International Journal of Computer Applications in Engineering Sciences
[VOL II, ISSUE III, SEPTEMBER 2012] [ISSN: 2231-4946]

Classification of Women Health Disease (Fibroid) Using Decision Tree algorithm

Girija D.K# ,M.S. Shashidhara*

#Department of Computer Science,GFGC, Pavagada.
**Department of Computer Application(MCA),The Oxford College of Engineering, Bangalore.*

[1]girijadk@rediffmail.com
[2]msshashidhara@gmail.com

Abstract **- Data Mining is the process of discovering meaningful new correlation, patterns and trends by sifting through a large amount of data stored in repositories, using pattern recognition technologies as well as statistical and mathematical techniques. (Larose 2005, 2.) Before Data mining i.e. data analysis methods involve manual work and interpretation of data which is slow, expensive, and highly subjective. So the achievement of data mining resolves these factors. In this paper they have investigate data mining techniques in health care. In particular, they have discuses data mining and its application in area where women's are affected seriously by uterine fibroids. This paper identifies the risk factors associated with the effect of fibroids in uterus. By using Data Mining Technique like classification algorithm they find meaningful hidden patterns which give meaningful decision making to this health disease.**

Keywords **- Data Mining, Fibroids affected in women, classification algorithm, Decision tree J 48.**

I. INTRODUCTION

A. Fibroids as Health Disease

Fibroids are muscular tumors that grow in the wall of the uterus (womb). Medical term for fibroids is" myoma". (figure 1) Fibroids are almost always benign (not cancerous). Fibroids can grow as a single tumor, or there can be many of them in the uterus. They can be small as an apple seed or as big as a grapefruit. In unusual cases they can become very large.[1]

Fibroids was considered to be a problem related to uterus these leads to affect of infertility and pregnancy and increase the risk of complication during delivery.

Figure 1: Fibroids

B. Who gets fibroids?

- **Age:** Fibroids become more common as women age, especially during the 30s and 40s through menopause. After menopause, fibroids usually shrink.
- **Family History:** Having a family member with fibroids incases our risk. If a women's mother had fibroids, her risk of having them is about 3 times higher than average.
- **Ethnic Origin**: African women re more likely to develop fibroids than white women.
- **Obesity**: Women who are overweight are at higher risk for fibroids. For very heavy women, the risk is two to three times greater than average.
- **Eating Habits**: Eating a lot of red meat (e.g., beef).

Their goal is to create a decision tree using WEKA so that they can classify the fibroids causes like normal or severe. From analysing their symptoms.

II. LITERATURE SURVEY OF THE PROBLEM

A uterine fibroid is the most common benign (not cancerous) tumor of a woman's uterus (womb). Fibroids are tumors of the smooth muscle that is normally found in the wall of the uterus. They can develop within the uterine wall itself or attach to it. They may grow as a single tumor or in clusters. Uterine fibroids can cause excessive menstrual bleeding, pelvic pain, and frequent urination; so even though they are termed "benign (not cancerous) tumors," fibroids potentially can cause many health problems.

These growths occur in up to 50% of all women and are one leading cause of hysterectomy (removal of the uterus) in the United States. An estimated 600,000 hysterectomies are performed in the US annually, and at least one-third of these procedures are for fibroids. Medications and newer, less invasive surgical treatments are now available to help control the growth of fibroids.

To understand these health problems in women we discuss with medical practitioners and specialists like Gynecologist, Radiologist and General Doctors. We also gathered details about the effect of fibroids in uterus from Gynecologist and Radiologist. By analyzing all these we came to know that the excessive menstrual bleeding and pelvic Paine can creates a women health problems like anemia, Infertility and rarely it can turn into a cancer called a leiomyosarcoma.

This happens to an estimated 1 to 1,000 women who have fibroids.

III. DATA PREPARATION

Based on information collected from Gynecologist and Radiologist. We paper questioners to get raw data from various women's. The women's with different age group with different symptoms were interview with the health of questioners we prepared.

Total data collected from Doctor's 25 records with all the criteria's. So based on the medical Doctor's advice while classifying the data, the degree of symptoms is placed in several compartments as follows:

None: No symptoms found

Mild: Those who are found with one to three symptoms are grouped under Mild

Severe: Those who are found with more than 3 high symptoms are grouped as severe.

IV. CLASSIFICATION AS THE DATA MINING TECHNIQUE

The goal of classification is to accurately predict the target class for each case in the data.

Classification is one of the major data mining tasks. Although this task is accomplished by generating a predictive model of data, interpreting the model frequently provides information for discriminating labeled classes in data. Decision trees provide a predictive model that is easy to interpret to provide a description of data.

In order to believe any predictive model, the accuracy of the model must be estimated. Several methods for evaluating the accuracy of models will be discussed during class lectures. For this assignment, 10-fold cross validation will be used for model assessment.

V. CLASSIFICATION ALGORITHMS

Data Mining provides the following algorithms for classification:

- Decision Tree: Decision trees automatically generate rules, which are conditional statements that reveal the logic used to build the tree.
- Naive Bayes: Naive Bayes uses Bayes' Theorem, a formula that calculates a probability by counting the frequency of values and combinations of values in the historical data.
- Generalized Linear Models (GLM): GLM is a popular statistical technique for linear modeling. Oracle Data Mining implements GLM for binary classification and for regression.
- Support Vector Machine: Support Vector Machine (SVM) is a powerful, state-of-the-art algorithm based on linear and nonlinear regression. Oracle Data Mining implements SVM for binary and multiclass classification.

The nature of the data determines which classification algorithm will provide the best solution to a given problem. The algorithm can differ with respect to accuracy, time to completion, and transparency. In practice, it sometimes makes sense to develop several models for each algorithm, select the best model for each algorithm, and then choose the best of those for deployment.

Typical Applications

- Credit approval
- Target marketing
- Medical diagnosis
- Treatment effectiveness analysis

So for our problem we consider the decision tree(C4.5 or J48) algorithm. The Decision Tree algorithm, like Naive Bayes, is based on conditional probabilities. Unlike Naive Bayes, decision trees generate rules. A rule is a conditional statement that can easily be understood by humans and easily used within a database to identify a set of records.

In some applications of data mining, the accuracy of a prediction is the only thing that really matters. It may not be important to know how the model works. In others, the ability to explain the reason for a decision can be crucial. For example, to treat a patient with a disease doctor may know complete description of the patient. The Decision Tree algorithm is ideal for this type of application.

TABLE I: CLASSIFICATION OF SYMPTOMS OF DISEASES

Heavy bleeding	**Pelvic Pain**	**Lower back pain**	**Pain During sex**	**Frequent urination**	**cause**
No	No	No	No	No	There is no Fibroid
High	High	High	High	No	Severe(Suffering from Fibroid)
No	No	No	High	High	Mild(Starting Stage)

Table 1

VI. WEKA AS DATA MINING TOOL

Weka(Waikato Environment for Knowledge Analysis) is a popular suite of machine learning software written in Java, development at the university of Waikato, New Zealand. Weka is free software.

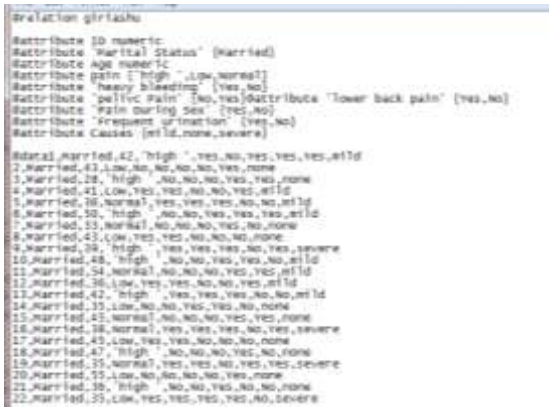


Figure 2: An arff file

The key features:

- It provides many different algorithms for data mining and machine learning ex. Classification, clustering, Associate etc.,
- It is open source and freely available like CAD/CAM, animation software etc.,
- It is platform independent because it is created by Java.
- It is easily useable by people who are not data mining specialists.
- It provides flexible facilities for scripting experiments.
- It has kept up-to date, with new algorithms being added as they appear in the research literature.
- 49 data preprocessing tools.
- 76 classification/regression algorithms.
- 8 clustering algorithms.
- 3 algorithms for finding association rules.
- 15 attribute/subset evaluators + 10 search algorithms for feature selection.

In this paper they have used WEKA a Data Mining tool for classification techniques. This software is able to provide the required data mining functions and methods effectively. So the suitable data format of WEKA data mining software is MS-Excel and ARFF(Attribute Relation File Format) formats respectively. [5]

VII. CLASSIFICATION IN WEKA

Classification creates a model based on which new instances can be classified into the existing classes or determined classes for example by creating a decision tree based on symptom's of diseases we can determine how a patient's condition is mild or severe.

In this paper we are using decision tree J48 algorithm to classify the data. It is,

- Divide and conquer algorithm
- Convert tree to classification rules
- J48 can handle numeric attributes.

To Start with classification we use or create a arff or csv (or any supported) file format. An arff file is a table. (figure 2)

Our goal is to create a decision tree using WEKA so that we can classify the fibroids causes like normal or severe.

There are three kind of patients they are 1) There is no Fibroid 2) Mild condition 3) Severe

Data File : We have a data file containing attribute values of 25 patients samples in arff format. Concept behind the classification is the condition of patient's data like heave bleeding, pelvic pain, lower back pain etc., help us to identify the causes. The data file contains all the 9 attributes. The algorithm I am going to use to classify is WEKA J48 decision tree learner.

After follow the step finally we get the "classifier output" show below.

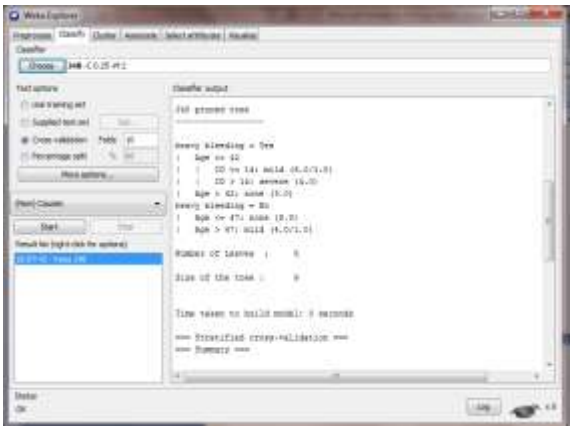


Figure 3: classifier output

Now we will see the tree structure. Right click on the entry in "Result list" and select Visualization tree. Then Decision tree will be visible in new window like this. (figure 4)

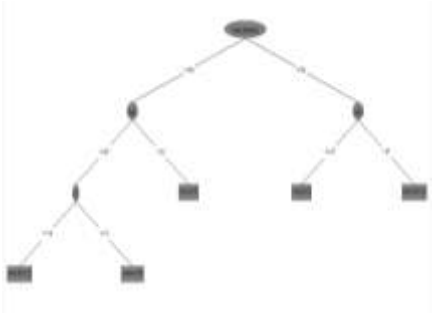
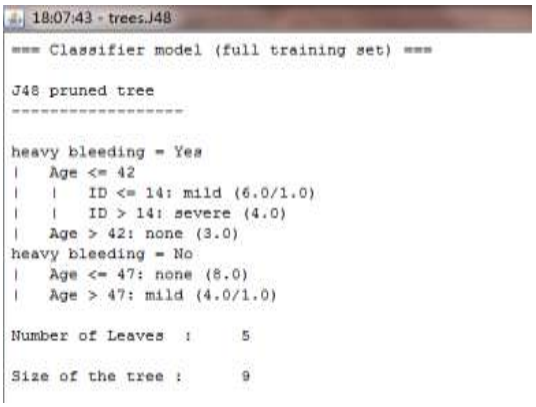


Figure 4: Decision tree

It give the decision structure or flow of process to be followed during classification. For Example if heavy bleeding is Yes , age <=42 , Id > 14 it implies the patient's condition is severe.

Now look at the classifier output box. The rules describing the decision tree is described as given in the picture. (figure 5)



```
18:07:43 - trees.J48
=== Classifier model (full training set) ===

J48 pruned tree
------------------

heavy bleeding = Yes
|   Age <= 42
|   |   ID <= 14: mild (6.0/1.0)
|   |   ID > 14: severe (4.0)
|   Age > 42: none (3.0)
heavy bleeding = No
|   Age <= 47: none (8.0)
|   Age > 47: mild (4.0/1.0)

Number of Leaves  :     5

Size of the tree :      9
```

Figure 5: The rules describing the decision tree

As we can see in the decision tree we don't require other attributes like pelvic pain, lower back pain etc.

Here we are using Test option as cross validation for all the data set because we use only 25 instances(figure 6).

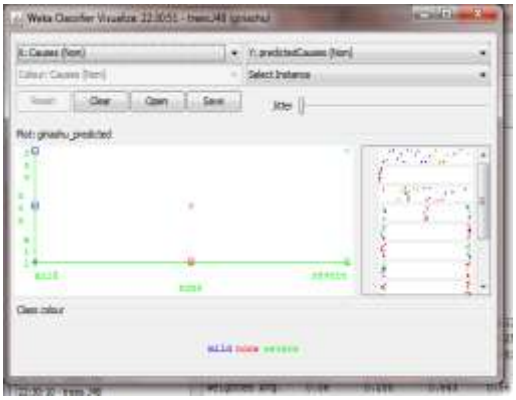


Figure 6: validation for all the data set because we use only 25 instances

The figure analysis that

1. WEKA took 25 samples out of which 14 are classified correctly and 11 are classified incorrectly.
2. If you look at the confusion matrix below in classifier output box. We see all out of 10 mild causes only 5 are correctly classified like None(7/12) and severe(2/5) are correctly classified
3. To find more information or to visualize how decision tree did on test samples. Right click on "Result list" and select "visualize classifier errors".
4. A new window will open. Now as our tree has used heavy bleeding and age to classify. We select causes of X axis and predictedcauses for Y axis. (Figure 7)

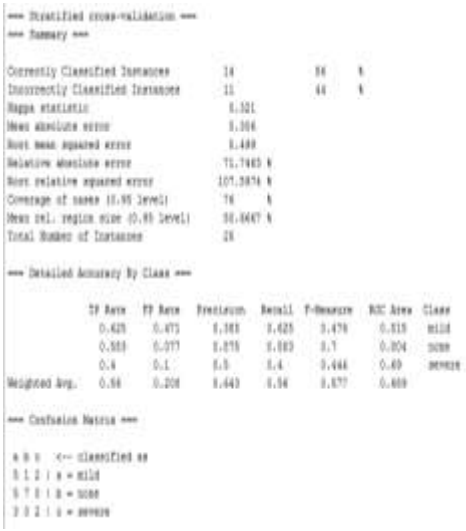


Figure 7: predicted causes

5. Here "x" or cross represents correctly classified sample and squares represents incorrectly classified samples.
6. Results of causes are display in different colors as blue for mild, red or none and Green for Severe.
7. As we can see why these are classified incorrectly as sample class identification error or attribute error or inappropriate classification algorithm.

VIII. CONCLUSION

Decision tree J48 algorithm is implemented using WEKA 3.7.5 data miner. Here it classifies the predictedcauses over the causes taken by the data. It classifies the data in to correctly and incorrectly instance.

Data mining applied in health care domain, by which the women get beneficial for their lives. As the analog of this research found the meaningful hidden pattern that from the real data set collect by the patients affected by fibroids. by which we can easily know that the women do not get awareness among themselves about the uterus fibroids. If it continues in this way it may lead to some health diseases like anemic, infertility, cancer.

REFERENCES

[1] Gynecology by Hricak,Akin,Sala,Acher, Levine, einhold
[2] Women's health.gov Healthoma.com
[3] http://www.emedicinehealth.com/uterine_fibroids/article_em.htm
[4] Peter Reutemann, Ian H. Witten," The WEKA Data Mining Software: AnUpdate - White paper", Pentaho Corporation. SIGKDD Explorations Volume 11, issue 1 pp. 10 - 18, 2005.
[5] Data Mining by Jan H witten, Eibe Kaufman Publisher.