International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2018 IJSRCSEIT | Volume 3 | Issue 5 | ISSN : 2456-3307

A Review on Breast Cancer Detection Using Data Mining Techniques

Ahmad Zubair Zarbi, Prof. Rajvi Parkh
Department of Information & Technology, GCET,Vidanager, Gujarat, India

ABSTRACT

In Today's Data Mining focuses on the discovery of previously unknown properties in the data It does not need a specific goal from the domain but instead focus on finding new and interesting knowledge. Mining provides useful information from the huge volume of the data stored in repositories the present study focus on implementing three different algorithms using the data mining WEKA. The Algorithms in the study include Naïve Byes, J48 Decision Tree and One R. All these well-known familiar Algorithms are used in classification rule mining Techniques. Dataset are collected also, these collected datasets are pre-processed and then used for implementing the Algorithm. The different types of Algorithms are executed using the collected datasets; the results are shown in separate window as graphical.

Keywords: Breast cancer, J48 Decision Tree, Naïve Byes, One R , Artificial Neural Network, Classification ,Weka.

I. INTRODUCTION

Breast Cancer is becoming a leading cause of death among women in the whole World; meanwhile, it is confirmed that the early detection and accurate diagnosis of this disease can ensure a long survival of the patients. This paper work presents a disease status prediction employing a hybrid methodology to forecast the changes and its consequence that is crucial for lethal infections.

Providing clinical predictions for cancer patients by analysing their genetic make-up is a difficult and very important issue. With the goal of identifying genes more correlated with the prognosis of breast cancer, we used data mining techniques to study the gene expression values of breast cancer patients with known clinical outcome. Focus of our work was the creation of a classification model to be used in the clinical practice to support therapy prescription.

Understanding what portions of the genome are involved in the development of cancer cells is a difficult and currently very important issue in medicine. Providing clinical predictions for cancer patients by analysing their genetic make-up is a central goal of many research groups. In this respect, our contribution here illustrated regarded the use of knowledge extraction techniques that are derived from artificial intelligence and globally known as knowledge discovery. It focused on cases of women suffering from breast cancer; in particular, we evaluated the possibility of predicting metastatic recurrence within five years from surgery.

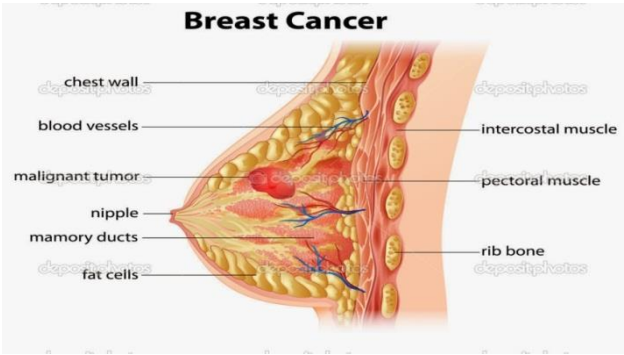

Figure 1. Breast Cancer

Classification is a data mining technique which involves the use of supervised machine learning

techniques which assigns labels or classes to different objects and groups.

It involves the process of model construction (analysis of training data for patterns) and model usage where the constructed model is used for classification. Classification accuracy is usually estimated as the percentage of test samples that are correctly classified.

This study aims at using data mining techniques to classify breast cancer Matjaz Zwitter & Milan Soklic (physicians) Institute of Oncology University Medical Center Ljubljana, Yugoslavia which contains the risk factors and the cancer classes (unlikely, likely and benign). The J48 decision trees and naïve bayes' classification of breast cancer was performed using the WEKA software.

Figure 2. Weka GUI Chooser PROPOSED ALGORITHM

A. BACKGROUND

Weka tools is used here for implementing the data mining techniques/algorithm for mining the knowledge from the huge amount of data.

B. WEKA

The aim of the weka tool is to deliver a complete group of data pre-processing and machine learning algorithms for exclusively for research scholars and the educationalist. This tool will help us to compare the different types of machine learning and data mining techniques. Weka tool is very easy to use, because, the tool was developed as a simple Application Programming Interface using the Java language.

Weka software also support the file formats include WEKA's own ARFF format, CSV, Lib SVM's format, and C4.5's format. All the above-mentioned file formats can be downloaded freely from the several URL's from the web, also developed for experimental comparison of algorithms which also exist in the WEKA. The task of the explorer are handled by the "Knowledge Flow".

The Knowledge Flow interface is substitute to the Explorer. It will work incrementally, on hypothetically unlimited data streams. different applications. The first application is called "Explorer", this will help us to open or retrieve the datasets from the web or any other storage devices. Next application is Excel file format can be converted to CSV. Primary window of the Weka Graphical User Interface applications contains four.[3]

C. Data Mining

Data mining focuses on the discovery of previously unknown properties in the data. It does not need a specific goal from the domain, but instead focuses on finding new and interesting knowledge. Data mining is concepts are used to analysis the data in sophisticated manner for finding the information which not known already with the effective patterns and to find the relationships among the huge volume of datasets.

D. Classification

Classification rule mining goals to determine a lesser set of rules in the database to form a precise classifier (e.g., Quinlan 1992; Bierman et al 1984). Very often the attributes of classification datasets are continuous. Mostly all the attributes are in the numeric type of data. Classification also one of the datamining technique which is used to forecast the relationship among the group of data instances.

Different types of classification methods are available. These are Bayesian networks, decision tree induction, case-based reasoning, fuzzy logic, genetic algorithm and K- nearest neighbour. In this paper, some of the algorithm from the classification methods are used which is available in the WEKA tools. Also, the output of these algorithms are compared and converted into the graphical representation also

II. RELATED WORKS

They are anther paper, we use different data mining algorithms to predicts all those cases of breast cancer that are recurrent using Wisconsin Prognostic Breast Cancer (WPBC) dataset from the UCI machine learning repository. In shorty, this research is to identify the most successful data mining algorithm that helps to predicts those cases of cancer, which cans recur. The objectives here is also to find critical attributes which play major role in determining and predicting in advance the possibility of recurrence of the breast cancer using C5.0 algorithm. They finally result of these algorithms are clearly they are showing outlined in this paper with necessary results. The classification algorithms, C5.0 and SVM have shown 81% accuracy in classifying the recurrence of the disease. This is found to be best among all. On the other hand, EM was found to be the most promising clustering algorithm with the accuracy of 68%. The research shows that the classification algorithms are better predictor than clustering Algorithms. The impact factors of various parameters responsible for predicting the occurrence/non-occurrence of the disease can be verified clinically.[2]

III. TRAINING DATASET DESCRIPTION

This is one of three domains provided by the Oncology Institute that has repeatedly appeared in the machine learning literature. (See also lymphography and primary-tumour.) This data set includes 201 instances of one class and 85 instances of another class. The instances are described by 9 attributes, some of which are linear and some are nominal.

Table 1. Description of Datasets Attributes

Attributes	Values
age	10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80-89, 90-99
menopause	lt40, ge40, premeno
tumor-size	0-4, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44,45-49, 50-54, 55-59
inv-nodes	0-2, 3-5, 6-8, 9-11, 12-14, 15-17, 18-20, 21-23, 24-26,27-29, 30-32, 33-35, 36-39
node-caps	yes, no
deg-malig	1, 2, 3
breast	left, right
breast-quad	left-up, left-low, right-up, right-low, central
irradiation	yes, no
class	no-recurrence-events, recurrence-events

IV. EXPERIMENTAL RESULT AND DISCUSSION

They are experimental finally result of this study and using the three classifiers are discussed using the WEKA tools and software data mining tools As discussed, breast cancer is classifieds as either unlikely, likely and benign. The performances evaluate results and the error rates are also discussed as follows.[5]

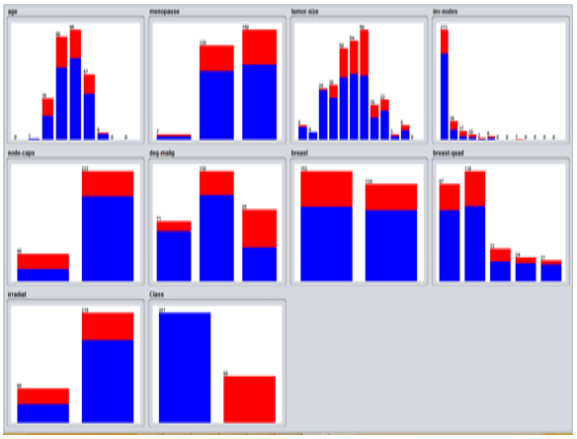

Figure 3. Distribution data set and Training

Table 2. Evaluation Methods

Evaluation Criteria	Classifiers		
	Naïve Bayes	J48	One R
Timing to build model (In sec)	0	0.08	0.01
Correctly Classified in stances	205	216	188
Incorrectly classified instances	81	70	98
Accuracy (%)	71.68%	75.52%	65.73%

V. EVALUATION METHODS

The have carried out some result in order to evaluate the performance of different and default Algorithms for prediction breast cancer detection in order to time to build a model, mostly classified instances , incorrectly and classified instances and percentage accuracy[7].

Table 3. Testing and Simulation Error

Evaluation Criteria	Classifiers		
	Naïve Bayes	J48	One R
Kappa Statistic (KS)	0.2857	0.2826	0.0936
Mean absolute error (MAE)	0.3272	0.3676	0.3427
Root mean squared error (RMSE)	0.4534	0.4324	0.5854
Relative absolute error (RAE)	78.21%	87.86%	81.89%
Root ralative squared error (RRSE)	99.19%	94.61%	128.07%

VI. TRAINING AND SIMULATION ERROR

Kappa statistic, mean absolute error and root mean squared error will be in numeric value only .they also show the relative mostly Error and Root relative squared error in percentage the references and evaluation. they results of the simulation are showing in this tables[3].

Table 4. Confusion Matrix

	Naïve Bayes	One R	J48
Sensitivty	0.8358	0.8258	0.9601
Specificity	0.4352	0.2588	0.2705
Accuracy	71.67%	65.73%	75.52%

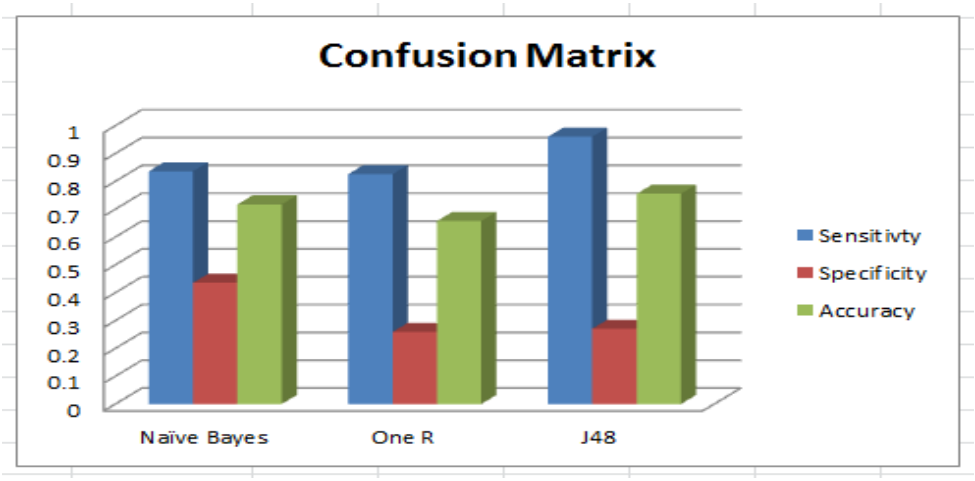

Figure 4. Confusion Matrix

VII. CONCLUSION

In this Report, the accuracies of classification techniques is evaluated based on the selected classifier algorithm. In this study three different data mining's classifications techniques was used for the Detection of breast cancer and their performance was compared in order to evaluate the best classifier. Experimental results shows that this J48 decisions trees is a better model then two other Algorithms in this case for the Detection of breast cancer for the values of accuracy, recall, precisions and error rates recorded for these models. Second, an efficient and effective classifier for breast cancer has been identified while the Number of attributes covered by the classifier can be increased by increasing the sample size of the training set and hence the development of a more accurate models[6].

VIII. REFERENCES

[1]. S.Muthuselvan, Dr.K. Soma Sundaram,Dr.Prabasheela ." Predication of Breast Cancer Using Classification Rule Mininig Techniques in Blood Test Datasets" International Conference on information Communication And Embedded System (ICICES 2016).

[2]. Abdelghani Bellaachia, Erhan Guven,"Predicting Breast Cancer Survivability Using Data Mining Techniques " .

[3]. Ahmed Iqbal Pritom,Shahed Anzarus Sabab,"Predicting Breast Recurrence Using effective Classification and Feature Selection technigue"19Th International conference on computer and information technoloygy,December 18-20-2016 ,North South University ,Dhaka,Bangladesh.

[4]. K.Sivakami, "Mining Big Data Breast Cancer Prediction Using DT-SVM Hybrid Model" International Journal of Scientific Engineering and Applied Science (IJSEAS)-Volume-1,Issue-5,August 2015 ISSN 2395-3470 www.ijseas.com.

[5]. Shomona Gracia Jacob, R.Geetha Ramani ,"Efficient Classifier for Classification of Prognostic Breast Cancer Data through Data Mining Techniques " Proceeding of the world congress on Engineering and Computer science 2012 Vol I WCECS 2012,October 24-26-2012 San Francisco, USA.

[6]. Gabriele Giarratana, Marco Pizzera ,Marco Masseroli ,Enzo Medico ,Pier Luca Lanzi,"Data Mining Techniques for the Identification of Genes with Expression Levels Related to Breast Cancer Prognosis" 2009 Ninth IEEE International Conference on Bioinformatis and Bioengineering.

[7]. Mr.Chintan Shah,Dr.Anjali G.Jivani "Comparison of Data Mining Classification Algorithms for Breast Cancer " 4th ICCCN 2013,IEEE-31661

[8]. Peter Adebayo Idown ,Kehinde Oladipo Williams, Jeremisah Ademola Balogun," Breast Cancer Risk Predication Using Data Mining Classification Techniques " TNC Transaction on Networks and Communications Volume 3 ,Issue 2 ISSN: 2054-7420.

[9]. Gayathri Devi.S,"Breast Cancer Predication System Using Feature Selection and Data Mining Methods" International Journal of Advanced Research in Computer Science ,Volume 2, No. 1, Jan-Feb 2011, ISSN No.0976-5697.

[10]. G.Ravi Kumar ,Dr.G.A.Ramachandra,K.Nagamani, "An Efficient Predication of Breast Cancer Data using Data Mining Techniques" International Journal of Innovations in Engineering and Technology (IJIET) Vol.2 Issue 4 August 2013 SSN: 2319-1058.