Volume 9, Special Issue No. 1, February 2018

ISBN: 978-93-5300-455-2

International Journal of Advanced Research in Computer Science
(ISSN: 0976-5697)

RESEARCH PAPER

SHRI SHANKARLAL SUNDARBAI SHASUN JAIN COLLEGE FOR WOMEN
ICRACC 2018
(23rd–24th February 2018)

Available Online at www.ijarcs.info

A Study on Efficient Classification Model for Breast Cancer Prediction Based on Feature Selection Techniques

B. Tamilvanan
Research and Development Centre
Bharathiar University,
Coimbatore-641046, TN, India,

Dr. V. MuraliBhaskaran
Principal,
Dhirajlal Gandhi College of Technology
Salem-636290, TN, India.

ABSTRACT

Classification algorithms are efficiently utilized in the area of general medical diagnosis applications in order to identify the disorders in advance. One such disease, breast cancer is the most prevalent and earnest quandary with women in most of the developing countries. Many attempts are made in order to identify this problem with the objective of high precision and better accuracy. In this paper, an attempt is made with the most popular and efficient classification algorithms namely Naive Bayes, Best First Search(BFS), Random Search (RNS) and Genetic Search (GNS) to amend the efficiency of the detection, accuracy for the breast cancer dataset. As an objective of improving accuracy, an efficient dimensionality reduction technique is incorporated in this work. The performances of these approaches are evaluated using the metrics such as the precision, recall, f-measure and accuracy. From these measures it is clearly observed that Naive Bayes with Best First Search algorithm is able to achieve high accuracy rate along with minimum error rate when compared to other algorithms. The review can be stretched out to draw the execution of other characterization systems on an extended information set with more particular ascribes to get more exact outcomes.

Keywords: Classification, Feature selection, Naive Bayes, Best first search, Random search and Genetics search.

INTRODUCTION

Data mining systems and programming are used in a substantial differ of fields, together with securities exchange, ERP, media transmission, endeavor enterprises, , climate, social insurance and sizably voluminous information [1] [2]. These days wellness mind industry creates a tremendous measure of data about patients, disease conclusion, and so on. Some exceptional sorts of procedures to developing right groupings have been proposed (e.g., NB,BFS-NB, GNS-NB, RNS-NB). In characterization, we give a Breast Cancer informational index of case report or the info information, called the check informational index, with each archive comprising of different attribute.

An attribute can be both a numerical attribute and categorical attribute. If values of an attributes belong to an authoritatively mandated domain, the attribute is referred to as numerical attribute(e.g. Tumor-size, Deg-Malig, Menopause, Age, Inv-nodes). A categorical attribute (e.g. Irradiant, Breast, Node-cape, Breast-Quad, Class).Classification is the process of splitting a dataset into mutually exclusive groups, called a class, based on suitable attributes.

In this world, individual sorts of Breast Cancer maladies are a typical type of disease influencing all ladies of various ages. Bosom disease influences the bosom tissue and lobules. The classification of breast cancer is resulted from its beginning, if breast cancer is originate from milk ducts then it is known as ductal carcinoma while cancer cells found in lobules makes cancer termed as “lobular carcinoma.” The viewing of bosom malignancy is an essential stride which sifts through the manifestations that can be utilized to analyze the patient's real obsessive condition. Breast cancer is the most continuous reason for death in more established ladies however in the meantime, it is critical to note that more youthful ladies who don't go under tumor screening process stay in risk hover of breast cancer.

In this paper is designed accordingly: the relates works and show of the focused parts of the utilized data mining methods in section 1. The information of the dataset for Breast Cancer in section 2.The experimentation outcome and conversation in section 3. And finally, conclude the paper and future enhancements.

LITERATURE REVIEW

A multinomial logistic-regression model with a hill-like estimator generalizes logistic regression by using more than two distinct outcomes between the categorical and multinomial distributions [3].This model is mainly designed to predict the probabilities of different outcomes when using categorically dependent and independent variables.

Best First Search Algorithm:

The Best First Search is an important AI search strategy that allows back tracking along the search path. Like the best first search moves through the search space by making local changes to the current feature subset. However, unlike hill climbing method, suppose path being explored begins to look less promising, the best first search method can back-track to a more promising previous subset and continue the search from there. A best first search will explore the entire search space for specified time, so it is common to use a stopping criterion. Normally this involves limiting the number of the fullyexpanded subset and that results in no improvement [4][5].

Genetic Search Algorithm:

Search techniques traverse the attribute space to locate a decent subset and the quality is estimated by the property subset evaluator through CFS subset evaluator and hereditary pursuit is being utilized as a search techniques. The parameters of the genetic algorithm area number of generations, population size and the probabilities of mutation and crossover. A member of the initial population generates by specifying a list of attribute indices as a search point. For generating progress reports, every so many generation can be used [6][7].

PROPOSED METHOD

proposed BFSCFS-NB algorithm:

Step 1: To start with an OPEN list containing the start state, the CLOSED listempty and BEST← start state.

Step 2: Let assign s = arg max e(x) (get the state from OPEN with the highestevaluation).

Step 3: Eliminate s from OPEN and add to CLOSED.

Step 4: If e(s) ≥ e(BEST), then BEST← s.

Step 5: For every child t of s that is not in the OPEN or CLOSED list, evaluateand add to OPEN.

Step 6: If BEST changed in the last set of expansions, go to 2

Step 7: Return BEST.

Step 8: Obtain the new data set.

Step 9: Construct both training and test data discrete.

Step 10: Estimate the prior probabilities P(Cj), j=1,... k from the training data,where k is the number of classes.

Step 11: Estimate the conditional probabilities P(Ai= aℓ│Cj), i= 1,....,D,j=1,....,k, ℓ= 1,....,d from the training data, where D is the number offeatures, d is the number of discretization level.

Step12: Estimate the posterior probabilities P(Cj│A) for each test example xrepresented by a feature vector A.

Step 13: Assign x to the class C* such that C*=arg max j=1,2 P(Cj│A).

The first half of the algorithm from step one to eight is used to select thesubset using Best First Search and then the second half of the algorithm from nineto thirteen are for classification using Naive Bayes.

BREAST CANCER DATASET

The performance of these classification algorithms namely Naive Bayes, Best First Search, Genetics Search and Random Search was tested in a medical database for Breast Cancer Disease dataset from UCI machine learning repository (available at http://archive.ics.uci.edu/ml/datasets/Breast+Cancer [8]. The data set has ten features of the attributes. Table- 1 describes the data for Breast Cancer. The medical dataset contains data from reviews conducted among patients, each of which has ten features. All features can be considered as on indicators of Breast Cancer disease for a patient. The dataset holds records of the following attributes.

Table 1: UCI Dataset of Breast Cancer

Attributes Name	Attribute Type	Description
Age	Numeric	Age (years)
Inv-Nodes	Numeric	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	Discrete	yes, no.
Menopause	Numeric	lt40, ge40, premeno
Deg-Malig	Numeric	1, 2, 3.
Tumor-Size	Numeric	0-4, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59
Breast	Discrete	left, right
Breast-Quad	Discrete	left-up, left-low, right-up, right-low, central.
Irradiat	Discrete	yes, no.
Class	Discrete	no-recurrence-events, recurrence-events

CONFUSION MATRIX

The confusion matrix shows how many occurrences have been given to each class and the elements of the matrix illustrate the number of test examples whose concrete class is the row and whose predicted class is the column. Tables 7, 8 and 8 illustrate the confusion matrix that is calculated for the Best First Search based NB, Genetic Search based NB and Random Search based NB algorithms.

Table 5 Different outcome of two class prediction

Actual Class	Predicated Class		
		a	b
	a	Ture Positive	False Negative
	b	False Positive	True Negative

Precision

It is utilized to speak to the portion of recovered information from associating datasets, which pertain to the search. Precision will be used to represent how many instance have been correctly classified in the confusion matrix table

(correct classified data is true positive and incorrect classified data is error positive).

$$\text{Precision} = \frac{tpA}{tpA + eBA}$$

Where tpA is represented as true positive for the class A and eBA are represented as false positive.

Recall
It is utilized to speak to the portion of recovered information from associating datasets; that are important to the inquiry that is successful. It is used to find out the ratio between the true positive and both true positive and false positive values.

$$\text{Recall} = \frac{tpA}{tpA + eAB}$$

Where tpA is represented as true positive for the class A and eAB are represented as error positive.

F-measure This is evaluated by the harmonic mean between precision and recall.

$$\text{F-Measure} = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Accuracy This is calculated as the proportion of true positive, true negatives and true results from all the given data.

$$\text{Accuracy} = \frac{tpA + tpB}{tpA + eAB + eBA + tpB}$$

EXPERIMENT RESULTS AND DISCUSSION

In this section, we explain the test database and investigational analysis and the current evaluation results for four algorithms namely NB, BFS-NB, GNS-NB, RNS-NB classifier.

In this experimental analysis, NB, BFS-NBwith 5 potential attributes namely (Tumor-Size, Inv-Nodes, Node-Caps, Breast-Quad, Irradiat) GNS-NB5 potentialattributes(Tumor-Size, Menopause, Deg-Malig, Node-Caps, Irradiat), RNS-NBwith 6 potential attributes namely (Tumor-Size,Menopause, Inv-Nodes, Node-Caps,Deg-Malig, Irradiat)Algorithms performance were compared based on their application in medical datasets. Weka tool is is utilized for research area, share markets, bankingsector, education institute and climate datasets. It helps in composed exercises in machine learning, data mining, and text mining. It supports all the mining process to get a valid and clear visualization of accurate results. ten-fold cross-validation with feature selection attributes were to the input datasets in the experiments

Experimental Step Up

A brief description of the classification process by all algorithms, NB, BFS-NB, GNS-NB, RNS-NBare given below:

Table 10: Performance analysis related to accuracy

Method	Accuracy (%)
NB	72
GNS-NB	73
RNS- NB	77
BFS - NB	82

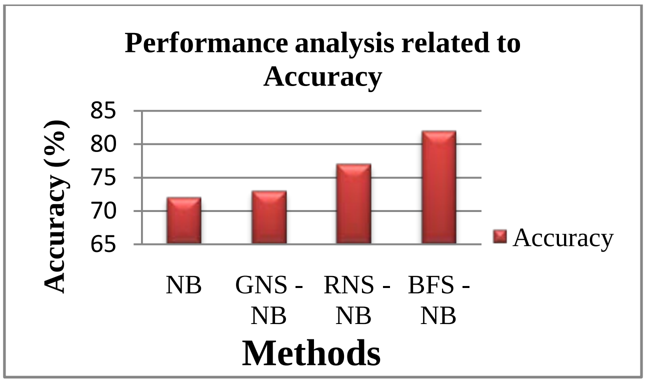

Fig 1 : Performance analysis related to accuracy

CONCLUSION

In this work popular classification algorithms along with feature selection method are used to calculate the breast cancer detection process more efficiently. The efficient classification algorithms namely NB, GNS-NB, RNS-NB, BFS-NB are used to develop the model and all are evaluated 10 fold cross-validation. The dimensionality reduction technique is able to select more efficient and relevant features from the ten original features and also observed that results obtained using relevant features are better than or equal to the results obtained using ten features with less effort. These classification algorithms are compared, and accuracy is evaluated for true positive and false positive rate. From the experiments, it is observed that Naive Bayes using Best First Search classification algorithm performs compare than other classification algorithms with 82% accuracy for both after feature selection and before feature selection using ten-fold cross validations.

REFERENCES

[1] B.Tamilvanan and Dr. V. MuraliBhaskaran, “A New Feature Selection Techniques Using Genetics Search and Random Search Approaches For Breast Cancer”, Biosciences and Biotechnology Research Asia, , vol. 14, no.1, pp. 409-414, March 2017.

[2] Sitar-Taut, V.A., et al, Using machine learning algorithms in cardiovascular disease risk evaluation. Journal of Applied Computer Science & Mathematics, 2009.

[3] You-Shyang Chen, Modeling hybrid rough set-based classification procedures to identify hemodialysis adequacy for end-stage renal disease patients, Computers in Biology and Medicine, 2013, vol. 43, pp. 1590–1605.

[4] Hall, Mark A. and Lloyd A.Smith., "Feature subset selection: a correlation based filter approach", 1997.

[5] Hall. M, “Correlation based feature selection for machine learning”,Doctoral dissertation,University of Waikato, Dept. of Computer Science, 1999.

[6] Pallabi Borah et al.,“A statistical feature selection technique”, Netw Model Anal Health Inform Bioinforma, pp-3-55, 2014.

[7] ShashikantGhumbre, Chetan Patil and Ashok Ghatol, “Heart Disease Diagnosis using Support Vector Machine”, International Conference onComputer Science and Information Technology (ICCSIT'2011) , pp. 84-88,December 2011.

[8] UCI Machine Learning Repository http://archive.ics.uci.edu/ml/datasets/Breast+Cancer