Breast cancer prediction using ensemble data mining methods

[1]R.PRASANNA KUMARI , [2]Dr.A. M. SOWJANYA
[1]M.TECH , [2]ASSISTANT PROFESSOR
Department of Computer Science & Systems Engineering,
Andhra University College of Engineering (A), Visakhapatnam, India.

***Abstract*:** Machine learning methods are an effective way to predict and analyze cancer data. Techniques like classification and clustering are used for this purpose. We have analyzed wisconsic diagnostic breast cancer dataset and developed a model to predict weather the cancer is malignant or begin. Preprocessing, Outlier Analysis, and Dimensionality reduction have been streamed upon before using classification. The results indicate that ensemble classifiers give high accuracy than compare to other classifiers.

I. INTRODUCTION

Machine learning has become an area of a great interest for clinical research and practice.This is due to the fact that it enables evident base medicine so has to apply the evidence gained from a scientific study of patients. Classification, is the supervised learning in datamining, Classication aims to classify unknown situation based on learning executing patterns and categories from the dataset to subsequently predict feature situations. In this paper we have compare different classifiers models like Naïve Byes classifier, Logistic Regression, K-nearest Neighbors classifier (KNN), Support vector classifiers(SVC), Decision tree, Ensemble classifiers like Random forest Classifier, and Ada boost.

Naïve Bayes : Bayesian classifiers are statistical classifiers.They can predict class membership probabilities such as the probability that a given tuple belongs to a particular class.Naive bayes classifier assume the effect of an attribute value on a given class is independent of the values and other attributes.

K-nearest Neighbors classifier (knn): K-nearest Neighbors classifier are based on learning by analogy, knn is an approach to data classification that estimates how likely a data point is to be a member of one group or the other depending on a group of data.

Logistic Regression:,Logistic Regression is a statistical analysis also known as logit model is often used for predictive analytics and modeling, and extends to applications in machine learning.

Support vector classifiers(svc): Support vector classifiers(svc) is a supervised machine learning algorithm which can be used for both classification or regression.The Support vector classifiers algorithm was proposed based on the advances of the statistical learning theory and algorithmically based on principally on optimization techniques.

Ada Boost: AdaBoost is a Ensemble method in machine learning.Ada Boost is an ensemble technique that attempts to create a strong classifier from a number of weak classifiers.

Random forest Classifier: Random forests is a supervised learning algorithm. It can be used both for classification and regression. It is also the most flexible and easy to use Random forests creates decision trees on randomly selected data samples, gets prediction from each tree and selects the best solution by means of voting.

II. RELATED WORK

[1] Sobin, L. H., Gospodarowicz, M. K., and Wittekind, C, Mainly focuse on the Classification of Malignant Tumors, 2011

[2] Delen, D., Walker, G. and Kadam, A, Mainly focus on "Predicting Breast Cancer Survivability: A Comparison of Three Data Minin Methods," 2005

[3] Choi, J. P., Han, T. H. and Park, R. W, Mainly focus on "A Hybrid Bayesian Network Model for Predicting Breast Cancer Prognosis," Journal of Korean Society of Medical Informatics, 2009.

[4] Bellaachia, A. and Guven E, Mainly focus on the "Predicting Breast Cancer Survivability Using Data Mining Techniques," 2006.

[5] Maimon, O., and Rokach, L, Mainly focus on the "Data Mining and Knowledge Discovery Handbook, Springer", New York,2005.

III. METHODOLOGY

Real world data is assemble to noisy,missing and inconsistent data,due to its huge size and also its origin, from different sources as such data needs to be preprocessed inorder to improve the quality of the data.Health care information consisting of outcomes of diseases,Patient information,treatment analysis,needs to be preprocessed so has to improved in the quality of the prediction results.The major steps involves in data perprocessing are:

1. **Data Cleaning:** Data cleaning is used to clean the data by handling missing values and identifying outliers.

2. **Missing Values:** Missing data Wisconsin Diagnostic Breast Cancer data set has 569 records with 32 attributes since,The data set has no missing values this step is not required.

3. **Outlier Analysis:** Outlier analysis is the process of detecting and removing outlier from the given data set for this purpose we have considered the following techniques has different outlier techniques have identified different number of outliers. we have considered in the consequences of all the outlier detection techniques and removed them from the dataset as such 14 records where identified has outliers and removed.

3.1 Angle-Based Outlier Detection (ABOD): It considers the relationship between each point and its neighbor(s). It does not consider the relationships among these neighbors. The variance of its weighted cosine scores to all neighbors could be viewed as the outlying score. ABOD performs well on multi-dimensional data. Pyod provides two different versions of ABOD

1. Fast ABOD: Uses k-nearest neighbors to approximate.
2. Original ABOD: Considers all training points with high-time complexity.

3.2 Cluster-based Local Outlier Factor (CBLOF): It classifies the data into small clusters and large clusters. The anomaly score is then calculated based on the size of the cluster the point belongs to, as well as the distance to the nearest large cluster.

3.3 Feature Bagging Feature Bagging: A feature bagging detector fits a number of base detectors on various sub-samples of the dataset. It uses averaging or other combination methods to improve the prediction accuracy. By default, Local Outlier Factor (LOF) is used as the base estimator. However, any estimator could be used as the base estimator, such as Knn and Angle-Based Outlier Detection. Feature bagging first constructs n sub-samples by randomly selecting a subset of features. This brings out the diversity of base estimators. Finally, the prediction score is generated by averaging or taking the maximum of all base detectors.

3.4 Histogram-base Outlier Detection (HBOS): It is an efficient unsupervised method which assumes the feature independence and calculates the outlier score by building histograms.It is much faster than multivariate approaches, but at the cost of less precision.

3.5 Isolation Forest: It uses the scikit-learn library internally. In this method, data partitioning is done using a set of trees. Isolation Forest provides an anomaly score looking at how isolated the point is in the structure. The anomaly score is then used to identify outliers from normal observations.Isolation Forest performs well on multi-dimensional data.

3.6 K Nearest Neighbors (KNN):

For any data point, the distance to its kth nearest neighbor could be viewed as the outlying score PyOD supports three Knn detectors:

Largest : Uses the distance of the kth neighbor as the outlier score.
Mean : Uses the average of all k neighbors as the outlier score.
Median : Uses the median of the distance to k neighbors as the outlier score.

IV.DIMENSIONALITY REDUCTION

Dimensionality reduction is the process of reducing the number of random variables or attributes. The most famous dimensionality reduction approach is principal component Analysis. Principal Component Analysis is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables. Principal Component Analysis has linear projections to capture the underlying variance of the data. PCA can be viewed as a special scoring method.

V.CLASSIFICATION

A classification model was constructed based on the Wisconsin Diagnostic Breast Cancer dataset.This dataset was spilt into training data and test data with 10fold cross validation.The training data set was used to build the prediction model like Naïve Byes classifier, Logistic Regression, K-nearest Neighbors classifier (KNN), Support vector classifiers(SVC), Decision tree, Ensemble classifiers like Random forest Classifier, and Ada boost.From the above classifiers the classifer with the highest accuracy was applied on the test data to predict the outcome.

VI.RESULTS AND DISCUSSION

In order to compare the performance of the classifier models used in this paper various matrix like prediction, recall, accuracy have been calculated.

S.NO	NaïveBayes Classifier	K-nearest Neighbors classifier	Support vector Classifier	Logistic regression
Precision	0.97	0.98	0.98	0.98
Recall	O.88	0.91	0.92	0.94
F -value	0.93	0.94	0.95	0.96
Accuracy	0.94	0.95804	0.96503	0.97202

Classifier	Accuracy
Decision tree	0.9649
AdaBoost	0.9718
RandomForest	0.9860

VII.CONCLUSIONS :

Machine learning classifiers help in prediction of diseases,in the paper we have analyzed the Wisconsin Diagnostic Breast Cancer dataset.Thorough preprocessing has been done mainly focusing on different outlier techniques, Dimensional Reduction was performs using Principal Component Analysis and different classifiers where used to predict whether the Cancer is malignant or benign from our results it can been seen that ensemble classifiers perform a vary good job compared to other classifiers.

VIII. REFERENCES:

[1] Sobin, L. H., Gospodarowicz, M. K., and Wittekind, C., TNM Classification of Malignant Tumors, John Wiley & Sons, 2011

[2] Delen, D., Walker, G. and Kadam, A., 2005, "Predicting Breast Cancer Survivability: A Comparison of Three Data Mining Methods," Artificial Intelligence in Medicine, 34(2), 113-127.

[3] Choi, J. P., Han, T. H. and Park, R. W, 2009, "A Hybrid Bayesian Network Model for Predicting Breast Cancer Prognosis," Journal of Korean Society of Medical Informatics, 15(1), 49-57.

[4] Bellaachia, A. and Guven E., 2006, "Predicting Breast Cancer Survivability Using Data Mining Techniques," Age, 58(13), 10-110

[5] Maimon, O., and Rokach, L, Mainly focus on the "Data Mining and Knowledge Discovery Handbook, Springer", New York,2005.

[6]. Fallahi, A. and Jafari S., 2011, "An Expert System for Detection of Breast Cancer Using Data Preprocessing and Bayesian Network," International Journal of Advanced Science and Technology, 34, 65-70

[7]. Pena-Reyes, C. A. and Sipper M., 1999, "A Fuzzy-genetic Approach to Breast Cancer Diagnosis," Artificial Intelligence in Medicine, 17(2), 131-155