest, CART, Accuracy measures, precision, recall, accuracy. From results it was shown clearly that random tree algorithm gives the best accuracy for the breast cancer dataset of 683 records. An efficient classifier is identified to determine the nature of the disease which is highly essential in a clinical investigation of life threatening disease like breast cancer.

Gouda I. Salama et al. [6] determined classification accuracy and confusion matrix based on 10-fold cross validation method.A fusion at classification level has been implemented between the classifiers to get the most suitable multi-classifier approach for each data set. The experimental results in WBC dataset show that the fusion between MLP and J48 classifiers with features selection (PCA) is superior to the other classifiers.

S. Singaravelan et al. [7] evaluate the performance in terms of classification accuracy of J48 and Sequential Minimal Optimization algorithms using various accuracy measures like TP rate, FP rate, Precision, Recall, F-measure and ROC Area.On iris data set SMO proved to be better than J48.He concluded that algorithm based on neural network has better learning capability hence suited for classification problems if learned properly.

Holmes et al. [8] used *WEKA* as a machine learning workbench that is intended to aid in the application of machine learning techniques to a variety of real-world problems. It is an interactive tool for data manipulation, result visualization, database linkage, and cross-validation and comparison of rule sets, to complement the basic machine learning tools.

Osisanwo F.Y. et al. [9] used Supervised Machine Learning (SML) algorithms that reason from externally supplied instances to produce general hypotheses, which then make predictions about future instances.

Ron Kohavi et al.[10] used feature subset selection problem, a learning algorithm is faced with the problem of selecting some subset of features upon which to focus its attention, while ignoring the rest. In the wrapper approach, the feature subset selection algorithm exists as a wrapper around the induction algorithm.

III. CLASSIFICATION TECHNIQUE

3.1 Naive Bayes Classifiers

Naive Bayes classifiers [5] are a collection of classification algorithms based on Bayes' Theorem. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e. every pair of features being classified is independent of each other.Bayes' Theorem finds the probability of an event occurring given the probability of another event that has already occurred. Bayes' theorem is stated mathematically as the following equation:

$$P(A|B)=\frac{P(B|A)P(A)}{P(B)} \quad \text{-----(3.1)}$$

where A and B are events.

- Basically, we are trying to find probability of event A, given the event B is true. Event B is also termed as **evidence**.
- P(A) is the **priori** of A (the prior probability, i.e. Probability of event before evidence is seen). The evidence is an attribute value of an unknown instance(here, it is event B).
- P(A|B) is a posteriori probability of B, i.e. probability of event after evidence is seen.

3.2 Sequential Minimal Optimization (SMO)

Sequential Minimal Optimization (SMO) is used for training a support vector classifier using polynomial or RBF kernels. It replaces all missing the values and transforms nominal attributes into binary ones. A single hidden layer neural network uses exactly the same form of model as an SVM [7].

Training a Support Vector Machine (SVM) requires the solution of a very large quadratic programming (QP) optimization problem. SMO breaks this large QP problem into a series of smallest possible QP problems. These small QP problems are solved analytically, which avoids using a time-consuming numerical QP optimization as an inner loop.

The amount of memory required for SMO is linear in the training set size, which allows SMO to handle very large training sets.

3.3 Attribute Selected Classifier

Dimensionality of training and test data is reduced by attribute selection before being passed on to a classifier.

3.4 Decision Strump

A decision stump is a machine learning model consisting of a one-level decision tree [2]. That is, it is a decision tree with one internal node (the root) which is immediately connected to the terminal nodes (its leaves). A decision stump makes a prediction based on the value of just a single input feature. Sometimes they are also called 1-rules. Decision stumps are often used as components (called "weak learners" or "base learners") in machine learning ensembletechniques such as bagging and boosting.

3.5 J48

J48 algorithm of Weka software is a popular machine learning algorithm based upon J.R. Quilan C4.5 algorithm. All data to be examined will be of the categorical type and therefore continuous data will not be examined at this stage. The algorithm will however leave room for adaption to include this capability. The algorithm will be tested against C4.5 for verification purposes [7].

In Weka, the implementation of a particular learning algorithm is encapsulated in a class and it may depend on other classes for some of its functionality. J48 class builds a C4.5 decision tree. Larger programs are usually split into more than one class. The J48 class does not actually contain any code for building a decision tree. It includes references to instances of other classes that do most of the work. When there are a number of classes as in Weka software they become difficult to comprehend and navigate.

IV. FEATURE SELECTION

Feature selection is a pre-processing step, used to improve the mining performance by reducing data dimensionality [1]. Figure 4.1 gives the process of feature selection. Even though there exists a number of feature selection algorithms, still it is an active research area in data mining, machine learning and pattern recognition communities. Many feature selection algorithms confront severe challenges in terms of effectiveness and efficiency, because of recent increase in data dimensionality.

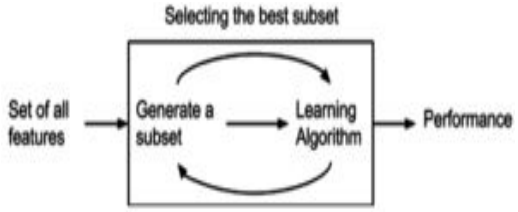

Figure 4.1: Feature selection process

A. WrapperSubsetEvaluation

The "wrapper" method wraps a classifier in a cross-validation loop: it searches through the attribute space and uses the classifier to find a good attribute set. Searching can be forwards, backwards, or bidirectional, starting from any subset. Wrapper methods evaluate subsets of variables which allows, unlike filter approaches, to detect the possible interactions between variables.

Best First: This search strategy searches the subsets from feature space by using greedy hill climbing amplified with backtracking. The intensity of backtracking may be controlled by locating an amount of successive non-improving nodes. This search method works both in SFS and SBE mode or may start from any random point and search bidirectional. In this paper we study the backward technique rather than forward. Threshold Set go 1 by which attributes can be discarded.

V. EXPERIMENTS

A. Attribute Name Description

Age: Patient's Age in years

Menopause: The period in a woman's life when menstruation ceases

Tumor-size: Patient's tumor-size on her breast

inv-nodes: Node size in main portion of the breast.

Node-caps: Node is present or not in cap of the breast(YES/NO)

Deg-malig :Stage of breast cancer

Breast: Left breast or Right breast or both breast

Breast-quad :Portion of the breast for example left-up, left-low, right-up, right-low, central.

Irradiate :Present or not (YES/NO)

Class:no-recurrence-events, recurrence-events (Reduce the risk of breast cancer).

B. Cross Validation: k folds

This approach involves randomly dividing set of observations into k groups of approximately equal size.The first fold is treated as a validation set and the method is fit on the remaining k-1 folds. The study is by choosing k=10 folds.

C. Problem Description

a. A comparison between five classifier algorithms on the dataset based on correctly and incorrectly classified instances.

b.Time taken to build each classifier model.

c.WrapperSubsetEvaluation has been carried out for each of the classifier algorithm.

VI. RESULT ANALYSIS

There are two class labeled as Recurrence(R) and Non-Recurrences (NR) class. Table 6.1 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. The columns represent the predictions, and the rows represent the actual class. Table 6.2 tabulates the comparison of the classifiers.

Table 6.1 Prediction of instances

Classifiers	Correctly classified instances	Incorrectly classified instances	Class
Naïve Bayes	171	30	NR
	48	37	R
Sequential Minimal Optimization	175	26	NR
	57	28	R
Attribute Selected Class	192	9	NR
	66	19	R
Decision Strump	161	40	NR
	40	45	R
J48	190	11	NR
	60	25	R

Table 6.2 Comparison analysis of classifiers

Classifiers	Correctly classified	Incorrectly classified	ROC Curve
Naïve Bayes	72.7273%	27.2727%	0.696
Sequential Minimal Optimization	70.979%	29.021%	0.600
Attribute Selected Class	73.7762%	26.2238%	0.637
Decision Strump	72.028%	27.972%	0.616
SJ48	75.1748%	24.8252%	0.641

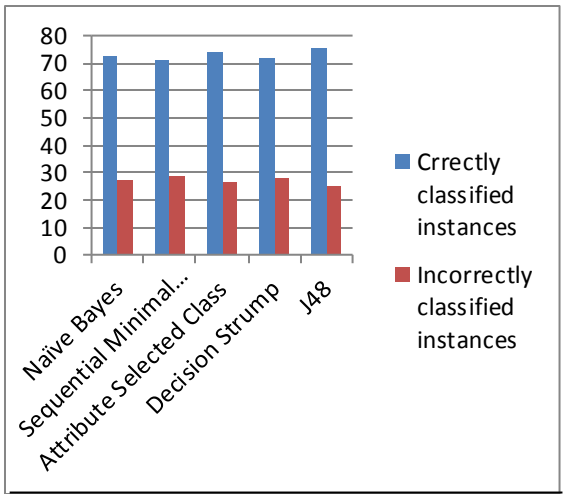

Figure 6.1: Comparison of different classifiers

From the figure 6.1, table 6.1 and table 6.2 of comparison analysis we observe that J48 algorithm has the highest number of correctly classified instances. Thus we conclude that it is more accurate classifier when compared to Naive Bayes, SMO, Attribute Selected Classifier and Decision Strump.

Figure 6.2 time graph, we observe that time taken to build SMO model is higher than other models.This result may be due to the training time being longer in small feature spaces than in larger ones and other algorithms time complexity depends on the size of the dataset.SMO is more efficient for larger dataset.

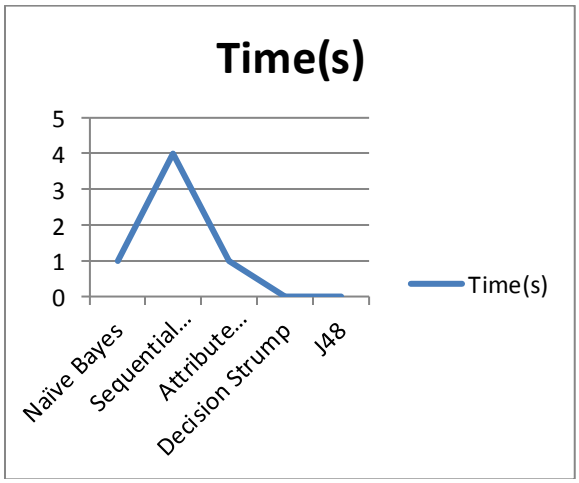

Figure 6.2 : Time graph

Table 6.3 describes accuracy of different Classifier Mode (before and after Selecting Attributes Using WrapperSubsetEval). By applying WrapperSubsetEvaluation method to each of the classifier to choose the highly ranked attributes while ignoring the others, we observe that there is a slight increase in the number of correctly classified instances thereby increasing the accuracy rate.

Table 6.3 Accuracy of algorithm before & after feature selection

Classifiers	Accuracy before	Attributes selected	Accuracy after
Naïve Bayes	72.7273%	Age,inv-nodes,deg-malig,breast,breast-quad	75.5245%
Sequential Minimal Optimization	70.979%	Age,menopause,tumor-size,node-caps,deg-malig	73.4266%
Attribute Selected Class	73.7762%	Invalid-nodes,deg-malig	74.1259%
Decision Strump	72.028%	Node-caps	72.3776%
J48	75.174%	Class,inv-node,node-caps,di-malig,breast,irradiat	75.5245%

VII. CONCLUSION

This paper implements five popular data mining classifiers: NaiveBayes, SMO, Attribute Selected Classifier, Decision Strump, J48. We observe that J48 algorithm outperforms all other classifiers on breast cancer data set. An important challenge in data mining is to build precise and computationally efficient classifiers considering the time constraint. Fewer attributes often yield better performance. In a laborious manual process starts with the full attribute set and remove the best attribute by selectively trying all possibilities, and carry on doing that Weka's select Attributes panel accomplishes this automatically.

VIII. FUTURE WORK

Ensembling different classifiers to obtain the most efficient multiple models combination [6]. The five basic algorithms can be applied to develop multiple models are Bagging, Random Forest, AdaBoost, Voting, Stacking. In future the work is proposed by using voting algorithm because of its simplicity in ensemble algorithm. Voting works by creating two or more sub models. Each sub-model makes predictions which are combined in certain way, such as taking the mean, mode or the probability of the predictions, allowing each sub-model to vote on what the outcome should be. The multiple model with highest accuracy can be used instead of a single classifier model.

IX. REFERENCES

[1] K.Sutha, Dr.J. Jebamalar Tamilselvi,"A Review of Feature Selection Algorithms for Data Mining Techniques", International Journal on Computer Science and Engineering (IJCSE), ISSN: 0975-3397 Vol. 7 No.6 Jun 2015

[2] 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 ISSN 2091-2730.

[3] Shelly Gupta, Dharminder Kumar, Anand Sharma,"Performance Analysis Of Various Data Mining Classification Techniques On Healthcare Data", International Journal of Computer Science & Information Technology (IJCSIT) Vol 3, No 4, August 2011.

[4] Vikas Chaurasia, Saurabh Pal," A Novel Approach for Breast Cancer Detection using Data Mining Techniques", International Journal of Innovative Research in Computer and Communication Engineering · January 2014.

[5] S. Syed Shajahaan, S. Shanthi, V. ManoChitra," Application of Data Mining Techniques to Model

[6] Gouda I. Salama, M.B.Abdelhalim , and Magdy Abd-elghany Zeid," Breast Cancer Diagnosis on Three Different Datasets Using Multi-Classifiers", International Journal of Computer and Information Technology (2277 – 0764) Volume 01– Issue 01, September 2012.

[7] S. Singaravelan, D. Murugan and R. Mayakrishnan," Analysis of Classification Algorithms J48 and Smo on Different Datasets", World Engineering & Applied Sciences Journal 6 (2): 119-123.

[8] Holmes, G., Donkin, A., Witten I.H., "WEKA a machine learning workbench", In: Proceeding second Australia and New Zeeland Conference on Intelligent Information System, Brisbane, Australia, pp.357-361, 1994.

[9] Osisanwo F.Y, Akinsola J.E.T., Awodele O, Hinmikaiye J.O, Olakanmi O,Akinjobi J, "Supervised Machine Learning Algorithms: Classification and Comparison", International Journal of Computer Trends and Technology (IJCTT) – Volume 48 Number 3 June 2017