International Journal of Engineering Science Invention (IJESI)
ISSN (Online): 2319 – 6734, ISSN (Print): 2319 – 6726
www.ijesi.org || PP. 06-09

Performance and Evaluation of Data Mining Ensemble Classifiers

Dr. V. Palaniyammal
Principal Sri SarathaMahavidyalayam Arts and Science College for Women / Pullur, Ulundurpet- 606 107

Abstract: *We analyze the breast Cancer data available from the WBC, WDBC from UCI machine learning with the aim of developing accurate prediction models for breast cancer using data mining techniques. Data mining has, for good reason, recently attracted a lot of attention, itis a new Technology, tackling new problem, with great potential for valuable commercial and scientific discoveries. The experiments are conducted in WEKA several data mining ensemble classification techniques were used on the proposed data. The data breast cancer data with a total 286 rows and 10 columns will be used to test and justify the different between the classification the ensemble methodology is to build a predictive model by integrating multiple classifier models. The ensemble methods can be used for improving prediction performance. In this regard this paper presents the important types of ensemble methods including boosting and bagging, stacking with WBDC Data set.*
Keywords: *Data mining Weka Ensemble classifiers Bagging, Boosting, stacking WBCD set*

I. Introduction

About 1 in 8 U.S women (just under 12%) will develop invasive breast cancer over the course of her life. In 2011, an estimated 230,480 new cases of invasive breast cancer were expected to be diagnosed in women in the U.S, along with 57,650 new cases of non-invasive (in situ) breast cancer. For women in the U.S breast cancer death rates are higher than those for any other cancer besides lung cancer. In 2011 there were more than 2.6million breast cancer survivor in the U.S.A women's risk of breast cancer approximately relative (mother, sister, daughter) who has been diagnosed with breast cancer. About 15% of women who get breast cancer have a family member diagnosed with it. About 85% of breast cancers occur in women who have no family history of breast cancer. These occur due to genetic mutations that happen as a result of the aging process and life in general.

II. Breast cancer (An overview)

Breast Cancerous the most common cancer diseases among women excluding no melanoma skin Cancers are divided into two types benign and Malignant. If the cancer is benign under the conditions of early diagnosis. Malignancy status includes the three basic measurements

1. Age
2. Longer tumor length
3. ADC or Apparent Diffusion coefficient (biopsy confirmed).

Attributes 1.patient id 2.diagnosis m=malignant ,b=benign 3.Real valued features. Cell uncles' 1.radius 2.texture 3.perimeter 4.Area 5.Smoothness.I Diagnosis Attributes are 1.Menopausal status a.Premenopausal, b.post menopausal. II.Basic diagnosis :-1.clinical examinations,2.mammography (y/n),3.MRI (y/n),4.Ultrasound (y/n),5.Fine Needle aspiration (y/n),6.Core biopsy (y/n),7.Open biopsy (y/n) ,8.Other (y/n),9.Unknown (y/n).III Clinical Trial Enrollment:- y/n/Not stated/inadequate.ivInitial Representation:- Screening-mammography, Screening MRI, Screening-Others, Symptomatic.V.Total extent of Lesion (DCIS AND INVASIVE) 1.Lesion in mm.VI.Lymphovascular Invasion (presence of tumor cells in endothelium- lined spaces) 1.present/absent/suspicious/not stated /unknown.

Stages of Breast cancer:-

There are 4 number stages of breast cancer, staging takes into various factors, including 1.The size of the tumor (tumor means either breast lymph's or Area of cancer cells found on a Scan or mammogram)

2. Cancer cells have spread into nearby lymph glands (lymph Node) 3.The tumor cells has spread to any other part of the body (Metastasized-TNM). Stage1 breast cancer is split into 2 Stages.Stage1A tumor is 2 cm or smaller and has spread outside the breast.Stage1B: Small areas of breast Cancer cells are found in the lymph node closed to the breast and either the tumor is 2 cm or smaller. Stage 2 breast cancer: stage 2A:Tumor 2cm or smaller in the breast and cancer cells are found in 1 to 3 lymph Node in the armpit or in the lymph Node near the breast bone. Stage 2B: The tumor is larger than 2 cm but not larger than 5 cm and small areas of cancer cells are in the lymph Node.2 to 5 cm spread 1 to 3 lymph nodes in the armpit or near the breastbones or the tumor is larger than 5cm and has not spread to the lymph node. Stage 3A breast cancer No tumor is seen in the breast or the tumor may be any size and cancer is found in 4 to 9 lymph glands under the arm or in the lymph

glands near the breast bone. The tumor is more than 5 cm and has spread into up to 3 lymph nodes near the breast bone.

III. Existing System

The data sets used are SEER data or Wisconsin data. Data pre-processing is applied before data mining to improve the quality of the data. Data pre-processing includes data cleaning, data integration, data transformation and data reduction techniques. The features used for classification purposes coincided with the Breast Imaging Reporting and Data System (BI-RADS) as this is how radiologists classify breast cancer. The BI-RADS features of density, mass shape, mass margin and abnormality assessment rank are used as they have been proven to provide good classification accuracy. 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. Feature selection with decision tree classification greatly enhances the quality of the data in medical diagnosis. CART algorithm with various feature selection methods to find out whether the same feature selection method may lead to best accuracy on various datasets of same domain. Artificial neural networks (ANNs) and support vector machines have been recently proposed as a very effective method for pattern recognition, machine learning and data mining. The discrimination capability of the features extracted from the sonograms was tested by using the SVM (support vector machine), ANN and KNN (Nearest neighbour) classifier. It was found that the SVM gave the greatest accuracy while the ANN had the highest sensitivity. Back- propagation neural network (BPNN) and radial basis function network (RBFN) were used for the training and testing of data.

IV. Proposed System

The term data mining is used to describe the process of knowledge discovery from data. This process involves analysis and summarization of a huge amount of data stored in a warehouse and extraction of non obvious and intricate patterns. The ensemble methodology is used to combine a set of models, each of which solves the same original task, in order to obtain a better composite global model, with more accurate and reliable estimates or decisions that can be obtained from using a single model [1, 2].

The predictive model is developed by integrating multiple models. Ensemble methods can also used for improving the quality and robustness of classification algorithm. Ensemble methods are very effective as it has various types of classifiers [1, 2]. There are several factors that differentiate between the various ensemble methods. The main factors are:

1. Inter-Classifiers relationship

Inter-classifiers describe how each classifier affects the other classifier. This ensemble factor can be divided into Sequential type and Concurrent type. In sequential approaches learning ensembles interaction happens in a sequential manner. It can be identified through learning runs. Thus, it is possible to take advantage of knowledge generated in previous iterations to guide the learning in the next iterations.

2. Combined Method

In combined method, the strategy of combining the classifiers generated by an induction algorithms. The simplest combiner determines the output solely from the output of the individual inducers. Stacking or arbitration is the sophisticated methods of this factor.

3. Diversity Generator

In order to make the ensemble effective, there should be some sort of diversity between the classifiers. Diversity may be obtained through different presentations of the input data, as in bagging variations in learner design or by adding a penalty to the outputs to encourage diversity.

4. Ensemble size

There number of classifier in the ensemble. Accurate result can be obtained by distributing various data sets.

Boosting

Boosting is a general method for improving the performance of any learning algorithm. The method works by repeatedly running a weak learner (such as, classification rules, decision rules). On various distributed training data set. The classifiers produced by the weak learners are combined into single composite strong classifier. In order to achieve a higher accuracy, the weak learner's have seen combined [3].

The main idea of this algorithm is to assign a weight in each example in the training set. In the beginning all weights are equal, but in every round, the weight of all misclassifier instances is increased while the weights of correctly classified instance are decreased.

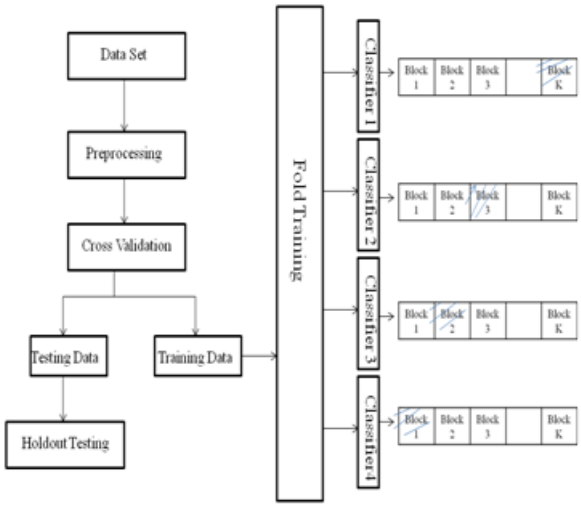

Figure.1 Architecture of an Ensemble Based System

Table captions appear centered above the table in upper and lower case letters. When referring to a table in the text, no abbreviation is used and "Table" is capitalized.

V. Boosting Algorithm

1. Initialize example weights Wi = 1/N, where i= 1 to N
2. For M = 1 to M //M number of iterations

a. Learn classifier Cm using the current example weights

b. Compute a weighted error estimate errm= ∑Wi of all incorrectly classifier ei/sum of weights

c. compute a classifier weight am = 0.5 log(1-errm/errm)

d. for all correctly classified examples ei:wi=wie-am

e. For all incorrectly classified examples ei:wi=wieam

f. Normalize the weights wi so that they sum to 1

3. For each test example

a. Try all classifiers cm

b. Predict the class that receives the highest sum of weights

In this algorithm weak learners are stacked on top of one another iteratively. Once a new weak algorithm is added, the data is reweighted. The tuples are misplaced gain importance over the correctly classified ones. When next classifier is added concentrated on the classification of the misplaced data. Hence eventually most of the data gets classified with great accuracy.

VI. Bagging

The most well known method that processes samples concurrently is bagging (Bootstrap aggregating). The method aims to improve the accuracy by creating an improved composite classifier. The multiple versions are formed by making bootstrap replicates of the learning set and using these as new learning sets. Tests on real and simulated data set using classification and regression trees and subset selection in linear regression show that bagging can give substantial gains in accuracy [1, 2].

Bagging Algorithm

1. For m = 1 to M // M number of iteration

a. Draw (with replacement) a Bootstrap Samples Sm of the data

b. Learn a classifier Cm from Sm

2. For each test example

a. Try all classifiers Cm

b. Predict the class that receives the highest numbers of votes

Bagging adopts the most popular strategies for aggregating the outputs of the base learners, which is voting for classification and averaging for regression. To predict a test instance, taking classification. For example, bagging feeds the instance to its base classifiers and collects all the outputs and then votes the labels and finally takes a winner label as predictions. Bagging algorithm can deal with binary classification as well as

multiclass classification [1].

H(x) = arg Max

VII. Stacking

Stacking is a general procedure where a learner is trained to combine the individual learners. Here the individual learners are called the first-level learners, while the combiner is called the second level learner or Meta learner [6].

The basic idea is to train the first-level learners using the original training dataset, and then generate a new dataset for training the second-level learners, where the outputs of the first-level learners are regarded as input futures, while original labels are still regarded as labels of the new training data[6].

Stacking algorithm [4, 5] is given in Figure

Stacking Algorithm

Input: Data set D = {(x1, y1), (x2, y2), … (xm, ym)};

First level learning algorithm ∑1 … ∑T

Process:

1. for t 1 …t; // Train a first level
2. Ht = ∑t (D);
3. End
4. D' = ø;
5. fori = 1, …m;
6. for t = 1, …m;
7. Zit = ht(xi);
8. End
9. D' = D'U((Zi1, … Zit), Yi);
10. End
11. H' = ∑(D') //Training the second level learner

Output: H(x) = h'(h1(x), … hT(x)).

In the following Table.1 shows the execution time of three classification algorithm. In this the ensemble algorithm stacking had to give the great accuracy

In the following Table.1 shows the execution time of three classification algorithm. In this the ensemble algorithm stacking had to give the great accuracy

Algorithm Total Instance 286	Correctly Classified Instance (%Value)	Incorrectly Classified Instances (%Value)	Time taken (Seconds)	Kappa Statistics
Boosting	(280) 97.9021	(6) 2.0979	0.16	0.9491
Bagging	(259) 90.5594	(27) 9.4406	0.10	0.7584
Stacking	(280) 97.9021	(6) 2.0979	0.10	0.9494

VIII. Conclusion

This paper presents effective classification Techniques. After investigation of different classification Algorithm we have chosen ensemble classifier based on our simulation performance and we have used stacking classifier achieved overall classification accuracy 94%, which is significant. In future work we propose to analyze Ensemble classifier for 100% accuracy.

References

[1] Breiman, Bagging Predictors, Technical Report 421, Department of Statistics, University of California, Berkeley, 1994.
[2] L. Breiman, "Bagging Predictors", Mach. Learning.24, 1996, pp. 123-140.
[3] Y. Freund, R. Schapire, Experiments with a new Boosting Algorithm. In Proceedings of the Thirteenth International Conference on Machine Learning, pp. 148156, Morgan Kaufmann, 1996.
[4] Zhi-Hua Zhou, "Ensemble Methods: Foundations and Algorithms" CRS press, 2012.
[5] L. Kuncheva, "Combining Pattern Classifier: Method and Algorithms", Willey, 2004.
[6] D. Nolpert, Stacked Generalization Neural Network, 5(2), pp: 241-259, 1992.
[7] Mitchell, T.M. 1997. Machine Learning. McGraw-Hill Science John, G., Cleary, E. & Leonard, E. 1995. K*: An Instance-based Learner Using an intropic
[8] Distance Measure. In: 12th International Conference on Machine Learning, 108 114