DISEASE DIAGNOSIS USING ASSOCIATION RULE MINING BASED KNOWLEDGE INFERENCE SYSTEM

M. Sudha*
Faculty, School of Information Technology & Engineering, VIT University.
Email: msudha@vit.ac.in

Received on 09-08-2016
Accepted on 28-08-2016

Abstract

Data mining is the process of knowledge extraction from a bulk of data so that we can use it in future. This research is to identification of convoluted disorder and render reason of “why the person suffering from this disease” so that doctors can provide best solution to the patients. To discover the recurrent data that have affinity with each other is handling with association rule. We primarily focus on breast cancer disease that has the second place in the list of major cancer in women, after the lung cancer. According to the WHO report in every eight women, one woman has breast cancer and breast cancer is finding more in developing country rather than developed countries. Apriori algorithm is proposed to find the frequent item set and rule generation from the breast cancer data so that we can help other patients and as well as doctors to making decision on what particular reason that a patient got the breast cancer. The model is designed and implemented in C# language to generate rule and developing secured system that is very easy to use and efficient.

Keywords: Disease Diagnosis, Association rule mining and Web service.

1. Introduction

In the world of online transaction every person perform a lot of activity through OLAP/OLTP and the after completion of each transaction we find huge amount of data to store in the database or data warehouse. From a small health check up to a long term disease, many entries collected by database. Keeping the old data can be help for the patient or doctors for taking decision easily. Medical data is not as simple as other data because it consists of text, images, videos, numerical etc. It is multidimensional data that help to solve or find the solution of disease. Medical data is collected from different type of test like ECG, MRI, X-ray, CT scan, TEE etc. each of the test includes hundreds of attribute. The breast cancer data is used for this study. The dataset are obtained from German breast cancer group database and this dataset consists of 16 instances. On these data some data pre-processing concept is applied to
remove noise and other errors from data. Breast cancer is the most common cancer that is finding more in the age between 50 and 60. Chance of survive is very minimum compare to other cancer but if the breast cancer is detected early, patient can be survive more and sometime cancer is removed from each cells of breast. Breast cancer is not only the problem for developing countries but also developed countries. For this reason governments provide many free treatments in India. Diagnosis of breast cancer is fully depends on the reason why the breast cancer arise. Because breast cancer is arise from many reasons such as age of the person, genetic problem, starting and ending of menopause etc.

This proposed approach makes use association rule for generating rule for disease diagnosis. Association rule of data mining help to find out interesting knowledge from a huge amount of data. Association rule simplifies the relation between the attribute and generate a rule from those attribute. Association rule is used to predicting customer behaviour. It is provides result based on confidence, support, lift and conviction values. There is many algorithms like apriori algorithm, FP-growth algorithm, node set algorithm, guha algorithm etc. to generate association rule. Association rule is mostly used in many industries such as marketing, health care, telecommunication, banking, business intelligence etc. to make a better decision.

2. Background

Breast Cancer Symptoms and Causes

Cancer is a group of distemper that occurs when irregular enlargement in the cell of a body. Today more than 200 different types of cancer present in our world and in every country this is most cause of cancer. According to ICMR report in India only more than 1300 people are facing death because of cancer and in 2014, 5 lack people were died because of cancer. So according to the report of ICMR breast cancer have second place in the list of all cancer in the term of death rate. Breast cancer is mostly finding in the older women after the age of 60. There is many reason behind arising the breast cancer is genetic problem, first birth of child of the age of 30, problem in liver, taking more alcohol’s etc.
Approximately 15% male patients have breast cancer other 85% are females. If the cells in the breast increase and make a tumor in it that is called breast cancer. Breast cancer is the cause of death in mostly cases but if it is detected earlier than the chance of saving patient life is more.

3. Data Mining and Association rule

Association rule is used for generating rules from a collection of dataset there one instance is depend or have some relationship with other instance and so on. It is used to find frequent item set. Association rule is used in many industries such as AI( artificial intelligence), medical, marketing, telecommunication etc. to find out frequent item and it help to grow the business and find out customer behaviour. For understanding the concept of association rule first we need to clear some important concept such as support count, confidence of item set, frequent item etc.

Support

Support is used to find out what is the frequency of instance in our data set. It depends on total number of transaction.

\[
\text{SUPPORT}(A \Rightarrow B) = \frac{\text{NUMBER OF TRANSACTION CONTAIN BOTH A AND B}}{\text{TOTAL NUMBER OF TRANSACTION}}
\]

Confidence

Confidence tells us how the one instance is associated with other. It is calculated with the help of support value.

\[
\text{CONFIDENCE}(A \Rightarrow B) = \frac{\text{SUPPORT}\{A,B\}}{\text{SUPPORT}\{A\}}
\]

3. Literature Survey

NNT for breast cancer

Author describe that the NNT algorithm work in parallel. It is used to generate the function that has relationship with other dataset. In this paper author used NNT for diagnosis of breast cancer and get effective result using back
propagation method. With this they developed a forward NN model so the output can be easily understand by the user of system. In this paper author describe advantage of using NNT and feed forward BP method. They used WBC dataset for performing algorithm and provide the classification algorithm. They describe other possible algorithm using NNT and compare the result of this algorithm with all other proposed algorithm. They get a conclusion about using two different topology the training time is different but result of both topology is same.

**KNN Application for Diagnosis**

The KNN algorithm application is suggested and discussed for Wisconsin-Madison BRCA diagnosis. Manish Sarkar purposed this system using KNN algorithm on Wisconsin-Madison data. They used KNN algorithm and find that it is very easily to develop and approximately 2% faster than other algorithm. They discussed many advantage of this algorithms such as it is not require any prior knowledge about data that were used for performing algorithm its work more efficient on small data set. They also provide the limitation about KNN algorithm and conclude that it is not works efficient every time and on large data set so using other algorithm may provide better result

**Breast cancer diagnosis using DTDM**

Here they used Decision Tree for diagnosis of breast cancer and develop a predictive model that can find out that invade neighbour tissues are present or not in the medical report that were used in the algorithm. They provide full description about the breast cancer and all other possible DM techniques for solving this disease. They used WEKA tool and J48 DT algorithm for generating rules. Using this techniques breast cancer is easily diagnosis and performance of the DTDM is also better compare to other techniques.

**NBC System**

NBC system is also used in diagnosis and prediction of different types of diseases here the author used this for breast cancer. They developed naïve Bayes classifier algorithm using java platform that takes input as medical data record of the patient and show probability statics that a patient can have breast cancer. This system is implemented using the concept of Bayes theorem. They used 699 data set and find 93% accuracy on that dataset. According to the author this system can be used in rural areas to find out probability that says that a person has a breast cancer or not. They designed a GUI system for taking medical record of a patient. They found out different class of breast cancer such as fatal and bonny. From all the data set of breast cancer only 34.5% cases were addressed the bonny breast cancer and other 65.5% were fatal breast cancer. Fatal breast cancer type more harmful than bonny because survival rate is higher in bonny breast cancer.
On these type of risk factor human do not have control and these are unchangeable.

1. **Age:** Mostly the breast cancer is rises after the age of forty. In 40% of cases mostly women were between 40 and 60’s.

![Age affected in breast cancer](image)

2. **Gender:** According to the report of WHO only 20% male are suffer from breast cancer. Commonly it is founded in females because of some genes factor that are available in female body and help in growing of breast cancer.

3. **Family History:** Before the age of 50, woman risk is approximately double when the level one degree relatives were diagnosis by breast cancer. In level one degree that can be patients mother, father, sister etc.

4. **Genetic Factor:** Alteration into the genes are also a main factor that help in growing of breast cancer. Mainly BRCA gene class increase the risk two times more compare to other genes.

**Avoidable Risk Factor**

1. **Overweight:** Overweight is mostly causes of any disease. Mainly a person becomes overweight due to lack of exercise and fast food habit etc. Those women weight is more compare to their age, they have higher risk of getting breast cancer.

2. **Life Style:** A women work at late night; having a habit of smoking, physical inactivity etc. can suffer from breast cancer. Change in life style is also increase the survival rate.

3. **HRT:** Hormone replacement therapy is considered as a main cause of breast cancer. Mainly women have HRT therapy for preventing the menopause effect and menopause is also a risk factor. Menopause means if women get periods before the age of 12 and after the age of 50 is considered in the menopause duration.

4. **Alcoholic:** Consumption of unlimited alcohol can affects DNA cells. Result of this is founded in the increasing the estrogen and hormones in the body.
Symptoms of Breast Cancer

Symptoms of breast cancer can be observed by the patient that shows the abnormality in the breast or behaviour into the human body. It can be observed by doctor during the check-up. On the first stage of breast cancer, it shows some abnormality or change in the shape, size of the breast.

1. Lump in the breast: Some time a person got injury, infection in the breast and do not feel any pain but the part that is infected is become hard. This is the first symptoms can be seen in the person who may or may not be have breast cancer.

2. Nipple changes: Nipple is the way from milk come out from the body. When the starting stage of breast cancer nipple turn inside or some time abnormally discharge of nipple.

3. Skin changes: skin of the breast is become red or sometime it gets swelling or dimple on the breast. Its look like infection on the skin.

4. Changes in breast: Breast size, shape, colour and touch is change from the normal breast.

4. Materials and methodology

Apriori Algorithm

Apriori algorithm is developed for finding frequent item set by Agarwal and Srikanth in 1994. It’s mainly use value of support, confidence to find out frequent item set from a very big collection of data. It was famous on marking industry to find out the customer purchasing behaviour such as which thing is purchased together so that the sales value can be increased. Now this algorithm is used in many industries such as medical, tele communication and other industries for finding out the frequent item set. Today, in medical industry thousands of transaction happens in an hour as the result Tera byte of data collected by database. Apriori algorithm help to generate rule and in diagnosis from these data. Bottom up approaches help to find out the candidate set from frequent item set and breadth first search algorithm also used in.

Pseudo code for Apriori algorithm

```plaintext
Ai: Candidate item set of size i
Bi: frequent item set of size i
B1= (frequent items);
for(i = 1; Bi!=Ø; i++) do begin
Ai+1= candidates generated from Bi;
for each transaction in database do
increment the count of all candidates in Ai+1 that are contained in t
Bi+1= candidates in Ai+1 with min_support end
return Ui Ai;
```
Sample\textsuperscript{[1]}

<table>
<thead>
<tr>
<th>Itemset</th>
<th>{KL,FJ,RF,KO}</th>
<th>{KL,FJ,KO}</th>
<th>{KL,FJ}</th>
<th>{FJ,RF,KO}</th>
<th>{FJ,RF}</th>
<th>{RF,KO}</th>
<th>{FJ,KO}</th>
</tr>
</thead>
</table>

Step 1: We will add all the value of the \{KL\},\{FJ\},\{RF\},\{KO\} and count each instance from all data.

<table>
<thead>
<tr>
<th>Item</th>
<th>{KL}</th>
<th>{FJ}</th>
<th>{RF}</th>
<th>{KO}</th>
</tr>
</thead>
<tbody>
<tr>
<td>Support</td>
<td>3</td>
<td>6</td>
<td>4</td>
<td>5</td>
</tr>
</tbody>
</table>

We will calculate assess of support is 3 so that all the instance should include at least 3 instance.

Step 2: Group the value of 2 instances together will be consider.

<table>
<thead>
<tr>
<th>Item</th>
<th>{KL,FJ}</th>
<th>{KL,RF}</th>
<th>{KL,KO}</th>
<th>{FJ,RF}</th>
<th>{FJ,KO}</th>
<th>{RF,KO}</th>
</tr>
</thead>
<tbody>
<tr>
<td>Support</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>3</td>
</tr>
</tbody>
</table>

The pairs \{KL,FJ\}, \{FJ,RF\}, \{FJ,KO\}, and \{RF,KO\} these all have at least 3 instance ,we will take these all instance in next step. The pairs \{KL,RF\} and \{KL,KO\} are not valid they don’t have 3 instance .Those who don’t have 3 instance that will not be go in next step and eliminate here.

<table>
<thead>
<tr>
<th>Item</th>
<th>{FJ,RF,KO}</th>
</tr>
</thead>
<tbody>
<tr>
<td>Support</td>
<td>2</td>
</tr>
</tbody>
</table>

So the result is \{FJ,RF,KO\} and have only 2 instance in it but we need at least 3 so there is no frequent instance that have relationship with each other

5. Proposed Data Model

We used apriori feature selection algorithm to find out the feature that are related to each other and count the value of support. After feature selection complete we generate rule for diagnosis of breast cancer and for this we again used apriori algorithm.

Input: SEER Breast Cancer data

Output: Diagnosis of breast cancer

Procedure

1. First of all we took the Germany breast cancer data from SEER database that include 16 variables with 700 cases.

2. Data pre-processing:-

2.1 Attribute Selection: Select attribute that are needed for diagnosis.
2.2 Data Discretization: Divide the attributes into the parts for making the attribute value in 0’s and 1’s only.

2.3 Data Transformation: convert all data into 0’s and 1’s form.

3. Perform apriori algorithm for frequent item set selection

4. Generate rule for breast cancer diagnosis using apriori algorithm

5. Provide solution and reason to the patient that why she/he suffering from breast cancer.

Attribute Selection:

From these data we will select only that attribute that are need and help in diagnosis and we will remove other attribute that does not have effects on diagnosis such as patient id, date of diagnosis etc. so from all 16 variables we choose only 10 variables.
After performing all the operation and algorithm we find a collection of rule with the confidence value. From these rules we extract some rules and put into the GUI system that will send email to the patient that were diagnosis by the breast cancer. As a result patient and doctor will find out why this person suffering from breast cancer. After finding the reason, doctor can provide best solution and can make decision easily. This is the list of all rules were generated by the apriori rule generation algorithm.

Table 5.2: Attribute selection for breast cancer diagnosis

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Description</th>
<th>Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Age at Diagnosis</td>
<td>Age</td>
</tr>
<tr>
<td>2</td>
<td>Menopausal Status</td>
<td>Menopause</td>
</tr>
<tr>
<td>3</td>
<td>Hormone Therapy</td>
<td>Hormone</td>
</tr>
<tr>
<td>4</td>
<td>Tumour Grade</td>
<td>Grade</td>
</tr>
<tr>
<td>5</td>
<td>Time to Recurrence</td>
<td>Censrec</td>
</tr>
<tr>
<td>6</td>
<td>Recurrence Censoring</td>
<td>Censlead</td>
</tr>
<tr>
<td>7</td>
<td>Number of Nodes</td>
<td>Node</td>
</tr>
<tr>
<td>8</td>
<td>Tumour Size</td>
<td>Size</td>
</tr>
<tr>
<td>9</td>
<td>Number of Progesterone Receptors</td>
<td>prog_recp</td>
</tr>
<tr>
<td>10</td>
<td>Number of Estrogen Receptors</td>
<td>estrg_recp</td>
</tr>
</tbody>
</table>

Rule generation with confidence value using apriori algorithm
6. Conclusion

This paper presents the application for diagnosis of breast cancer and the proposed data model will be useful for doctors as well as patient to find out the reason why the patient suffering from BRCA. For generating rules and feature selection for this system we develop apriori algorithm of data mining. Those algorithms provide us the solution about which of the element is affecting or have relationships with other element. For apriori algorithm we used SEER Germany breast cancer data and applied data prepossessing. So using this system doctor will send mail to the patient “Why She/he suffered from Breast cancer”.

Reference


**Corresponding Author:**

M.Sudha*

**Email:** msudha@vit.ac.in