Chemotherapy Prediction of Cancer Patient by using Data Mining Techniques

Reeti Yadav
Invertis University, Bareilly

Zubair Khan
Invertis University, Bareilly

Hina Saxena
Invertis University, Bareilly

ABSTRACT

Breast cancer is one of the prominent diseases for women in developed countries including India. It is the second most frequent cause of death in women. The identification of breast cancer patients for whom chemotherapy could prolong survival time is considered here as a data mining problem. We prescribe a procedure that uses support vector machines (SVMs) and Decision tree for classifying 100 breast cancer patients into two classes which are the two types of breast cancer diseases. It then compares the performance of both the classification techniques to find the better technique among them and use the appropriate technique for the next stage i.e. clustering. The identification is achieved by making clusters of above two classes into three prognostic groups: Good, Intermediate and Poor with the help of K-Means clustering technique. The result suggests that the patients in the Good group do not require chemotherapy. Chemotherapy is not of much importance in an Intermediate class while the Poor group is the most crucial group where chemotherapy can possibly enhance their survival.

General Terms

Chemotherapy, Classification, Cancer.

Keywords

Clustering, SVM, decision tree, k-means, classification, diagnosis, data mining.

1. INTRODUCTION

With ever increasing growth in science and technology, quality of human life is improving day by day. Health becomes a major concern for everyone. . Breast cancer has become the primary reason of death in women in developed countries. Today, about one in eight women over their lifetime have been affected by breast cancer in the United States. 5-10% of cancers are due to an abnormality which is inherited from the parents and about 90% of breast cancers are due to genetic abnormalities that happen as a result of the aging process [1]. The most effective way to reduce breast cancer deaths is detect it earlier. Early diagnosis requires an accurate and reliable diagnosis procedure that allows physicians to distinguish benign breast tumors from malignant ones without going for surgical biopsy [2] [3]. The objective of these predictions is to assign patients to either a "*benign*" group that is noncancerous or a "*malignant*" group that is cancerous. The prognosis problem is the long-term outlook for the disease for patients whose cancer has been surgically removed. In this problem a patient is classified as a 'recur' if the disease is observed at some subsequent time to tumor excision and a patient for whom cancer has not recurred and may never recur. The motive of these predictions is to handle cases for which cancer has not recurred (censored data) as well as case for which cancer has recurred at a specific time. As the use of computers powered with automated tools, large volumes of medical data are being collected and made available to the medical research groups. As a result, Knowledge Discovery in Databases (KDD), which includes data mining techniques, has become a popular research tool for medical researchers to identify and exploit patterns and relationships among large number of variables, and made them able to predict the outcome of a disease using the historical cases stored within datasets. Thus breast cancer diagnostic and prognostic problems are mainly in the scope of the widely discussed classification problems.

2. DATA MINING

The data mining consists of various methods. Different methods serve different purposes, each method offering its own advantages and disadvantages .Classification and clustering are the two most common techniques of data mining which are used in field of medical science.[4] However, most data mining methods commonly used for this review are of classification category as the applied prediction techniques assign patients to either a "benign" group that is non- cancerous or a "malignant" group that is cancerous and generate rules for the same. Hence, the breast cancer diagnostic problems are basically in the scope of the widely discussed classification problems. In data mining, classification is one of the most important task. It maps the data in to predefined targets. It is a supervised learning as targets are predefined. The aim of the classification is to build a classifier based on some cases with some attributes to describe the objects or one attribute to describe the group of the objects. Then, the classifier is used to predict the group attributes of new cases from the domain based on the values of other attributes. The commonly used methods for data mining classification tasks can be classified into the following groups.

3. RELATED WORK

In this section, we review the related work on breast cancer diagnosis using data mining techniques. In [5] to classify the medical data set a neural network approach is adopted. The neural network is trained with breast cancer data base by using feed forward neural network model and back propagation learning algorithm with momentum and variable learning rate. The performance of the network is evaluated. The experimental result shows that by applying parallel approach in neural network model yields efficient result. In

[6] the Authors Abdelghani Bellaachia & Erhan Guven have performed an analysis of the prediction of survivability rate of breast cancer patients using three data mining techniques the Naïve Bayes, the back-propagated neural network, and C4.5 decision tree algorithms using the Weka toolkit . In [7] Author proposed a parallel approach which used neural network to help in the diagnosis of breast cancer. The neural network is trained with breast cancer data by using feed forward neural network model and back propagation learning algorithm with momentum and variable learning rate. The performance of the network is evaluated. In [8] the authors have explored the applicability of decision trees to do find a group with high susceptibility of suffering from breast cancer. The goal was to find one or more leaves with a high percentage of cases and small percentage of controls. Delen et al, in their work, have developed models for predicting the survivability of diagnosed cases using SEER breast cancer dataset [8] two algorithms artificial neural network (ANN) and C5 decision tree were used to develop prediction models.

4. METHODOLOGY

This work consists of two phases. In the first phase SVM and Decision Tree algorithm have been used which classify the breast cancer's patients into two classes Malignant and Benign. In the second phase K-Means clustering technique has been used which partitions the above two classed of patients into three clusters i.e. Poor, Intermediate and Good. Poor group is the most crucial group where chemotherapy can possibly enhance their survival.

4.1 Proposed Architecture

This proposed work is a two stage process. In this stage SVM algorithm and decision tree algorithms are applied to form two classes i.e. Benign and Malignant and then performance is evaluated of both the data mining classification techniques of accuracy. The better technique will be used for the next stage to take place and then k-means clustering technique has been used which partitions the above two classed of patients into three clusters i.e. Poor, Intermediate and Good.

4.2 Support Vector Machine

Support Vector Machine is an important classification approach which is very popular in present days. Classification is an organized approach for differentiating among the data items for handling large data volume. Using classification we can compare between similar and dissimilar properties of the data set. Classification is a technique for data mining that is used to forecast the membership of data items in a group. In 1995, Vapnik V and his colleagues presented Support Vector Machine (SVM). It is used for both classification and regression approaches. SVM is excellent in handling non-linear problems .It uses nonlinear map from original data input space to feature space. Due to structural risk minimization principle, SVM has a good generalization performance [10] [11]. SVM is a supervised learning method used to examine dataset and identify patterns. In the present era healthcare monitoring becomes important issue, so a model has been constructed that provides better quality of life [13].

Steps for Training and Testing Our Classification with the Help of Svm

1. First is loading of data set.
2. Then model the whole training data.
3. Testing of the training data
4. Get classified dataset

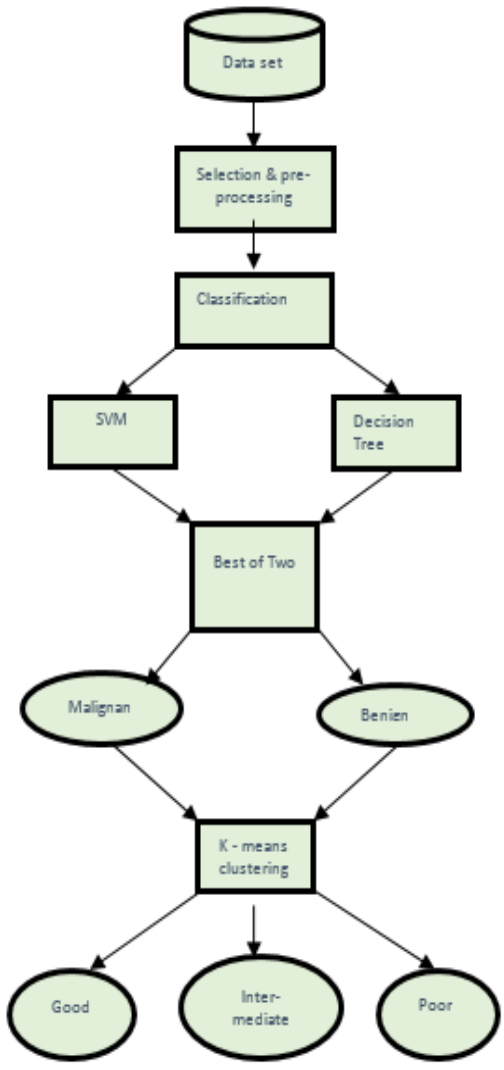

Fig 1: Block Diagram

4.3 K-Means Clustering

K-Means algorithm partitions a set of data objects into k clusters so that the inter-cluster similarity is low but the inter-cluster similarity is high [15]. The algorithm is shown as

1. Randomly select k data objects as the initial centroids.

2 .Assign all data objects to the closest (the most similar) centroid.

3. Recomputed the centroid of each cluster.

4. Repeat steps 2 and 3 until the centroids do not change.

K-Means algorithms are relatively efficient and scalable. The complexity of both algorithms is linear in the number of documents. In addition, they are so easy to implement that they are widely used in different clustering applications. K-Means is easy to implements and has several other flexibilities but still it has many disadvantages due to which, it is not used for large documents datasets [13].

4.4 Database Representation

Breast cancer becomes one of the leading causes of death of women in the world. s. In this research, a medical data based on breast cancer attributes was used for the purpose of classification between two types of cancers, benign and malignant.

4.4.1 Breast Cancer Database

The database used in our study is the UCI breast cancer database which has been taken from UCI repository. The same database has been used by researchers for the purpose of classification and testing algorithms in the world of data mining. 699 patients form the total available database.11 features or attributes represent the data for each patient, which are nine cytological characteristics of breast fine- needle aspirates and two other attributes contain the id number of each patient and the class label, that correspond to the type of breast cancer (benign or malignant). The values of cytological characteristics are in a range from one to ten.

5. TECHNICAL ANALYSIS

In technical analysis, Support Vector Machines (SVM) and Decision tree has been applied on pre-processed data which will classify it into two classes i.e. Benign and Malignant. Best technique among two will be considered on the basis of accuracy. Later k –means clustering technique is applied on the pre-processed data sets of two classes Benign and Malignant to get three clusters Poor, Intermediate and Good. By applying this clustering technique, the patients who require chemotherapy can be easily identified.

Phase 1

In this stage SVM algorithm and decision tree algorithms are applied to form two classes i.e. Benign and Malignant and then performance is evaluated of both the data mining classification techniques of accuracy. The better technique will be used for the next stage to take place.
Section one: In this section of phase one we applied the SVM algorithm to classify the dataset in to two classes i.e. Benign and Malignant by applying SVM.

Confusion Matrix:

	Benign	Malignant
Benign	58	1
Malignant	1	41

On the basis of this confusion matrix we have calculated the accuracy of svm algorithm and the accuracy of this classification, is 98%.
Section two: In next section of this step we applied the decision tree algorithm to classify the dataset of cancer patients.

Confusion Matrix:

	Benign	Malignant
Benign	55	1
Malignant	3	41

Accuracy of this classification is 96%. Above two figure shows that the accuracy given by SVM is 98% and accuracy given by Decision Tree is 96%. Hence on the basis of accuracy, SVM is the best technique to classify the breast cancer patients in to two class i.e. Benign and Malignant.

Table: Database Parameters Information

ATTRIBUTE	DOMAIN
1.Sample code number	Id number
2.Clump thickness	1-10
3.Cell size	1-10
4.Cell shape	1-10
5.Marginal Adhesion	1-10
6.Single Epithelial Cell Size	1-10
7.Bare Nuclei	1-10
8.Bland Chromatin	1-10
9.Normal Nucleoli	1-10
10.Mitoses	1-10
11.Class	2 for benign & 4 for malignant

Phase 2

In the second phase k-means clustering technique has been used which partitions the above two classed of patients into three clusters i.e. Poor, Intermediate and Good. Poor group is the most crucial group where chemotherapy can possibly enhance their survival.

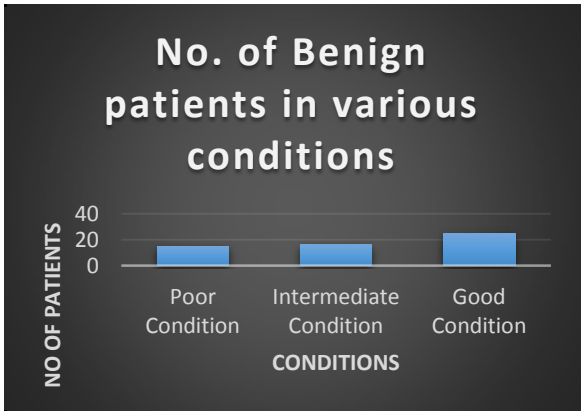

Fig 2: Benign Patients lie in certain group

Fig 2 shows the number of patients of Benign Class which lies in the following groups.

Fig 3 shows the number of patients of Malignant Class lies in the following groups.
This research work provides solution for the medical practitioners to identify the patients which are in urgent need of chemotherapy before wasting lot of time in conducting tests and then get the final diagnosis.

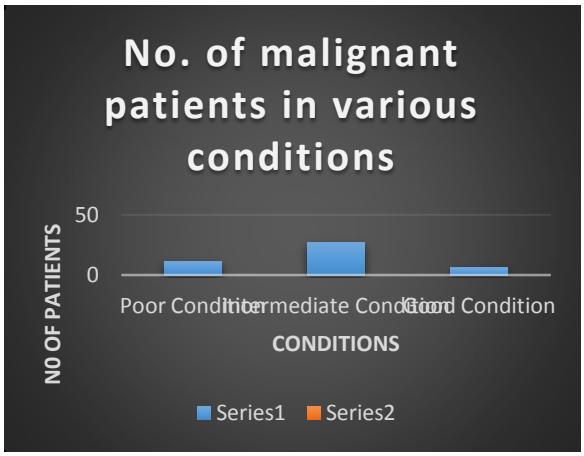

Fig 3: Malignant Patients lie in certain group

6. CONCLUSION AND FUTURE WORK

This research work provides solution for the medical practitioners to identify the patients which are in urgent need of chemotherapy before wasting lot of time in conducting tests and then get the final diagnosis. This two stages research work is highly successful in both stages. In first stage, SVM and Decision Tree have been used to classify the patients into two classes Benign and Malignant. On the basis of Accuracy of both the techniques, best one is used i.e. SVM which gave the 98% accuracy. And in second stage final clusters have been made with the help of k-means algorithm i.e. Poor, Intermediate and Good to determine whether the patient is in urgent need of chemotherapy with respect to the survival time of the patient.

REFERENCES

[1] Breast Cancer statistics from Centers for Disease Control and Prevention,

[2] http://www.cdc.gov/cancer/breast/statistics/.

[3] D. M. Parkin, F. Bray, J. Ferlay, "Global cancer statistics 2002," CA Cancer J Clin, vol.55, pp. 74-108, 2005.

[4] http://www.breastcancerindia.net/bc/statistics/stat_global.htm.

[5] Goharian & Grossman, (2003) "Data Mining Classification", Illinois Institute of Technology,http://ir.iit.edu/~nazli/cs422/CS422-Slides/DM-Classification.pdf.

[6] Lundin M, Lundin J, Burke HB, Toikkanen S, Pylkkanen L, Joensuu H. Artificial neural networks applied to survival prediction in breast cancer. Oncology 1999; 57:281—6.

[7] Abdelghani Bellaachia, Erhanguven, Predicting Breast cancer survivability using Data Mining Techniques.

[8] Abdelaal Ahmed Mohamed Medhat and Farouq Wael Muhamed, "Using data mining for assessing diagnosis of breast cancer," in Proc. International multiconference on computer science and information Technology, 2010, pp. 11-17.

[9] Maria-Luiza Antonie, Osmar R. Za¨ıane, Alexandru Coma, .Application of Data Mining Techniques for Medical Image Classification. Proceeding of second International workshop on Multimedia data mining (MDM/KDD'2001), in conjunction with ACM SIGKDD conference. SAN FRANCISCO, USA, AUG 26, 2001.

[10] Orlando Anunciac¸ ˜ao and Bruno C. Gomes and Susana Vinga and Jorge Gaspar and Arlindo L. Oliveira and Jos´e Rueff , A Data Mining Approach for the detection of High-Risk Breast Cancer Groups.

[11] V. Vapnik, (1998), Statistical Learning Theory, Wiley.

[12] V. Vapnik, (1998). "The support vector method of function estimation".

[13] X. Rui, "Survey of Clustering Algorithms", IEEE Transactions on Neural Networks, 16 (3), pp. 634-678, 2005.

[14] N. Chistianini & J. Shawe-Taylor, (2000). "An Introduction to Support Vector Machines, and other kernel-based learning methods", Cambridge University Press, 2000.

[15] N. Cristianini, & Shawe-Taylor, J. (2000) "An Introduction to Support Vector Machines". Cambridge University Press.

[16] L. Kaufman and P. J. Rousseeuw, "Finding Groups in Data: An Introduction to Cluster Analysis", John Wiley and Sons, March 1990.

[17] Delen, G. Walker, A. Kadam, "Predicting breast cancer survivability: comparison of three data mining methods," Artificial Intelligence in Medicine, vol. 34, pp. 113-127, 2005.

[18] D. Delen, G. Walker, A. Kadam, "Predicting breast cancer survivability: comparison of three data mining methods," Artificial Intelligence in Medicine, vol. 34, pp. 113-127, 2005.