International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
Since 2012

http://www.ijcsjournal.com Volume 5, Issue 1, No 14, 2017 ISSN: 2348-6600
Reference ID: IJCS-221 PAGE NO: 1379-1385

Sri Vasavi College, Erode Self-Finance Wing 3rd February 2017
National Conference on Computer and Communication NCCC'17
http://www.srivasavi.ac.in/ nccc2017@gmail.com

COMPARATIVE STUDY OF CANCER PATIENT BY USING DATAMINING TECHNIQUES

M.karthika[1], S.Muruganandam[2],
M.Phil Full time Research Scholar, PG and Research Department of Computer Science[1]
Asst. Professor, Department of Computer Science [2]
Vivekanandha College of Arts & Sciences for Women (Autonomous),
Tiruchengode, Namakkal-637 205,
Tamil Nadu, India
kannankars07@gmail.com
mr.smanand@gmail.com

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. 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. We have investigated more of data mining techniques: the Naïve Bayes, the back-propagated neural network, and the C4.5 decision tree algorithms. Several experiments were conducted using these algorithms. The achieved prediction performances are comparable to existing techniques. However, we found out that C4.5 algorithm has a much better performance than the other techniques.

KEYWORDS- Clustering, SVM, decision tree, k-means, classification, data mining, Breast cancer survivability, data mining, SEER, Weka.

1. INTRODUCTION

Day by day increasing growth in science and technology, quality of human life is improving. 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.

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. 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. 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.

The discovery of the survival rate or survivability of a certain disease is possible by extracting the knowledge from the data related to that disease. One of these data sources is SEER (Surveillance Epidemiology and End Results), which is a unique, reliable and essential resource for investigating the different aspects of cancer. The SEER database combines patient-level information on cancer site, tumor pathology, stage, and cause of death.

In this paper, we present data mining techniques to predict the survivability rate of breast cancer patients. In our study, we have used the SEER data and have introduced a pre-classification approach that take into account three variables: Survival Time Recode (STR), Vital Status Recode (VSR), and Cause of Death (COD).

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. 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 tasks. 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.

3. RELATED WORK

In this section, we review the related work on breast cancer diagnosis using data mining techniques. In 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 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 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 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. Dalen et al, in their work, have developed models for predicting the survivability of diagnosed cases using SEER breast cancer dataset two algorithms artificial neural network (ANN) and C5 decision tree were used to develop prediction models.

In our discussion with the authors of, we found out that the pre-classification process was not accurate in determining the records of the "not survived" class. They did take into consideration neither the Vital Status Recode (VSR), nor the Cause of Death (COD). They assume that all patients are dead with cancer, which is not always true. In our study, we have used a newer version of SEER database (period of 1973-2002 with 482,052 records) and, unlike, we have included two other fields in the pre-classification process: ƒ Survival Time Recode (STR), Vital Status Recode (VSR), ƒ Cause of Death (COD) The next section presents our pre classification process.

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 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.

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. 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. Good, Poor group is the most crucial group where chemotherapy can possibly enhance their survival.

4.1 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. The algorithm is shown as

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

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 centroid 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.

4.2 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.

4.3 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.7 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
58	1
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
55	1
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 classes i.e. Benign and Malignant.

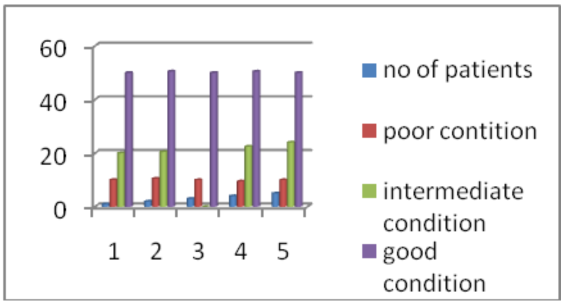
Table: Database Parameters Information

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

Phase 2

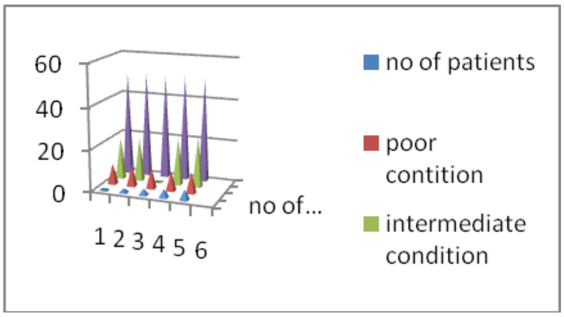
In the second phase k-means clustering technique has been used which partitions the above two classed of patients into three clusters i.e. Poor, Intermediate and Good. Poor group is the most crucial group where chemotherapy can possibly enhance their survival

No. of Benign patients in various conditions



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

No. of malignant patients in various conditions



6. EXPERIMENTAL RESULTS

In this study, the accuracy of three data mining techniques is compared. The goal is to have high accuracy, besides high precision and recall metrics. Although these metrics are used more often in the field of information retrieval, here we have considered them as they are related to the other existing metrics such as specificity and sensitivity. These metrics can be derived from the confusion matrix and can be easily converted to true-positive (TP) and false-positive (FP) metrics. The experimental results of our approach as presented in Table

Table 4: Combined Results (our study)

Classification Technique	Accuracy (%)	Class	Precision	Recall
Naïve Bayes	84.5	0 1	0.70 0.88	0.57 0.94
Artificial Neural Net	86.5	0 1	0.84 0.87	0.52 0.97
C4.5	86.7	0 1	0.80 0.88	0.56 0.96

Table 5: Results for C4.5 (dataset as in Table3)

Classification Technique	Accuracy (%)	Class	Precision	Recall
C4.5	81.2	0 1	0.86 0.76	0.81 0.81

7. CONCLUSIONS AND FUTURE WORK

This paper has outlined, discussed and resolved the issues, algorithms, and techniques for the problem of breast cancer survivability prediction in SEER database. Unlike the pre-classification process used in our approach takes into consideration, besides the Survival Time Recode (STR), the Vital Status Recode (VSR) and Cause of Death (COD). The experimental results show that our approach outperforms the approach used in.

This study clearly shows that the preliminary results are promising for the application of the data mining methods into the survivability prediction problem in medical databases. Our analysis does not include records with missing data; future work will include the missing data in the EOD field from the old EOD fields prior to 1988. This might increase the performance as the size of the data set will increase considerably. Finally, we would like to try survival time prediction of certain cancer data such as respiratory cancer where the survivability is seriously low. We think of discrediting the survival time in terms of one year and then classifying using the aforementioned data mining algorithms.

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] Delen D, Walker G, Kadam A. Predicting breast cancer survivability: a comparison of three data mining methods. Artificial Intelligence in Medicine. 2005 Jun; 34(2):113-27.
[10] Ian H. Witten and Eibe Frank. Data Mining: Practical machine learning tools and techniques, 2nd Edition. San Fransisco:Morgan Kaufmann; 2005
. [11] J. R. Quinlan, C4.5: Programs for Machine Learning. San Mateo, CA: Morgan Kaufmann; 1993.
[12] Weka: Data Mining Software in Java, http://www.cs.waikato.ac.nz/ml/weka/