Analysis of Breast Cancer Dataset Using Big Data Algorithms for Accuracy of Diseases Prediction

Ankita Sinha(✉), Bhaswati Sahoo, Siddharth Swarup Rautaray, and Manjusha Pandey

KIIT Deemed University, Bhuneshwar, India
ethosankita@gmail.com, {bhaswati.sahoofcs, siddharthfcs,manjushafcs}@kiit.ac.in

Abstract. Data Mining Techniques easily handle and solve the problem of handling the massive amount of data due to heterogeneous data, missing data, inconsistent data. HealthCare is one of the most important applications of Big Data. Diagnosis of diseases like cancer at an early stage is also very crucial. This paper focuses on the prediction model analysis for the breast cancer diagnosis either benign or malignant at an early stage as it increases the chances for successful treatment So predicting breast cancer at benign increases the survival rate of women. Data mining classification algorithm like SVM, Naive Bayes, k-NN, Decision Tree compares a variety of statistical techniques like accuracy, sensitivity, specification, positive prediction value, negative predictive value, area under curve and plotted ROC curve in R analytical tool which is promising independent tool for handling huge datasets is proven better in a prediction of the breast cancer diagnosis.

Keywords: Big data · Cancer · Breast cancer · Data mining classification algorithm · R analytical tool · Prediction

1 Introduction

Breast cancer mostly endows in women usually sustain in women breast's duct or glands. Breast cancer diagnosis is of two types Benign and Malignant. In benign the cancerous cell spread slowly throughout the other body parts and malignant is the dangerous one as it spread very rapidly throughout the other body parts and effects lungs, kidney, brain, skin, etc. So the diagnosis of breast cancer at preliminary increases the chances of overcomes from the cancerous cells. For the analysis and extraction of information from the large massive amount of datasets is quite challenging for analyst and also faces the problem regarding the prediction analysis of diseases because of noisy data, inconsistent data, missing values in large datasets. So the Data Mining techniques with Big Data Analytical tool come into the pictures. Data Mining Classification Algorithm always play a vital role in Prediction analysis, which compares a variety of statistical techniques for different Data Mining Classification

S. Smys et al. (Eds.): ICCNCT 2019, LNDECT 44, pp. 271–277, 2020.
https://doi.org/10.1007/978-3-030-37051-0_31

Algorithm that analyzes the current and historical facts to make predictions about the unknown events (either benign or malignant).

This paper is organized into section as follows. Section 2 summarizes the background followed by a brief discussion of breast cancer-related research work. The experimental setup steps are discussed in Sect. 3. Section 4 is all about the brief introduction of the classification algorithms like naive bayes, decision tree, SVM, k-NN. Problem definition for Predictive analysis of breast cancer followed by dataset collection attributes information, comparative analysis is discussed in Sect. 5. Section 6 is the conclusion, summarizes a brief overview of the proposal followed by and future work.

2 Background

Breast cancer is very different and complicated disease occur in one's breast's ducts and gland with multiple symptoms like pain in breast part or may be a pain in nipples, rosiness or decrease in the density of breast or nipple, irritation in nipples and many more. It can be easily analyzes in blood tests, MRI test, mammogram test or in CT scan. Result gives the details of effective biopsy tissues and that area of breast goes for advanced treatment like surgery, chemotherapy, radiation, hormone therapies. When the breast cancer is diagnosed in benign stage it can be easily cure within 5 years but if it is diagnoses as malignant it is very different to recurred it. Since to focus on the breast cancer related research, some related work on breast cancer is provided. Sakri et al. [1], evaluated the naive bayes, k-NN and fast decision learner in order to increase the accuracy of breast cancer recurrences prediction model by using feature selection i.e. Particle swarm optimization with an objective to reduce the number of attributes and the naive bayes generates the best output among all the other classifier with PSO and without PSO. Alwidian et al. [2], proposed a new prediction technique with new pruning methodology i.e. WCBA for weighting and pruning with an objective to produces more accurate association rule based on a statistical measure in order to enhance the accuracy level of association classifier. Asri et al. [3], compares the Wisconsin breast cancer datasets performances on different classifiers like the random forest, naive bayes, idk on WEKA which gives best and high accuracy results for naive Bayes. Tripathy et al. [4], introduces a new technique i.e. Parallel SVM for risk analysis with an objective to generate best and an efficient way to work on large datasets based on map reduce. Bhardwaj and Tiwari [5], introduces a new genetically optimized neural network algorithm to classify breast tumor benign or malignant with an objective to deal with classification issues and compares the performances like accuracy, sensitivity, specificity, confusion matrix, ROC curve, area under curve. Gupta et al. [6], presented the overview of all the latest research on breast cancer using a data mining algorithm to upgrade the breast cancer detection and prediction.

3 Experimental Setup

The systematic diagram of the proposed prediction analysis model as shown in Fig. 1, sketched after studying number of papers. Initially, breast cancer data are collected from Kaggle and then datasets are subjected to data pre-processing in order to remove noise, inconsistent, outliers and missing values. Then data mining classification algorithms is applied and obtain the output as performances statistical results, we compared the all statistical measures like accuracy, sensitivity, specificity, auc, ROC cure, ppv, npv for different data mining classification algorithms. And in the final step, we will do a prediction based on result analysis.

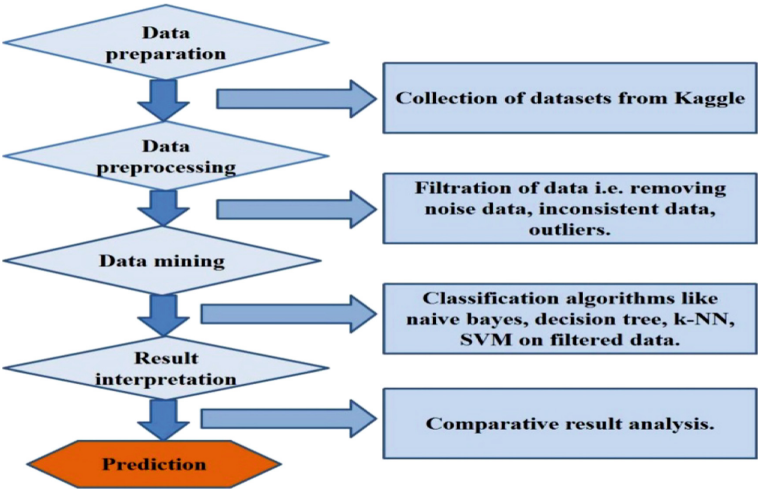

Fig. 1. Schematic diagram for prediction analysis [1–5]

4 Classification Algorithm

In the research area, data mining has appealed a lot of consideration because of the excessive amount of data and for extracting the new information and insight pattern for unknown events. Data mining is also known as Knowledge Discovery Databases (KDD), used for exploring new and appropriate information from a massive amount of datasets [7].

4.1 Decision Tree

It is a very simple classifier represent in the form of a tree diagram that exhibits the range of possible outcomes and after the initial decision, the subsequent decision comes into the picture (Tables 1, 2 and 3).

Table 1. Advantages and disadvantages of decision tree

Advantages	Disadvantages
Easy to understand and generates rules	Training cost is high, over fitting
Reduces problem complexity	Document connected to one branch

Table 2. Advantages and disadvantages of naive bayes

Advantages	Disadvantages
Fast in train and classify the data	Assume independence of features
Handle discrete, streaming data, missing value	Outcome is very biased results

Table 3. Advantages and disadvantages of support vector machine

Advantages	Disadvantages
High accuracy & less over-fitting	Training time is high for large data
Uses small and clean datasets	Less effective for noisy data

4.2 Naive Bayes Classifier

Naive Bayes classifier is the collection of classifier family where all the pair of feature shares the common principle but independent to each other based on Bayes Theorem.

4.3 Support Vector Machine

SVM is used for text classification like assignment, detecting spam and sentiment analysis. It is mainly used for image reorganization as well as in aspect-based classification and also in colour based classification [8–10].

4.4 k-Nearest Neighbour

k-Nearest Neighbour is also known as the lazy learning algorithm which classifies the datasets based on their similarity measures with a neighbour and k stands for the number of datasets items that are considered for classification [9] (Table 4).

Table 4. Advantages and disadvantages of k-NN

Advantage	Disadvantage
Non-parametric & best performing text classifier	Difficult for similarity measure
Handle large amount of predictors	Computationally expensive

5 Problem Definition of Predictive Analysis of Breast Cancer

5.1 Data Source

To classify all the classification algorithm, we have used Kaggle Wisconsin Breast Cancer datasets. The datasets consists of 31 attributes and one class attribute i.e. diagnosis with 699 instances. Figure 2 presents the attribute specification of datasets of breast cancer.

Sl.no	Attribute	Description
1	id	ID number
2	radius_mean	mean of distances from center to points on the perimeter
3	texture_mean	standard deviation of gray-scale values
4	perimeter_mean	mean size of the core tumor
5	area_mean	mean area inside the boundary of core tumor
6	smoothness_mean	mean of local variation in radius lengths
7	compactness_mean	mean of perimeter^2 / area - 1.0
8	concavity_mean	mean of severity of concave portions of the contour
9	concave points_mean	mean for number of concave portions of the contour
10	symmetry_mean	mean of similar area of tumor parts that matches
11	fractal_dimension_mean	mean for "coastline approximation" – 1
12	radius_se	standard error for the mean of distances from center to points on the perimeter
13	texture_se	standard error for standard deviation of gray-scale values
14	perimeter_se	standard error for mean size of the core tumor
15	area_se	standard error for mean area inside the boundary of core tumor
16	smoothness_se	standard error for local variation in radius lengths
16	smoothness_se	standard error for local variation in radius lengths
17	compactness_se	standard error for perimeter^2 / area - 1.0
18	concavity_se	standard error for severity of concave portions of the contour
19	concave points_se	standard error for number of concave portions of the contour
20	symmetry_se	standard error for mean of similar area of tumor parts that matches
21	fractal_dimension_se	standard error for "coastline approximation" – 1
22	radius_worst	"worst" or largest mean value for mean of distances from center to points on the perimeter
23	texture_worst	"worst" or largest mean value for standard deviation of gray-scale values
24	perimeter_worst	"worst" or largest mean value for mean size of the core tumor
25	area_worst	"worst" or largest mean value for mean area inside the boundary of core tumor
26	smoothness_worst	"worst" or largest mean value for local variation in radius lengths
27	compactness_worst	"worst" or largest mean value for perimeter^2 / area - 1.0
28	concavity_worst	"worst" or largest mean value for severity of concave portions of the contour
29	concave points_worst	"worst" or largest mean value for number of concave portions of the contour
30	symmetry_worst	"worst" or largest mean value for similar area of tumor parts that matches
31	fractal_dimension_worst	"worst" or largest mean value for "coastline approximation" – 1
32	diagnosis	The diagnosis of breast tissues (M = malignant, B = benign)

Fig. 2. Detail description of dataset [11]

5.2 Comparative Analysis

A Confusion Matrix helps in finding the comparison between classifier by computing Accuracy, Sensitivity, Specificity, Area under curve and ROC curve [12]. Confusion Matrix table for breast cancer diagnosis is shown in Table 5.

Table 5. Confusion matrix for breast cancer

Diagnosis	Classified as benign	Classified as malignant
Benign	TP	FN
Malignant	FP	TN

Table 6 gives the performance result of classifier algorithms. According to performance table the SVM gives the highest accuracy i.e. 97% as compared to other classification algorithms.

Table 6. Performance table of classifier algorithms

Classifier	Acc (%)	Sen (%)	Spec (%)	PPV (%)	NPV (%)	AUC
Decision tree	91.15	93.85	87.50	91.04	91.30	0.9516
Naive bayes	95.33	97.49	66.51	83.06	94.00	0.9649
SVM	97.35	61.07	41.99	63.93	39.04	0.9888
k-NN (Roc)	85.94	97.48	66.51	83.05	94.00	0.8572
k-NN (Acc)	74.99	81.23	22.16	63.73	41.22	0.8147

True Positive: Probability of (+) given the individual has the Benign stage.
False Negative: Probability of a Benign stage tests negative (-).
True Negative: Probability of (-) given the individual does not have the benign stage, have Malignant stage;
False Positive: P(+|M) = Probability of a Malignant stage tests positive (+).

Figure 3 represents the ROC curve for the different classification classifiers and SVM classification techniques is the superior algorithm as compared to other classifier. SVM gives higher accuracy and higher area under curve.

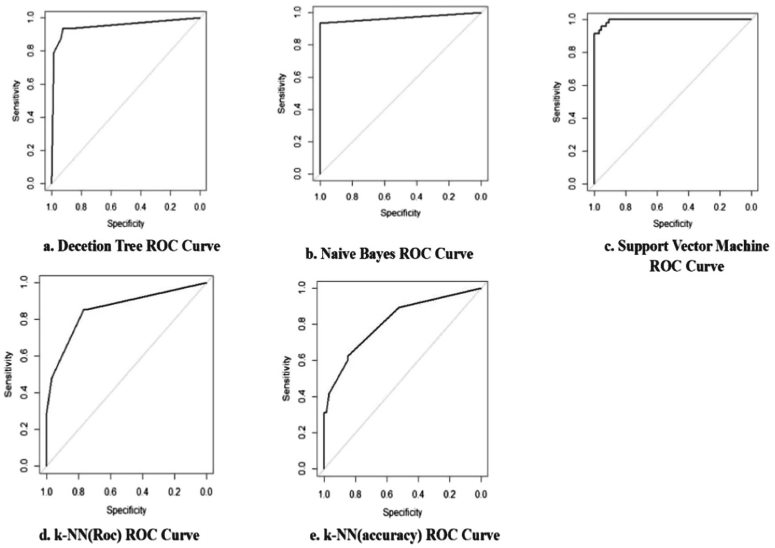

Fig. 3. Roc curve of different classifier

6 Conclusion and Future Work

Data mining classification algorithm enhances the work of predictive analysis we have presented the results of our experiments on popular classifying algorithms, NB, DT, SVM and k-NN and SVM generates better output in both field accuracy and ROC curve as it produces highest area under curve as compared to other classification techniques. In future work we will try to implement model, in association rule mining techniques on Breast Cancer Wisconsin-datasets.

References

1. Sakri, S.B., Rashid, N.B.A., Zain, Z.M.: Particle swarm optimization feature selection for breast cancer recurrence prediction. IEEE Access **6**, 29 (2018)
2. Alwidian, J., Hammo, B.H., Obeid, N.: WCBA: weighted classification based on association rules algorithm for breast cancer disease. Appl. Soft Comput. **62**, 536–549 (2018)
3. Asri, H., Mousannif, H., Al Moatassime, H., Noel, T.: Using machine learning algorithms for breast cancer risk prediction and diagnosis. Procedia Comput. Sci. **83**, 1064–1069 (2016)
4. Tripathy, P., Rautaray, S.S., Pandey, M.: Parallel support vector machine used in map-reduce for risk analysis. In: 2017 Second International Conference on Electrical, Computer and Communication Technologies (ICECCT), pp. 1–4. IEEE (2017)
5. Bhardwaj, A., Tiwari, A.: Breast cancer diagnosis using genetically optimized neural network model. Expert Syst. Appl. **42**(10), 4611–4620 (2015)
6. Gupta, S., Kumar, D., Sharma, A.: Data mining classification techniques applied for breast cancer diagnosis and prognosis. Indian J. Comput. Sci. Eng. (IJCSE) **2**(2), 188–195 (2011)
7. Witten, I.H., Frank, E.: Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, San Francisco (2005)
8. Zaki, M.J., Meira Jr., W.: Data mining and analysis: fundamental concepts and algorithms. Cambridge University Press, New York (2014)
9. Jonsdottir, T., Hvannberg, E.T., Sigurdsson, H., Sigurdsson, S.: The feasibility of constructing a Predictive outcome model for breast cancer using the tools of data mining. Expert Syst. Appl. **34**(1), 108–118 (2008)
10. Han, J., Pei, J., Kamber, M.: Data mining: concepts and techniques. Elsevier, New York (2011)
11. Database. http://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(diagnostic)
12. Shah, C., Jivani, A.G.: Comparison of data mining classification algorithms for breast cancer prediction. In: 2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT), pp. 1–4. IEEE (2013)