ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 3, March 2015

Data Mining Techniques For Diagnosis And Prognosis Of Cancer

Jaimini Majali[1], Rishikesh Niranjan[2], Vinamra Phatak[3], Omkar Tadakhe[4]

Department of Computer Engineering P.E.S. MCOE, Pune[1,2,3,4]

Abstract: In this paper we are using data mining techniques for diagnosis and prognosis of cancer. Cancer is the most important cause of death for both men and women. The early detection of cancer can be helpful in curing the disease completely. So the requirement of techniques for the detection of cancer in early stage is increasing. Breast cancer is one of the leading cancers for women in developed countries including India. It is the second most common cause of cancer death in women. The high incidence of breast cancer in women has increased significantly in the last years. The malignant tumor develops when cells in the breast tissue divide and grow without the normal controls on cell death and cell division. Hence, cancer on breast tissue is called breast cancer. Worldwide, it is the most common form of cancer in females that is affecting approximately 10% of all women at some stage of their life. With early diagnosis, 97% of women survive for 5 years or more years. In this paper we present a system for diagnosis and prognosis of cancer using Classification and Association approach in Data Mining. We are using FP algorithm in Association Rule Mining (ARM) to conclude the patterns frequently found in benign and malignant patients. We are also using Decision Tree algorithm under classification to predict the possibility of cancer in context to age.

Keywords: Classification, Association, Frequent Pattern growth, Decision tree, Breast cancer, benign, malignant.

I. INTRODUCTION

In this paper we intend to present a system for diagnosis and prognosis of cancer disease using data mining techniques. We are using Frequent Pattern [1] algorithm under Association Rule Mining (ARM) and ID3 algorithm in decision tree under Classification in this system. Diagnosis of cancer is very important as detection of cancer at early stage can help in proper treatment for the cancer patient. Thus this system is very helpful in medical research. The aim is to assist doctors in diagnostic decisions.

Fp-growth algorithm is used for generation of frequency itemset without candidate generation thus improves performance of algorithm. This method make used of Divide and Conquer strategy. This take place in 2 steps:-

Step 1: It compresses the input database showing frequency item-set into fp-tree.Fp-tree is build using 2 passes on dataset.
Step 2: It then divide fp-tree into set of conditional dataset and mines them separately. Thus extract the frequency item set from fp-tree.

A decision tree is a kind of flowchart where each internal and node (non-leaf node) denotes a test on an attribute, each branch represents an outcome of the test and each leaf node (terminal node) holds a class label. Decision trees are most commonly used as the construction of decision tree classifiers does not require any domain knowledge or parameter setting. They can handle multidimensional data and are simple and fast. There are many Decision tree algorithms such as HUNTS algorithm (this is one of the earliest algorithm), CART, ID3, C4.5 (a later version ID3 algorithm), SLIQ, SPRINT.

Over the centuries, new medical developments and techniques have changed the face of healthcare. The medical field has always brought together the best and brightest of society to help those in need. From treating cancer and delivering babies to dealing with heart attacks, doctors have developed technology and improved techniques.

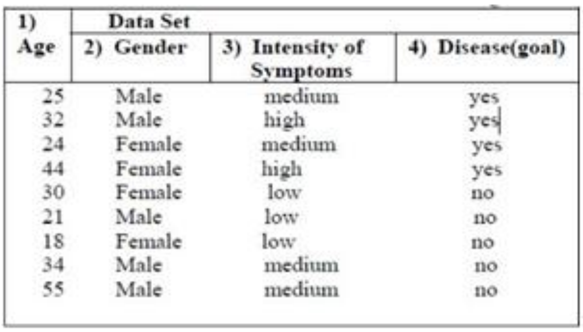
The most important thing in medical field is early diagnosis of any disease which helps to cure it in early stage and increase the life expectancy chances. Cancer is one of such disease where early diagnosis can reduce the death rate in cancer patients. There are generally two types of cancer:-
1) Benign cancer and
2) Malignant Cancer.

If cancer is detected in benign phase life expectancy of a patient increases.Breast cancer is one of the leading cancers for women in developed countries including India. It is the second most common cause of cancer death in women. The malignant tumor develops when cells in the breast tissue divide and grow without the normal controls on cell death and cell division. Hence, cancer on breast tissue is called breast cancer. Worldwide, it is the most common form of cancer in females that is affecting approximately 10% of all women at some stage of their life. Although scientists do not know the exact causes of most breast cancer, they do know some of the risk factors that increase the likelihood of a woman developing breast cancer. These factors include such attributes as age, genetic risk and family history. Although breast cancer is the second leading cause of cancer death in women, the survival rate is high. With early diagnosis,97% of women

DOI 10.17148/IJARCCE.2015.43147

survive for 5 years or more years. The two main type's whereas chemotherapy and hormone therapy are systematic therapies[2].

II. LITERATURE SURVEY



1) Age	Data Set		
	2) Gender	3) Intensity of Symptoms	4) Disease(goal)
25	Male	medium	yes
32	Male	high	yes
24	Female	medium	yes
44	Female	high	yes
30	Female	low	no
21	Male	low	no
18	Female	low	no
34	Male	medium	no
55	Male	medium	no

Figure 2.1: A decision tree built from the data in Table 1

Data mining techniques can be used to predict cancer in a patient using various symptoms data from previous results. Valuable knowledge can be discovered through this data mining techniques. In this paper we are using Association Rule Mining (ARM) and Classification for diagnosis and prognosis of cancer. Under ARM we are using FP growth algorithm which is applied on following attributes of patient data to predict benign or malignant. There are 699 records in this database. Each record in the database has nine attributes. The nine attributes detailed in Table 1 are graded on an interval scale from a normal state of 1–10, with 10 being the most abnormal state[3][6].

Attribute	Domain
Sample Code Number	**Id**
Clump Thickness	1-10.
Uniformity of cell size	1-10.
Uniformity of cell shape	1-10.
Marginal adhesion	1-10.
Single epithelial Cell Size	1-10.
Bare nuclei	1-10.
Bland Chromatin	1-10.
Normal nucleoli	1-10.
Mitosis	1-10.
Class	(2 for benign, 4 for malignant)

Under classification, we are using decision tree for prognosis of cancer. Following is an example to understand the decision tree approach.The decision tree shown in Fig 2 is built from the very small training set. In this table each row corresponds to a patient record. We will refer to a row as a data instance. The data set contains three predictor attributes, namely Age, Gender, Intensity of symptoms and one goal attribute, namely disease whose values (to be predicted from symptoms) indicates whether the corresponding patient have a certain disease or not.

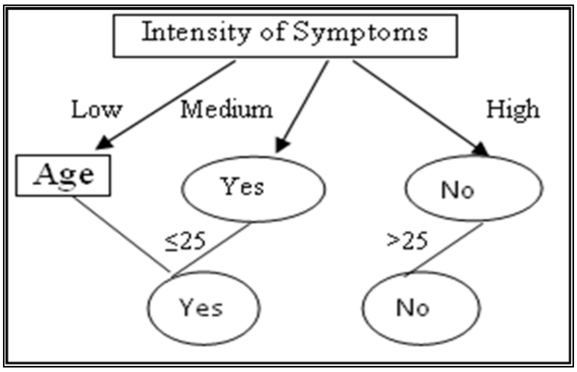


Figure 2.1: A decision tree built from the data in Table 1

Decision tree can be used to classify an unknown class data instance with the help of the above data set given in the Table 2. The idea is to push the instance down the tree, following the branches whose attributes values match the instances attribute values, until the instance reaches a leaf node, whose class label is then assigned to the instance. For example, the data instance to be classified is described by the tuple (Age=23, Gender=female, Intensity of symptoms = medium, Goal =?), where "?" denotes the unknown value of the goal instance. In this example, Gender attribute is irrelevant to a particular classification task. The tree tests the intensity of symptom value in the instance.

If the answer is medium; the instance is pushed down through the corresponding branch and reaches the age node. Then the tree tests the Age value in the instance. If the answer is 23, the instance is again pushed down through the corresponding branch. Now the instance reaches the leaf node, where it is classified as yes.

III. PROPOSED SYSTEM

We intend to develop a system using data mining techniques for diagnosis and prognosis of cancer disease. We are developing our system with two approaches of Data Mining association rule mining and classification techniques. We are applying frequent pattern growth algorithm for diagnosis of cancer. Frequent pattern mining searches for recurring relations in a given data set. We are using Wisconsin data set which consists of attributes as mentioned above. These attributes form the input data for FP-Growth algorithm. The problem of finding association rule is divided into two sub problems [3]:-

1. Support.
2. Confidence.

- Support (s): It is an indication of item how frequently it occurs or finding the frequency item sets for ex: -

DOI 10.17148/IJARCCE.2015.43147

Consider the rule A=>B it support if it include A and B together. (I.e. A U B.)

$$Support(A => B) = Support = P(AUB)$$

- Confidence (c): No of times the statement is found to be true. For ex: - Consider the rule A=>B it Confidence if it include the above A together with B.

$$Confidence(A => B) = \frac{Support(AUB)}{Support(A)} = P\left(\frac{B}{A}\right)$$

In our system Fp growth algorithm scans whole data for different values of support and confidence and frequently found patterns are mapped to form the rules. These rules can be helped to analyse which attributes with what values can classify a tumor as benign or malignant.

Also we are developing a dynamic application based on decision tree as the classification technique which can be useful for the normal citizens to get awareness regarding cancer and its causes and what habits can lead to cancer as well as the prevention measures that can be taken so that cancer can be detected at early stage. As if cancer is detected at an early stage chances of survival are high. In this, we will be presenting a questionnaire to people, each question related to cancer causing habits. Questions are provided with certain options which will help to identify the nature of the user. These answers are applied as input data to ID3 algorithm in decision tree technique to classify the user as having possibility yes or no. A statistical property, called information gain, is used in ID3 algorithm. Gain measures how well a given attribute separates training examples into targeted classes. The one with the highest information (information being the most useful for classification) is selected. In order to define gain, we first borrow an idea from information theory called entropy. Entropy measures the amount of information in an attribute.

Given a collection S of c outcomes
Entropy(S) = S -p(I) log2 p(I)
Where p(I) is the proportion of S belonging to class I. S is over c. Log2 is log base 2.
Note that S is not an attribute but the entire sample set. Gain(S, A) is information gain of example set S on attribute A is defined as
Gain(S, A) = Entropy(S) - S (($|S_v|$ / |S|) * Entropy(S_v))
Where:

S is each value v of all possible values of attribute A
S_v = subset of S for which attribute A has value v
$|S_v|$ = number of elements in S_v
|S| = number of elements in S
By applying decision tree we will be able to predict whether the person may have risk of getting cancer or the person is fit.

IV. CONCLUSION

This paper provides a study of various technical and review papers on breast cancer diagnosis and prognosis problems and explores that data mining techniques offer great promise to uncover patterns hidden in the data that can help the clinicians in decision making. From the above study it is observed that the accuracy for the diagnosis analysis of various applied data mining classification techniques is highly acceptable and can help the medical professionals in decision making for early diagnosis and to avoid biopsy.

We have applied Wisconsin data set to fp growth algorithm and obtained the rules which indicate the general behavior and range of values for malignant and benign tumor. The comparison between ranges of values for malignant and benign tumor are as specified below.

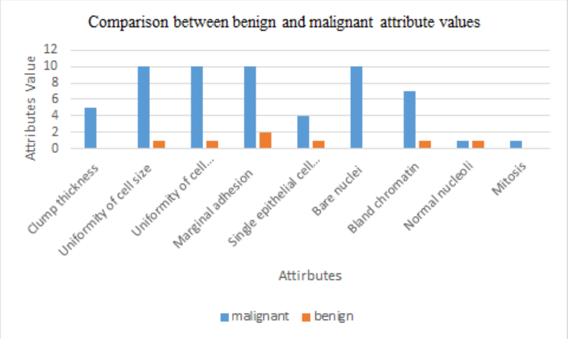


Figure 4.1: Comparison between benign and malignant attributes values

Following graph shows variation in execution time with change in support value for fixed value of confidence.

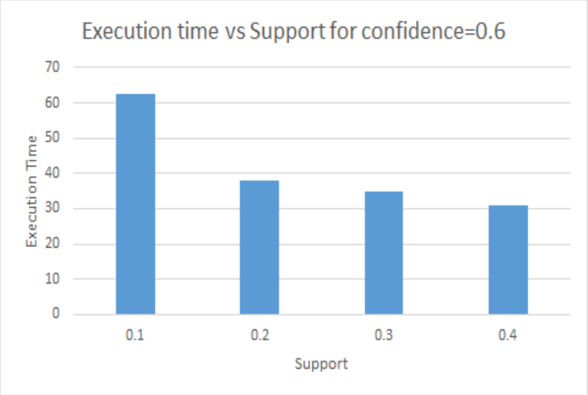


Figure 4.2: Execution time vs Support for confidence=0.6

Among the various data mining classifiers and soft computing approaches, Decision tree is found to be best predictor on Wisconsin dataset. With this algorithm applied to Wisconsin dataset 94% class-labels were predicted correctly.

REFERENCES

[1] Jaimini Majali, Rishikesh Niranjan, Vinamra Phatak, Omkar Tadakhe, "Data Mining Techniques For Diagnosis And Prognosis Of Breast Cancer", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (5) , 2014, 6487-6490

[2] V.Krishnaiah, Dr.G.Narsimha, Dr.N.Subhash Chandra, "Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques", International Journal of Computer Science and Information Technologies, Vol. 4 (1) , 2013, 39 – 45

[3] Shweta Kharya, "Using data mining techniques for diagnosis and prognosis of cancer", International Journal of Computer Science,

Engineering and Information Technology (IJCSEIT), Vol.2, No.2, April 2012.

[4] Rakesh Agrawal, Ramakrishnan Srikant, “Fast Algorithms for Mining Association Rules”, Proceedings of the 20th VLDB Conference Santiago,Chile,1994

[5] Kuldeep Malik, Neeraj Raheja, Puneet Garg, “ENHANCED FPGROWTH ALGORITHM”, (IJCEM) International Journal of Computational Engineering and Management, Vol.12, April 2011.

[6] M. Karabatak, M.Cevdet, “An Expert System for detection of breast cancer based on association rule and neural network”, Expert Systems with Applications 36 (2009) 3465–3469

[7] D.Lavanya, Dr. K.Usha Rani, “Performance Evaluation of Decision Tree Classifiers on Medical Datasets”, International Journal of Computer Applications (0975 – 8887) Volume 26– No.4, July 2011.

[8] Shomona Gracia Jacob, R. Geetha Ramani, “Efficient Classifier for Classification of Prognostic Breast Cancer Data through Data Mining Techniques”, Proceedings of the World Congress on Engineering and Computer Science 2012 Vol I WCECS 2012, October 24-26, 2012, San Francisco, USA

[9] Shelly Gupta, Dharminder Kumar, Anand Sharma, “Data mining classification techniques applied for breast cancer diagnosis and prognosis”, Indian Journal of Computer Science and Engineering (IJCSE).

[10] J Han, M Kamber, “DATA MINING: - Concepts and Techniques.

[11] Yangon Kim, Yuncheol Baek, “Analysis of breast cancer using data mining & statistical techniques”, Pages 82-87, 23-25 may 2005.

[12] O.L. Mangasarian, W.N. Street and W.H. Wolberg. Breast cancer diagnosis and prognosis via linear programming. Operations Research, 43(4), pages 570-577, July-August 1995

DOI 10.17148/IJARCCE.2015.43147