Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 132 (2018) 371–376

Procedia Computer Science

www.elsevier.com/locate/procedia

International Conference on Computational Intelligence and Data Science (ICCIDS 2018)

Breast Cancer Prediction system

Madhu Kumari[a], Vijendra Singh[b]

b a Department of Computer Science and Engineering, The NorthCap University, Sector 23A,Gurugram, Haryana, 122017, India

Abstract

Breast cancer became the major source of mortality between women. The accessibility of healthcare datasets and data analysis promote the researchers to apply study in extracting unknown pattern from healthcare datasets. The intention of this study is to design a prediction system that can predict the incidence of the breast cancer at early stage by analyzing smallest set of attributes that has been selected from the clinical dataset. Wisconsin breast cancer dataset (WBCD) have been used to conduct the proposed experiment. The potential of the proposed method is obtained using classification accuracy which was obtained by comparing actual to predicted values. The outcome confirms that the maximum classification accuracy (99.28%) is achieved for this study.

Keywords: WCBD, classification, knowledge mining, prediction system.

1. Introduction

Most common type of medical hazard found in middle aged women is, breast cancer. Mortality rate of women due to breast cancer can be reduced if can be detected at a relatively early stage. With the help of latest, efficient and advanced screening methods, the majority of such cancers are diagnosed when the disease is still at a localized stage [1]. The utility of machine learning techniques in healthcare analysis is growing progressively. Certainly analysis of patient's clinical data and physician's judgment are the most considerable features in diagnosis. Most of the possible medical flaws can be avoided by the using classification systems, and also offer healthcare data to be analyze in lesser time and in more exhaustive manner. Accurate and timely prediction of breast cancer allows physicians and healthcare providers to make most favorable decision about the patient treatment.

This paper is organized into various sections which mainly focus only on predicting breast cancer. Section 2 allows the reader to have basic knowledge on breast cancer study and earlier pertinent literature. Section 3 includes the

Corresponding Author: madhunain@gmail.com

10.1016/j.procs.2018.05.197

details of the approach used. Section 4, measures of performance evaluation are given and followed by section 5, which shows the prediction results of all three. We summarize the research findings, and discuss the limitation in section 6.

2. Previous research

Lots of breast cancer research has been reported in the literature of medical data analysis, and most of them turn up with good classification accuracies. Polat et al , proposed LS-SVM classifier algorithm for the diagnosis of breast cancer [2] and achieved the classification accuracy of 98.53% using 10-fold cross validation. Akay, proposed a new method for the breast cancer diagnosis using support vector classification algorithm on the most predictive features and obtain the classification accuracies of 99.02% without cross-validation [3].

Yeh et al. present a innovative technique for breast cancer detection, by using statistical methods in combination with swarm optimization and reported the accuracy of 98.71%[4].

Marcano-Cede~no et al. proposed a new method AMMLP for the classification of breast cancer datasets by using an Artificial Neural Network over the biological metaplasticity property and acquire classification accuracy of 99.26 [5].

Kaya and Uyar [6] in their work presented a hybrid approach for the detecting hepatitis disease by means of rough set and extreme machine learning algorithm. The selected hepatitis disease dataset was from UCI repository. 20 reducts containing three to seven attributes were produced by using rough set theory. The reducts are selected and then the records with missing value are removed from each reduct. Classification is done on selected reducts by using back propagation neural network and obtained the accuracy of 98.6%. All of the above listed works are just a small representative of the existing huge number of research in utilizing machine learning and data mining techniques to a range of healthcare domains for forecasting and pattern recognition purposes.

3. Proposed Framework

In vision of the problem statement described in the introduction section, a classification model is proposed with boosted accuracy to predict the breast cancer patient. The framework is composed of the following important phases:

- Dataset Selection
- Data Preprocessing
- Learning by Classifier (Training) i.e. SVM, Linear Regression and KNN.
- Achieving trained model with highest accuracy
- Using trained model for prediction

The detail description of the components and the activities performed against each component is mentioned below.

3.1. Data Source

The dataset was obtained from UCI repository and is a benchmark dataset. Breast Cancer Wisconsin (Original) Dataset contains 699 instances out of which 16 instances suffer from missing values. The dataset is distributed over of 65.0% cancerous samples and 35.0% of non cancerous samples. The complete details of all the eight features are shown below in table 1.

Table 1. Description of the WBCD

Attribute numbers	Attribute description	Values of attribute	Mean	Standard deviation
1	Clump thickness	1–10	4.44	2.83
2	Uniformity of cell size	1–10	3.15	3.07
3	Uniformity of cell shape	1–10	3.22	2.99
4	Marginal adhesion	1–10	2.83	2.86
5	Single epithelial cell size	1–10	2.23	2.22
6	Bare nuclei	1–10	3.54	3.64
7	Bland chromatin	1–10	3.45	2.45
8	Normal nucleoli	1–10	2.87	3.05
9	Mitoses	1–10	1.60	1.73

3.2. Data understanding and data preparation

WBCD consisted of 699 instances and 11 features. These 11 features provide precise information pertaining to the occurrence of breast cancer. Moreover, the dataset was scrutinized for unknown values, inconsistency and erroneous data. Unknown values can have a consequential effect on the interpretations that can be derived from the data. 16 instances with missing values are present which are denoted by"?" in the Breast cancer dataset. Each missing value is replaced with random number constant and is ignored during analysis Understanding of data distribution among dataset is very critical task and must be done effectively. Analysis of data distribution uncovers various interesting relationship and insights which can be useful in selecting best predictive feature.

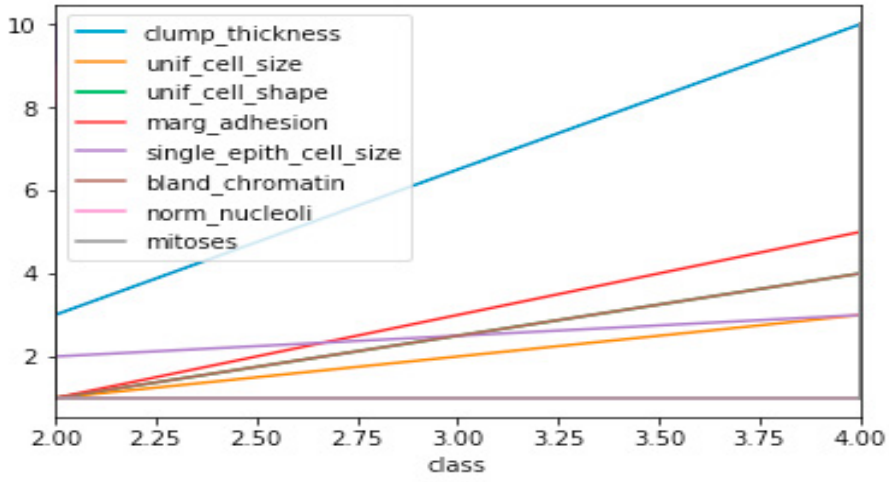

Figure 1: Collective data variability with respect to class

3.3. Feature Selection

Here we have used filter method to select most relevant features from the dataset. Features are selected on the basis of their scores in statistical tests for their correlation with the outcome variable. A good feature subset is one that contains features highly correlated with (predictive of) the class. Correlation-Based Measures:For feature X with values x and classes C with values c, where X, C are treated as random variables, Pearson's linear correlation coefficient is defined as

$$r=\frac{\sum_{i=1}^{n}(xi-x)(yi-y)}{\sqrt{[\sum_{i=1}^{n}(xi-\bar{x})^2][\sum_{i=1}^{n}(yi-\bar{y})^2]}} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1)$$

where *xi* and *yi* is the i[th] value of x and y respectively.
% of (r) = ±1 if X and Y are linearly dependent and zero if they are completely uncorrelated.

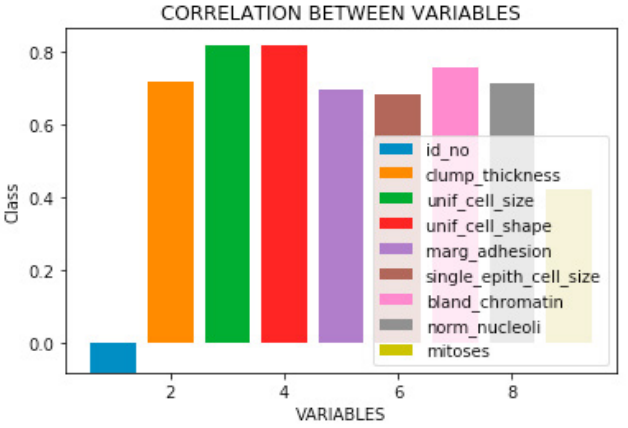

Figure2: Correlation between WBCD variables.

3.4. Training and Classification

Classifications of the data sets are done on the basis of specific properties posses by the sample variable that the capable to classify them, and each sample variable is assigned a malignant or benign class. Classification is principally done by making predictions based on known sample data that has been learned from training data. Designed algorithm is first trained on the known data labels and further uses this learning to predict the class labels for the new unknown set of data sample. The classification objective set for this study is to achieve enhanced accuracy by using LR, SVM and KNN classifiers and determine which one suits the most for diabetes classification technique.

We train the classifier with known sample data in a training dataset and check its performance by examining the test dataset, which consists of the unknown sample used to predict its class label KNeighborsClassifier is a supervised, instance-based learning classifier which learns from the labled data samples. The pseudo code for the KNN classifier is given in Algorithm 1. K folds cross validation technique is used for training data. In this technique, the original sample is divided into k equivalent size subsamples and one subsample is used for validating the model, while the k-1 remaining subsamples are utilized as training data. After that this cross-validation process is recurring k times (called the folds), with every of the k subsamples just used one time as the validation data. It works in loop manner. In this study we set the value of k=10.

KNeighborsClassifier(X,y, x) // X: training data, y: class labels of X, x: unknown sample
i=1
do
Calculate distance D(Xi, x)
while(i<=m)
Compute set Z having indices for the k smallest distances D(Xi, x).
Return label with {yi where $i \in$ Z}

Algorithm 1: K-Neighbors Classifier Pseudo code

4. Measure for Performance Evaluation

Performance of the proposed system is evaluated by considering the actual and predicted classification. Accuracy of the system is calculated by using the confusion matrix obtained for the classifier used. Table 2 shows the confusion matrix for a two class classifier. Classification accuracy, sensitivity, specificity, positive predictive value and negative can be defined by using the elements of the confusion matrix as.

Classification accuracy: Accuracy of the classification is obtained by using the given equation:

$$Accuracy = \frac{TP+TN}{\mathrm{TP+TN+FP+FN}} \times 100 \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2)$$

Where TP: Correctly classified as having breast cancer
TN: Correctly classified as not having breast cancer.
FP: Classified as having breast cancer but actually they don't have (Error of type I)
FN: Classified as not having breast cancer but actually they have cancer. (Error of type II)

Table 2 Confusion matrix representation

Actual	**Predicted**	
	Positive	Negative
Positive	True Positive(TP)	False Negative(FN)
Negative	False Positive(FN)	False Positive(FP)

5. Comparison of Results

We compared the results achieved in this study with the results reported by other researcher in the existing literature. We mainly focused on the method used and the accuracy achieved by the other studies. Table 3 shows that KNN perform better than other classifier used in the study.

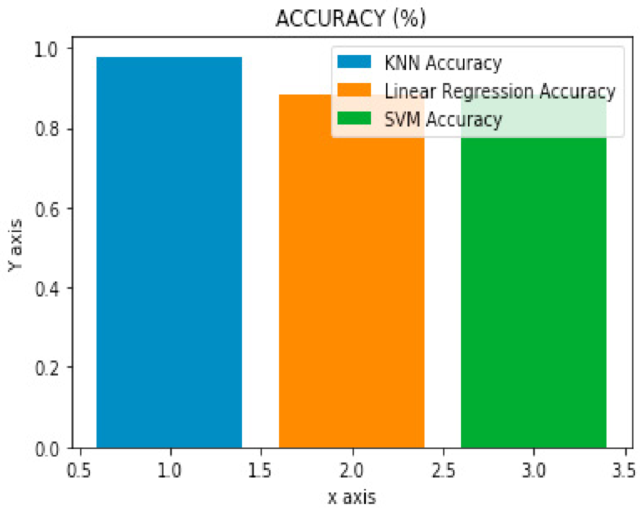

Figure3: Comparison of LR, SVM and KNN accuracy.

Many studies have been carried out, that uses number of different classifiers and approach for the prediction of breast cancer. Performance of our proposed method in this study shows best result as compared to the approach used by other author on the same dataset. Summary of performance comparison has been shown in table 3.

Table 3: Performance comparison of this study with approaches used by other authors

Dataset	Paper Title	Approach	Accuracy	This Study Accuracy
Breast Cancer Wisconsin (Diagnostic) Data Set	Knowledge Mining from Clinical Datasets Using Rough Sets and Backpropagation Neural Network[7]	Rough Sets and Backpropagation Neural Network	98.60%	99.28%
	Breast cancer diagnosis based on feature extraction using a hybride of K-mean and support vector machine algorithms[8]	Feature extraction using a hybride of K-mean and support vector machine algorithms	97.38%	
	Predicting breast cancer survivability a comparission of three data mining methods[9]	LR, Neural network and Decision tree	89.2%	

6. Conclusion and Future Work

A decision support system for predicting breast cancer helps and assist physician in making optimum, accurate and timely decision, and reduce the overall cost of treatment. Different classifiers have been used to conduct experiments on the standard WBCD. It is been observed KNN classifier yields the highest classification accuracies when used with most predictive variables. The proposed system greatly reduces the cost of treatment and improves the quality of life by predicting breast cancer at early stage of development. The future work will focus on exploring more of the dataset values and yielding more interesting outcomes. This study can help in making more effective and reliable disease prediction and diagnostic system which will contribute towards developing better healthcare system by reducing overall cost, time and mortality rate.

References

[1] Jemal A, Murray T, Ward E, Samuels A, Tiwari RC, Ghafoor A, Feuer EJ, Thun MJ. Cancer statistics, 2005. CA: a cancer journal for clinicians. 2005 Jan 1;55(1):10-30.
[2] Polat K, Güneş S. Breast cancer diagnosis using least square support vector machine. Digital Signal Processing. 2007 Jul 1;17(4):694-701.
[3] Akay MF. Support vector machines combined with feature selection for breast cancer diagnosis. Expert systems with applications. 2009 Mar 1;36(2):3240-7.
[4] Yeh WC, Chang WW, Chung YY. A new hybrid approach for mining breast cancer pattern using discrete particle swarm optimization and statistical method. Expert Systems with Applications. 2009 May 1;36(4):8204-11.
[5] Marcano-Cedeño A, Quintanilla-Domínguez J, Andina D. WBCD breast cancer database classification applying artificial metaplasticity neural network. Expert Systems with Applications. 2011 Aug 1;38(8):9573-9.
[6] Kaya Y, Uyar M. A hybrid decision support system based on rough set and extreme learning machine for diagnosis of hepatitis disease. Applied Soft Computing. 2013 Aug 1;13(8):3429-38.
[7] Nahato KB, Harichandran KN, Arputharaj K. Knowledge mining from clinical datasets using rough sets and backpropagation neural network. Computational and mathematical methods in medicine. 2015;2015.
[8] Liu L, Deng M. An evolutionary artificial neural network approach for breast cancer diagnosis. InKnowledge Discovery and Data Mining, 2010. WKDD'10. Third International Conference on 2010 Jan 9 (pp. 593-596). IEEE.
[9] Chen HL, Yang B, Liu J, Liu DY. A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis. Expert Systems with Applications. 2011 Jul 1;38(7):9014-22.