Data Mining Techniques for Prediction of Breast Cancer

Abstract

Breast cancer is one of the common cancers in Women and early detection of the disease can find a potential importance in this area. The main criteria is understand how well these data mining algorithms predict the probability of recurrence of disease. Experiments are conducted and indicate that classification algorithms are better than clustering algorithms in this scenario.

Introduction

Breast cancer is the most common cancer in the world among women according to World Health Organization's(WHO) report. As per the report, Indian women are most affected by this disease and,therefore, it is the most common cause of death too.Many biological techniques can be used for early detection of breast cancer so that preventive measures can be taken.

Different clustering and classification algorithms of data mining techniques have been used to find the performance of these prediction models Four clustering algorithms (K means, EM, PAM and Fuzzy c-means) and four classification algorithms (SVM, C5.0, Naive bayes and KNN). We use different data mining algorithms to predict all those cases of breast cancer that are recurrent using Wisconsin Prognostic Breast Cancer (WPBC) dataset from the UCI machine learning repository

R programming tool is used for the implementation purpose that provides free software environment for data analysis

Survey

Delen et al, used the SEER data (period of 1973-2000 with 202,932 records) of breast cancer to predict the survivability of a patient using 10-fold cross validation method. The result indicated that the decision tree (C5) is the best predictor with 93.6% accuracy on the dataset, artificial neural network (ANN) also showed good performance with 91.2% accuracy.The logistic regression model was less successful with 89.2% accuracy as compared to other two.

Jacob et al, who compared various classifier algorithms on Wisconsin Breast Cancer diagnosis dataset. Their results demonstrate that Random Tree and C4.5 classification algorithm produce 100% accuracy.

Pendharkar et al, used several data mining algorithms for discovering patterns in breast cancer. They showed that data mining could be used in discovering similar patterns in breast cancer cases, which could be a great help in early detection and Prevention of this disease.

Procedure:

Data Source

In WPBC dataset, Out of 35 attributes,the 'Outcome' is the target attribute (class label); and, all other 32 attributes (except ID) are decisive attributes whose value helps in predicting the recurrence of the disease.This data set consists of 198 records of patients out of which, the

value of the attribute 'Lymph node' status was missing in 4 records. Since lymph node value is an important factor in determining the breast cancer status.Thus the records containing the missing data of this attribute were removed from the dataset rather than removing this attribute itself. Thus the final dataset contains 194 records in which 148 were non recurrent and 46 were recurrent cases.

Prediction Models

A short description of these algorithms are as follows

Clustering Algorithms

We have used four clustering algorithms namely K-means, EM, PAM, Fuzzy c-means. The K-means clustering algorithm works by partitioning n observations in to k sub-classes defined by centroids, where k is chosen before. EM (Expectation–maximization) is a statistical model that depends on unobserved latent variables to estimate the parameters using maximum -likelihood. The PAM (Partitioning around Medoids)is similar to K-means except that here partitioning is based on K-medoids method that divides data into a number of disjoint clusters.

Classification Algorithms

Four classification algorithms used are KNN, SVM, Naive Bayes and C5.0. In KNN (K Nearest Neighbor), object is classified by a majority vote of its neighbors, with the object being assigned to class most common among its k nearest neighbours. In SVM (Support Vector Machines),data is first converted in to a set of points and then classified in to classes that can be separated linearly. Naive Bayes model works by estimating the probability of a dataset that can belong to class using Bayes' rule. C5.0 algorithm is a decision tree that recursively separates observations in branches to construct a tree for the purpose of improving the prediction accuracy.

Results

Clustering Algorithms					Classification Algorithms				
Algorithms	Confusion Matrix	Accuracy	Sensitivity	Specificity	Algorithms	Confusion Matrix	Accuracy	Sensitivity	Specificity
K-Means	N R N 100 48 R 23 23	0.6340	0.8130	0.3239	C5.0	N R N 47 0 R 11 0	0.8103	1.0	0.0
EM	N R N 117 31 R 31 15	0.6804	0.7905	0.3260	KNN	N R N 47 0 R 11 0	0.7068	0.8297	0.2
PAM	N R N 64 84 R 29 17	0.4175	0.4324	0.1683	Naïve Bayes	N R N 47 0 R 11 0	0.5344	0.5319	0.2142
Fuzzy c-means	N R N 50 98 R 24 22	0.3711	0.3378	0.1833	SVM	N R N 47 0 R 11 0	0.8103	1.0	0.0
Mean		0.5257	0.5934	0.2504	Mean		0.7154	0.8404	0.1036

The above result shows that on comparison, classification algorithms are better predictors than clustering algorithms. The classification algorithms were 0.7154 accurate as compared to the accuracy of 0.5257 of clustering algorithms.

Conclusion 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%. Classification algorithms are better predictor than clustering algorithms.