Data Preprocessing-Based Comparative Study on Decision Tree Induction Algorithm

Khaing Myint, Win Mar Oo
Computer University (Taunggyi)
khaingmyint.mdy08@gmail.com, winmaroo.ucs@gmail.com

Abstract

Data preprocessing is an important step in the knowledge discover process and quality decisions such as, real-world data which tend to be incomplete, noisy and inconsistent data. The system uses the discretization and concept hierarchy generation, that transforms the numeric data to categorical values in the data preprocessing techniques.Moreover, this paper addresses the speed and scalability issues by proposing a data classification method which attributes selection measure and the induction of decision tree .Myanmar is a country that its' economy is based on agriculture. Crop production plays an important role in the session of agriculture. So, this system provides business application areas for crop information. This system compares decision tree generation time and the accuracy of the data preprocessing with and without preprocessing data.

1. Introduction

Crop information is pivotal in today's business application areas. The users are actively looking for new technologies that will assist them to decide whether should buy or should not buy crop become more profitable and competitive.

Data mining also called knowledge discovery in database (KDD) is one of the fastest growing fields in the computer science and technology. It is automated extraction of patterns representing knowledge implicitly stored in large database and other massive information repositories. It aims at discovering useful information from large collections of data.

Data mining techniques can be applied to the problem of business process reengineering, in which the goal is to understand interactions and relationships among practices and organizations. Large scale data mining application involving complex decision making can access billions of bytes of data. Hence, the efficiency of such application is paramount.

Classification is a key data mining techniques whereby database tuples, acting as training samples, are analyzed in order to produce in model of the given data. Classification has numerous applications including credit approval, product marketing and medical diagnosis.

One method of classification is the induction of decision trees. A decision tree is a flow-chart-like tree structure, where each internal node denotes a test on an attribute, each branch represent the different classes in which the several records can be classified. The knowledge can be extracted and represented in the form of classification IF-THEN rules. Therefore, speed and scalability become issue of concern when this algorithm is applied to the mining of very large real-world databases.

2. Data Preprocessing

Data preprocessing transforms the data into a format that will be more easily and effectively processed for the purpose of the user .The techniques to preprocess data include data cleaning, data integration and transformation, data reduction and discretization and concept hierarchy generation [5].

2.1 Discretization and Concept Hierarchy Generation for Numeric Data

Concept hierarchies for numeric attributes can be constructed automatically based on data distribution analysis. There are five methods for numeric concept hierarchy generation:

- Binning
- Histogram Analysis
- Cluster Analysis
- Entropy-Based Discretization
- Segmentation by Natural Partitioning [4].

2.2 Binning

These methods are also forms of discretization. For example, attribute values can be discretized by distributing the values bins, and replacing each bin value by the bin mean or median, as in smoothing by

bin means or smoothing by bin medians, respectively. These techniques can be applied recursively to the resulting partitions in order to generate concept hierarchies.

Smoothing by binning methods

- Partition into equi-depth bins
- Smoothing by bin means
- Smoothing by bin boundaries [6].

2.3 Classification

Classification is a system used to identify, describe, organize, and evaluate the different kinds of work performed in an organization. Similar positions group into classes based on the different kind and level of work performed. Data classification is a two-step process. The classification techniques are:

- Decision Tree Induction
- Naive Bayesian Classification
- Neural Networks
- Genetic Algorithms
- The k-Nearest Neighbor Classifier
- Case-Based Reasoning [7].

2.3.1 Decision Tree Induction

A decision tree is a representation of a decision procedure for determining the class of a given instance. Each node of the tree specifies either a class name or a specific test that partitions the space of instances at the node according to the possible outcomes of the test. Each subset of the partition corresponds to a classification sub-problem for that subspace of the instances, which is solved by a sub-tree. A decision tree can be seen as a divide-and-conquer strategy for object classification [1].

2.3.2 Algorithm for Decision Tree Induction

1. Basic algorithm (a greedy algorithm)

- Tree is constructed in a top-down recursive divide- and- conquer manner.
- At start, all the training examples are at the root.
- Attributes are categorical (if continuous-valued, they are discretized in advance).
- Examples are partitioned recursively based on selected attributes.
- Test attributes are selected on the basic of a heuristic or statistical measure (e.g., information gain).

2. Conditions for stopping partitioning

- All samples for a given node belong to the same class.
- There are no remaining attributes for further partitioning majority voting is employed for classifying the leaf .
- There are no samples left [2].

2.3.3 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 is chosen as the test attribute for the current node [3].

The attribute selection measure is performed as follows. 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 (for i = 1,.., m). Let s_i be the number of samples of S in class C_i. The expected information needed to classify a given sample is given by

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

Let attribute A have v distinct values, $\{a_1,a_2,...,a_v\}$. Attribute A can be used to partition S into subsets, $\{S_1, S_2,...,S_v\}$, where S_j contains those samples in S that the 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 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

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

Thus, the information gained by branching on A is,

$$Gain(A) = I(s_1, s_2, ..., s_m) - E(A). \quad (3)$$

2.3.4 Estimating Classifier Accuracy

In the k-fold cross validation, the initiate data are randomly partitioned into k mutually exclusive subsets or "folds", S_1, S_2,..., S_k, each of approximately equal size. Training and testing is performed k times. In iteration i, the subsets 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 $S_1,S_3,..,S_k$ 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 initiate data [3].

3. The Proposed System Design

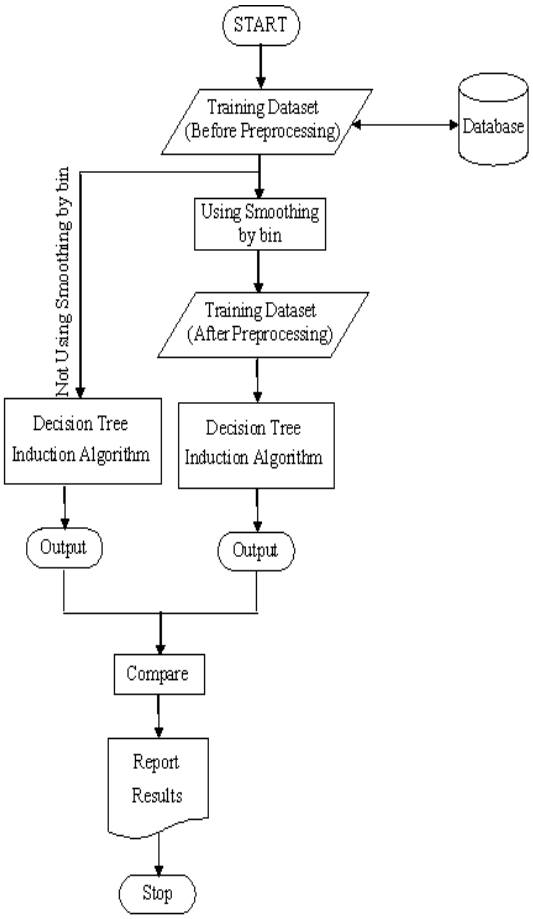

Figure 1. The Proposed System Design

According to Figure 1, the training dataset (before preprocessing) are fed to the system. This raw data are stored in large database and can be extracted. It is found that there are both using and not using smoothing by bins.

Without smoothing by bins, when the raw data can be mined with the decision tree induction algorithm that the results are the decision tree and rules generation time.

Alternatively, the training dataset can get by using smoothing by bins. This approach causes to be less the result of entropy gain and fast the time complexity in the system .Moreover, the data can be reduced. The data after preprocessing can be mined with the decision tree induction algorithm in the previous case. Again, it can get the time of decision tree and rules.

These two outputs can be compared according to their time complexity of speed and scalability. Therefore, the system obtains experimental result as shown in Figure 2, 3, and 4.

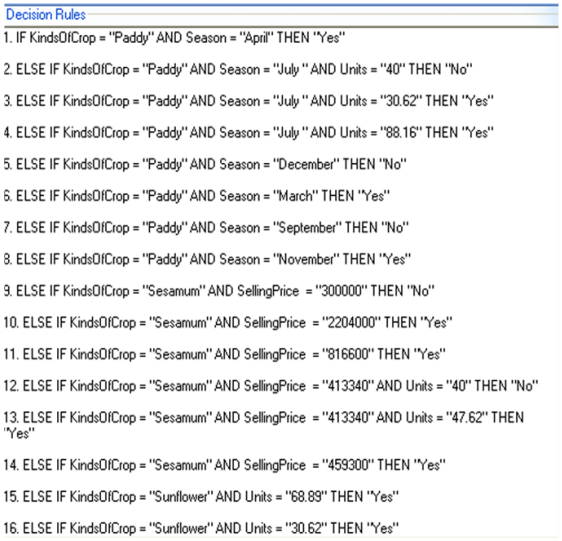

Figure 2. Decision Rules for Imported Dataset

According to Decision Rules for the Imported Dataset shown in Figure 2, it can get Decision Tree shown in Figure 3.

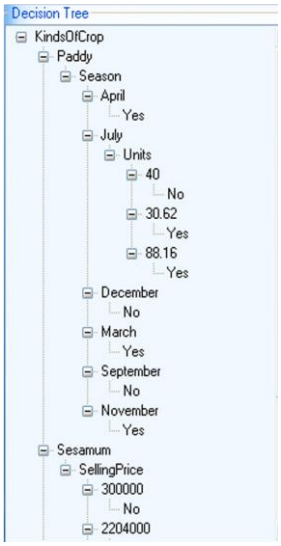

Figure 3. Decision Tree for Imported Dataset

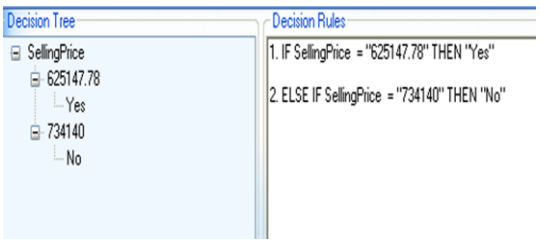

Figure 4. Decision Rules and Tree for Preprocessed Dataset

ID	SellingPrice	Units	Season	QualificationOfCrop	KindsOfCrop	BuysCrop
1	190480	47.62	April	Nonmoisturewntent	Paddy	Yes
2	190480	40	July	Moisturecontent	Paddy	No
3	300000	88.16	April	Nonmoisturecontent	Sesamum	No
4	2204000	88.16	May	Moisture content	Sesamum	Yes
5	416340	68.89	July	Freepest	Sunflower	Yes
6	612400	88.16	February	Freepest	Pulses	Yes
7	300000	30.62	December	Normoisturecontent	Yellowcorn	Yes
8	2204000	88.16	August	Moisturecontent	Groundnut	No
9	816600	40.83	October	Sanitation	Groundnut	No
10	612400	30.62	August	Nonmoisturecontent	Sunflower	Yes
11	190480	40	August	Freepest	Wheat	No
12	2204000	30.62	February	Sanitation	Groundnut	No
13	413340	40.83	August	Moisturecontent	Groundnut	Yes
14	816600	68.39	February	Sanitation	Seasamum	Yes
15	612400	30.62	October	Moisturecontent	Pulses	Yes

Figure 5. Training dataset

In figure 5, the attributes are selling price, units, season, qualification of crop and kinds of crop, tuple is sample that belongs to a pre-defined class label; it is the result of the classification and the training set is the set of samples used to build a model which is represented as classification rules, decision trees or mathematical formula.

The system includes 300 record sets and 5 attributes. Among the attributes, the kind of crop is the highest information grain.

4. Performance Analysis

Table 1. Decision Tree and Rule Generation Time

Number of record sets	Imported data set (millisecond)	Preprocessed data set (millisecond)
1-50	147	10
1-100	210	17
1-150	431	24
1-200	841	32
1-250	1251	40
1-300	1760	46

The running time for the number of record sets 1 to 50 to generate decision tree and rules time is 147 milliseconds for Imported dataset and 10 milliseconds for Preprocessed dataset. The running time for the number of record sets 1 to 100 to generate decision tree and rules time is 210 milliseconds and 17 milliseconds. The running time for the number of record sets 1 to 150 to generate decision tree and rules time is 431 millisecond and 24 milliseconds. The running time for the numbers of record sets 1 to 200 to generate decision tree and rules time is 841 milliseconds for Imported dataset and 32 milliseconds for Preprocess dataset. Similarly, the running time for the number of record sets 1 to 250 to generate decision tree and rules time is 1251 milliseconds and 40 milliseconds and the running time for the number of records set 1 to 300 to generate decision tree and rules time is 1760 millisecond and 46 milliseconds for Imported dataset and Preprocessed dataset.

According the facts explained above, the generating decision tree and rules time for preprocessed dataset is fast. By using the table 1, we have got this bar chart as shown in Figure 6.

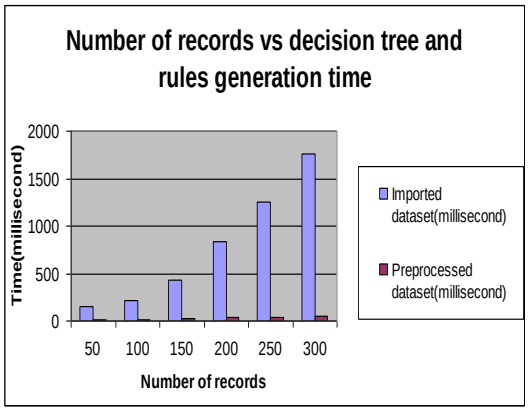

Figure 6. Performance Evaluations for Imported Dataset and Preprocessed Dataset

In figure 6 shows the difference performance for numbers of record sets and decision tree and rules generation time according the data of the Table 1 in chart. Y axis shows milliseconds for decision tree and rules generation time and X axis shows the numbers of record sets.

It also calculates the accuracy of the system by measuring the formula.

$$\text{Accuracy (\%)} = (\text{Total Correct Count} / \text{Initial Sample Record sets Count}) * 100 \quad (4)$$

Accuracy (imported) =(115/300)*100
=38%

Accuracy (preprocessed)=(300/300)*100
=100%

Therefore, the result of accuracy for the imported data set is 38 % and the preprocessed data set is nearly 100%.

5. Conclusion

Today's computers and corresponding software tools support the processing of data sets with millions of samples and hundreds of features.

Classification is the form of data analysis that can use to exact model describing important data classes. Preprocessing of the data in preparation for classification can involve data cleaning to reduce noise, relevance analysis to remove irrelevant attributes, etc.

Decision Tree Induction Algorithm is nonparametric approach for building classification models. This function analyzes a set of training data, constructs a model for each class based on the features in the data, and adjusts the model based on the test data. The algorithm includes ID3, C4.5, and SLIQ.

The completeness of the system can compute the time complexity by measuring milliseconds and the accuracy of the system is nearly 100%.

6. References

[1] Avigdor Gal, Aviv Segev, *"The Past Thirty Years Angent Oriented Data Integration"*, Technion.Israel Institute of Technology.

[2] C.E.Brodley and P.E.Utgoft, *"Induction of decision trees. Machine Learning"*, 1995.

[3] Jiawei Han, Micheline Kamber, *"Data Mining Concept and Techniques"*, ISBN 1-55860-489-8.

[4] Kalyani Anumalla, "*Data Preprocessing Managernent System"*, December, 2007.

[5] Lori Bowen Ayre, "*Data Mining for Information Prefessionals"*, June, 2006.

[6] Nick Taylor (N.K.Taylor@hw.ac.uk), "*Data Mining & Machine learning".*

[7] Selim Aksoy (saksoy@cs.bilkent.edu.tr), "*K-Nearest Neighbor Classifier and Distance Functions"*, Department of Computer Engineering Bilkent University, CS 551, spring 2008.