Malaria Diagnosis System by Using ID3 Classification Algorithm

Khaing Mar Thuai, Moe Thant
Computer University (Kalay)
khaingmarthuai@gmail.com , umoethant@gmail.com

Abstract

Decision Tree algorithms are the most popular algorithms for classification in data mining field. The main goal of classification is prediction of the categorical labels (classes). In this system, ID3 algorithm is used to predict infection of malaria disease on patients by selecting training data (patients' medical records), constructing decision model and adjust the model based on testing data (part of patients' medical records). The constructed model is represented in the form of decision tree and classification rules. The choice of suitable model to predict malaria infection on patient can decide against the correctness of model (classifier accuracy). To get the best classifier accuracy, this system permits selecting no of records to train the system and remove unnecessary braches of tree.

Keywords: Classification Rule, Classifier, Classes, Data Mining, Decision Tree.

1. Introduction

Abilities of both generating and collecting data have been growing rapidly in the last several decades. Contributing factors include the common use of bar codes for most commercial products, the computerization of many business, scientific, government transactions, and advances in data collection tools ranging from scanned text and image platforms to satellite remote sensing systems. In addition, fashionable use of the World Wide Web as a global information system has flooded us with a tremendous amount of data and information. This explosive growth in stored data has generated an urgent need for new techniques and automated tools that can intelligently support us in transforming the vast amounts of data into useful information and knowledge.

In this paper, we choose to demonstrate the malaria diagnosis system by using ID3 classification Algorithm. It is a simple system for user who needs to classify the malaria infected patient without laboratory blood test. And also, it allows user to choose the best classification rules to determine malaria infected or not. The resulted classifier's accuracy can examine repeatedly. The best classifier may be used as classification rules that can be determined the next unknown case. To classify malaria infected patients, the fifteen significant symptoms are chosen for every patients. And then, these data are stored in database.

The rest of the paper is organized as follows: Section 2 presents Related Work of the system. Section 3 describes Theoretical Background. Propose System and Implementation of our approach is described in Section 4. We list the concluding remarks in section 5.

2. Related Work

USAMA M. FAYYAD and KEKI B. IRANI presented a result applicable to classification learning algorithms that generate decision trees or rules using the information entropy minimization heuristic for discretizing continuous-valued attribute [6]. In this system, discretizing step for continuous-valued attributes can be extend as preprocessing task before generating decision tree.

3. Background Theory

3.1. Data mining

Data mining is the process of extracting hidden patterns from data. As more data is gathered, with the amount of data doubling every three years [5], data mining is becoming an increasingly important tool to transform this data into information. It is commonly used in a wide range of profiling practices, such as marketing, surveillance, fraud detection and scientific discovery [3].

A primary reason for using data mining is to assist in the analysis of collections of observations of behaviour.

3.2. Classification

Data mining creates classification models by examining already classified data (cases) and inductively finding a predictive pattern. These existing cases may come from an historical database, such as people who have already undergone a particular medical treatment or moved to a new long distance service. They may come from an experiment in which a sample of the entire database is tested in the real world and the results used to create a classifier. For example, a sample of a mailing list would be sent an offer, and the results of the mailing used to develop a classification model to be applied to the entire database. Sometimes an expert classifies a sample of the database, and this classification is then used to create the model which will be applied to the entire database [8].

In this paper, we use decision tree algorithm known as ID3 (Iterative Dichotomiser). This algorithm does not require any domain knowledge or parameter setting, and

therefore is appropriate for class prediction. This algorithm can handle discrete value. And also the learning and classification of ID3 algorithm are simple and fast. In general, ID3 classifier has good accuracy. However, successful use may depend on the data at hand. In this system, ID3 algorithm is used for generating decision tree. Generated decision tree is used to predict the patient who is infected with malaria or not.

3.2.1. Process of Classification

Data classification is a two-step process. In the first step, a model is built describing a predetermined set of data classes or concepts. The model is constructed by analyzing database tuples described by attributes. Each tuple 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. The data tuples analyzed to build the model collectively form the training data set. The individual tuples making up the training set are referred to as training samples and are randomly selected form the sample population. Since the class label of each training sample is provided, this step is also known as supervised learning. It contrasts with unsupervised learning, in which the class label of each training sample is not known, and the number or set of classes to be learned may not be known in advance.

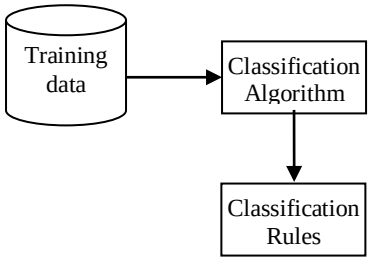

Figure 1 Learning Process in Classification

In the second step, the model is used for classification. First, the predictive accuracy of the model is estimated. The holdout method is a simple technique that uses a test set of class-labeled samples. These samples are randomly selected and are independent of the training samples. The accuracy of a model on a given test set is the percentage of test set samples that are correctly classified by the model.

If the accuracy of the model is considered acceptable, the model can be used to classify future data tuples or objects for which the class label is not known[3,Page 286].

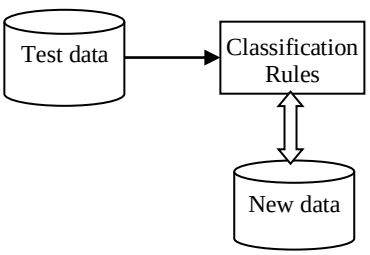

Figure 2 Classification Process in Classification

3.3. Decision Tree Learning

Decision tree learning, used in data mining and machine learning, uses a decision tree as a predictive model which maps observations about an item to conclusions about the item's target value. More descriptive names for such tree models are classification trees or regression trees. In these tree structures, leaves represent classifications and branches represent conjunctions of features that lead to those classifications.

In decision analysis, a decision tree can be used to visually and explicitly represent decisions and decision making. In data mining, a decision tree describes data but not decisions; rather the resulting classification tree can be an input for decision making.

3.3.1. Decision Tree

A decision tree (or tree diagram) is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. Decision trees are commonly used in operations research, specifically in decision analysis, to help identify a strategy most likely to reach a goal. Another use of description trees is as a descriptive means for calculating conditional probabilities [4].

3.3.2. ID3 Classification Algorithm

The basic algorithm for decision tree induction is a greedy algorithm that constructs decision trees in a top-down recursive divide-conquer manner. The algorithm, summarized, is a version of ID3, a well-known decision tree induction algorithm[3]. During the late 1970s and early 1980, J. Ross Quinlan, a researcher in machine learning, developed a decision tree algorithm known as ID3 (Iterative Dichotomiser).

The ID3 algorithm is used to build a decision tree, given a set of non-categorical attributes C_1, C_2, .., C_n, the categorical attribute C, and a training set T of records.

function ID3 (R: a set of non-categorical attributes,
C: the categorical attribute,
S: a training set) returns a
decision tree;
begin
If S is empty, return a single node with value Failure;
If S consists of records all with the same value for the categorical attribute,
return a single node with that value;
If R is empty, then return a single node with as value the most frequent of the values of the categorical attribute that are found in records of S; [note that then there will be errors, that is, records that will be improperly classified];
Let D be the attribute with largest Gain (D,S) among attributes in R;

Let {dj| j=1,2, .., m} be the values of attribute D;
Let {Sj| j=1,2, .., m} be the subsets of S consisting respectively of records with value dj for attribute D;
Return a tree with root labeled D and arcs labeled d1, d2, .., dm going respectively to the trees
ID3(R-{D}, C, S1), ID3(R-{D}, C, S2), .., ID3(R-{D}, C, Sm);
end ID3;

3.3.3. 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 an 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, C_i. The expected information needed to classify a given sample is given by

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

Where p_i is the probability that an arbitrary sample belongs to class C_i and is estimated by s_i/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, $\{s_1, s_2, \ldots s_v\}$, where s_j 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 C_i in a subset S_j. The entropy, or expected information based on the partitioning into subsets by A, is given by

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

The term $\frac{s_{1j} + \ldots + s_{mj}}{s}$ acts as the weight of the j^{th} subset and is the number of samples in the subset (i.e., having value a_j 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 for a given subset S_j,

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

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

The encoding information that would be gained by branching on A is

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

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 the 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,Page 296-298].

3.3.4. Estimating Classifier Accuracy

Using training data to derive a classifier and then to estimate the accuracy of the classifier 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 assessing classifier accuracy, based on randomly sampled partitions of the given data.

In the holdout method, the given data are randomly partitioned into two independent sets, a training set and a test set. Typically, two thirds of the data are allocated to the training set, and the remaining one third is allocated to the test set. The training set is used to derive the classifier, whose accuracy is estimated with the test set[3,Page364]. In this paper, holdout method is used for classifier accuracy.

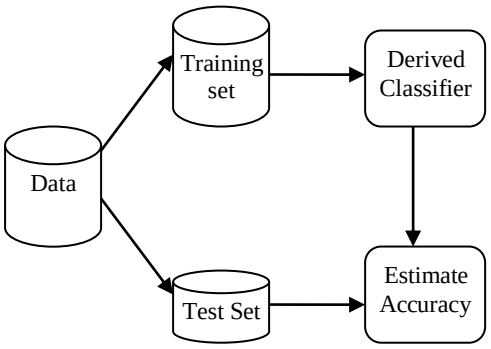

Figure 3 Holdout Method

4. System Implementation

4.1. Overview of the System

Users of this system may be anyone who wants to classify malaria infection on patients without knowing laboratory test results. In this system, user must separate patients' medical records, so system can recognize training data and testing data.

System is firstly trained with training data that are medical records of malaria infected patients including other facts to need to record patients. These records are stored in database. In the next step, user can draw the decision tree and then extract classification rules by training the system. In this step, user can prunch the tree node that does not support the result. The resulted rules are used to test with testing data in next step. In testing step, testing data are classified with rules and the resulted target class values are matched with known values. And then, percentage of correctness of system is shown to user. Therefore, user can decide the decision that another training and testing are need or not for more precise accuracy. System can be also presented the training data records and testing data records to user.

4.2. Database Design

Malaria is an infectious disease mainly found in tropical areas such as Sub-Saharan Africa, Central and South America, the Indian subcontinent, South East Asia and the Pacific islands which are known as malarious regions. In this paper, possible symptoms that cause malaria are used as attributes of malaria patient record. The attribute values are discrete values. Attributes and data types are shown in Table 1. Patient records' database is built in Microsoft Access.

Table 1. Design of Patient Records Database

No	Symptoms	Data Types	Values
1	ID	Integer	1,2,3,…..
2	TEMPERATURE	String	High, Normal
3	MALARIA_HISTORY	String	Yes, No
4	PRIMARY_FEVER	String	Yes, No
5	REGULAR_FEVER	String	Hot stage, Sweating stage, Normal, cold Stage
6	CHILLnRIGOR	String	Yes, No
7	HEADAGE	String	Yes, No
8	VOMITING	String	Yes, No
9	DELIRIUM	String	Yes, No
10	DYSPNEA	String	Yes, No
11	WEAKNESS	String	Yes,No
12	DROWNESS	String	Yes, No
13	OVERSWEATING	String	Yes, No
14	JAUNDICE	String	Yes, No
15	FITSnCOMA	String	Yes, No
16	CCE_SHOCK	String	Yes, No
17	RESULT	String	Malaria, Not Malaria

4.3. System Architecture & Design

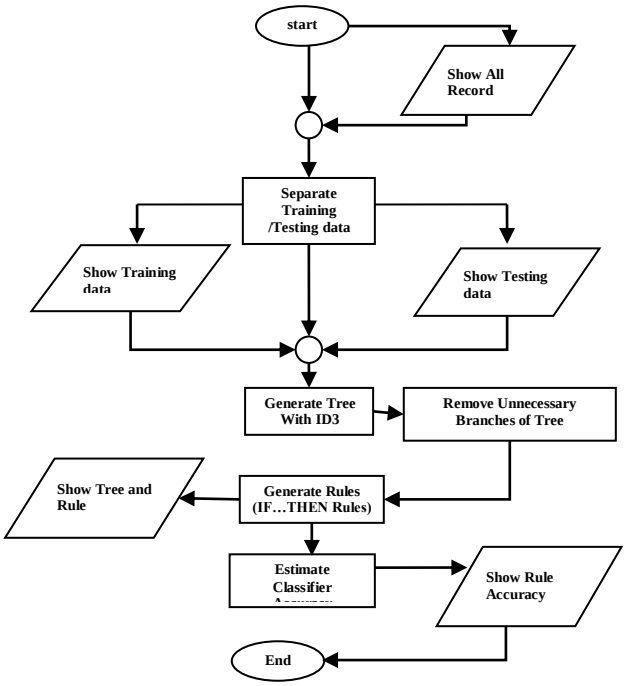

Figure 4 System Flow Chart for Malaria Diagnosis System

In Malaria diagnosis system, firstly user can view all sample records or user can separate training and testing data directly without viewing sample records. Separated training and testing data records can view after separating process. Figure 5 and 6 show the output of training data and testing data records.

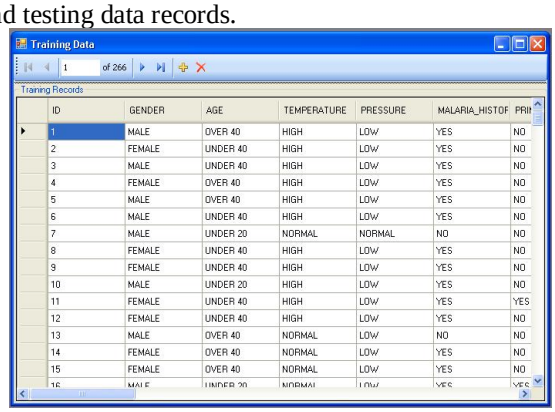

Figure 5 Training Data Records

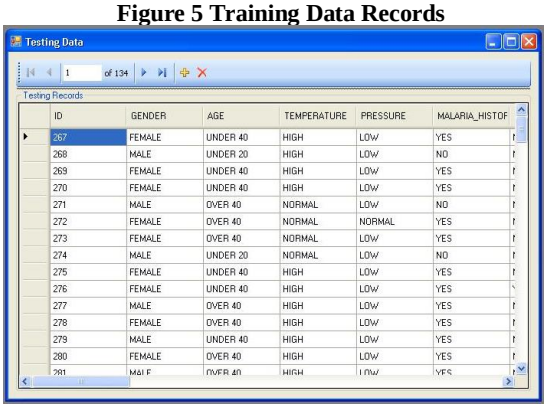

Figure 6 Testing Data Records

In the next step, user can generate decision tree by ID3 algorithm and then, extract the classification rules from generated tree. User can see generated tree and extracted rules in this step. Figure 7 shows decision tree output that is generated by ID3 algorithm. In this output view, user can remove unnecessary branches of tree by selecting each leaf node and click Delete Node button. System will automatically redraw decision tree. Normally, a leaf node that has small path length (for example, less than 3 path length) should remove from decision tree.

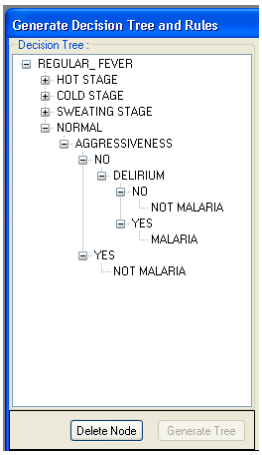

Figure 7 Decision Tree(Generated by ID3 Algorithm)

After removing unnecessary branches, user can generate classification rules for resulted decision by clicking Generate Tree button and then user must click Used button to use these rules for estimating rule accuracy with testing data. To extract rules from a decision tree, one rule is created for each path from the root to a leaf node. Each splitting criterion along a given path is logically ANDed to form the rule antecedent ("IF" part). The leaf node holds the class prediction, forming the rule consequent ("THEN" part). In this system, the resulted rules are 21 rules. Figure 8 presents output of classification rules.

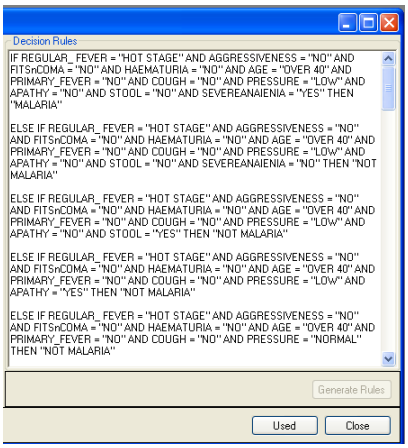

Figure 8 Generated Rules from Decision Tree

The last step is estimating the classifier's accuracy. According to holdout method, one third of sample data set may be tested with extracted classifier. The percentage of accuracy is showed on window form. In this system, accuracy is calculated as follows:

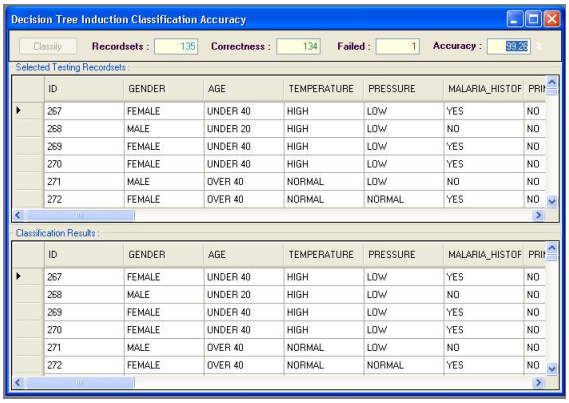

Figure 9 Estimating Classifier's Accuracy with Testing Data

4.4. Analysis of System Performance

To determine Malaria Diagnosis System's performance, the system was tested with different amount of sample data. The system's classifier accuracy increased with the number of sample data amount. The result of our test is described with the following table 2.

Table 2. Result of Classifier Accuracy with Different Amount of Sample Data

No of Sample Records	No of Test Data Records	No of Corrected Records	No of Failed Records	Classifier's Accuracy (%)
400	133	121	13	90.3
300	100	89	11	89
200	67	57	10	85.07
100	34	28	6	82.36

5. Conclusion

In this paper, we have demonstrated the ID3 classification algorithm in Malaria diagnosis system. The extracted classification rules from decision tree are tested with various amounts of test data. As a result of this approach, we can observe the impact of large amount of sample data set on classifier's accuracy. The greater amount of sample data set, the higher the classifier's accuracy. And also, we can notice that the noise of sample data may change and decrease the classifier's accuracy.

. This system used 17 attributes that are essential symptoms to classify malaria infected or not. System will work successfully, if the attributes are less than these 17 attributes. And also, system can have high performance in accuracy. But system's predictions will not uncertain for prediction on malaria disease.

6. Reference

[1] Goharian and Grossman, "*Lecture Notes*", Illonois Institute of Technology (2003).

[2] Hyafil, RL Rivest. "*Information Processing Letters*", Vol. 5, No. 1. (1976)

[3] Jiawei Han and Micheline Kamber "*Data Mining and Concepts*" ISBN 978-81-312-0535-8.

[4] Kantardzic, Mehmed, "*Data Mining: Concepts, Models, Methods, and Algorithms*." ISBN 0471228524. OCLC 50055336. John Wiley & Sons

[5] Lyman, Peter; Hal R. Varian "*How Much Information*" (2003). http://www.sims.berkeley.edu/how-much-info-2003. Retrieved on 2008-12-17.

[6] Usama M. Fayyad and Keki B. Irani "*On the Handling in Decision Tree of Continuous-Valued Attributes Generation*" http://www.springerlink.com/index/H8P82R5213473T36.pdf

[7] http://datamining.eruditionhome.com/classification

[8] http://www.wikipedia.com