Suggestion System for Matriculation Exam Result using Decision Tree Induction

Khwar Nyo Khine, Myint Myint Yee
Computer University (Sittway)
khwarnyo88@gmail.com, myintmyintyii@gmail.com

Abstract

This paper implements suggestion for matriculation examination result using decision tree induction algorithm. To be a well-developed human society, there must not be a great gap among the youths and so it is very necessary to carry out various measures for knowledge skills in Myanmar. Many basic education high schools are opened to promote educational Standard in all states and divisions, and setting the effective guardianship plans for increasing matriculation students' exam result, our society as well as our country will be developed. The progress makes the basic education to be developed. This paper intends to develop effective preplanning process for matriculation exam result based on data mining approach. Decision tree is a flow-chart-like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and the leaf nodes represent classes or class distributions. Depending on the attribute values of the data set, this system can classify the result of the matriculation students' exam whether it is in serious or normal conditions. Thus, the user can test a student's conditions for a student's matriculation exam result according to monitoring board of supervision committee. This system is based on Guardianship System for Basic Education Schools, Ministry of Education, Myanmar.

Keywords: Data mining, Classification, Decision tree induction, Matriculation exam.

1. Introduction

Data mining is the process of analyzing data from different perspectives and summarizing it into useful information. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases. Data are any facts, numbers, or text that can be processed by a computer.

Data mining can be performed on data represented in quantitative, textual, or multimedia forms. Data mining applications can use a variety of parameters to examine the data. They include association, sequence or path analysis, classification, clustering and forecasting. It is used for a variety of purposes in both the private and public sectors.

Data mining is the process of using tools such as classification, association rule mining, clustering, etc. Decision tree induction algorithm is one on the most popular algorithms in the mining classification. The primary intent of the system is to design and develop an efficient approach for extracting decision rules.

Decision tree are one of the most attractive and easy to use tools in decision-making activities. Decision tree is commonly used for gaining information for the purpose of decision-making. Classification rules represent the classification knowledge as IF-THEN rules and are easier to understand for users.

2. Related Work

Decision tree classifier is a simple yet widely used classification techniques. Jiawei Han and Micheline Kamber describes the concept buy-computer, that is predict using decision tree induction algorithm whether or not a customer at AllElectronics is likely to purchase a computer [5]. Khaing Nay Kyi also describes classification of industry test by using decision tree induction [7]. Soe San Oo applied diagnosis of acute diarrhea in children by using decision tree induction at [11]. Myo Myo Than Naing describes decision making for Poultry diseases using decision tree algorithm [9]. Minos.G, Dongjoon.H, Rajeev R. and KyuseokS used decision tree in "Efficient Algorithms for constructing Decision Tree with constraints" [8]. Kamber, L.Winstone, W.Gong, S.Cheng, J.Han applied decision tree induction algorithm in

"Generalization and Decision Tree Indction: Efficient Classification in Data Mining", 1997 [6]. Naing Naing Khin, discusses Risk Level Prediction for Heart Disease Using Decision Tree Induction" [10]. Decision tree can be applied to a wide variety to business and medical field.

3. Data Mining

Data Mining is the process of discovering meaningful, new correlation patterns and trends by sifting through large amounts of data stored in repositories, using pattern recognition techniques as well as statistical and mathematical techniques. The primary goals of data mining in practice tend to be prediction and description. Data mining serves as an essential step in the process of knowledge discovery in the databases. Diverse fields such as marketing, customer relationship management, engineering, medicine, crime analysis, expert prediction, Web mining, and mobile computing, besides others utilize data mining. It combines statistical analysis, machine learning and database technology to extract hidden patterns and relationships from large databases [5].

3.1. Classification

Classification is the action of assigning an object to a category according to the characteristics of the object. In data mining, classification refers to the take of analyzing a set of pre-classified data objects to learn a model that can be used to classify an unseen data object into one of several predefined classes. A data object is described by a set of attributes or variables. One of the attributes describes the class that an example belongs to and is thus called the class attribute or class variable. Other attributes are often called independent or predictor attributes. The set of examples used to learn the classification model is called the training data set. Classification belongs to the category of supervised learning, distinguished from unsupervised learning. In supervised learning, the training data consists of pairs of input data, and desired outputs, while in unsupervised learning there is no a priori output.

Classification has various applications, such as learning from matriculation database to train students' features, such as attendance, oral_ response, monthly_test, pre_test, IQ, handwork, health and uses suitable classification algorithm to get suggestion for specified student's exam result in real world education section. There are many practical situations in which classification is of immense use. Basic techniques for data classification are decision tree induction, Bayesian classification and Bayesian belief networks, and neural networks [1].

3.2. Decision Tree Algorithm

Decision trees are commonly used in operations research, specifically in decision analysis, to help identify a strategy most likely to reach a goal. Decision tree can be used to classify an example by starting at the root of the tree and moving through it until a leaf node, which provide the classification of the instance. The information gain measure is used to select the test attribute at each node in the tree. Such a measure is used to as an attribute selection measure or a measure of goodness of split. The attribute with the highest information gain is chosen as test attribute for the current node [3],[4].

A decision tree is a flow-chart-like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and leaf nodes represent classes or class distributions. The top most node in a tree is the root node. The basic algorithm for decision tree induction is a greedy algorithm that constructs decision trees in a top-down recursive divide-and-conquer manner.

Let S be a set consisting of a 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 the class Ci. The expected information needed to classify a given sample is given by

$$I(S_1, S_2, \ldots, S_m) = -\sum_{i=1}^{m} p_i \log_2(p_i), \ldots \text{equation1}$$

where p_i is the probability that an arbitrary sample belongs to the 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 partitions 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 S_i in class C_i in a S_j. The entropy, or expected information based on the partitioning into subsets by A, is given by

$$E(A) = \sum_{j=1}^{v} \frac{S_{1j} + \ldots + S_{mj}}{S} I(S_{1j} + \ldots + S_{mj}), \ldots \text{Equation2}$$

The term $\frac{(S_{1j} + \ldots + S_{mj})}{S}$ act as the weight of the jth subset and is the numbers of samples in the subsets divide 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,

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

where $Pij = \frac{S_{1j}}{|S|}$ and P is the probability

that a sample in Sj, belongs to class Ci. The encoding information that would be gained by branching on A is

$$Gain(A) = I(S_1, S_2, ..., S_m) - E(A),$$

…. Equation 3

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 labeled with the attribute, branches are created for each value of the attribute, and the samples are partitioned accordingly.

As example calculation;

Information for attendance attribute:

I(s1,s2,s3,s4)= I(5,5,4,6) =

$$-5/20\log_2 5/20 - 5/20\log_2 5/20 - 4/20\log_2 4/20 - 6/20\log_2 6/20$$

=1.9855

For attendance=">95%"

$I(s11,s12,s13) = -1/5\log_2 1/5 - 1/5\log_2 1/5 - 3/5\log_2 3/5$ = 1.3709

For attendance = "80%-90%"

$I(s12,s22,s32) = -2/5\log_2 2/5 - 2/5\log_2 2/5 - 1/5\log_2 1/5$ =1.5219

For attendance ="75%-80%"

$I(s13,s23,s33) = -0/4\log_2 0/4 - 1/4\log_2 1/4 - 3/4\log_2 3/4$ = 0.8113

For attendance ="<75%"

$I(s14,s24,s34) = -0/6\log_2 0/6 - 1/6\log_2 1/6 - 5/6\log_2 5/6$ = 0.65

Entropy for attendance attribute:

E (attendance) =(5/20 * 1.3709)+ (5/20*1.5219)+(4/20*0.8113)+ (6/20 * 0.65) = 1.0805

The gain information will be calculated :

Gain (attendance) = **I (s1,s2,s3,s4) – E** (attendance) = 1.9855 – 1.0805 = 0.905

Similarly; the rest of other attributes' gains can be calculated.

Gain(oral_response)	= 0.7340
Gain(doing_exercise)	= 0.7271
Gain(doing_homework)	= 0.6986
Gain(interesting)	= 0.2477
Gain(overtime_attendance)	= 0.8043
Gain(discipline)	= 0.0986
Gain(monthly_test)	= 0.6485
Gain(pretest)	= 0.3848
Gain(IQ)	= 0.0653
Gain(supporting)	= 0.4007
Gain(byheart_rate)	= 0.058
Gain(health)	= 0.0538

Gain value of attendance attribute is highest information gain among the attributes. It is selected as the root for tree structure.

Root node for decision tree is created with attendance, and branches are grown for each of attribute's values respectively.

4. Decision Tree Construction

Decision Tree uses the gain ratio criterion selects, from among those attributes with an average or better gain, the attribute that maximizes the ratio of its gain divided by its entropy. The algorithm is applied recursively to form sub-trees, terminating when a given subset contains instances of only one class. It is an approximation discrete function method and can yield lots of useful expressions. The decision tree induction algorithm is as follow:

Input: Data partition, *D*
attribute_list
Attribute_selection_method

Output: A decision tree.

Method:

(1) Create a node *N*;
(2) **if** tuples in *D* are all of the same class, *C* **then**
(3) return *N* as a leaf node labeled with the class *C*
(4) if *attribute_list* is empty then
(5) return *N* as a leaf node labeled with the majority class in *D*;
(6) apply **Attribute_selection_method** (*D, attribute_list)* to find the "best" splitting_ criterion:
(7) label node *N* with *splitting_criterion;*
(8) **if** splitting_attribute is discrete -valued and multiway splits allowed **then**
(9) *attribute_list* ← *attribute_list – splitting_ attribute;*
(10) **for** each outcome *j* of *splitting_criterion*
(11) let D_j be the set of data tuples in *D*satisfying outcome *j;*
(12) let D_j is *empty* then
(13) attach a leaf labeled with the ajority class in *D* to node *N;*
(14) **else** attach the node returned by **Generate_decision_tree**(D_j,*attribute_list*) to node *N;* **endfor**
(15) retrun *N;*

5. Attribute Information

The attributes which are important for matriculation examination suggestion system are described. This system can be used attributes selection for matriculation student supervision data. In training data set many students' records to evaluate the performance of the classification system. These records consist of 13 attributes for three classes. The detailed description of the parameters and their corresponding values are given as follows:

No	Parameter	Values
1	attendance	>95% Between80%and95% Between75%and80% <75%
2	monthly_test	regular passed so often passed always failed
3	pretest	pass fail
4	IQ	higher middle lower
5	byheart_rate	good fair bad
6	oral_response	regular so often
7	doing_exercise	regular so often
8	doing_homework	regular so often rarely
9	interesting	regular so often
10	overtime_attend	regular so often
11	discipline	good bad
12	supporting	regular rarely
13	health	good bad

6. Flow of the System

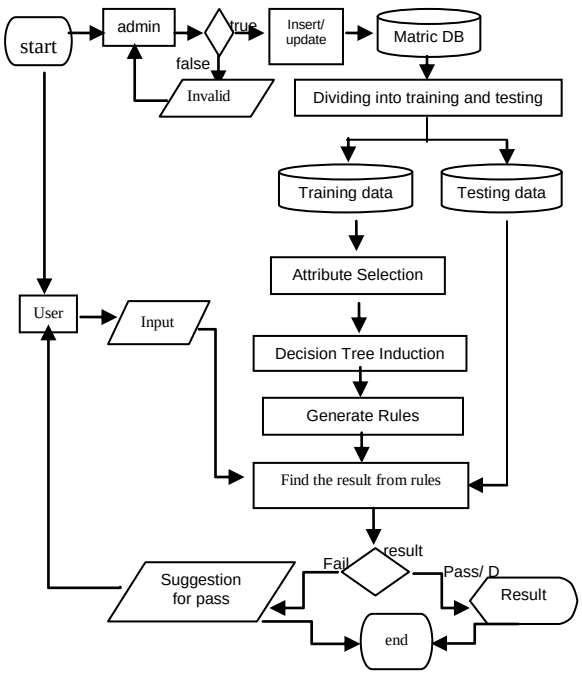

Figure 1: System Flow Diagram

7. System Implementation

The matriculation student data set consists of attributes to classify for three examination result suggestion (such as D, pass, fail). The user who wants to know the student's result that will get D, pass or fail in coming matriculation examination. The training data are managed from supervision committee, Basic Education High Schools, Rakhine State, Ministry of Education. The system computes the information gain of each attribute by using Equation 1 and computes the entropy or expected information of each attribute by using Equation 2. By using equations, the highest information gain among the attribute is selected and created as root node. Finally, the system can generate the decision tree by using decision tree induction algorithm. The knowledge represented in decision trees can be extracted and represented in the form of classification IF-THEN rules. After generating the rules, the system can classify the matriculation examination result as figure 3. The system can be estimated the classifier accuracy test result using Holdout method. The new student can also be tested the conditions of his features in this system. This system is developed on Microsoft Access 2003 for database and implemented using Java programming language, JDK 1.6 in Eclipse platform.

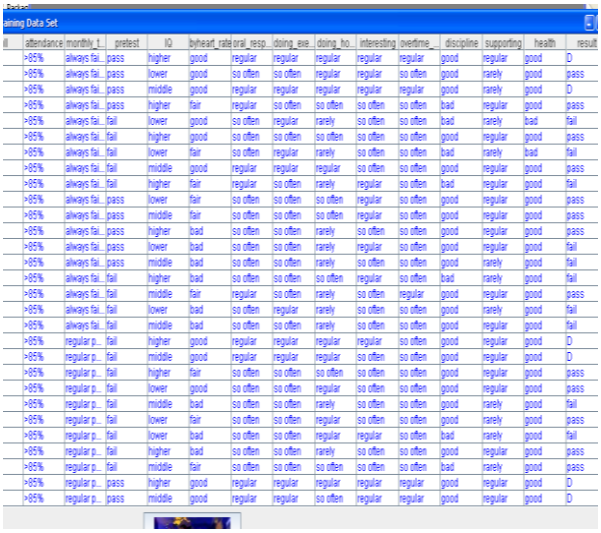

Figure 2: Student training data set

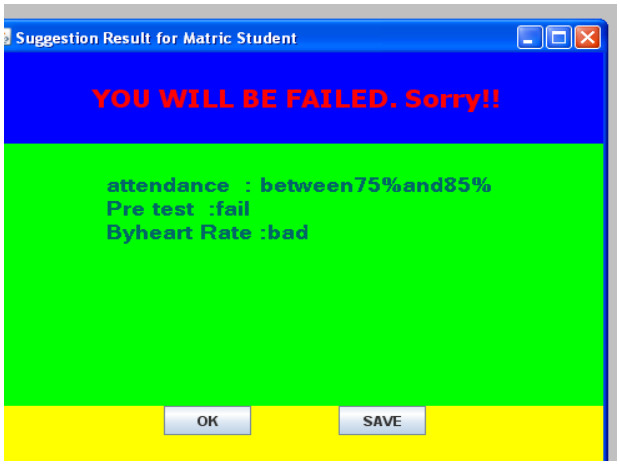

Figure 3: Suggestion Result for Matriculation Examination System

8. Experimental Result

The proposed system introduces the result of applying decision tree induction algorithm to get the suggestion of the result for matriculation exam. When the user enters into the system, the user selects the 13 attributes are going to be chosen for new student's result. Then, the user request for the result by clicking on result button, the system displays the result as figure 3.

9. Conclusion

This system focuses on the classification rules mining that based on decision tree induction algorithm. This system is intended to develop an effective result for matriculation examination and to get positive effective suggestion when the result appears fail. As a student, when he gets fail result from the system, he should obey the suggestion from system. The decision statements can be generated as the rules by using decision tree induction algorithm. 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 decision trees is as a descriptive means for calculating conditional probabilities. This system is mainly support for suggestion for matriculation students' qualification result to make a valid prediction. User can get specific result and detail information for a student's success for coming examination.

8. References

[1] Aijun ,A "Classification Methods", New York University , Canada . http://en.wikipedia.org/wiki/Granular-computing

[2] D.A. Keim, "Knowledge Discovery and Data Mining New Port Beach , USA, 1997".

[3] http://decision tree learning applet.htm

[4] http:// decisiontree.net

[5] Han, Jiawei & K.Micheline "Data Mining Concept and Techniques".

[6] Kamber, L.Winstone, W.Gong, S.Cheng, J " Generalization and Decision Tree Indction: Efficient Classification in Data Mining ", 1997.

[7] Khaing Nay Kyi, "Classification of Industry Test by Decision Tree Induction".

[8] Minos.G, Dongjoon.H, Rajeev R. and Kyuseok S. "Efficient Algorithms for Constructing Decision Tree with constraints".

[9] Myo Myo Than Naing, "Decision Making for Poultry Diseases using Decision Tree Induction Algorithm".

[10] Naing Naing Khin, "Risk Level Prediction for Heart Disease by using Decision Tree Induction algorithm".

[11] Soe San Oo, "Diagnosis of Acute Diarrhoea in Children by using Decision Tree Induction".