Systemic Lupus Erythematosus (SLE) Disease by Using ID3 Algorithm

Thin Aye Phyu, Nan Si Kham
University of Computer Studies, Yangon
thinayephyu@gmail.com,nansikham@gmail.com

Abstract

Decision tree algorithm is a method for approximating discrete valued target functions, in which the learned function is represented by a decision tree. A major challenge facing healthcare organizations (hospitals, medical centers) is the provision of quality services at affordable costs. Most hospitals today employ some sort of hospital information systems to manage their healthcare or patient data. Quality service implies diagnosing patients correctly and administering treatments that are effective. Poor clinical decisions can lead to disastrous consequences which are therefore unacceptable. Hospitals must also minimize the cost of clinical tests. The main task performed in these systems is using inductive methods to the given values of attributes of an unknown object to determine appropriate classification according to decision tree rules. It is shown how the basic rules based on medical data and represented in computer system based on ID3 algorithm.. In this paper, the system implements basic ID3 algorithm. Our results illustrate a truly complementary effort of human and computers for early diagnosis of SLE from symptoms.

Key Words : Decision Tree, ID3 Algorithm, Systemic Erythematous prediction

1. Introduction

A 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 are commonly used for gaining information for the purpose of decision -making. 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 its outcome. Decision tree learning is a method for approximating discrete-valued target functions, in which the learned function is represented by a decision tree.

Decision tree learning is one of the most widely used and practical methods for inductive inference'. Decision tree learning algorithm has been successfully used in expert systems in capturing knowledge. The main task performed in these systems is using inductive methods to the given values of attributes of an unknown object to determine appropriate classification according to decision tree rules. Decision trees classify instances by traverse from root node to leaf node. We start from root node of decision tree, testing the attribute specified by this node, then moving down the tree branch according to the attribute value in the given set. This process is the repeated at the sub-tree level.

ID3 is a simple decision learning algorithm developed by J. Ross Quinlan (1986)[1]. ID3 constructs uses statistical property call *information gain* to select which attribute to test at each node in the tree. Information gain measures how well a given attribute separates the training examples according to their target classification.

Minos Garofalakis, Dongjoon Hyun, Rajeev Rastogi and Kyuseok Shim [6] proposed that classification is an important problem in data mining. A number of popular classifiers construct decision trees to generate class models. Frequently, however, the constructed trees are complex with hundreds of nodes and thus difficult to comprehend, a fact that calls into question an often cited benefit that decision trees are easy to interpret. At there, they addressed the problem of constructing "simple" decision trees with few nodes that are easy for humans to interpret. By permitting users to specify constraints on tree size or accuracy, and then building the" best" tree that satisfies the constraints, and ensured that the final tree is both easy to understand and has good accuracy.

In this paper, Decision tree algorithm is commonly used for gaining information for the purpose of decision–support. In this algorithm, an entropy- based attributes selection measure is used to select the test attribute at each node in the tree. After a training period, it is able to retrieve material of interest to the user. Here, the system presents decision support system for SLE disease level by using symptoms.

2. ID3 Algorithm

2.1 Decision Tree

Decision tree algorithm is a data mining induction techniques that recursively partitions a data set of records using depth-first greedy approach (Hunts et al, 1966) or breadth-first approach (Shafer et al, 1996) until all the data items belong to a particular class. A decision tree structure is made of root, internal and leaf nodes. The tree structure is used in classifying unknown data records. At each internal node of the tree, a decision of best split is made using impurity measures (Quinlan, 1993). The tree leaves is made up of the class labels which the data items have been group.

Decision tree classification technique is performed in two phases: tree building and tree pruning. Tree building is done in top-down manner. It is during this phase that the tree is recursively partitioned till all the data items belong to the same class label (Hunts et al, 1966). It is very tasking and computationally intensive as the training data set is traversed repeatedly. Tree pruning is done is a bottom-up fashion. It is used to improve the prediction and classification accuracy of the algorithm by minimizing over-fitting (noise or much detail in the training data set) (Mehta et al, 1996). Tree pruning is less tasking compared to the tree growth phase as the training data set is scanned only once. In this study we will review Decision tree algorithms implemented in a serial pattern, identify the algorithms commonly used and compare their

classification accuracy and execution time by experimental analysis

Decision trees classify instances by sorting them down the tree from the root to some leaf node, which provides the classification of the instance. Each node in the tree specifies a test of some attribute of the instance and each branch descending from that node corresponds to one of the possible values for this attribute.

The reasons for decision learning tree algorithms to be attractive are:

1. They generalize in a better way for unobserved instances, once examined the attribute value pair in the training data.
2. They are efficient in computation as it is proportional to the number of training instances observed.
3. The tree interpretation gives a good understanding of how to classify instances based on attributes arranged on the basis of information they provide and makes the classification process self-evident.

There are various algorithms in this area like ID3, C4.5, ASSISTANT etc. This paper selected ID3 algorithm to evaluate because it builds tree based on the information (information gain) obtained from the training instances and then uses the same to classify the test data. ID3 algorithm generally uses nominal attributes for classification with no missing values. ID3 can even work well on datasets with missing attribute values to certain extent. The basics of the algorithm are explained in brief and then implementation and evaluation part is elaborated.

2.2 Attribute Selection Measure

The information gain measure is used to select the test attribute at each node in the tree. Such a measure is referred to as an attribute selection measure or a measure of the goodness of split. The attribute with the highest information gain (or greatest entropy reduction) is chosen as the test attribute for the current node. This attribute minimizes the information needed to classify the samples in the resulting partitions and reflects the least randomness or "impurity" in these partitions. Such and information-theoretic approach minimizes the expected number of tests needed to classify an object and guarantees that a simple (but not necessarily the simplest) tree is found.

Let S be a set consisting of s data samples. Suppose the class label attribute has m distinct values defining m distinct classes, Ci (for i = 1, … , m). Let si be the number of samples of S in class Ci. The expected information needed to classify a given sample is given by

$$I(s_1,s_2,\ldots,s_m) = -\sum_{i=1}^{m} p_i \log_2 (p_i), \quad . \ (1)$$

where pi is the probability that an arbitrary sample belongs to class Ci and is estimated by si /s. Note that a log function to the base 2 is used since the information is encoded in bits.

Let attribute A have v distinct values,$\{a_1, a_2,\ldots a_v\}$.Attribute A can be used to partition S into v subsets, $\{ S_1, S_2,\ldots, Sv\}$, where Sj contains those samples in S that have value a_j of A. If A were selected as the test attribute (i.e., the best attribute for splitting) then these subsets would correspond to the branches grown from the node containing the set S.Let s_{ij} be the number of samples of class Ci in a subset S_j.The entropy, or expected information based on the partitioning into subsets by A, is given

$$E(A) = \sum_{j=1}^{v} (s_{1j} + \ldots + s_{mj}) / s \; I (s_{1j} , \ldots , s_{mj}) \ldots \ (2)$$

The term $(s_{1j} + \ldots + s_{mj}) / s$ acts as the weight of the jth subset and is the number of samples in the subset (i.e., having aj of A) divided by the total number of samples in S .The smaller the entropy value, the greater the purity of the subset partitions. Note that for a given subset Sj,

$$I (s_{1j} ,s_{2j} ,\ldots,s_{mj}) = -\sum_{i=1}^{m} p_{ij} \log_2 (p_{ij}) \quad (3)$$

where pij = sij / | sj | and is the probability that a sample in Sj belongs to class Ci.The encoding information that would be gained by
branching on A is

$$Gain(A) = I (s_1,s_2,\ldots,s_m) - E(A) \quad (4)$$

In other words, Gain (A) is the expected reduction in entropy caused by knowing the value of attribute A. The algorithm computes the information gain of each attribute. The attribute with the highest information gain is chosen as the test attribute for given set S. A node is created and labeled with the attribute, branches are created for each value of the attribute, and the samples are partitioned accordingly.[3].In this digital library system, there are 12 classes (m=12). Let class C1 correspond to class medical, C2 correspond to class technology, C3 correspond to class religious and so on. si mean that total count of each class from training data set and s is total count of training data set.

To compute the all class of expected information, we first use Equation(1). Next we need to compute the entropy of each attribute. Let's start with the attribute age, we need to look at the distribution of 12 classes samples for each value of age. We compute the expected information of age attribute for each of these distributions by using Equation (2). Using Equation (2), the expected information needs to classify a given sample if the samples are partitioned according to age attribute value. To compute information gain of age attribute we use Equation (4). The rest of attribute are also computed like this. Let age has the highest information gain among the attributes, it is selected as the test attribute. A node is created and labeled with age and branches are grown for each of the attribute's value. The samples are then partitioned accordingly each of the age attribute's value. The process is continued with recursive with each partition.

2.3. Tree Induction

The knowledge represented in decision trees can be extracted and represented in the form of classification IF-THEN rules. One rule is created for each path from the root to a leaf node. Each attribute value pair along a given path forms a conjunction in the rule antecedent ("IF" part). The leaf node holds the class prediction, forming the rule consequent ("THEN" part). The IF-THEN rules may be easier for humans to understand, particularly if the given tree is very large. Some parts of classification rule for SLE Disease Decision System are following:

IF (ESR == High AND PH == Positive) THEN decision = Serious;

IF (ESR == High AND PH == Negative AND MR == Negative) THEN decision = Serious;

IF (ESR == High AND PH == Negative AND MR == Positive) THEN decision = Moderate;

IF (ESR == Normal AND MR == Positive) THEN decision = Initial;

IF (ESR == Normal AND MR == Negative) THEN decision = Serious;

IF (ESR == Low) THEN decision = Moderate;

In this system, train the ID3 algorithm with following data attributes.

Table 1 the attribute and value of the system

Attributes	Values
Malar Rash	positive, negative
Discoid Rash	positive, negative
Photosensitivity	positive, negative
Arthritis	positive, negative
CNS disorder	positive, negative
Immunological	positive, negative
Hameoglobin	low,normal,high
Platelets	low,normal,high
ESR	low,normal,high
Oral Ulcer	positive, negative
antinuclear antibody	positive, negative
Renal disorder	low,normal,high
Fever	very high, high
White blood cells	low,normal,high
Class Labels	I,II,III

Table 2 Class Label

Class Label	Description
I	Initial
II	Moderate
III	Serious

3. Implementation and Prediction of the System

This system will ask expert to provide training data. From training data, a classification schema will be derived by using ID3 decision tree induction algorithm. The ID3 decision tree induction algorithm will produce decision tree and rules from training dataset. With the rules produced by ID3 algorithm, user can calculate the accuracy by providing testing dataset. In this system, fourteen attributes (symptoms) are used as input and three classes (disease level) are used as output result.

The system will calculate accuracy by matching produced rules form ID3 algorithm. With the SLE input training and testing data set, this system will be able to provide prediction for new patient data. With the symptoms provided by user this system will predict the condition of patient's disease. The classification process requires two set of data**:** training data and testing data. There are fourteen attributes and three classes in the system. The design of this system is described in Figure -1

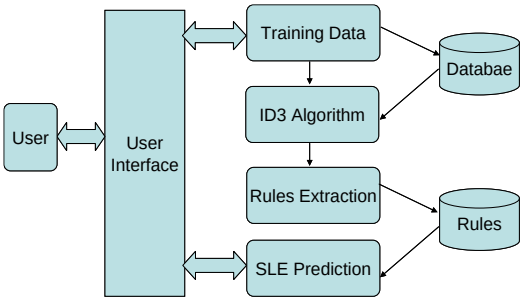

Figure. 1 SLE Diagnosis Architecture

4. Accuracy Assessment

Estimating classifier accuracy is important in that it allows one to evaluate how accurately a given classifier will label feature data, that is, the data on which the classifier has not been trained. In this paper, confusion matrix is used for the system accuracy.

An n*n confusion matrix[9] is created with each element c (i,j) representing the number of testing data classified as class i but the reference data classified the testing as belonging to the class j. Here, n is the number of classes. Clearly, the diagonal elements of the confusion matrix represent the correctly classified testing data of class i.

From the confusion matrix we calculate the various accuracy parameters.

a. Testing data Accuracy: defined as the number of correctly classified testing data of class i divided by the total number of testing data of class i from reference data.

b. Training data Accuracy: defined as the number of correctly classified testing data of class i divided by the total number of testing data of class i given by the classified rules.

c. Overall Accuracy is the ratio of correctly classified testing data to the total number of testing data, given by $\sum x_{ii} / n$, where n is the total number of testing data.

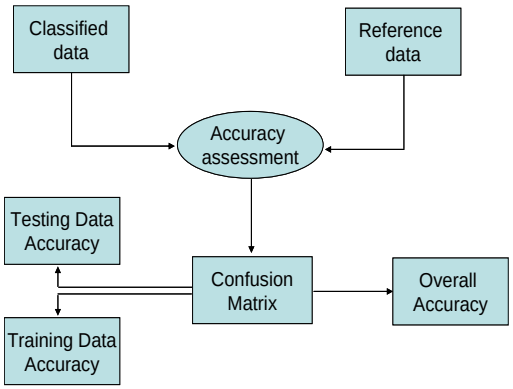

Figure. 2 Accuracy Assessment Figure

5. Conclusion

Decision support system are powerful tools integrating scientific methods for supporting complex decisions with techniques developed in medical application. In this paper, we proposed a decision tree for SLE disease. Our results demonstrate the power of having a large image feature set combined with effective feature reduction tools, and simple classifiers to reduce a massive feature set to a low dimensional, practically useful, and set of biomarkers for future MedicalRecord (MR) image classification. A prototype SLE disease prediction system is developed using three data mining classification modeling techniques. The system extracts hidden knowledge from a historical SLE disease database. The models are trained and validated against a test dataset. So the user will be received correct decision for SLE disease in short time and the system assist junior doctors for SLE disease.

6. References

[1] Chang HL., Kim JW., Lee BH., and Michael JS., "*Application of Decision-Tree Induction Technique to Personalized Advertisement on Internet Storefronts*". International Journal of Electronic Commerce 2001, Vol 5, No3. pp 45-62.

[2] Duda, R & Hart, "*Pattern classification and scene analysis"*, New York: John Wiley & Sons, 1973.

[3] David McG. Squire "The *ID3 Decision Tree Algorithm*" August , 2004

[4] Han J., and Kamber M., "*Data Mining: Concepts and Techniques*", Academic Press, USA, 2001

[5] Khet Khet Oo Tha "*Décision Support System for sustainable agricultural land management in Myanmar*"(Asian Institue of Technology-school of Advanced Technology, Bangkok , Thailand) December 2000

[6] Maeve Cumminges, Stephen Haag "*Information System Essential*"

[7] Mahar, Faizullah and Baiochistan, Khuzdar "Information Technology for Management and Decision Support Systems" *Pakistan Journal* of Information and Technology 2(1):58-60, 2003

[8] Marek Druzdzel J. and R.Flynn, Roger "*Decision support system*"

[9] Marco Bassetti, Massimiliano Bernabe, *"Valildation of CFS Classification with different data sources"* February, 2009

[10] Wei Peng, Juhua Chen, Haiping Zhou "*An Implementation of ID3---Decision Tree Learning algorithm*"

[11] Zin Lin Than "*Decision Support System For Sales Effective System Of Cosmetic*" May, 2009