PREDICTION OF BREAST CANCER USING STACKING ENSEMBLE APPROACH

[1]VALLURI RISHIKA, M.TECH

COMPUTER SCENCE AND SYSTEMS ENGINEERING, ANDHRA UNIVERSITY

[2]A. MARY SOWJANYA, Assistant Professor

COMPUTER SCENCE AND SYSTEMS ENGINEERING, ANDHRA UNIVERSITY

ABSTRACT:

Breast cancer is the most common cancer in women and thus the early stage detection in breast cancer can provide potential advantage in the treatment of this disease. Early treatment not only helps to cure cancer but also helps in its prevention of its recurrence in women. It occurs when the growth of the cells in breast tissue become out of control. Cells are the building blocks for the organs and tissues in the body. When the growth of new cells is uncontrolled then they build-up mass of tissue called tumor. The tumors are categorized in to benign and malignant tumors. Early diagnosis needs an accurate diagnosis procedure that can be used by physicians to classify whether the tumor is benign or malignant. Classification algorithms are used to classify whether it is benign or malignant tumor. The main objective of this paper is to compare the results of supervised learning classification algorithms and combination of these algorithms using stacking classifier technique. Stacking is one of the ensemble approach where multiple models are combined for the better classification. The dataset is taken from Wisconsin University database.

Index terms: Decision tree, Neural Network, Naive Bayes, Stacking approach.

INTRODUCTION:

Cancer occurs when changes called mutations take place in genes that regulate cell growth. The mutations let the cells divide and multiply in an uncontrolled, chaotic way. The cells keep multiplying, producing copies that get progressively more abnormal. In most cases, the cell copies eventually form a tumor. Breast cancer

became one of the deadliest cancers in women. When the growth of cells in breast tissue became uncontrollable it forms a mass tissue called tumor. These tumors are mainly classified in to benign and malignant tumors which are cancerous and non-cancerous. Benign tumors are not harmful which do not spread to the other parts of the body. They can be removed completely and they do not grow back again. Malignant tumors are threat to life and they can spread to other parts of the body and reappearing of malignant tumors can be seen often even when they are removed. Breaking away from the breast tumors, cancer cells can travel through lymph vessels and blood vessels to reach other parts of the body. It may attach to other tissues of the body parts and grow to form new tumors that can cause damage to the entire function. Several tests are included to diagnose the patient, it includes surgical biopsy where patient need to go through operation. Nowadays data mining application has been increased in medical field. There are a few arguments that can support the use of data mining in health sector for breast cancer like early detection, early avoidance, and indication based medication, rectifying hospital data errors. Many machine learning algorithms are used for the better treatment to the patient. Supervised algorithms such as classification and unsupervised such as regression and clustering which are helpful for the diagnosis of the patient. There are many techniques to predict and classification breast cancer pattern. This paper compares performance of three classification algorithms and their combination using ensemble approach that are suitable for direct interpretability of their results. This paper presents a new model by combining three classifiers that enhances the accuracy in recognizing breast cancer patients. Stacking ensemble technique is used for classification of benign and malignant tumor. In this paper we inspected the generalization performance of Decision tree, Neural Network, Naive Bayes in order to boost the prediction models for decision-making system in the prediction of breast cancer survivability. In this paper, a stacking ensemble approach is used where all three classification algorithms are combined for the prediction of breast cancer.

Literature Survey:

Data mining classification algorithms are used on large set of breast cancer data to classify whether the cell is benign or malignant. The dataset which we are used is taken from the Wisconsin breast cancer data set. Even though some of the papers included several individual algorithms for the correct classification of the benign or

malignant tumor. We have taken the three best classification algorithms trained with dataset individually and combined the three algorithms using voting approach for the best accuracy and correct predictions. Several individual algorithms have its own drawn backs and its own strength. In the recent trend Stacking is the one of the best approach for the combination of two or more algorithms for the accurate classification. The power of the three classifiers combined and can be predict the type of tumor precisely.

Chithanshah, et al., compared Decision tree, K-Nearest Neighbor and Naïve Bayes algorithms in his paper. They performed all the experimental work in Weka tool. The results showed that Naïve Bayes classifier algorithm is the best classification algorithm to the breast cancer. For the experimental work, the Naïve Bayes is chosen as one of the classification algorithm.

Naïve Bayes classification is based on Bayes theorem with an assumption of independence between features. Bayesian classifier is a statistical classifier as well as a supervised learning method. Basically, it classified in to three categorizes, Gaussian, multinomial, Bernoulli Naïve Bayes. It is particularly used for large data sets. Naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable. It performs well in case of categorical input variables compared to numerical variable.

Neural networks are capable of modeling extremely complex, typically non-linear functions. 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. An Artificial Neural Network (ANN), generally known as "Neural Network"(NN), is a mathematical model or computational model based on biological neural networks. It consists of an interconnected group of artificial neurons and processes information using a connectionist approach to computation. In most cases, an ANN is a robust system that changes its structure based on external or internal information that flows through the network during the learning phase. ANN has confirmed as a powerful method for cancer prognosis. It is also superior to conventional methods employed for breast cancer prediction such as TNM (Tumor, Node and Metastasis) staging system and logistic regression. In this paper neural network is used for predictions.

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. 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. Once the tree or rules are derived in learning phase to test the accuracy of a classifier, test data is taken randomly from training data. After verification of accuracy, unlabeled data is classified using the tree or rules obtained in learning phase. Both theoreticians and practitioners are continually seeking techniques to make the process more efficient, cost-effective and accurate.

Stacking is the abbreviation of Stacked Generalization. The basic idea of Stacking is to combine different classifiers from different classification algorithms, such as decision tree, multilayer propagation, naive Bayes, to generate a higher level classification system. As mentioned in the introduction, diversity of base-level classifiers is important to generate an ensemble. The algorithms to generate classifiers apply different hypothesis, thus the errors and bias of them differ from each other. And the differences of classifiers are considered as not correlated, which could explain the diversity of the base-level classifiers in Stacking. Stacking uses the meta-level classifier to map the outputs of the base-level classifiers to the final decision. Once all the base-level classifiers are trained, their outputs of each training instance will be taken as the independent attributes and the real class labels of the instances will be taken as the dependent attribute. For all the training instances, the new training set is generated to train the meta-level classifier. When all the training processes of the baselevel classifiers and the meta-level classifier are finished, a Stacking ensemble is obtained. To classify a new instance, the meta-level classifier takes the predictions of the baselevel classifiers as its input and give its prediction as the decision. Since the proposition of Stacking, the question of how to obtain a right configuration to optimize its performance is asked. Some effort and research is devoted to solve this problem. Most effort of Stacking is spent on the selection of meta-level data or algorithm to generate the meta-level classifier.

PROPOSED WORK:

In our proposed work achieving highest accuracy is aimed by taking combination of three classifiers such as Decision Tree, Naïve Bayes and neural network. Each classifier has its own disadvantages which will be beneficial to the other classifier. By combining two or more of these classifiers we can achieve a strong ensemble

model based on Stacking strategy. The data used in this study are obtained from the University of Wisconsin Hospitals, Madison from Wolberg. The data set provides required amount of information for prediction of cancer.

EXPERIMENTAL SETUP :

An ensemble is a process of combining predictions of each base classifiers to classify new examples. Stacking method involves two types of learners i.e. first-level learners and second-level learner. The individual learners are called as first-level learners and the combiner is called as the second level learner, or meta-learner. The original training data set is given as input to the first-level learners, and then output of first-level learners are given input to train the second-level learner. The new dataset is generated after applying first-level learners. A leave-one-out or a cross validation procedure is applied to generate a training set for learning the meta-level classifier. In leave-one-out approach, each of the base-level learning algorithms are applied to the entire dataset except one example. Predictions of the base-level classifiers are considered as the feature for meta-learner. Then meta-learner combines these predictions and gives the final output.

Step 1: In this step all the data is pre-processed and replace the missing values with the modes and means of the training dataset.

Step 2: Splitting the dataset into the Training set and Test set.

Step 3: The train set is split into 10 parts.

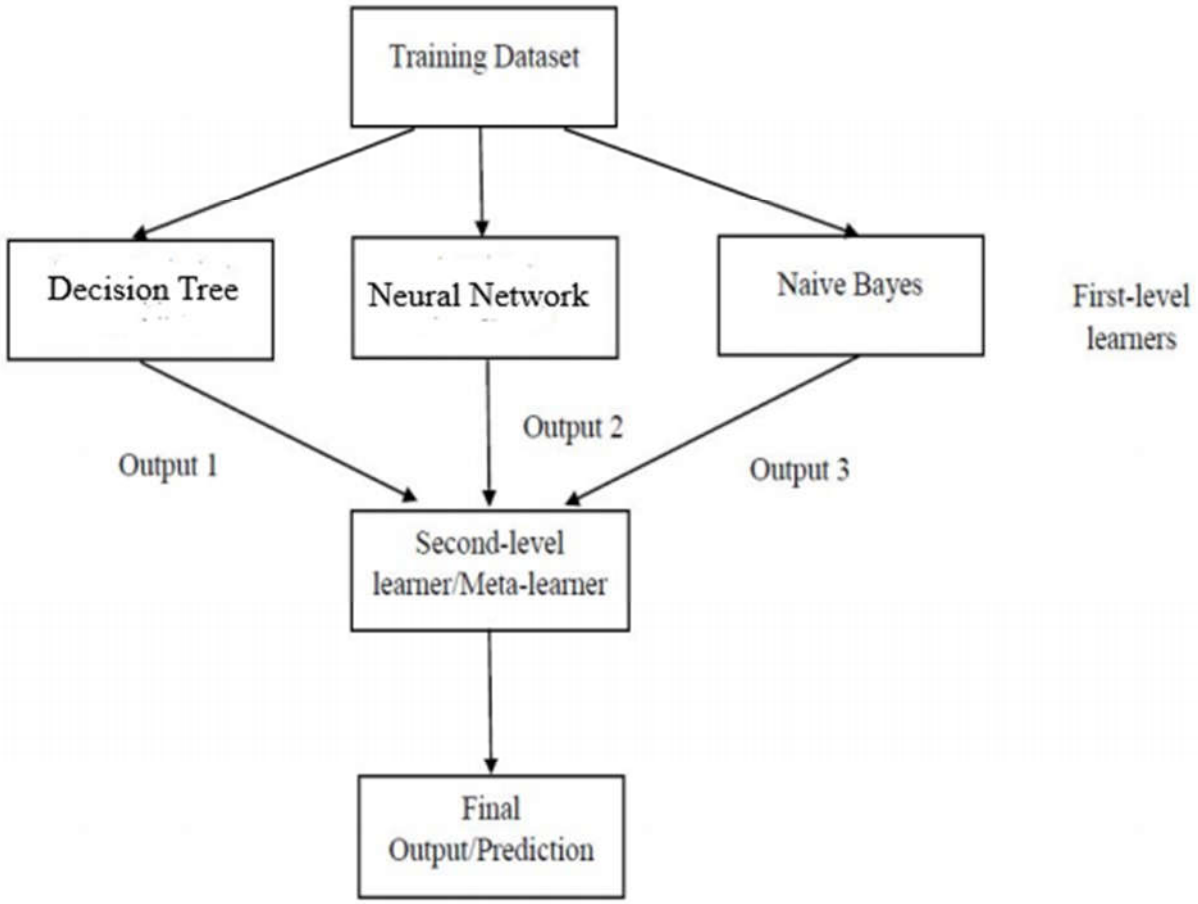
Step 4: A base model (suppose a decision tree) is fitted on 9 parts and predictions are made for the 10th part. This is done for each part of the train set.

Step 5: The base model (in this case, decision tree) is then fitted on the whole train dataset.

Step 6: Using this model, predictions are made on the test set. Steps 3 to 5 are repeated for another base model (say neural network and naïve Bayes) resulting in another set of predictions for the train set and test set.

Step 7: The predictions from the train set are used as features to build a new model.

Step 8: This model is used to make final predictions on the test prediction set that increases the percentage of accuracy.



A. **Performance evaluation**:

In machine learning confusion matrix is also called as error matrix kind of contingency table having two dimensions predicted and actual. It mainly reports true positive, true negative, false positive, false negative. Accuracy is used as a common evaluation measure to evaluate the performance of classifier. But this seems to be insufficient while dealing with imbalanced data. So in this paper, ROC curve, Kappa static, sensitivity and specificity and relative error are used as our evaluation criteria. From the confusion matrix all these values are derived.

Accuracy is calculated as ratio of the sum of correct classification as benign and malignant to the total no of instances, and is calculated by using the equation:
Accuracy = (TP+TN)/TOTAL

Misclassification rate is the overall how often our prediction is wrong, it is proportion of negatively identified cases to the total, and is calculated by using the equation:
Misclassification Rate = (FP+FN)/Total

ROC curve is a graphical plot of the true positive rate against the false positive rate for the various thresholds of a diagnostic test. ROC curve is used to quantify the performance of a classifier, and to give a higher score from the previous classifier. The true positive rate also known as sensitivity and the false positive rate known as specificity Sensitivity is also called true positive rate, how often it is predicted as true when it is true, and is calculated by using the equation:
Sensitivity = TP / (TP + FN)

Specificity is also called as true negative rate, how often it is predicted as false when it false and is calculated by using the equation:
Specificity = TN / (TN + FP)

Kappa statistic is a measure of how well the classifier performed when compared to how well it would have been performed by chance. If a model having high score then there having big difference between accuracy and null rate.
K=(Po-Pe)/(1-Pe)
Where P0 is observed accuracy and Pe is the expected.

Classifier	Decision Tree	Naive Bayes	Neural network	Stacking
Accuracy	91.95%	93.56	89.12	97.18
Kappa Statics	0.92	0.91	.88	0.78
Mean Absolute Error	0.04	.03	.06	.02
ot Mean Squared Error	0.197	0.181	0.22	0.167
lative Absolute ror	7.89	8.73	9.12	6.78
ot Relative uared Error	41.78	39.98	43.74	35.95

Table: individual classifiers accuracy

B. **Results and Discussion**

The dataset is trained and tested for every individual Classification algorithm. Cross validation is used for accurate prediction and it is also called rotation estimation. We are using 10-fold cross-validation to limit problems like over fitting and this method uses over repeated random sub-sampling is that all observations are used for both training and validation, and each observation is used for validation exactly once. All the evaluation results are shown in the table.

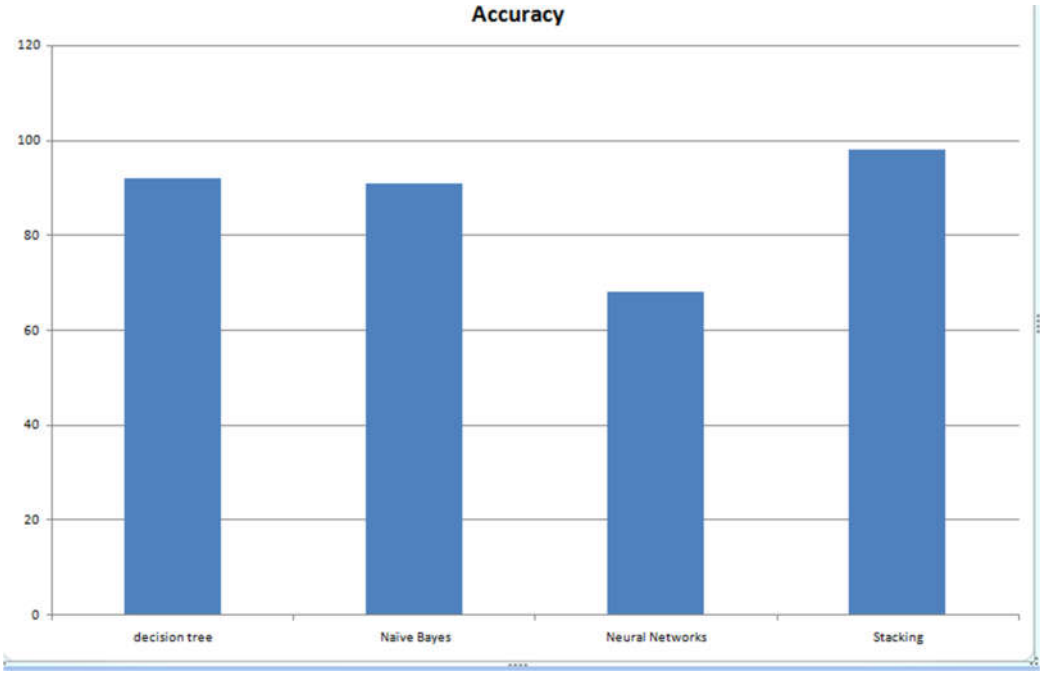


figure: Comparison of different algorithms.

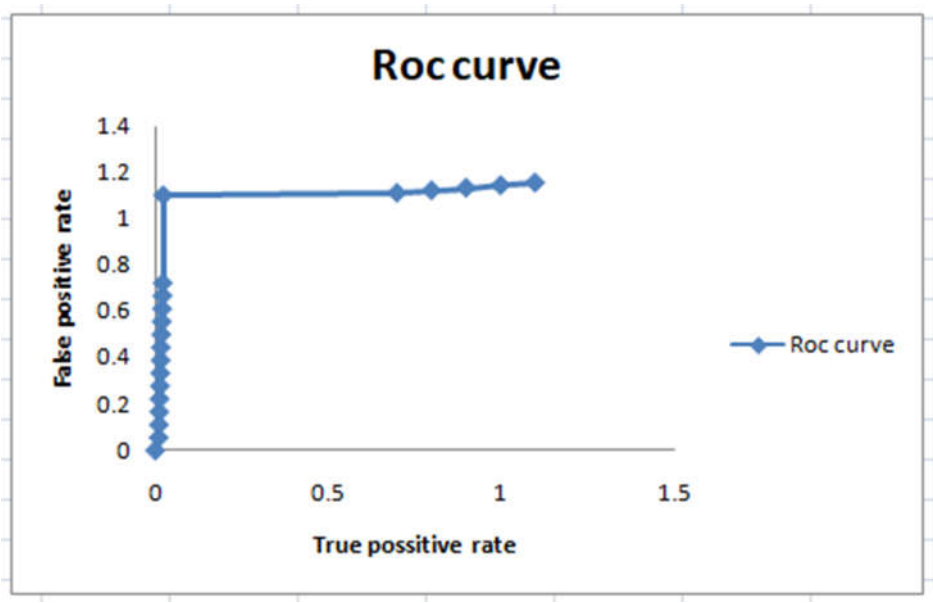


Figure: ROC curve for benign and malignant using stacking ensemble technique.

ROC curve in Figure for benign and malignant using stacking classifier are very close to the Y-Axis which is false positive rate shows stacking technique has the highest accuracy rate.

we can say that combination of naïve Bayes, neural network and decision tree gives the highest accuracy. This shows we can make accurate prediction by classifying the tumors whether it is cancerous or Non-cancerous by using stacking of these three algorithms. Hence we can conclude that by using stacking ensemble approach of these three classification algorithms can be optimum combination to classify Cancerous or Non-Cancerous which are benign or malignant tumor.

Conclusion:

Several data mining techniques are used for the classification of benign and malignant tumor. In this paper the best of three supervised learning classification algorithms are used for prediction of breast cancer and compared on different parameters. here mainly concentrated on which classifier has the better accuracy for prediction and combined two or more algorithms for the highest accuracy using one of ensemble approach. Instead of using one classification power we are using combinational power of rest of algorithms. The model we induced by combining three multiple class will be more reliable and it is sophisticated. From the

malignant. Above all 235 instances correctly classified as malignant and 6 instances are incorrectly classified as benign. For all the experimental results based on accuracy and other parameters we can conclude that combining all three algorithms using stacking approach is the best approach for predicting breast cancer.

REFERENCES:

[1] V. Karthikeyani, I. Parvin, K. Tajudin, and I. Shahina Begam, "Comparative of Data Mining Classification Algorithm in Diabetes Disease Prediction". International journal of computer application 2012.12.26-31

[2] F. Mutaher and Ba-Alwi, "Comparative Study for Analysis the Prognostic in Hepatitis Data: Data Mining Approach", International Journal of Scientific and Engineering Research, Vol 4, Issue 8, August 2013 680ISSN 2229-5518.

[3] C. Shah and A. Jivani, "Comparison of Data Mining Classification Algorithms for Breast Cancer Prediction".4th ICCCNT 2013, july 4-6, India

[4] http://www.cs.waikato.ac.nz/ml/weka/

2017 IEEE International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials114

[5] B. Padmapriya, T. Velmurugan, "A Survey on BreastCancer Analysis Using Data Mining Techniques",Computational Intelligence and Computing Research (ICCIC), 2014 IEEE International Conference on.IEEE, 2014.

[6] S. Gosh, S. Mondal, B. Ghosh, "A Comparative Studyof Breast Cancer Detection Based on SVM And MLP BPN Classifier", Automation, Control, Energy and Systems (ACES), 2014 First International Conference on. IEEE, 2014.

[7] N. Rathore, D. Tomar, S. Agarwal, "Predicting the Survivability Of Breast Cancer Patients Using Ensemble Approach." Issues and Challenges in Intelligent Computing Techniques (ICICT), 2014 International Conference on. IEEE, 2014.

[8] J. Han, M. Kamber, "Data Mining Concepts and Techniques", third edition, Morgan Kaufmann Publishers an imprint of Elsevier.

[9] https://www.analyticsvidhya.com/blog/2018/06/comprehensive-guide-for-ensemble-models/

[10] W. H. Wolberg and O. L. Mangasarian, "Multisurface Method of Pattern Separation for Medical Diagnosis Applied To Breast Cytology", Proc. of the National Academy of Sciences, U.S.A., Volume 87, December1990, pp 9193-9196.

[11] M. Naib, A. Chhabra, "Ensemble Vote Approach For Predicting Primary Tumours Using Data Mining", Confluence The Next Generation Information Technology Summit

[12] (Confluence), 2014 5th International Conference. IEEE, 2014

[13] http://www.dataschool.io/simple-guide-to-confusion- matrix-terminology

[14] C Vikas, Saurabh Pal, "Data mining techniques: To predict and resolve breast cancer survivability", International Journal of Computer Science and Mobile Computing 3.1 (2014): 10-22.

15]http://www2.islab.ntua.gr/attachments/article/86/Ensemble%20method%20%20Zhou.pdf