Developing Decision Tree Using ID3 for Depression Testing

Theingi Aye, Dr. Aye Lelt Lelt Phyu
theingiayee@gmail.com.mm,allphyu@gmail.com.mm
University of Computer Studies, Yangon

Abstract

Data classification is the process of building a model from available data called the training dataset and classifying object according to their attributes. A derived model may be represented in various forms, such as decision tree, mathematical formula and neural networks. Decision tree algorithm has been used for classification in a wide range of application domains. In this paper, we propose a new system that tests the level of depression by using decision tree induction classification algorithm. Depending upon the data tuples of patient's symptom dataset, the system can classify the level of depression within a minimum of time. This system will be implemented by using C#(2005) and SQLSever 2000.

Keywords: Classification, Decision Tree, ID3 algorithm

1. Introduction

Nowadays, people are busy for their related workplace. Every person gets a lot of stress in their workplace because they try to progress their level of jobs. They will probably get a depression concerning with their jobs. So, this proposed system supports for physicians to take the most appropriate decision concerning whether the users are suffering depression or not within a minimum of time.

Clinical decisions are often made based on doctors' intuition and experience rather than on the knowledge-rich data hidden in the database. This practice leads to unwanted biases, errors and excessive medical costs which affects the quality of service provided to patients. Wu, et al proposed that integration of clinical decision support with computer-based patient records could reduce medical errors, enhance patient safety, decrease unwanted practice variation, and improve patient outcome [6]. This suggestion is promising as data modeling and analysis tools, e.g., data mining, have the potential to generate a knowledge-rich environment which can help to significantly improve the quality of clinical decisions.

This paper proposes a system that could assist patients to decide the level of depression based on Decision Tree Classification algorithm. This paper is organized as follows: Session 1 introduces the paper. The theoretical background will be discussed in Session2. Session 3 will be discussed the algorithm of ID3. The design and implementations are presented in Session 4. Session 5 discusses the experimental results and concludes the paper in Session 6.

2. Theoretical Background

In this session, the background theory of Classification and Decision Tree probabilistic model will be discussed.

2.1 Classification

Classification is the process of finding a set of models that describe the model and also distinguish data classes or concepts, for the purpose of being able to use the model to predict the class of object whose class label is unknown. The derive model is based on the analysis of a set of training data. [5]

The derived model may be represented in various forms, such as classification (IF-THEN) rules, decision tree, mathematical formulae or neural networks. Classification can be used for predicting the class label of data objects; user may wish to predict some mission or unavailable data values.

A model is built describing a predetermined set of data classes. The model is constructed by analyzing data tuples described by attributes. Each topple is assumed to belong to a predefined class, as determined by one of the attributes, called the class label attribute in the context of classification, data tuples are also referred to as samples, examples or objects.

2.2 Decision Tree probabilistic model

Decision tree is a tree in which each branch node represents a choice between a number of alternatives, and each leaf node represents a decision. Decision tree starts with a root node on which it is for users to take actions. From this node, users split each node recursively according to decision tree learning algorithm. The final result is a decision tree in which each branch represents a possible scenario of decision and outcome. Many types of Decision tree algorithm such as ID3, C4.5 and so on.

3. Iterative Dichotomiser

ID3 based on a well known and influential algorithm for the task of decision tree learning. ID3 are widely used in real clinical applications.

ID3 is a simple decision tree learning algorithm. The basis idea of ID3 is to construct decision tree by employing top-down, greedy search through the given sets, to test each attribute at every tree node. In order to select the attribute that is most useful for classifying a give sets, requires a metric— information gain.

3.1. Information Gain

In information theory, entropy is a measure of the uncertainty about a source of messages. The more uncertain a receiver is about a source of messages, the more information that receiver will need in order to know what message has been sent [5].Information gain for whole database is computed by Equation (1):

$$Info\ (\mathbf{D}) = -\sum_{i=1}^{m} p_i \ Log_2\ (p_i) \qquad (1)$$

Where D is the whole recordset in database, m is the number of class label, p_i is the probability that an arbitrary tuple in **D** belongs to class C_i and is estimated by $|Ci, D| / |D|$. **D** is the Database. A log function to the base is used, because the information is encoded in bits. Info (**D**) is just the average amount of information needed to identify the class label of a tuple in **D**. *Info* (**D**) is known as the entropy of **D**.

Partition the tuples in **D** on some attribute A having v distinct values, $\{a_1. a_2 \ldots a_v\}$, observed from the training data. Attribute A can be used to split D into v partitions, $\{D_1, D_2, \ldots, D_v\}$, where D_j contains those tuples in D that has outcome a_j of A. These partitions correspond to the branches grown from node A.

Entropy for each attribute may be classified by Equation (2):

$$\text{info}_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \text{info}(D_j) \qquad (2)$$

where $\frac{|D_j|}{|D|}$ is the weight of the *j*th partition. $\text{info}(D_j)$ is the expected information required to classify a tuple from D based on the partitioning by A.

Gain for each attributes may be classified by Equation (3):

$$Gain(A) = Info(D) - \text{info}_A(D) \qquad (3)$$

Gain (A) is how much would be gained by branching on A. The algorithm computes the information gain of each attribute. The attribute A with the highest information gain is chosen as the splitting attribute at node N. A node is created and labeled with the attribute, branches are created for each value of the attribute, and the samples are partitioned accordingly.

4. Design and Implementation

The overall system design can be seen in Figure 1. This system composed with one database: Training and Testing Database which stores patients' records. This database consists of two tables: Training and Testing table. Training table is used for producing rule and Testing table are used for calculating accuracy.

This system requires input data from the user. These input data are match with Rule Table that is generated by classifying the training data of the patient using ID3 algorithm. Finally, level of depression for the user can be generated. Testing data are used for calculating accuracy of the generated rules.

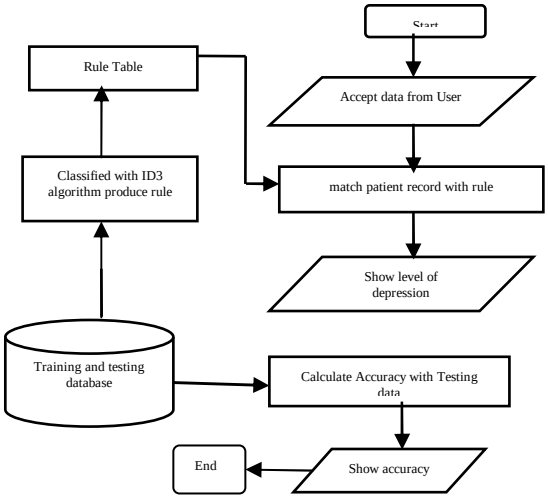

Figure 1 System Design

4.1 Attribute Information

In this system, the training data set consists of eleven attributes and four classes. Table 1 described attributes and possible values of those attributes. Class label of the system is described in Table 2.

Table 1 Attribute List

Attribute	Description
Loss of interest(LOI)	Yes, No
Depress mood(DM)	Yes, No
Sleeping condition(SC)	Insomnia, Normal, Oversleep
Eating condition(EC)	Appetite, Normal, Overeating
Body condition(BC)	Retarded, Normal, Agitated
Constipation(C)	Yes, No
Decrease Libido(DL)	Yes, No
Helplessness(H)	Yes, No
Guilt feeling(GF)	Yes, No
Delusion(D)	Yes, No
Hallucination(H)	Yes, No

Table 2 Class label of the system

Class label	Description
1	No Depression
2	Mild Depression
3	Moderate Depression
4	Severe Depression

4.2 Induction of Decision Tree using Information gain

This session describes how the decision tree for depression test will be built using the sample dataset. The sample dataset for the system is shown in Table 3.

Table 3 Sample dataset for system

LOI	DM	SC	EC	BC	C	DL	H	GF	D	H	Diseases
Yes	Yes	Insomnia	Normal	Retarded	No	No	No	No	Yes	No	Severe
Yes	No	oversleep	Appetite	Agitate	Yes	Yes	Yes	Yes	No	Yes	Severe
No	yes	Normal	Normal	Normal	Yes	No	No	No	No	No	No
No	No	Normal	Overeating	Normal	No	No	Yes	No	No	No	No
Yes	Yes	Oversleep	Normal	Normal	Yes	No	Yes	No	No	No	Moderate
Yes	No	Insomnia	Appetite	Normal	Yes	Yes	No	Yes	No	No	Moderate
Yes	No	Normal	Normal	Normal	Yes	Yes	No	No	No	No	Mild
Yes	No	Normal	Normal	Agitate	No	Yes	No	No	No	No	Mild
No	No	Oversleep	Normal	Normal	Normal	No	No	No	No	No	No
No	Yes	Insomnia	Appetite	Retarded	Yes	Yes	Yes	Yes	Yes	Yes	Severe

In sample database (D), the class attribute Diseases, has four distances, namely, mild, moderate, severe and no depression. So, m = 4. Let class C1 correspond to mild, C2 corresponds to moderate, C3 corresponds to severe and C4 corresponds to no depression. A root node is created for the tuple in D. First, the expected information is computed to classify a tuple in D according to Equation (1):

$$\text{Info}(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

$$= -\frac{3}{10}\log_2\frac{3}{10} - \frac{2}{10}\log_2\frac{2}{10} - \frac{3}{10}\log_2\frac{3}{10}$$

$$= 0.5211+0.4644+.4644+0.5211$$

$$= 1.9701$$

Next, the expected information for each attribute has to be computed according to Equation (2). So, expected information for attribute (Loss of Interest) is:

$$\text{info}_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \text{info}(D_j)$$

$$Info_{LOI}(\mathbf{D}) = \frac{6}{10}\left[-\frac{2}{6}\log_2\frac{2}{6} - \frac{2}{6}\log_2\frac{2}{6} - \frac{2}{6}\log_2\frac{2}{6}\right] + \frac{4}{10}\left[-\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4}\right]$$

$$= 1.2755$$

Then, the gain information for Loss of interest is computed in Equation (3):

Gain (loss of interest) = 1.9701 – 1.2755 = 0.6946

Similarly, the gain information for each attribute is computed. They are Gain (Depress mood) = 0.2192, Gain (Sleeping condition) = 0.8191, etc.

Body condition has the highest information gain among attributes, it is selected as these splitting attribute. So, Body condition is defined as root node and branches are grown for each of the attribute's values. The tuples are then partitioned accordingly in Sample Dataset. The tuples falling into the partition for Body condition=Retarded all belong to the same class. Because they belong to class 'Severe', a leaf should be created at the end of this branch and labeled with 'Severe'. The tuples falling into the partition for Body condition=Agitated and Body condition=Normal are not belong to the same class. So, tuples are again partitioned.

Finally the final decision tree will be produced that can be seen in Figure 2:

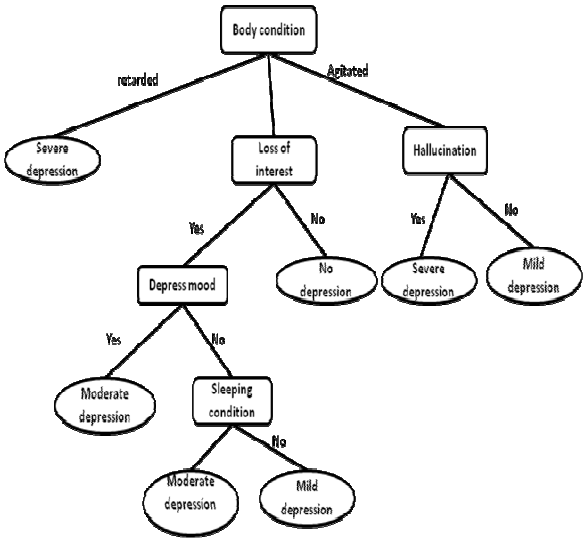

Figure 2 Final Decision Tree for sample dataset

4.3 Rule Extraction from sample Decision Tree

Decision tree classifier is popular method of classification. It is easy to understand how decision trees work and they are known for their accuracy. Decision tree can become large and difficult to interpret. So, built the rule based classifier by extracting IF-THEN rules from a decision tree. IF-THEN rules may be easier for humans to understand.

The rules extracted from the decision tree shown in Figure 2 are:

Rule 1: IF Body condition = Normal AND Loss of interest = Yes AND Depress mood = Yes THEN Diseases = Moderate depression

Rule 2: IF Body condition = Retarded THEN Diseases=Severe depression

Rule 3: IF Body condition = Normal AND Loss of Interest = Yes AND Depress mood = No AND Sleeping condition =Yes THEN Diseases=Moderate depression

Rule 4: IF Body condition = Normal AND Loss of Interest = Yes AND Depress mood = No AND Sleeping condition = No THEN Diseases = Mild depression

Rule 5: IF Body condition = Normal AND Loss of Interest=No THEN Diseases = No depression

Rule 6: IF Body condition = Agitated AND Hallucination = Yes THEN Disease = Severe depression

Rule 7: IF Body condition = Agitated AND Hallucination = No THEN Diseases = Mild depression

5. Experimental Results

The system is implemented over 3008 patients and their records are saved in Training table. Decision Tree will be produced using those Training data. In Testing table data of 1000 patient are stored and that are used for calculating accuracy of the generated rules.

5.1 System Accuracy

To evaluate the classifier on testing data, we defined the accuracy as follows:

Accuracy=correctly classified diseases for each class label / Total diseases in each class label * 100 (4)

The accuracy is calculated by according to Equation 4 by using the data records in Testing table. Accuracy for the system can be seen in Table 4.

Table 4 Accuracy for the system

	No of Records in database		
	Training	**Testing**	**Accuracy**
No depression	432	250	98.8%
Mild depression	135	250	98.6%
Moderate depression	108	250	98.6%
Severe depression	2333	250	100%
Total	3008	1000	99%

6. Conclusion

The mind and the body are two parts of the whole person and should never through of separately. Every person gets a lot of stress in their workplace because they try to progress their level of jobs. They will probably get a depression concerning with their jobs. This proposed system is intended for the physician to decide whether the patient is suffering depression or not. Using this system, Physicians can take the most appropriate decision concerning whether the users are suffering depression or not within a minimum of time.

References

[1] A.Seime, "Web Mining; Application and Techniques", State University of New York College at Brockport, USA.

[2] H.Lu.R.Setino, and H.Liu, "Neurorule: A connectionist approach to data mining" VLDB, Swizerland, 1995

[3] J.Ham and M.Kamber, "Data Mining Concepts and Techniques", Kaufmann, 2001

[4] J.Han and M.Kamber,"Data Mining Concept and Techniques", Second Edition

[5] L.D.Radet, "Principle of Data Mining and Knowledge Discovery", ISBN 3540425349, Oct 1,2001 by Springer

[6] R.Wu,., W.Peters, M.W Morgan,.: "The Next Generation Clinical Decision Support: Linking Evidence to Best Practice", Journal Healthcare Information Management. 16(4), 50-55, 2002