Breast Cancer Detection in Early Stage using Machine Learning

K. Arputha Ajitha Rose[1] A. Haseena Beevi[2]
[1]PG Scholar [2]Assistant Professor
[1,2]Department of Computer Science & Engineering
[1,2]Pet Engineering College, Vallioor, India

Abstract— Breast cancer represents one of the diseases that make a high number of deaths every year. It is the most common type of all cancers and the main cause of women's deaths worldwide. Classification and data mining methods are an effective way to classify data. Especially in medical field, where those methods are widely used in diagnosis and analysis to make decisions. In this paper, Random Forest Algorithm on the Wisconsin Breast Cancer (original) datasets is conducted. The main objective is to assess the correctness in classifying data with respect to efficiency and effectiveness of the algorithm in terms of accuracy, precision, sensitivity and specificity. Experimental results show that Random Forest Algorithm gives the highest accuracy with lowest error rate.

Key words: Breast Cancer Detection, Machine Learning

I. INTRODUCTION

The second major cause of women's death is breast cancer (after lung cancer) 246,660 of women's new cases of invasive breast cancer are expected to be diagnosed in the US during 2016 and 40,450 of women's death is estimated. Breast cancer represents about 12% of all new cancer cases and 25% of all cancers in women. Data mining approaches, for instance, applied to medical science topics rise rapidly due to their high performance in predicting outcomes, reducing costs of medicine, promoting patients' health, improving healthcare value and quality and in making real time decision to save people's lives. There are many algorithms for classification and prediction of breast cancer outcomes. Random Forest Algorithm which is most influential data mining algorithms in the research community and among the top 10 data mining algorithms. Our aim is to evaluate efficiency and effectiveness of this algorithms in terms of accuracy, sensitivity, specificity and precision.

II. EXISTING SYSTEM

Support Vector Machine (SVM) is a supervised machine learning algorithm which can be used for classification. In this algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiate the two classes very well. SVMs map the input vector into a feature space of higher dimensionality and identify the hyperplane that separates the data points into two classes. The marginal distance between the decision hyperplane and the instances that are closest to boundary is maximized. The resulting classifier achieves considerable generalizability and can therefore be used for the reliable classification of new samples. It is worth noting that probabilistic outputs can also be obtained for SVMs. SVM might work in order to classify tumors among benign and malignant based on their size and patients' age. The identified hyperplane can be thought as a decision boundary between the two clusters

III. PROPOSED SYSTEM

Random forest algorithm is a supervised classification algorithm. As the name suggest, this algorithm creates the forest with a number of trees. In general, the more trees in the forest the more robust the forest looks like. In the same way in the random forest classifier, the higher the number of trees in the forest gives the high accuracy results. In a normal decision tree, one decision tree is built and in a random forest algorithm number of decision trees are built during the process. A vote from each of the decision trees is considered in deciding the final class of a case or an object, this is called ensemble process. This is a democratic process. Since many decision trees are built and used in a process of Random Forest algorithm, it is called a forest. Combining predictions from multiple models in ensembles works better if the predictions from the sub-models are uncorrelated or at best weakly correlated. Random forest changes the algorithm for the way that the sub-trees are learned so that the resulting predictions from all of the subtrees have less correlation.

IV. SYSTEM DESCRIPTION

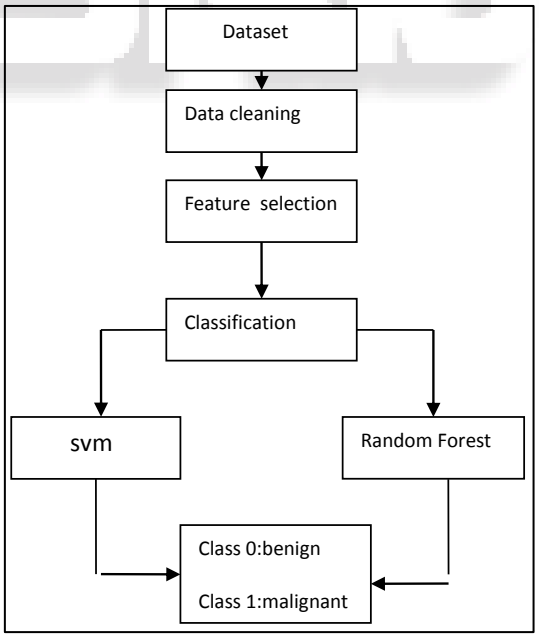

Fig. 4.1: Block Diagram

Figure 4.1 describes, at data cleaning stage data is imputated by using median calculation. A well accepted method is N-Fold cross validation, in which you randomize the dataset and create N (almost) equal size partitions. Then choose Nth partition for testing and N-1 partitions for training the classifier. Within the training set you can further employ another K-fold cross validation to create a validation set and find the best parameters. And repeat this process N times to

get an average of the metric. At next stage specified attributes are extracted. Using IG, we want to determine which attribute in a given set of training feature vectors is most useful. Finally classified into benign and malignant by random forest algorithm.

V. MODULES

A. Data Imputation

Missing Values issue must be tackled by attributing values by utilizing effective imputation method. Mean consists of replacing the missing data for a given variable by the mean of all known values of that variable. Mean imputation technique is a standout amongst the most commonly used strategies. Mean substitution replaces missing values on a variable with the mean estimation of the observed values.

B. Splitup Dataset

Train the classifier using 'training set', tune the parameters using 'validation set' and then test the performance of your classifier on unseen 'test set'. A well accepted method is N-Fold cross validation, in which you randomize the dataset and create N (almost) equal size partitions. Then choose Nth partition for testing and N-1 partitions for training the classifier. Within the training set you can further employ another K-fold cross validation to create a validation set and find the best parameters. And repeat this process N times to get an average of the metric. Since we want to get rid of classifier 'bias' we repeat this above process M times (by randomizing data and splitting into N fold) and take average of the metric. Cross-validation is almost unbiased, but it can also be misused if training and validation set comes from different populations and knowledge from training set is used in the test set. In k cross fold validation method

1) Randomly split your entire dataset into k folds.
2) For each k folds in your dataset, build your model on k – 1 folds of the data set. Then, test the model to check the effectiveness for kth fold.
3) Record the error you see on each of the predictions.
4) Repeat this until each of the k folds has served as the test set.
5) The average of your k recorded errors is called the cross-validation error and will serve as your performance metric for the model.

C. Attribute Selection

Using a decision algorithm, we start at the tree root and split the data on the feature that results in the largest information gain. We repeat this splitting procedure at each child node down to the empty leaves. Gini Index is a metric to measure how often a randomly chosen element would be incorrectly identified. It means an attribute with lower gini index should be preferred.

We want to determine which attribute in a given set of training feature vectors is most useful. In other words, IG tells us how important a given attribute of the feature vectors is. We will use it to decide the ordering of attributes in the nodes of a decision tree.

D. Random Forest Implementation

Many decision trees are built and used in a process of Random Forest algorithm, it is called a forest. for some number of trees *T*:

1) Sample *N* cases at random with replacement to create a subset of the data (see top layer of figure above). The subset should be about 66% of the total set.
2) At each node:

a) For some number m (see below), m predictor variables are selected at random from all the predictor variables.
b) The predictor variable that provides the best split, according to some objective function, is used to do a binary split on that node.
c) At the next node, choose another m variables at random from all predictor variables and do the same.

When a new input is entered into the system, it is run down all of the trees. The result may either be an average or weighted average of all of the terminal nodes that are reached, or, in the case of categorical variables, a voting majority.

Key advantages of using Random Forest

– Reduce chances of over-fitting
– Higher model performance or accuracy

VI. OUTPUT

A. Data Imputation

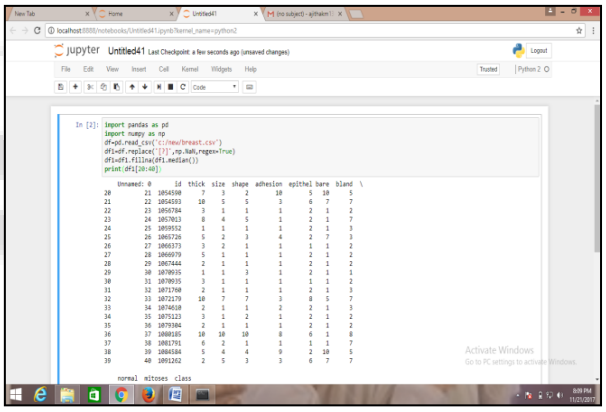

Fig. 6.1: Data Imputation

Data is cleaned by median value calculation and replace NaN value by median value. Then only we can pass it into algorithm. Mean consists of replacing the missing data for a given variable by the mean of all known values of that variable

B. Attribute Selection

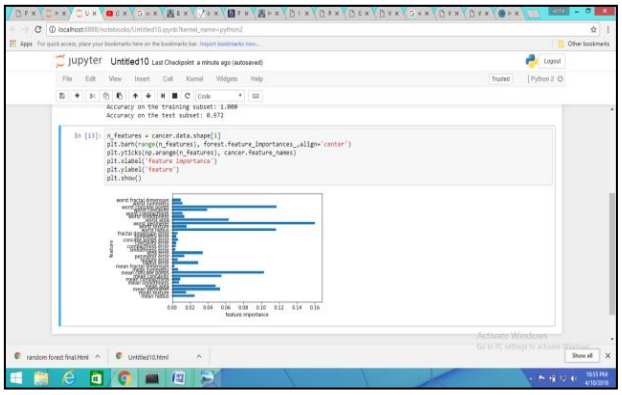

Fig. 6.2: Attribute Selection

Attributes selected based on information gain and gini index.

C. Random Forest Implementation

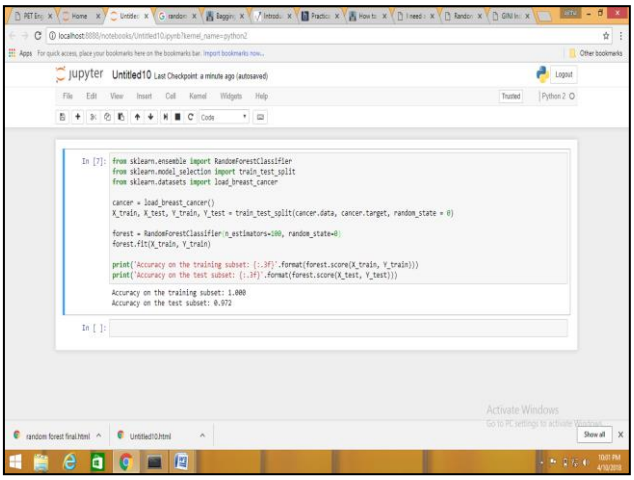

Fig. 6.3: Random Forest implementation

RANDOM FOREST algorithm is implemented in python language and the accuracy is achieved 97.2%.this was better one from SVM and decision tree

VII. CONCLUSION

To analyze medical data, various data mining and machine learning methods are available. An important challenge in data mining and machine learning areas is to build accurate and computationally efficient classifiers for Medical applications. In this study, we employed three main algorithms: SVM and Random Forest on the Wisconsin Breast Cancer (original) datasets. We tried to compare efficiency and effectiveness of those algorithms in terms of accuracy, precision, sensitivity and specificity to find the best classification accuracy. Random Forest reaches the high accuracy and outperforms, therefore, all other algorithms. In conclusion, Random Forest has proven its efficiency in Breast Cancer prediction and diagnosis and achieves the best performance in terms of precision and low error rate.

In future, K-Means algorithm will be used. It is a type of unsupervised algorithm which solves the clustering problem. Its procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters). Data points inside a cluster are homogeneous and heterogeneous to peer groups.

REFERENCES

[1] U.S. Cancer Statistics Working Group. United States Cancer Statistics: 1999–2008 Incidence and Mortality Web-based Report. Atlanta (GA): Department of Health and Human Services, Centers for Disease Control and Prevention, and National Cancer Institute; 2012.
[2] Siegel RL, Miller KD, Jemal A. Cancer Statistics, 2016. 2016; 00(00):1-24. doi:10.3322/caac.21332.
[3] "Globocan 2012 - Home." [Online]. Available: http://globocan.iarc.fr/Default.aspx. [Accessed: 28-Dec-2015].
[4] Asri H, Mousannif H, Al Moatassime H, Noel T. Big data in healthcare: Challenges and opportunities. 2015 Int Conf Cloud Technol Appl. 2015:1-7. doi:10.1109/CloudTech.2015.7337020.
[5] Noble WS. What is a support vector machine? Nat Biotechnol. 2006; 24(12):15651567.doi:10.1038/nbt1206-1565.
[6] Rish I. An empirical study of the naive Bayes classifier. IJCAI Work Empir methods Artif Intell. 2001; 3(November):41-46.
[7] Quinlan JR. C4.5: Programs for Machine Learning. 2014:302. https://books.google.com/books?hl=fr&lr=&id=b3ujBQAAQBAJ&pgis=1. Accessed January 5, 2016.
[8] Larose DT. Discovering Knowledge in Data. Hoboken, NJ, USA: John Wiley & Sons, Inc.; 2004.
[9] X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. Mclachlan, A. Ng, B. Liu, P. S. Yu, Z. Z. Michael, S. David, and J. H. Dan, Top 10 algorithms in data mining. 2008, pp. 1–37.
[10] Datafloq - Top 10 Data Mining Algorithms, Demystified. https://datafloq.com/read/top-10-data-mining-algorithms- demystified/1144. Accessed December 29, 2015.
[11] V. Chaurasia and S. Pal, "Data Mining Techniques: To Predict and Resolve Breast Cancer Survivability," vol. 3, no. 1, pp. 10– 22, 2014.

