Diagnosis of Hypothyroid Disease by Using Decision Tree Induction

May Zin Htut, Thin Zar Win
Computer University, Kyainge Tong
mayzinhtut777@gmail.com,thinzarwin07@gmail.com

Abstract

Data mining is the process of discovering useful information underlying the data. Powerful techniques are needed to extract patterns from large data because traditional statistical tools are efficient enough any more. Classification and prediction are two forms of data analysis that can be used predict future data trends. In this paper, we proposed a diagnosis system for Hypothyroid disease by using decision tree induction algorithm. Decision tree induction algorithms have been used for classification in a wide range of application domains. Decision trees are potentially powerful predictors and explicitly represent the structure of a dataset. In this system, C4.5 classification algorithm is analyzed to build the model from a given set of training data and then produced classification rules. It examines the patient is positive or negative in hypothyroid. The performance evaluation of our system is also discussed in this paper.

Keywords: Data mining, Classification, Decision Tree Induction, Cross Validation method.

1. Introduction

Hypothyroid is an under active thyroid gland. Hypothyroid means that the thyroid gland can't make enough thyroid hormone to keep the body running normally. People are hypothyroid if they have too little thyroid hormone in the blood. Hypothyroid is more common than you would believe, and millions of people are currently hypothyroid and don't know it.

Classification is one of data analysis that can be used to extract models to predict future data trends. Decision Tree Induction is one of the most effective and efficient classification algorithms. The aim to use classification algorithm especially C4.5 method is to deal with Hypothyroid data for new patients. We modeled these diagnostics system using Decision Tree Induction technique to provide an estimation of the probability of hypothyroid or negative.

2. Related Work

Hypothyroid occurs when the thyroid gland does not produce enough thyroid hormone to meet the body's needs. Without enough thyroid hormone, many of body's functions slow down. About 5 percent of the U.S. population has Hypothyroid. Women are much more likely than men to develop Hypothyroid. Data mining technique plays an important role in knowledge discovery process for extracting the most useful data from several records. Decision tree induction algorithm is one of the most popular algorithms in the mining classification. Discovering all classification rules from very large dataset. Data mining system, which is able to mine classification rule directly from set-valued data of hypothyroid disease [3].

Classification is a well-studied important problem. It has many applications. It has been used in the insurance for industry, for tax and credit card fraud detection, for medical diagnosis and so forth [1]. The aim of these techniques is to classify the data records based on the class label and predicts the class of new data. Depending on the nature of data interaction, there are different classification techniques such as Nave Bayesian Classifier [5], Back Propagation techniques [3], case based reasoning approach [2].

Decision Trees are powerful and popular tools for classification. Rules for classifying data using attributes. The tree consists of decision nodes and leaf node. A decision node has two or more branches, each representing values for the attributes are tested. A leaf node attribute produces a homogenous result (all in one class), which does not require addition classification testing. In the prepruning approach, a tree is "pruned" by halting its construction early. Upon halting, the node becomes a leaf. The leaf may hold the most frequent class among the subset samples or the probability distribution of those samples. Stratified k-fold cross-validation is a recommended method for estimating classifier accuracy [3].

3. Background Theory

In classification process, variable selection is different and important problem in machine learning. For classification task, it can lead to increase accuracy or to reduce computational costs. C4.5 mainly classifies the data by variable selection and builds the decision tree by recursively selecting attribute on which to split.

It is also the best method to implement continuous attributes C4.5 approach calculates the information gain for continuous attribute by sorting the values of this attribute in increasing order and then finding midpoint between each pair of adjacent values the point with the minimum expected information requirement is selected as the split point for this attribute >split_point1 and attribute<=split_point2.

The critical steps for variable for selection process in C4.5 are as follows. Firstly, it groups the values of class label attribute and computes the information conveyed or entropy of class label attribute by using the equation

$$Info(T) = -(p_1 * \log(p_1) + p_2 * \log(p_2) + ... + p_n * \log(p_n)).....Equation(1)$$

Where, T represents class label attribute and P_n represents the probability of each group of class label attributes. Secondly, it calculates the expected information of other remaining attributes based on the groups of class label attributes by using the equation (2):

$$SplitInfo(att, T) = \sum((\frac{att_i}{att}) * Info|att_i|.....Equation(2)$$

Where, att represent other remaining attributes. Thirdly, it calculates the quantity gain of each attribute by subtracting the value of expected information from the value of entropy of class label attribute by using the equation(3)

$$Gain(att, T) = Info(T) - SplitInfo(att, T).....Equation(3)$$

In which, information gain measure is used to select the variable at each node in containing the decision tree. C4.5 uses gain ratio to overcome the problem (normalization to information). So, finally it calculates the gain ratio by using the equation (4):

$$GainRatio(att, T) = Gain^{(att,T)}\big/_{SplitInfo(att,T)}.....Equation(4)$$

The attribute with the maximum gain ratio is selected as the best attribute to split the data

Classification is the process of finding a set of models (or function) that describe and distinguish data classes or concepts, for the purpose of being able to predict the class of object whose is unknown. The derived model is based on the analysis of a set of training data. In decision tree algorithm, this attribute is used as a root node and formulate a logical test on that attribute. Each branch represents of the test and other remaining attributes of training set are placed as child nodes of this attribute. And then the same process runs recursively on each child node. Finally, termination rule specifies when to declare a leaf node.

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

Data classification is a two-step process. In the first step, a model is built describing a predetermined set of data class. The model is constructed by analyzing database tuples described by attributes. Each tuples 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.

In the second step, the model is used for classification. First, the predictive accuracy of the model or classifier is estimated. If the accuracy of the classifier is considered acceptable, the model can be used to classify future data tuples for which the class label is not known. Such data are also referred to as "unknown" or "previously unseen" data. [4]

The classifier design can be performed with labeled or unlabeled data. Using a supervised learning method the computer is given a set of objects with known classification and is asked to classify an unknown object based on the information acquired by it during the training phase.

3.1 Decision Tree Induction

A decision tree classifier is a simple yet widely used classification techniques. The tree has three types of nodes. A root node that has no incoming edges and zero or more outgoing edges. Internal nodes, each of which has exactly one incoming edge and two or more outgoing edges. Leaf r terminal nodes, each of which has exactly one incoming edge and no outgoing edges. In a decision tree, each leaf node is assigned a class label. The non-terminal nodes, which include the root and other internal nodes, contain attribute test condition to separate records that have different characteristics. Classifying a test record is straightforward once a decision tree has been constructed. Starting from the

root node, apply the test condition true the record and follow the appropriate branch based on the outcome of the test. This will lead to another internal node, for which a new test condition is applied, or to a leaf node. The class label associated with the leaf node is then assigned to the record [7].

4. Proposed System

Our proposed system has two basic steps. In the first step, a model is built describing a predetermined set of data from hypothyroid patient. In the second, the model is used for classification and the predictive accuracy of the model (or) classifier is estimated. The capability and practical use of this system was proved in testing the hypothyroid patients.

If a new patient wants to know whether he or she has hypothyroid or not, the system can diagnose the disease according to the patient's symptoms. The system will predict whether the patient is positive or not (Hypothyroid). Our proposed system's architecture is shown in figure (1).

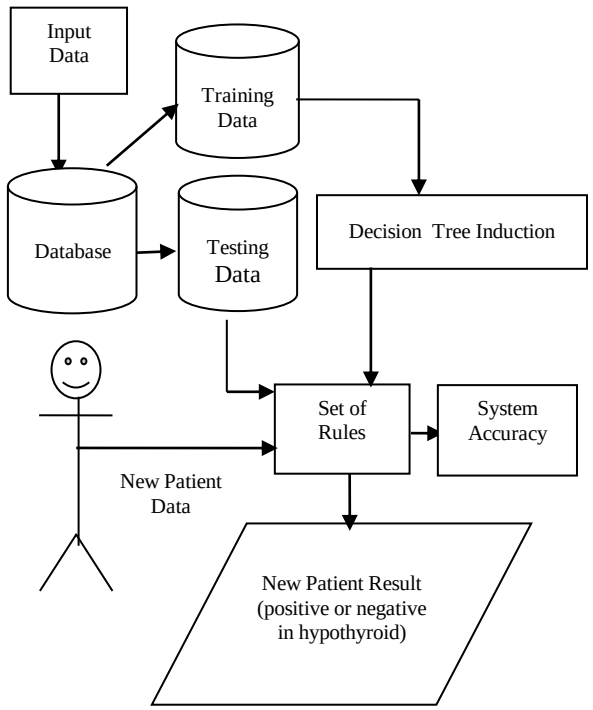

Figure 1.Proposed system Architecture

5. Dataset of the system

Many of the attributes in the data may be irrelevant to the classification. Furthermore, some attributes may be redundant. Hence, relevance analysis may be performed on the data with the aim of removing any relevant or redundant attributes from learning process. This system can be used attributes selection for hypothyroid data. The data can be generalized to higher-level concepts. This is particularly useful for continuous-valued attributes [6]. In training data set, there are 13 attributes. There are 2 class labels, positive or negative in hypothyroid. The following table (1) describes name and value of each attribute.

Table 1. Attribute of the system

ATTRIBUTE NAME	ATTRIBUTE VALUE
Age	Over50, Between40and51, Between3oand41, Between19and31, Under20,Anyage
Sex	M,F
TSH_measured	y,n
TSH	Zero0,Overll, 0.3to3.0, 0.5to5.90,
T3_measured	y,n
T3	2.0to4.80,0.4to3.10,
TT4_measured	y,n
TT4	Over39,13to39, 2to12, 0.03to0.5,0.15to1.50
T4U_measured	y,n
T4U	1.01to3.90, 0.2to0.98
FTI_measured	y,n
FTI	1to30, Over70,31to50, 51to70, Zero0,

6. System Implementation

This system is intended to develop the Diagnosis of Hypothyroid, by using C4.5 decision tree algorithm. The hypothyroid dataset contain 13 attributes and 2 class labels. The system can predict positive or negative in hypothyroid based on the patients' symptoms. The user who wants to know the symptoms of hypothyroid can choose the desire attributes in the hypothyroid dataset. Equation (1) is used to compute the expected information needed to classify a tuple in T. Equation (2) is applied to compute the expected information of each attribute.

By using the equation (3), the highest information gain among the attribute is selected and created as root node. According to our training dataset, "FTI" attribute got highest information gain and it is chosen as root node as in figure (2). By using Equation (4), the potential information is generated by splitting the training data set.

Finally, the decision tree is generated as in figure (2). The knowledge represented in decision tree can be extracted and represented in the form of classification IF-THEN rules. One rule is created for each path from root to a leaf. Each (attribute, value) pair along a path forms a conjunction and leaf node holds the class prediction. The sample rules of our system are as follows:

RULES(1)-IF FTI=Over70 THEN Result="Negative"
RULES(2)-IF FTI=Zero THEN Result="Negative"
RULES(3)-IF FTI=31 to 50 AND TT4=y AND TSH=Zero THEN Result="Negative"
RULES(4)-IF FTI=31to50 AND TT4=y AND TSH=o.3to3.0 THEN Result="Positive"
RULES (5)-IF FTI=31to50 AND TT4=y AND TSH=4.1to6.90 THEN Result="Positive"

After generating the rules, the system can predict the type positive or negative in hypothyroid according to new patients' symptoms.

```
FTI = Over70: Negative (774.0/3.0)
FTI = Zero0: Negative (2.0/1.0)
FTI = 31to50
|  TT4_measured = y
|  |  TSH = Zero0: Negative (4.0)
|  |  TSH = 0.3to3.0: Positive (3.0)
|  |  TSH = 4.1to6.90: Positive (0.0)
|  |  TSH = Over11: Positive (22.0)
|  |  TSH = 0.5to5.90: Positive (0.0)
|  TT4_measured = n: Negative (8.0)
FTI = 1to30
|  T3_measured = y: Positive (34.0/1.0)
|  T3_measured = n: Negative (27.0/1.0)
FTI = 31to51: Negative (2.0)
FTI = 51to70
|  TSH = Zero0: Negative (1.0)
|  TSH = 0.3to3.0: Positive (0.0)
|  TSH = 4.1to6.90: Positive (0.0)
|  TSH = Over11: Positive (14.0/1.0)
|  TSH = 0.5to5.90: Negative (9.0)
```

Figure 2. Decision tree for hypothyroid disease

7. Performance Evaluation

Evaluating the performance of learning the algorithms was a fundamental aspect of machine learning. Estimating classifier accuracy was important in that allows one to evaluate how accurately a give classifier will label future data, that is, data on which the classifier has not been trained and also help in the comparison different classifier.

Using the training set to measure accuracy will typically provide an optimistically biased estimate and can result in misleading overoptimistic estimates due to over specialization of the learning algorithm to the data. Holdout and cross-validation are two common techniques for accessing classifier accuracy, based on randomly sampled partitions of the given data. Our system is based on cross-validation method.

In k-fold cross-validation, the initial data are randomly partitioned into k mutually exclusive subsets or "folds," $S_1, S_2, \ldots S_k$ each of approximately equal size. Training and testing is performed k times. In iteration i, the subset S_i is reserved as the test set, and the remaining subsets are collectively used to train the classifier. That is, the classifier of the first iteration is trained on subsets $S_2, \ldots, S_k$ and tested on S_1; the classifier of the second iteration is trained on subsets $S_1, S_3, \ldots, S_k$ and tested on S_2; and so on.

The accuracy estimate is the overall number of correct classifications from the k iterations, divided by the total number of samples in the initial data. In stratified cross validation, the folds are stratified so that the class distribution of the samples of the samples in each fold is approximately the same as that in the initial data.

In our proposed system, we tested with 900 records of hypothyroid data. Among then 400 records are used as testing data and 500 records are used as training from general description of thyroid disease databases and related files. When we used C4.5 classification approach to build the model, it took 0.16 seconds to build the model and took 0.02 seconds to test the model on training and testing data. Result show that our system provides an accuracy of about 99.2% and the error rate is 0.82%.

8. Conclusion

In this paper, the decision tree induction algorithm is used to diagnose hypothyroid disease. Our proposed system can predict positive or negative in hypothyroid based on the patient's symptoms. The mention symptoms and disorder categories of assessment levels in this system can be successfully differentiated with a certain amount of accuracy. In future, we intend to undertake training with larger data sets, using different patients' attributes and various types of conditions. We expect that our system can be exploited as a useful tool for medical field although it needs further modifications.

9. References

[1] A.Scime, “Web Mining; Applications and Techniques”, State University of New Yor Collage at Brockport, USA.

[2]P.J. Camp, H. Groves, and J. L. Kolodner, “Modeling and Case-Based Reasoning in Support of the Reflective Inquiry in Earth.”, Science College of Computing, Georgia Institute of Technology, Atlanta.

[3]J. Han and M. Kamber, “Data Mining: Concepts and Techniques”, Morgan Kaufmann,2001.

[4] J. Han, and M. Xin,” Discovering web Access patterns and trends by applying OLAP and data mining technology on Web”.

[5] Y. Tsuruoka and J. Tsujii, “Training a Naïve Bayes Classifier via the EM Algorithm with a Class Distribution Constraint”, Department of Computer Science.

[6] P.Ning, T.M.Steinbach, V.Kumar, “Introduction to Data Mining”.

[7] L.D.Radet, "Principles of Data Mining and Knowledge Discovery", ISBN 3540425349, Oct 1, 2001 by springer.