Research of Feature Selection Methods to Predict Breast Cancer

K. Venkateswara Rao, L. Mary Gladence, V. .Raja Lakshmi

Abstract:Human health is most important than anything in the world, one should take care of it. Among various disease, cancer is the most terrible and deadly disease, so it is necessary to predict such disease in early stage. In this paper different feature selection methods used for feature extraction with different feature classification methods to identify the breast cancer. Breast cancer data is taken from UCI repository and is processed using WEKA tool and proposed techniques are applied to classify data accurately. This study well defines that data mining approach is suitable for predicting breast cancer.

Keywords: cancer; feature selection; classification; WEKA.

I. INTRODUCTION

Data mining is the process of transforming data into useful data by using various methods. In data mining, medical data mining plays key role in diagnosis of various deceases. It is very complicated task to diagnosis decease efficiently and accurately. In data mining process large data sets are taken and they are pre-processed and sorted then relationships are identified to analyse the taken data and solved various types of problems [1-5]. In machine learning data and feature classification is the most important thing because it has variety of applications in various fields like forecasting, bio informatics and multimedia etc.

In olden days various data classification algorithms are used like lazy learning, decision tree, back propagation, rule based learning, nearest neighbouretc [6-8]. Moreover, there are so many advanced technologies in medical imaging like image annotation, content based image retrieval, image segmentation, computer aided diagnosis etc .Hence medical imaging also plays important role and lot of medical data is available and accessible to the public now a days so that by developing new technologies and various classification methods it is possible to diagnosis decease accurately and effectively [9-11].

II. PROPOSED WORK ARCHITECTURE

The proposed methodology uses breast cancer data sets and is undergoing through various process to predict breast cancer effected samples. Flow chart for the classification of breast cancer data set with and without attribute selection is given in fig.1 below.

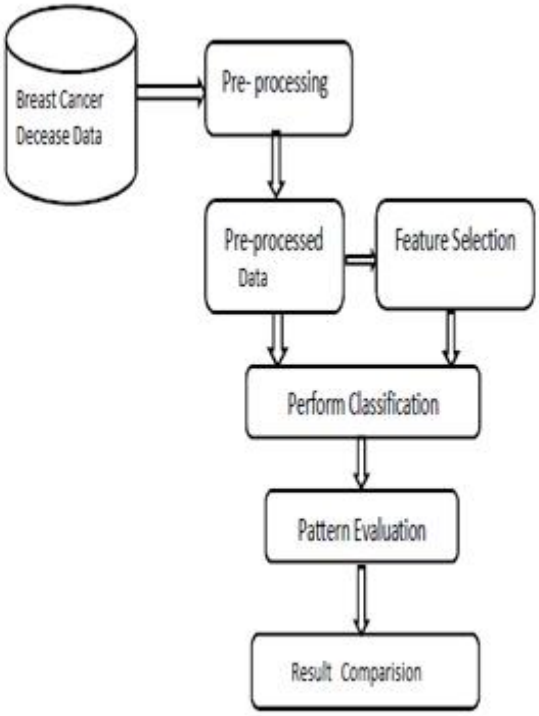

Fig.1.proposed work architecture.

Dataset collection:

Initially raw dataset is collected and considered data set is breast cancer (original) data set having 11 attributes including class label.

Data pre-processing:

Data available for analysis is raw data which may be in different forms. Such data is not having consistency means some data may be missing, may have some irrelevant attributes or it may be noisy data so such data need to be processed first later it can be used to apply for any data mining algorithm for the analysis purpose.

Data processing is one of the important and critical task in data mining which transforms the given initial data sets. This pre-processing of raw data is done by the following ways such as data cleansing, transformation, integration and data reduction.

Performance Evaluation:

Performance evaluation of each attribute is carried out on the basis of recall, precision, F-measure, ROC Area etc.

Revised Version Manuscript Received on 10 September, 2019.

K.VenkateswaraRao, Research Scholar, Sathyabama Institute of Science and Technology, Chennai, Tamilnadu,India

Dr.L.MaryGladence, Associate Professor,Dept.of IT, Sathyabama Institute of Science and Technology, Chennai, Tamilnadu,India

Dr.V. .Raja Lakshmi, Associate Professor,Dept.of CSE, Sathyabama Institute of Science and Technology, Chennai, Tamilnadu,India

III. METHODOLOGY

In the proposed work dataset related to breast cancer has been taken from UCI repository and WEKA tool used for the classification.

Various attributes of dataset are classified using five different classification algorithms on WEKA interface. Proposed work done with WEKA Explorer to evaluate the Classification algorithm accuracy. Accuracy of methodology is evaluated for each attribute.

Performance of algorithm is analysed based on various parameters such as FP rate, TP rate, recall, and precision etc. Whole data set is divided into training and testing datasets.

This work is carried out in the following steps.

There are five different classification algorithms used to predict breast cancer.

- Select various attributes of data set.
- Pre-process all attributes of taken data set by applying filter.
- Select proper classifier to classify the data set.
- Train and test all data sets for classification.
- Analyse the performance of classifier based on taken parameters.

IV. PROPOSED WORK

The experiment conducted using Weka machine learning tool. This dataset contains. The dataset has 699 instances and 10 attributes.

S.NO	Attribute Names
1	Code Number
2	Thickness of clump
3	Cell Size uniformity
4	Uniformity of Cell Shape
5	Marginal Adhesion
6	Single Epithelial Cell Size
7	Bare nucleoli
8	Bland Chromatin
9	Normal nucleoli
10	Mitoses

Case I:

In this scenario, five different classification techniques such as SVM, Bagging, Naïve Bayesian, classification via regression and J48 decision tree are applied on pre-processed data. Each classification algorithms performs differently on same data. Individually, chosen algorithms are classifies the heart disease dataset with the class label presence and absence.

Case II:

Prediction of dataset by applying feature selection method experiment was conducted with many attribute selection methods using WEKA tool. They are, CfsSubsetEval, Information Gain, Gain Ratio and Wrapper method. Each feature selection techniques calculate the value of every attributes and assign the ranking using ranker method. And some feature selection techniques use the search method to search the best attribute from dataset.

V. RESULT ANALYSIS

The Main stay of this Paper is to identify the appropriate classification techniques and features for breast disease prediction. To determine the finest classification algorithm an experimentation was conducted on the breast disease dataset by applying various classification methods. The experimental results showed computation time and accuracy for all the methods. The accuracy of each model is as shown in the below table.1.

Table.1. Comparison of classification algorithms with and without feature selection

Classification Techniques	Accuracy (%)					
	Scenario--1	Scenario -2				
		Cfs (6)	Info Gain (11)	Gain Ratio (9)	Correlation (9)	Wrapper (10)
SVM	99.7	86.4	99.7	98.4	98.0	99.4
Bagging	91.7	89.6	91.9	93.8	91.7	91.6
Naïve Bayes	92.7	87.1	92.6	92.2	91.4	92.6
Cls. Regression	99.7	89.6	99.7	99.4	97.0	99.4
J48	92.0	89.3	79.0	76.8	98.8	85.6

Table.2. Analysis of performance measures for different classifiers

Performance Parameters	Naive Bayes	Bagging	Cl.reg	J48	SVM
Precision	0.750	0.771	0.811	0.831	0.841
Recall	0.761	0.772	0.802	0.834	0.840
F-Measure	0.761	0.765	0.802	0.833	0.835
ROC area	0.820	0.714	0.871	0.882	0.883

Table.3 The Performance of Classification Algorithms on Breast Cancer Dataset.

Data set	Naive bays	Bagging	Cl.Regression	SVM	J48
Breast cancer	72.7%	65.3%	60.03%	82.53%	79.8%

VI. CONCLUSION

WEKA tool is considered as one of the best and accurate tool for data classification in data mining. SVM gives accurate results compared to other algorithms on breast cancer data set. We have evaluated 286 instances and 10 attributes for breast cancer. So we conclude that SVM is the best classification algorithm for classifying the breast cancer data set.

REFERENCES

1. RohitAroraand Suman "Comparative Analysis of Classification Algorithms on Different Datasets using WEKA" 2012 International Journal of Computer Applications (0975 - 8887) Volume 54- No.13, September 2012.
2. K.Sivakami et al., "Mining Big Data: Breast Cancer Prediction us-ing DT - SVM Hybrid Model", International Journal of Scientific Engineering and Applied Science, volume 1, 2015.
3. KashishAraShakilShadmaAnisMansafAlam "Dengue disease prediction using weka data mining tool" ar Xiv preprint arXiv:1502.05167< 2015"
4. HibaAsri, "Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis", The 6th International Symposium on Frontiers in Ambient and Mobile Systems, pp.1064-1069.
5. Canlas, R. D. "Data mining in healthcare: Current applications and issues." School of Information Systems &Management,Carnegie Mellon University, Australia (2009).
6. MA Jabbar, "Prediction of heart disease using k-nearest neighbor and particle swarm optimization", Biomedical Research , volume 28, 2017.
7. G.L. Pappa and A.A. Freitas, Automating the Designof Data Mining Algorithms. An EvolutionaryComputation Approach, Natural Computing Series, Springer, 2010.
8. G. Sumalatha et al., "A Study on Early Prevention and Detection of Breast Cancer using Data Mining Techniques", International Jour-nal of Innovative Research in Computer and Communication Engi-neering, volume 5,2017.
9. HibaAsri, "Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis", The 6th International Symposium on Frontiers in Ambient and Mobile Systems, pp.1064-1069.
10. BangsukJantawan et al., "A Comparison of Filter and Wrapper Approaches with Data Mining Techniques for Categorical Varia-bles Selection", International Journal of Innovative Research in Computer and Communication Engineering, Volume 2, 2014.
11. Animesh et al., "Study and analysis of Breast cancer Cell Detec-tion using Naïve Bayes, SVM and Ensemble Algorithms", Intenational Journal of Computer Applications, vol.2, 2016.