Experimentation Based on Decision Tree Induction Analysis for Hepatitis Patients

Min Min Shwe Sin, Khin Sandar, May Phyo Oo
Computer University (Pathein, Myanmar)
minminshwesinn@gmail.com, drkhinsandar@gmail.com,mayphyooo@gmail.com

ABSTRACT

This paper is about Hepatitis medical diagnosis using decision tree induction algorithm in machine learning. This paper describes experimentation of decision tree induction analysis for hepatitis patients. A chain reaction process is taken with the help of an algorithm if so decision trees are produced, in turn producing rules to analysis for hepatitis patients. This system shows the stages of how the rules are produced and the factors that will influence the end rules. This paper includes data mining information from a patient record using a domain to symptom and processor for creating a patient model of diagnostic process. This system used ID3 decision tree learning algorithm.

1. Introduction

Decision tree is one of the most important techniques in machine learning. Most decision tree learning algorithm grows the tree step by step. A recursive algorithm is used that decides for a given dataset of training examples. Data mining refers to extracting or mining knowledge from large amounts of data. There are many other terms carrying a similar or slightly different meaning to data mining. Data mining techniques are used to mine different rules and patterns. This include model of summarization, classification, clustering or regression. Many classification and prediction methods have been proposed by researchers in machine learning, expert system, statistics, and neurobiology.

Decision tree models are simple and easy to understand. Moreover, trees can be easily converted into SQL statements that can be used to access database efficiently. Decision tree classifier obtains similar and sometimes better accuracy. We can have focused on building ID3 algorithm.

A decision tree is a class discriminator that recursively partitions the training set until each partition consists entirely or dominantly of examples from one class. Each non-leaf node of the tree contains a split point which is a test on one or more attributes and determines how the data is partitioned. Random sampling is often used to handle large datasets when building a classifier. Previous work on building tree-classifier from large data sets.

2. Related Work

Classification is an important data mining problem classification methods are decision tree induction, Bayesian classification and Bayesian belief networks and neural networks, k-nearest neighbor classifier, case-based reasoning, genetic algorithms, rough sets and fuzzy logic techniques [5]. Machine learning investigates the mechanisms by which knowledge is acquired through experience. Research at UCI (University of California at Irvine) spas the spectrum of models for machine learning including those based on statistics, logic, mathematics, neural structures, information theory and heuristic search algorithms [6].

The development and analysis of decision tree algorithms identify patterns in observed data in order to make predictions about other data set. Machine learning algorithms often result from research into the effect of problem properties on the accuracy and run-time of existing algorithms.

Learning is investigated from structured databases (for application such as screening loan applicants), image data (for applications such as locating relevant sites on the World Wide Web. UCI also maintains the international machine learning database repository, an archive of databases used specifically for evaluating machine learning algorithms [6].

Researchers [Jiang Su & Harry Zhang] implemented the title of a fast decision tree learning Growing Algorithm to research the complexity and classification accuracy base on Decision tree learning algorithm [4].

3. Proposed System

Proposed System can serve as a diagnosis tool for Hepatitis disease prediction extracted from U.C.I medical datasets. This system helps guide health care regarding preventive and predictive action. Disease related information is the goal through each stage of analysis. Data mining techniques of decision tree learning algorithm analysis for Hepatitis diseases prediction.

In order to test the patients , a set of symptom and investigation of Hepatitis are required. When the new symptom values are chosen, the system produce DIE or LIVE by using decision tree rules. Decision tree are build ID3 algorithm and the best probability, highest information gains to use recursive function. All dataset patients are positive Hepatitis diseases.

This system includes two classes DIE or LIVE and it intends to develop the decision support system for classification. The process flow of the proposed system design is show in figure 1.

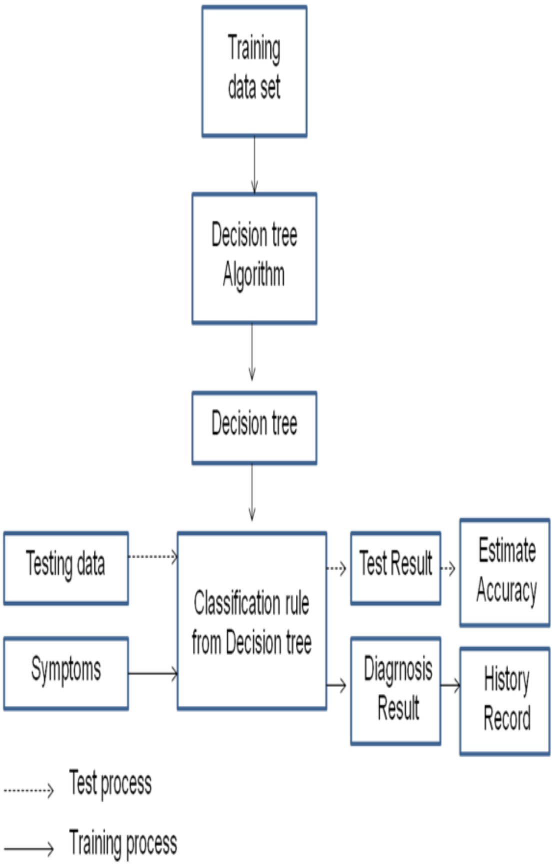


Figure 1. Design of the System

3.1 Decision Tree Classification

Decision tree is a classifier in the form of a tree structure figure-2. Decision trees are powerful and popular tools for classification and prediction. It is a predictive model that is a mapping of observations about on item to conclusions about the item's target value. A decision tree is a flow-chart-like-tree structure, each internal node denotes a test on item to conclusions about the item's target value an attribute, each branch represents an outcome of the test on an attribute, each branch represents an outcome of the test and leaf nodes represent class distributions. [5]

Leaf node indicates the value of the target attribute (class) of examples or a decision node-specifies some test to be carried out on a single attribute-value, with one branch and sub-tree for each possible outcome of the test. C4.5 is a software extension of the basic ID3 algorithm designed by Quinlan to address the following issues not dealt with by ID3

- o Avoiding over fitting the data
- o Reduced error purning
- o Rule-post-purning
- o Handling continuous attributes
- o Choosing an appropriate attribute selection measure.
- o Handling training data with missing attribute values.
- o Handling attributes with differing costs.
- o Improving computational efficiency.

3.2 Decision Tree Learning ID3 Algorithm

In classification, we are given a set of examples records, called a training set, where each record consists of several fields or attributes. Attributes are continuous, coming from an ordered domain or categorical, coming from an unordered domain. One of the attributes, called the classifying attribute. The highest information gain or greatest entropy reduction attribute is chosen.

$$I(S_1, S_2, \ldots, S_m) = -\sum_{i=1}^{class} P_i \log_2(P_i) \quad \ldots (1)$$

S be a set consisting of s data samples.

S_i be the number of samples of S in class C_i

$$E(A) = \sum_{j=1}^{v} \frac{S_{1j} + \ldots + S_{mj}}{S} I\,(S_{ij}, \ldots, S_{mj}) \ldots (2)$$

S_{ij} be the number of samples of class C_i in a subset S_j.

$P_{ij} = \frac{S_{ij}}{|S_j|}$ and is the probability that a sample in S_j to class C_i.

Gain (A) = I $(S_1, S_2, \ldots, S_m)$ – E (A) … (3)

Gain (A) is the expected reduction in entropy caused by knowing the value of attribute A.

3.3 Attribute Information

1.	AGE	AGE	10,20,30,40,50,60,70,80
2	SEX	SEX	male, female
3	ANT	ANTIVIRALS	no, yes
4	STE	STEROID	no, yes
5	FAT	FATIGUS	no, yes
6	MAL	MALAISE	no, yes
7	ANO	ANOREXIA	no, yes
8	LIB	LIVER BIG	no, yes
9	LIF	LIVER FIRM	no, yes
10	SPL	SPLLEN PALPABLE	no, yes

In this system, knowledge base with the UCI medical experts is used. These data sets are all positive hepatitis diseases. The number of attributes is 20 (including the class attribute)

3.4 Sample Data Set

In this section Sample Data Sets of UCI are described as follows. There are 19 sample attributes, 1 class attribute and 250 records of training data in this system.

@<30, male, no, no, no, no, no, no, no, no, no, no, no, 1, 85, 18, 4, ?, no, LIVE>

@<50, female, no, no, yes, no, no, no, no, no, no, no, no, 0.9, 135, 42, 3.5, ?, no, LIVE>

@<78, female, yes, no, yes, no, no, yes, no, no, no, no, no, 0.7, 96, 32, 4, ?, no, LIVE>

@<38, female, ?, yes, no, no, no, yes, no, no, no, no, no, 0.4, 243, 49, 3.8, 80, yes, DIE>

@<34, female, yes, no, no, no, no, yes, no, no, no, no, no, 1, ?, 200, 4, ?, no, LIVE>

@<34, female, yes, no, no, no, no, yes, no, no, no, no, no, 0.9, 95, 28, 4, 75, no, LIVE>

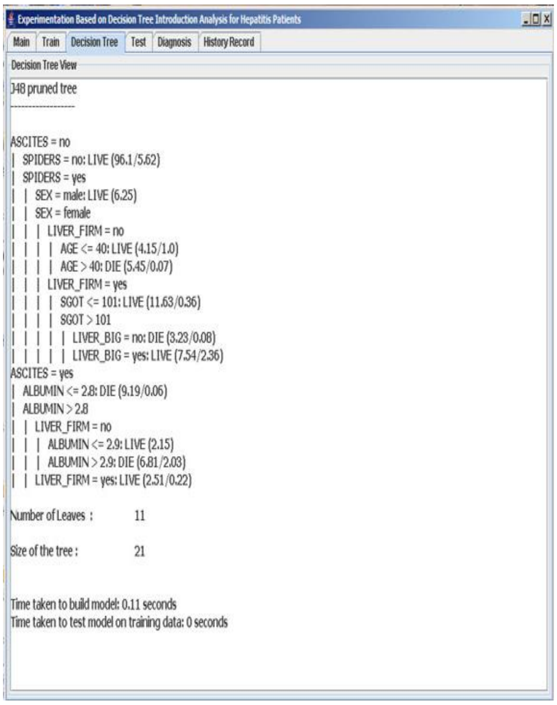
@<51, female, no, no, yes, no, yes, yes, no, yes, yes, no, no, 1.7, 86, 220, 2.1, 46, no, DIE>

@<23, female, yes, no, no, no, no, yes, no, no, no, no, no, 4.6, 56, 16, 4.6, 54, no, LIVE>

@<39, female, yes, no, yes, no, no, yes, yes, no, no, no, no, 0.7, ?, 48, 4.4, ?, yes, LIVE>

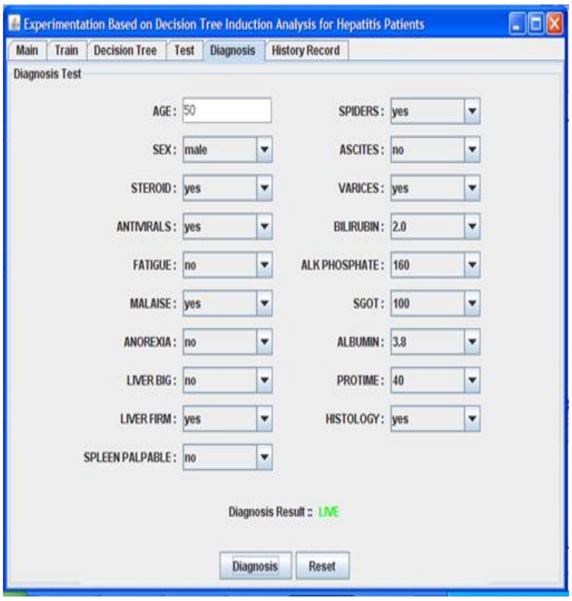
3.5 Implementation of the System

Implementation of the System can be seen in figure 2 and figure3. Classification rules can be got from Decision tree to use the system diagnosis. Figure 2 shows decision tree of the poultry diseases system to extract rules. According to figure 2, ASCITES is the highest information gain.


Experimentation Based on Decision Tree Introduction Analysis for Hepatitis Patients
Main
Train
Decision Tree
Test
Diagnosis
History Record
Decision Tree View

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

ASCITES = no
|  SPIDERS = no: LIVE (96.1/5.62)
|  SPIDERS = yes
|  |  SEX = male: LIVE (6.25)
|  |  SEX = female
|  |  |  LIVER_FIRM = no
|  |  |  |  AGE <= 40: LIVE (4.15/1.0)
|  |  |  |  AGE > 40: DIE (5.45/0.07)
|  |  |  LIVER_FIRM = yes
|  |  |  |  SGOT <= 101: LIVE (11.63/0.36)
|  |  |  |  SGOT > 101
|  |  |  |  |  LIVER_BIG = no: DIE (3.23/0.08)
|  |  |  |  |  LIVER_BIG = yes: LIVE (7.54/2.36)
ASCITES = yes
|  ALBUMIN <= 2.8: DIE (9.19/0.06)
|  ALBUMIN > 2.8
|  |  LIVER_FIRM = no
|  |  |  ALBUMIN <= 2.9: LIVE (2.15)
|  |  |  ALBUMIN > 2.9: DIE (6.81/2.03)
|  |  LIVER_FIRM = yes: LIVE (2.51/0.22)

Number of Leaves  :      11

Size of the tree :       21


Time taken to build model: 0.11 seconds
Time taken to test model on training data: 0 seconds
```

Figure 2. Decision Tree


Experimentation Based on Decision Tree Induction Analysis for Hepatitis Patients
Main
Train
Decision Tree
Test
Diagnosis
History Record
Diagnosis Test
Diagnosis
Reset

AGE : 50
SEX : male
STEROID : yes
ANTIVIRALS : yes
FATIGUE : no
MALAISE : yes
ANOREXIA : no
LIVER BIG : no
LIVER FIRM : yes
SPLEEN PALPABLE : no
SPIDERS : yes
ASCITES : no
VARICES : yes
BILIRUBIN : 2.0
ALK PHOSPHATE : 160
SGOT : 100
ALBUMIN : 3.8
PROTIME : 40
HISTOLOGY : yes

Diagnosis Result :: LIVE

Figure 3. Diagnosis of the system

Figure 3 shows the new patient Hepatitis diagnosis system. We are going to select the 19 attribute list box for patient's conditions. Click on top-down list and to choose one value on click. These steps are selected all 19 attribute on click. After choosing the 19 attribute, click on diagnosis button. We can see the result of diagnosis system “DIE” or “LIVE”. If we start the new symptom selection, click on Reset button.

4. Classifier Accuracy

Estimating classifier accuracy is important in that it allows on to evaluate how accurately a given classifier will label future data. Accuracy estimates also help in the comparison of different classifier. Accuracy percentage is checking the improved quality of our system. Holdout and cross validation are two common techniques. We use cross validation method. Accuracy estimate is taken as the average of the accuracies obtained from data set. Applying cross validation , the testing estimate accuracy is generated as 93.8537%.

$$specificity = \frac{TN}{N}$$

$$sensitivity = \frac{TP}{P}$$

$$precision = \frac{TP}{(TP + FP)}$$

$$accuracy = sensitivity\frac{P}{(P+N)} + specificity\frac{N}{(P+N)}$$

where

TP = true positive, P = positive

TN = true negative, N = negative

Above mentioned Sensitivity is the positive patients can express correctly. And Specificity is the negative patients can express correctly.

5. Strengths of Decision Tree Methods

The strengths of decision tree method are:

- Decision trees are able to generate understandable rules.
- Decision trees perform classification without requiring much computation.
- Decision trees are able to handle both continuous and categorical variables.
- Decision trees provide clear indications of with fields are most important for prediction or classification.

6. Conclusion

Decision tree induction analysis has been experimented using ID3 algorithm for hepatitis patients. We regard the classification test as constructing a diagnosis and estimate accuracy of search results. In addition, this system intends to retrieve the document more quickly. There is growing interest in scaling up the widely used decision tree learning algorithm to very large data set. Hepatitis patients can be predicted without the help of physicians if we have chemical test and laboratory test data. This system can explain the stages of how the rules are produced and the factors that will influence the end rules. Time and costs can be reduced and it is easy to use for users without requiring much computer skill due to the good result of this system. This system can be applied for not only Hepatitis Patitients but also other medical experts in the future.

REFERENCES

[1] C.wallace and J.Patick. “Coding decision trees. M.L”, 11:7-22, 1993

[2] H.Hamilton. E.Gurak, L.Findlater-W.Olive, “Overview of Decision Trees” http://www.cs.wegina.ca/~dba/cs831/notes/ml/dtrees/4dtrees1.html

[3] J.R. “Quinlan-Improved use of continuous attributes in C4.5 Journal of Artifical Intelligence Research” 4:77-90, 1990

[4] Jiang Su and Harry Zhong, “A fast Decision tree learning Algorithm”: Faculty of Computer Science, University of New Brunswick, NB, Canada, E3B5A3 {jiang, SU, h2hang} @ unb.ca

[5] Jiawei Han and Micheline Kamber, “Data Mining Concepts and Techniques (Second Edition)”, University of lllinois at Urbana. Champaign

[6] John Shafer*, Rakesh Agrawal, Manish Mehta, “SPRINT: A Scalable Parallel Classifier for Data Mining”, (IBM Almaden Research Center, 650 Harry Road, San Jose, CA 95120)

[7] Mc Graw Hill, Tom M.Mitchell, “Lecture slide for text book Machine Learning”, 1997.

[8] “ML / Inductive Inference / DT / Construction” http://www.w.cs.uregina.ca/~dbd/cs831/notes/ml/dtrees/4-dtrees2.html

[9] Morgan Kaufmann Publisher, “Qunlan J.R.C45: Programs for Machine Learning”; 1993

[11] Swe Swe Aung, “ACUTE DIARRHEA DIAGNOSIS SYSTEM USING D.T.L.A”, M.C.Sc, 2004 July