A Comparative Analysis of Breast Cancer Data Set Using Different Classification Methods

M. Navya Sri, J. S. V. S. Hari Priyanka, D. Sailaja and M. Ramakrishna Murthy

Abstract Patterns or models in data can be found using data mining algorithms. This is a knowledge discovery process in which data mining is involved. It is a scientific method which is intended to examine massive data, so as to find out the systematic relationships and consistent patterns among variables and further check for the accuracy of the findings. This can be done by taking new subsets of data and applying the detected patterns to them. The core part of the data mining techniques is classification. In classification, in order to develop a model which will categorize the population of records, we make use of a set of pre-classified examples. The techniques of classification use the model which is built on basis of training data and apply it to test data. "Breast cancer Wisconsin data set is used as a training set." There is an open source data mining tool named WEKA, which consists of implementation of data mining algorithms. By making use of WEKA we have compared the well-known classification algorithms that are decision tree and Bayesian algorithms. It is concluded that decision tree classification algorithm got high accuracy compared to Bayesian classification algorithm.

M. N. Sri (✉) · J. S. V. S. H. Priyanka · D. Sailaja · M. Ramakrishna Murthy
Department of Information Technology, Anil Neerukonda Institute of Technology and Sciences, Sangivalasa, Bheemunipatnam, Visakhapatnam
Andhra Pradesh, India
e-mail: navyasrimullapudi@gmail.com

J. S. V. S. H. Priyanka
e-mail: priyapatnaik.hari@gmail.com

D. Sailaja
e-mail: dadla.sailaja25@gmail.com

M. Ramakrishna Murthy
e-mail: ramakrishna.malla@gmail.com

S. C. Satapathy et al. (eds.), *Smart Intelligent Computing and Applications*,
Smart Innovation, Systems and Technologies 104,
https://doi.org/10.1007/978-981-13-1921-1_17

1 Introduction

In this new era, we have data everywhere. Data are the facts which are collected for analysis. In the past twenty years, a massive enlarges in the amount of data being accumulated. The user aims to have more sophisticated data. This rises up and about the demand of data mining. Data mining refers to dig out or "mining" knowledge from huge amount of data [1]. The purpose of data mining is to concentrate on mining of information from bulk collection of data and make over it into an effortless understandable formation for further purpose. The various methods of data mining are clustering, classification, association rules, prediction. In the course of classification algorithms, the well-known classification algorithms are decision tree classification and Bayesian classification. Data mining has a wide range of applications in the real life; those are customer segmentation, financial banking, criminal investigation, research analysis, bioinformatics, fraud detection, intrusion detection, and customer relationship management [2]. Improvement of health systems is one of the significant applications of data mining. It uses data and analytics to recognize superlative practices that reduce expenses and develop care. In urbanized nations, breast cancer has turned into the crucial reason of death in females. Breast cancer is a heterogeneous disease to one in every eight women which affects during their lives. It forms in the cells and develops from breast tissues with "high degree of diversity between and within tumor's as well as among cancer-bearing individuals" being a factor together to identify the risk of disease progression and therapeutic resistance [3]. Accurate prediction of breast cancer cells at earliest possible could help in decreasing the cost of health and enhances the time that is required for a patient to take the required treatment as well. Therefore, this paper presents an advanced approach which allows analyzing and classifying tumors at an unprecedented depth. A lot of research has been made for becoming aware of survivability of cancer. In this paper, several research mechanisms carried out using Bayesian classification and decision tree classification. These algorithms play an important role to identify the patient functioning.

The enduring of this manuscript is ordered as follows. Section 2 is regarding the survey of the authors experimented with various data mining techniques that are applied to the examination of breast cancer. The classification algorithms are used for the analysis of breast cancer and the methodology are explained in Sect. 3. Section 4 shows the experimental results of comparative study between decision tree and Bayesian classification. The entire survey work is concluded in Sect. 5.

The remaining portions of this paper are structured as follows: Section 2 specifies the literature survey. Section 3 specifies different classification algorithm. Section 4 provides the outcomes of this experiment. Section 5 concluded and mentions the future work of this paper. Finally, lists the references used in this paper.

2 Literature Survey

According to B. Padmapriya [4], a comparative analysis has been made between ID3 and C4.5 classification algorithms to detect breast cancer in female and proved that C4.5 classification algorithm gives efficient results to diagnose breast cancer compare to ID3 classification algorithm. They used SEER data set for breast cancer analysis and discussed the various applications of breast cancer applications. Rostemmennor [5] discovered a drug to use machine learning algorithms for big data analytics on top of map reduce for the breast cancer protein receptor. In this, they have concentrated on protein of breast cancer, so they have selected 4JLU receptor which is a critical structure. Chandresharya [6] discussed different expert systems designed for last twenty years. Especially, these systems are used to detect breast cancer tumor in human. They experimented on mammogram images to find out the abnormality in patient. In this, Wisconsin breast cancer data set is used and experimented with different classification approaches; those are support vector machine algorithm, neural network algorithm. Xinagchumet all [7] analyzed various statistical procedures to detect breast cancer tumor in women. In this research work also mammogram images are applied and got the precision from 68 to 79%. In this, a statistical technique that is principal component analysis is used to identify abnormalities in patient.

As per the above, a lot of research work is accomplished on the detection of breast cancer tumor in women. This paper provides the best algorithm and its accuracy to find out the abnormality tumor in women. In this, the comparative analysis has been made between the well-known classification algorithms which are the decision tree and Bayesian algorithms.

3 Classification Algorithms

A progression of building a class model from a data set which encloses class labels is described as classification. In this, the well-known UCI machine learning repository that is breast cancer Wisconsin (Diagnostic) data set is used for experimentation. This data set features are extracted from digitized representation of a fine needle aspirate (FNA) of a breast dimension. These extracted features are applied as input for both decision tree and Bayesian classification algorithms.

3.1 Decision Tree Classification (J48)

A researcher named J. Ross Quinlan in 1980 build up an algorithm known as decision tree classification also known as ID3 (Iterative Dichotomiser 3) [8]. An extension of ID3 algorithm is j48 implemented by the WEKA project team. This

algorithm builds the classification model in tree structure form. The data set is divided into subsets up to leaf nodes. The leaf nodes of a tree are the decisions or class labels. This algorithm is implemented as j48 in WEKA tool. The following steps are fundamental steps of this algorithm which are (i) the instances which belong to same class of the tree labeled with the same class name. (ii) For every attribute, the gain (potential information) has to be calculated, and (iii) Calculate entropy which is a measure of the data disorder. In the data sets, a number of instances are not well defined from remaining instances. So to decrease the classification errors the pruning is applied for reducing classification errors [9]. This is one of the reasons that decision tree algorithm gives better accuracy because of pruning technique. After pruning applied, then the data set is given as input for decision tree algorithm, so that the classification errors can be reduced.

3.2 Bayesian Classification

It is a statistical and supervised learning method for training the data sets. It is a probabilistic model to determine the likelihood of outcomes. This algorithm is used to solve diagnostic health diseases [10]. This algorithm is proposed on Bayes theorem. It provides the prior knowledge to assess many challenges. Many algorithms are not capable of classifying noise data or null data exist in the data set, but even though some data sets have noise data, Bayes algorithm is robust to train noise in input instances [11]. To train the text documents in sentiment analysis or customer feedback analysis, Bayesian algorithm is one of the best-known algorithms to classify the text documents. In this paper, we selected Bayes algorithm because of its robust nature.

4 Experimental Results

The experiments are conducted on WEKA tool. The input breast cancer data set is given as input for decision tree and Bayesian classification algorithms. This data set has 32 attributes with the number instances 286. The characteristic of this data set is multivariate.

Methodologies:

1. Correctly classified instances (CCI): Correctly classified instances show the accuracy of decision tree classification algorithm compared to Bayesian.
2. Kappa statistic (KS): It computes to examine the same data by two independent tools.
3. Mean absolute error (MAE): It is a quantity used to measure how close estimates or predictions are to the eventual outcomes.

4. Root mean square error (RMS): It is one of the most widely used statistics in GIS. We use RMSE in a variety of applications when we need to compare two data sets.
5. Relative absolute error: It is calculated as the mean absolute error separated by the error of the ZeroR classifier.
6. Root relative squared error: It is considered in relation to what it would have been if a simple predictor had been used. More absolutely, this simple predictor is just the average of the actual values. Thus, the error takes the total squared error and normalizes it by separated by the total squared error of the simple predictor. By considering the square root of the relative squared error one makes smaller the error to the same dimensions as the quantity being predicted.

4.1 Performance Evaluation of Decision Tree Classification

Correctly classified instances (CCI): 217
Incorrectly classified instances (ICI): 69
Kappa statistic (KS): 0.2899
Mean absolute error (MAE): 0.3658
Root mean squared error (RMS): 0.4269
Relative absolute error (RAE): 87.4491
Root relative squared error (RRS): 93.4017
Total number of instances (TI): 286

4.1.1 Confusion Matrix of Decision Tree Classification

Classified as	A	B
non-recurrence events	194	7
recurrence events	62	23

4.2 Performance Evaluation of Bayesian Classification

Correctly classified instances (CCI): 214
Incorrectly classified instances (ICI): 71
Kappa statistic (KS): 0.3693
Mean absolute error (MAE): 0.3012
Root mean squared error (RMS): 0.4278
Relative absolute error (RAE): 72.0082
Root relative squared error (RRS): 93.6095
Total number of instances (TI):286

4.2.1 Confusion Matrix of Bayesian Classification

Classified as	A	B
non-recurrence events	174	27
recurrence events	44	41

4.3 Comparative Analysis

	CCI	ICI	KS	MAE	RMS	RAE	RRS	TI
J48	75.875%	24.12%	0.2899	0.3658	0.4269	87.4491	93.4017	286
Bayesian	75.17%	24.82%	0.3693	0.3012	0.4278	72.0082	93.6095	286

4.3.1 Detailed Accuracy by Class

	True positive rate	False positive rate	Precision	Recall	F-measure	Roc area
J48	0.759	0.523	0.76	0.759	0.716	0.639
Bayesian	0.752	0.404	0.74	0.752	0.743	0.76

The decision tree classification algorithm identified correctly 217 instances, whereas Bayesian classification identified 214 classified instances. Particularly, mean absolute error is the probability estimates which is high for decision tree compared to Bayesian classification. From the confusion matrix, we can say that one instance of class is recurrence event and other is non-recurrence events. In decision trees confusion matrix, 194 instances belong to class non-recurrence event and in Bayesian confusion matrix, 174 instances belong to non-recurrence event.

5 Conclusion and Future Work

Various techniques are reviewed on breast cancer tumor problems in this study. This scenario is mostly investigated under decision tree and Bayesian classification algorithms and decision tree got high accuracy in assessment to Bayesian. A diverse research has been done on breast cancer diagnosis, and it is notified that source and symptoms associated to each occurrence with the evidence correlated to each patient so that up to some extent it is possible to prevent from the cancer. The executions of both algorithms are investigated on WEKA tool. In future, the expert systems of breast cancer may be useful for accomplishing high classification rate.

References

1. Tomar, D., Agarwal, S.: A survey on data mining approaches for healthcare. Int. J. Bio-Sci. Bio-Technol. **5**(5), 241–266 (2013)
2. Sujatha, G., Usha Rani, K.: A survey on effectiveness of data mining techniques on cancer data sets. Int. J. Eng. Sci. Res. **04**(1), 1298–1304 (2013)
3. Utomo, C.P., Kardiana, A., Yuliwulandari, R.: Breast cancer diagnosis using artificial neural networks with extreme learning techniques. Int. J. Adv. Res. Artif. Intell. **3**(7) (2014). http://dx.doi.org/10.14569/IJARAI.2014.030703
4. Padmapriya, B.: A survey on breast cancer analysis using data mining techniques. IEEE. ISBN No: 978-1-4799-3975-6, 2014
5. Mennour, R.: Drug discovery for breast cancer based on big data analytics techniques. IEEE (2015)
6. Arya, C.: Expert system for breast cancer diagnosis: a survey. ICCCI. ISBN No: 978-1-4673-6680-9 (2016)
7. Xiangchun, K.X.: Analysis of breast cancer using data mining & statistical techniques. IEEE. ISBN No: 0-7695-2294-7 (2005)
8. Sivagami, P.: Supervised learning approach for breast cancer classification. Int. J. Emerg. Trends Technol. Comput. Sci. **1**(4), 115–129 (2012)
9. Rajesh, K., Anand, S.: Analysis of SEER dataset for breast cancer diagnosis using C4.5 classification algorithm. Int. J. Adv. Res. Comput. Commun. Eng. **1**(2), 72–77 (2012)
10. Babagholami-Mohamadabadi, B., Jourabloo, A., Zarghami, A., Kasaei, S.: A bayesian framework for sparse representation-based 3-d human pose estimation. IEEE Signal Process. Lett. **21**(3), 297–300 (2014)
11. Ramakrishna Murty, M., Murthy, J.V.R., Prasad Reddy, P.V.G.D: Text document classification based on a least square support vector machines with singular value decomposition. Int. J. Comput. Appl. (IJCA) (indexed by DOAJ, Informatics, ProQuest CSA research database) **27**(7), 21–26 (2011). ISBN 978-93-80864-56-6, https://doi.org/10.5120/3312-4540