Classification of Rank for Distributors of Multi-Level Marketing Company by Using Decision Tree Induction

Htet Htet Aung, Myint Myint Yee
University of Computer Studies, Yangon
htethtetaung@ucsy.edu.mm,myintmyintyee110@gmail.com

Abstract

Data Mining is a step in the knowledge discovery process consisting of particular data mining algorithms that, under some acceptable computational efficiency limitations, find patterns or models in data [1]. Classification and prediction are two forms of data analysis that can be used to extract models describing important data classes or to predict future data trends. Decision tree is commonly used for the gaining information for the purpose of decision making. This paper intends to apply the Decision tree induction ID3 algorithm on the distributor's performance data to generate the classification model and this model can be used to predict the promotion of rank for the distributors. The dataset is used from ***Zhulian****Multi-level Marketing Company. 10-fold Cross Validation accuracy method is used to approve the system accuracy.*

Keywords: *data mining, classification, decision tree, ID3 algorithm, K-fold Cross Validation accuracy method.*

1. Introduction

Classification is the process of finding a set of models that describes the model and also distinguishes 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. Decision Tree is one of the most popular classification algorithms in current use in data mining. Decision tree is commonly used for the gaining information for the purpose of decision making. The input to a classifier is a training set of records, each of which is a tuple of attribute values, one of the attributes is the class label. There are nine attributes are used in this system and Iterative Dichotomizer3(ID3) is commonly used to have gaining information for the purpose of decision making to give the right rank for the distributors of Multi-level Marketing Company. The input to the classifier (ID3) is preprocessed dataset, each of which is a tuple of eight attributes values with a class label (DS, SE, SM, SSM, DSM, CDM, RCM, RCD). The main task performed in this system is to determine the appropriate classification of rank for the distributors in the network marketing/multilevel marketing company by matching the user's inputs with the rules by ID3 algorithm.

The rest of this paper is organized as follows. Section 2 presents related work of the system. Section 3 describes theoretical background. The system implementation and accuracy performance is described in Section 4. Section 5includes conclusion of the paper.

2. Related Work

The snake classification system implemented by Khin Kyi Than also use ID3 Algorithm to classify and research process of snake venom. By using this system, doctors can get more accurate to give what kind of anti-venom injection within a minute by using the ID3 classifier of Snake Venom [5].

In the quality control system implemented by Su Mon Khine, the decision tree induction ID3 algorithm is used to classify the soft drink pass or fail. It has presented the evaluation of classifier accuracy using holdout method [8].

In IT professional level classification system implemented by May Thu Zin, determines IT professional level using rules given by I3 algorithm and predicts future level by using historical data in the database. In this system, fourteen classes are classified with six attributes by supervised learning [2].

3. Background Theory

3.1. Data Preprocessing

Data preprocessing is an important step in knowledge discovery process. Databases are highly susceptible to noisy, missing and inconsistent data

due to their typically huge size. Attributes of interest may not always be available. Therefore, preprocessing is always a necessity whenever the data to be mined is noisy or incomplete and this process significantly improves the effectiveness of the data classification. Data preprocessing techniques are data cleaning, data integration, data transformation data reduction.

Data cleaning: routines work to clean the data by filling in missing values, smoothing noisy data, identifying or removing outliers and resolving inconsistencies.

Data integration: data analysis task will involve data integration, which combines data from multiple sources into a coherent data store, as in data warehousing. These sources may include multiple databases, data cubes, flat files.

Data transformation: data transformation such as normalization, may improve the accuracy and efficiency of the mining algorithms involving distance measurements. Discretization and concept hierarchy generation are powerful tools for data mining in that they allow data mining at multiple abstraction levels. Normalization, data discretization, and concept hierarchy generation are forms of data transformation.

Data reduction: can reduce the data size by aggregating, eliminating redundant features, or clustering, for instance [3].

3.2. 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.

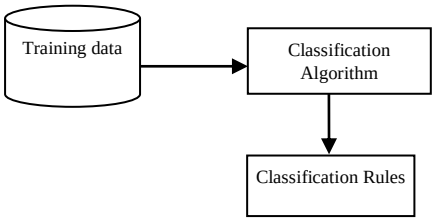

Figure 1. Learning Process in Classification

In the second step, the model is used for classification. First, the predictive accuracy of the model is estimated. The 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].

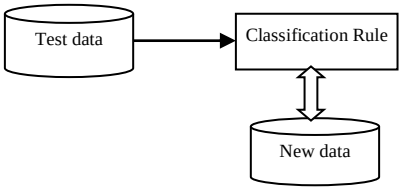

Figure 2. Classification Process in Classification

3.3. Decision Tree

Decision tree is a method of approximating discrete valued target functions. The learned functions represented by a tree structure. Most existing tree induction systems adopt a greedy (i.e. non-backtracking) top-down recursive divide and conquer manner. Learned tree represented as a set of IF…THEN rules to improve human readability. A decision tree is a flow-chat like tree structure, where each internal node (non-leaf nodes) denote a test on an attributes, each branch represents an outcome of test and leaf nodes (terminal nodes) represent class label. Decision tree can be easily converted to classification rules[1].

3.4. Iterative Dichotomiser 3 (ID 3)

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).

Informal formulation of ID3 algorithm

- Determine the attribute that the highest information gain on training set.
- Use the attribute as the root of the tree; create a branch for each of the values that the attribute can take.
- For each of the branches, repeat this process with the subset of the training set that is classified by this branch [3].

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 [4].

FunctionID3 (R: a set of non-categorical attributes,
C: the categorical attribute,
S: a training set)
Returnsa decision tree;

Begin

1) **If** S is empty, **return** a single node with value Failure;
2) **If** S consists of records all with the same value for the categorical attribute, **return** a single node with that value;
3) **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];
4) Let D be the attribute with largest Gain (D, S) among attributes in R;
5) Let {d_j| j=1,2, ., m} be the values of attribute D;
6) Let {S_j| j=1, 2, .., m} be the subsets of S consisting respectively of records with value dj for attribute D;
7) **Return** a tree with root labeled D and arcs labeled d_1, d_2, ..,d_m going respectively to the trees
8) ID3(R-{D}, C, S_1), ID3(R-{D}, C, S_2), .., ID3(R-{D}, C, S_m);
End ID3;

3.4.1. 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.

Let Sbe 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

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

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,….,a_v}. Attribute A can be used to partition S into v subsets, {s_1, s_2,…..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

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

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,

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

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,….,s_m) – E (A).**(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 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].

3.4.2. Data selection and transformation

In this phase, the data are put into a form suitable for the modeling phase. If required some selected variables are combined, transformed or used to create a new variable. For example, Qualified Sale Executive, Qualified Sale Manager, Qualified Diamond Sale Manager were used to generate a variable labeled as “Qualified Line”.

Before and after preprocessing of attributes or variables and their values are shown in Table 1 and 2.

Table 1. Related Variables before Preprocessing

No	Attributes	Possible Values
1	Current Rank	{ DS, SE, SM, SSM, DSM, CDM, RCM }
2	Personal PV	Numeric values { 100, 200, 400, … }
3	Personal Group PV	Numeric values {500, 3000, 2000,… }
4	Accumulated Personal Group PV	Numeric values { 16000, 8000, 18000, …}
5	Self Qualified	Numeric values { Yes, NO }
6	Qualified Sale Executive	Numeric values { 0, 2, 3, 4, … }
7	Qualified Sale Manager	Numeric values { 0, 2, 3, 5, … }
8	Qualified Diamond Sale Manager	Numeric values { 0, 2, 3, 6, … }
9	Consecutive Qualified	Numeric values { 0, 1, 2, 3, 4, … }
10	Next Rank	{ DS, SE, SM, SSM, DSM, CDM, RCM, RCD }

Table 2. Related Variables after Preprocessing

No	Attributes	Attribute Values
1.	Current Rank	{ DS, SE, SM, SSM, DSM, CDM, RCM }
2.	Personal PV	{UnQualified, Qualified }
4.	Accumulated Personal Group PV	{ good, bad }
5	Self-Qualified	{Yes, NO
6.	Qualified Line	{ zero, two, four ,six ,eight ,ten }
7.	Consecutive Qualified	{ zero, one, two, three }
8.	Next Rank	{ DS, SE, SM, SSM, DSM, CDM, RCM, RCD }

3.5. Advantages of Using ID3 Algorithm

- Understandable prediction rules are created from the training data.
- Builds the fastest tree.
- Only need to test enough attributes until all data is classified.
- Finding leaf nodes enables test data to be pruned, reducing number of tests.
- It has good accuracy, however, successful use may depend on the data at hand. Whole dataset is searched to create tree.

3.6. Classifier Accuracy

Estimating classifier accuracy is important in that it allows one to evaluate how accurately a given classifier has not been trained. Accuracy is measured using a test set of objects for which the class labels are known.

3.6.1. K-fold Cross Validation method

The system use 10-fold cross validation to split training set and test set to evaluate accuracy. The system use 10-fold cross validation to be able to both train with all the data and then indirectly test with all the data as well. In k-fold cross validation, the initial data are randomly partitioned into k mutually exclusive subsets or "folds", D_1, D_2,…,D_k, each of approximately equal size. Training and testing is performed k times. In iteration i, partition D_i is reserved as the test set, and the remaining partitions are collectively used to train the model. For classification, the accuracy estimate is the overall number of correct classification, from the k-iteration, divided by the total number of tuples in the initial data [3].

$$\text{Accuracy} = \frac{no.\ of\ correctly\ classified\ objects}{total\ number\ of\ tuples\ in\ the\ initial\ data} \quad \textbf{(5)}$$

The system use 10 fold cross validation on 1003 records to split training and testing set on 10 iteration. The overall accuracy is the total number of correct classification from 10 iteration. By using 10-fold cross validation, the classifier accuracy is 97.11%.

4.System Implementation

4.1. Overview of the System

This system is built as a computer-based classification system using decision tree induction. In this system, there are two parts, Admin and Classifier or user. Discretizing step for Continuous-valued attributes can be extended as preprocessing task before calculating rules using ID3 algorithm [9]. Rules are generated from the decision tree using ID3 decision tree algorithm. There are nine attributes are used in system to classify eight types of rank for users. Only eight features or attributes of distributor (such as current rank, personal point value, personal group point value and so on [7]) will be input to the system. The incoming input information match with rules and determine rank promote or not for the distributor (DS, SE, SM, SSM, DSM, CDM, RCM, RCD[7]).This system uses 1003 data records from Zhulian Company and classifier accuracy is 97.11%.

4.2. System Architecture and Design

The architecture and design of the system is shown in Figure 3.

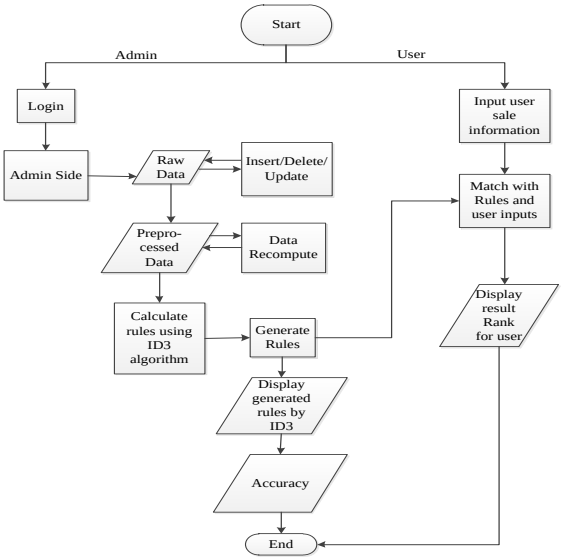

Figure 3. System flow Diagram for the Rank Classification System

There are two parts in this system. They are administrator and classifier or user. In the administrator part, administrator can see existing data records, insert new data record and can delete or update the existing data records after passing the login. After inserting, deleting and updating raw data records stored in the database, admin must recompute the preprocessed data and rule lists. Decision rules are produced by using preprocessed data and ID3 algorithm. The classifier accuracy is calculated by using 10-fold Cross Validation method. In user part, users input only eight attributes. These eight inputs/attributes match with the rules produced. And then, the system will produce the final result of distributor's rank.

4.3. Implementation of the System

In this system, there are four parts those are implemented. The first part is implemented for raw data entry; the second part is for preprocessing of the data; the third part is for decision rule generation and the final part is implemented for calculation accuracy. Each processing part can be seen in the respective figures as shown in below.

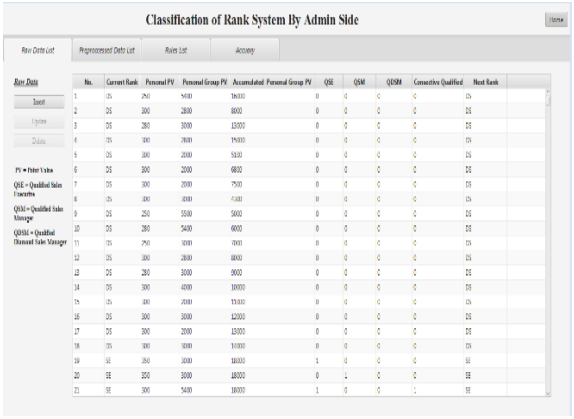

Figure 4. Raw Data of the System

Figure 4 shows the raw data list of the system. The raw data includes the old cases of the dataset that include nine attributes values and the related rank of distributor result.In raw data, admin can see the existing data records, can insert new data record and can delete or update the existing data records that admin wants to delete or update.

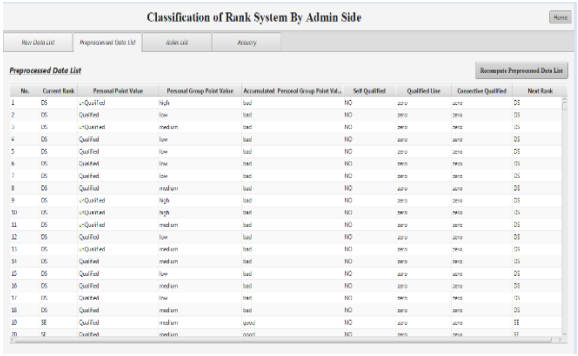

Figure 5. Preprocessed Data of the System

Figure 5 shows the preprocessed data of the system. In preprocessed data, seven attribute values and related rank for distributor's result are included. After inserting, deleting and updating records in the raw data, admin must recomputed the preprocessed data by clicking Recompute Preprocessed Data button.

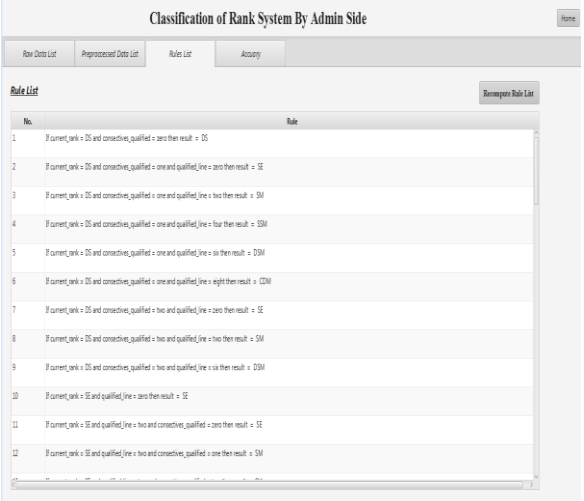

Figure 6. Rule lists of the System

From the rule lists shown in Figure 6, the generated decision rules can be seen. It is easy 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 57 rules. After inserting, deleting and updating the data records in the raw data, admin must recompute the rule list by clicking Recompute Rule Lists button.

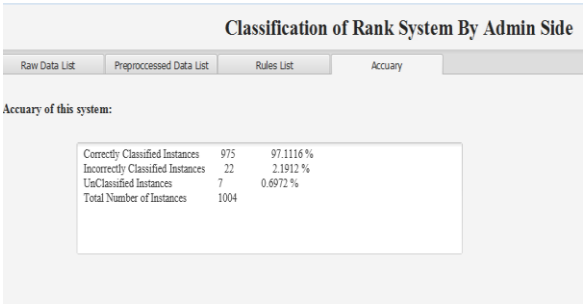

Figure 7. Classifier Accuracy of the System

Figure 7 shows the classifier accuracy of the system. The accuracy is based on the data records. The greater amount of sample data records, the higher the classifier's accuracy. Now, the system uses 1003 data records and accuracy is 97.11%. If the rule list changes, the classifier accuracy will do automatically.

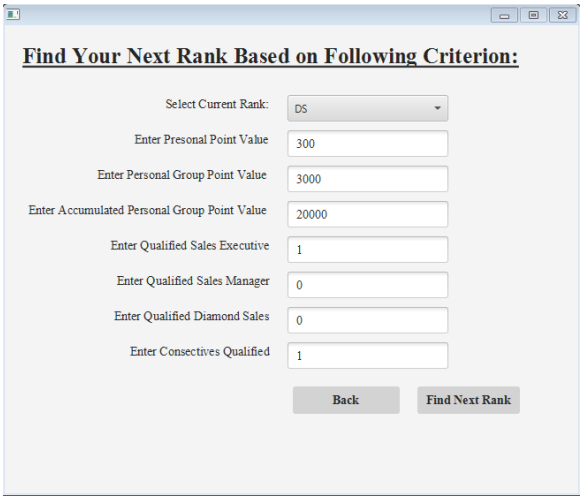

Figure 8. User Interface of the System

In Classifier or User part, user can classify the type of rank for distributors of multi-level marketing company by filling the sale information of a distributor. Figure 8presents user interface of the System. After filling user information to find next rank, Click Find Next Rank button. And then the result output is showed in figure 9:

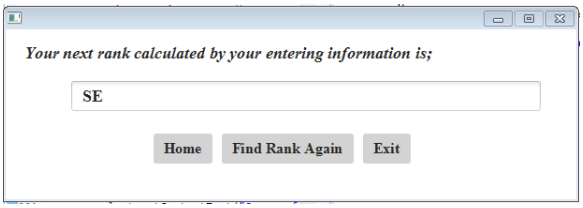

Figure 9. Prediction result of the System

4.4. Analysis of System Performance

To determine Rank Classification System's performance, the system was tested with different amount of sample data records. The accuracy of rank classification system has been increased when the amount of trained data is increased. The result of our test is described in figure 10.

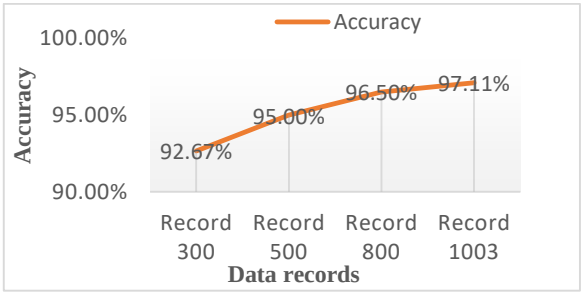

Figure 10. Accuracy comparison on different amount of data records

As in Figure 10, the percentage of accuracy is 92.67% if 300 data records are trained. When the amount of records is increased to 500, the accuracy is

92.67%. For, the 800 data records, the accuracy have been increase to 96.50%. Similarly, the accuracy percentage is increased to around 97.11% and when data amount is increased to 1003. By learning the accuracy, it can be seen that the percentage of system accuracy will be increased when the amount of data records have been increased.

5. Conclusion

In this paper, the ID3 classification algorithm is used in Rank Classification. The extracted rules are tested with the user inputs of monthly sale performance by distributor. According to the result of this approach, the impact of large amount of sample data records on classifier accuracy is observed. The greater the amount of sample data records, the higher the classifier's accuracy. And then, it can be seen that the noise of sample data may change and decrease the classifier's accuracy. This system used nine attributes that are essential information and also users only need to fill eight attributes to classify whether rank promotes or not. This system will help the multi-level marketing company to classify the rank of distributors easily and the distributors can also know their performances. The system can have high performance in accuracy and predict the right rank within a short time according to the user input.

References

[1] Alex Freitas "Datamining and Knowledge Discovery (CO832) Decision Tree Induction"

[2] Determining IT professional level using decision tree induction for classification with ID3 Algorithm, March,2010/ May Thu Zin.

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

[4] Malaria Diagnosis System by Using ID3 Classification Algorithm, Khaing Mar Thuai,Moe Thant ,Computer University(kalay)/2009.

[5] Snake Classification by using ID3 Algorithm .June , 2009/ KhinKyi Than.

[6] Suggesting mode of delivery by using Iterative Dichotomiser 3 (ID3) Algorithm, Yin Mon Aye, Khine Moe New/2009.

[7] The golden business plan of Zhulian.

[8] Test Center for Quality Control System Using Induction Method. May 2009/Su Mon Khine.

[9] 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
https://en.wikipedia.org/wiki/Discretization