ISSN:0975-9646

Data Mining Techniques For Diagnosis And Prognosis Of Breast Cancer

Jaimini Majali
Department of Computer Engineering
P.E.S. MCOE, Pune.

Rishikesh Niranjan
Department of Computer Engineering
P.E.S. MCOE, Pune.

Vinamra Phatak
Department of Computer Engineering
P.E.S. MCOE, Pune.

Omkar Tadakhe
Department of Computer Engineering
P.E.S. MCOE, Pune.

Abstract– **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 tumour 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 FP (frequent pattern mining) growth algorithm. We are also using Decision Tree algorithm to predict the possibility of cancer in context to age. We are using FP algorithm to conclude whether the tumour is malignant or benign tumour.**

Keywords: - Frequent Pattern growth, Apriori, Decision tree, Breast cancer, benign cancer, malignant cancer.

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] and decision tree algorithms 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.

The early diagnosis will need an accurate and reliable diagnosis methodology for the physicians to distinguish between benign tumour and malignant tumour without doing a surgical biopsy.

Thus the objective is to assign the patients with either benign category (noncancerous) or malignant category (cancerous).

Knowledge discovery and data mining have found numerous applications in business and scientific domain. Valuable knowledge can be discovered from application of data mining techniques in healthcare system. In this study, we briefly examine the potential use of classification based data mining techniques such as Rule based, Decision tree, Naïve Bayes and Artificial Neural Network to massive volume of healthcare data. Cancer research is generally clinical and/or biological in nature, data driven statistical research has become a common complement. Predicting the outcome of a disease is one of the most interesting and challenging tasks where to develop data mining applications. 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 case stored within datasets. The objective of this study is to summarize various review and technical articles on diagnosis of breast cancer. It gives an overview of the current research being carried out on various breast cancer datasets using the data mining techniques to enhance the breast cancer diagnosis [1].

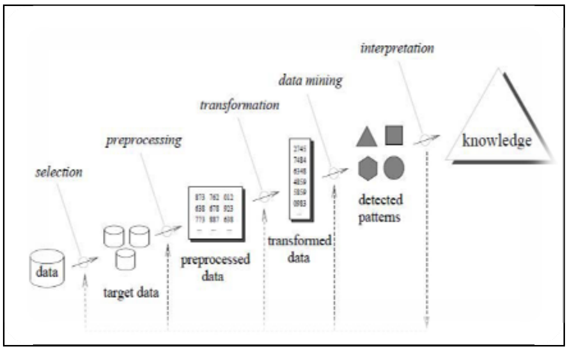


Figure 1: Steps in KDD

II. ASSOCIATION RULE MINING

Association rule mining aims at discovering associations between items in a transactional database. Given a set of transactions $D = \{T1, \dots\dots, Tn\}$ and a set of items $I = \{i1, \dots\dots, im\}$ such that any transaction T in D is a set of items in I, an association rule is an implication of the form A→ B where the antecedent A and the consequent B are subsets of a transaction T in D, and A and B have no common items. For the association rule to be acceptable, the conditional probability of B given A has to be higher than a threshold called minimum confidence. Association rules mining is a two-step process, in the first step frequent item-sets are generated (i.e. item-sets whose support is no less than a minimum support) and in the second step association rules are derived from the frequent item-sets obtained in the first step.

Association rule of data mining is used to find the relationship between data and database.ARM concept was introduce by Agrawal. It consist of two parts “Antecedent” and “Consequent”. Antecent is the item which is found in database and Consequent is found in relation with first [2]. For ex: $\{bread\} => \{jam\}$ here the left hand side is called Antecedent and right hand side is called as Consequent. This rule states that in sales suppose a customer buys a product (I.e. bread) then he is likely to buy another product (i.e. jam).this is a marketing strategy. Association rule is also used in bioinformatics, web usage mining.

The problem of finding association rule is divided into two sub problems:-

1. Support.
2. Confidence.

- Support (s): it is an indication of item how frequently it occurs or finding the frequency item sets for ex: - 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.
  $$Confidence(A => B) = \frac{Support(AUB)}{Support(A)}$$
  $$= P\left(\frac{B}{A}\right)$$

A. Apriori algorithm

Apriori is Latin word which mean “from what comes before”. It uses bottom up strategy. It uses Breadth first search Method (B.F.S).It is used in various treatment of disease like Cancer. It is the most commonly used Method in Frequent pattern mining. This method find frequency itemset using candidate generation method. In this itemset are sorted in lexicographic manner.Apriori property is any subset of Frequency itemset is always frequent.

Pseudo Code for Apriori algorithm:

1. Join Step: Ck is generated by joining Lk-1 with itself.
2. Min support: It is the minimum support used for searching frequency pattern that satisfy this constraint.
3. Min Confidence: It is used for finding the strong association rule that satisfy this threshold.
4. Prune Step: If (k-1) is not frequent itemset then the subset of (k) is also not frequent.

```
  Ck: - Candidate itemset of size k.
  Lk :-Frequency itemset of size k.
  L1 ={Frequency items};
  For (k=1;Lk!=0;k++) do begin
    Ck+1 =Candidate generated from Lk.
    For each transaction‘d’ in database do increment the count of candidate in Ck+1 that contained in d.
    Lk-1 = candidate in Ck-1 with min support
    End
Return L1 U L2………Lk; [3]
```

B. Fp-growth algorithm

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 itemset into fp-tree.fp-tree is build using 2 passes on dataset.

Step 2: It then divide fop-tree into set of conditional dataset and mines them separately. Thus extract the frequency item set from fp-tree.

Fp-Tree Structure is as shown below:-

i. One root labelled as “null” with a set of item-prefix sub trees as children, and a frequent-item-header table.
ii. Each node in the item-prefix sub tree consists of three fields:-
    1. Item-name: registers which item is represented by the node.
    2. Count: the number of transactions represented by the portion of the path reaching the node.
    3. Node-link: links to the next node in the FP-tree carrying the same item-name, or null if there is none.
iii. Each entry in the frequent-item-header table consists of two fields:-
    1. Item-name: as the same to the node;
    2. Head of node-link: a pointer to the first node in the FP-tree carrying the item-name.

```
PSEUDO CODE FOR FP-GROWTH ALGORITHM
Input: constructed FP-tree.
Output: complete set of frequent patterns.
Method: Call FP-growth (FP-tree, null).
Procedure FP-growth (Tree, α)
  {
  1) if Tree contains a single path P then
  2) for each combination do generate pattern β U α with support = minimum support
  Of nodes in β.
  3) Else for each header ai in the header of Tree do {
  4) Generate pattern β = aiUα with support = ai.support;
  5) Construct β.s conditional pattern base and then β.s conditional FP-tree Tree β
  6) If Tree β = null
  7) Then call FP-growth (Tree β, β)}
  } [4]
```

C. COMPARISION BETWEEN APRIORI AND FP-GROWTH

Table 1: Comparison between FP growth and Apriori algorithms

Parameter	Apriori Algorithm	FP-Growth Algorithm
Technique	It uses Apriori Property, join and Prune Property	It construct Conditional frequent Pattern Tree and Conditional pattern Base from database which satisfy minimum support
Memory utilisation	Due to candidate generation memory utilization is more.	As no Candidate are generated so memory utilisation is less.
No of Scans	Multiple scans for generating candidate sets.	Scan the database twice only.
Time Required	Time require is more as for generating candidate set require more time	Time require is less.
Efficiency	Less efficient	More Efficient

III. Data Mining Classification Methods

Data classification is a two-step process which consists of learning step and a classification step. The aim of classification is to construct classifiers that predict categorical class labels. Thus the classifier is build based on some cases with some attributes to describe the objects or one attribute to describe the group of the objects. The basic data mining classification techniques are: IF-THEN Rule, Decision tree, Bayesian classifiers and Neural Networks [5].

A. If-then rule

This rule consists of two parts. The IF part, the rule antecedent contains one or more conditions about value of predictor attributes and the other part the THEN part, rule consequent contains a prediction about the value of a goal attribute. An accurate prediction of the value of a goal attribute will improve decision-making process. IF-THEN prediction rules are very popular in data mining and they represent discovered knowledge at a high level of abstraction.

B. Neural networks

Neural networks (NN) are those systems modelled based on the human brain working. As the human brain consists of millions of neurons that are interconnected by synapses, a neural network is a set of connected input/output units in which each connection has a weight associated with it. The network learns in the learning phase by adjusting the weights so as to be able to predict the correct class label of the input.

C. Bayesian classifiers

In Bayesian technique, a classification problem can be written as the problem of finding the class with maximum probability given a set of observed attribute values. Such probability is seen as the posterior probability of the class given the data, and is usually computed using the Bayes theorem. Estimating this probability distribution from a training dataset is a difficult problem, because it may require a very large dataset to significantly explore all the possible combinations. Conversely, Naive Bayesian is a simple probabilistic classifier based on Bayesian theorem with the (naive) independence assumption. Based on that rule, using the joint probabilities of sample observations and classes, the algorithm attempts to estimate the conditional probabilities of classes given an observation. The Naive Bayes classifier is a robust method, which shows on average good performance in terms of classification accuracy, also when the independence assumption does not hold.

D. Decision tree algorithms

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.

The basic decision tree algorithm is as follows:

This algorithm is called with three parameters D(data partition), attribute_list and attribute_selection_method.

1. Create node N;
2. If tuples in D are all of the same class C,then
3. Return N as a leaf node labelled with the class C;
4. if attribute_list is empty then
5. return N as a leaf node labelled with the majority class in D;
6. apply attribute_selection_method(D, attribute_list) to find the “best” splitting_criterion;
7. label node N with splitting_criterion;
8. if splitting_attribute is discrete-valued and multiway splits allowed then
9. attribute_list-attribute_list – splitting_attribute;
10. for each outcome of j of splitting_criterion
11. let Dj be the set of data tuples in D satisfying outcome j;
12. if Dj is empty then
13. attach a leaf labelled with the majority class in D to node N;
14. else attach the node returned by Generate_decision_tree (Dj,attribute_list) to node N; endfor
15. return N;[6]

IV. Methodology

One of the common malpractices in medical field is failure in early diagnosis of any disease. 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 the leading cancer in developed countries including India. It is second most cause of cancer death in women. 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 [2][5].

	Attribute	Domain
1)	Sample code number	id
2)	Clump thickness	1-10.
3)	Uniformity of cell size	1-10.
4)	Uniformity of 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)	Mitosis	1-10.
11)	Class	(2 for benign, 4 for malignant).

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.

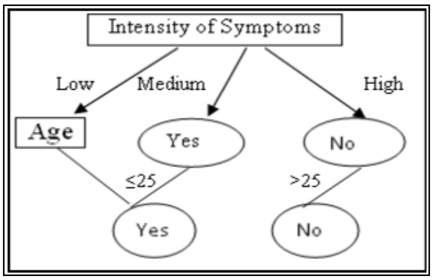
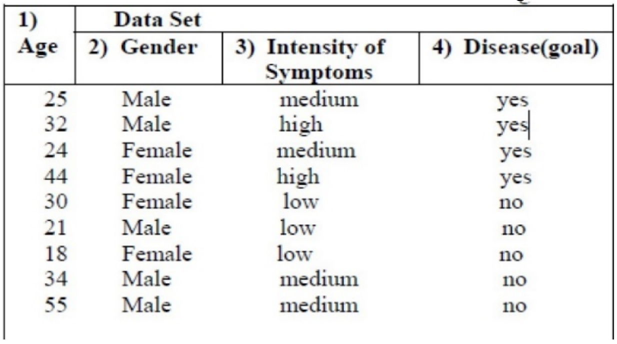


Fig. 2: A decision tree built from the data in Table 2

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 [1].

Table 2: Data set used to build decision tree of Fig. 2



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

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.

V. 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.

As compared to apriori algorithm fp-growth is more efficient as time required to execute is less than apriori and also memory utilisation is less in fp-growth. So we find that fp-growth is more efficient to use.

Among the various data mining classifiers and soft computing approaches, Decision tree is found to be best predictor on benchmark dataset (UCI machine learning dataset). In future the predictor can be used to design a web based application to accept the predictor variables and Automated system Decision Tree based prediction can be implemented in remote areas like rural regions or country sides, to imitate like human diagnostic expertise for prediction of ailment. In future we intend to design and implement such system for web based applications.

References

[1] 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

[2] 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.

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

[4] Kuldeep Malik, Neeraj Raheja, Puneet Garg, "ENHANCED FP-GROWTH ALGORITHM", (IJCEM) International Journal of Computational Engineering and Management, Vol.12, April 2011.

[5] 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

[6] 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.

[7] 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

[8] 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).

[9] J Han, M Kamber, "DATA MINING: - Concepts and Techniques.

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

[11] 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