ANALYSIS OF A DATA TRANSFORMATION METHOD BY USING DECISION TREE

Cho Zin Win, Yin Ko Latt
University of Computer Studies, Magway
chozinwin86@gmail.com

ABSTRACT

Data preprocessing needs to make data cleaning which routines work to clean the data by filling in missing values, smoothing noisy data, identifying, removing outliers, and resolving inconsistencies. Data transformation operations are additional data preprocessing procedures that would contribute towards the success of the mining process. Data normalization transforms data values for different database attributes into a uniform set of units or into a uniform scale range. Classification is the process of finding the common properties among different entities and classifying them into classes. A decision tree is built from a set of training data having attribute values and a class name. In this paper, the original dataset is transformed using three different normalization methods such as min-max, z-score and decimal scaling normalization. The normalized dataset is evaluated using decision tree and analyses number of leaf nodes and accuracy. Comparisons between the three normalization methods were discussed in this paper.

1. INTRODUCTION

Data mining tools predict behavior and future trends, allowing businesses to make proactive, knowledge-driven decisions. Data mining tools can answer business questions that traditionally were too time-consuming to resolve. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations.

Preprocessing is a necessity whenever the data to be mined is noisy or incomplete and this process significantly improves the effectives of the data classification. Data cleaning refers to processing of data in order to treat missing values and attempt to fill in missing values.

Data transformation such as normalization may be applied and may improve the accuracy and efficiency of classification algorithms. Data transformation the data can be generalized to higher level concepts. Concept hierarchies may be used for this purpose. The data may also be normalizing, particularly when neural networks or methods involving distance measurements are used in the learning step [1]**.**

Decision tree is a method which comes from the machine learning community and explores data. Such a method is able to give a summary of the data (which is easier to analyze than the raw data) or can be used to build a tool to help a user for any different decision making tasks. Decision trees are usually much faster in the construction (training) phase and they also tend to be faster during the application phase [2].

The result of the process is represented as a tree which nodes specify attributes and branches specify attribute values. Leaves of the tree correspond to sets of examples with the same class or to elements in which no more attributes are available [3].

2. PREPOSED SYSTEM OVERVIEW

In this paper, the data set to make testing can be extracted from database. The extracted dataset is transformed by using the tree normalization methods such as min_max normalization, z_score normalization and decimal scaling normalization.

The normalized dataset is evaluated by using decision tree approach. This evaluation can generate number of leaf nodes and accuracy. The user can add or modify the desired datasets to database. The system can be made testing by using different datasets.

The main objectives of this paper are to study concepts of data transformation theory under data mining**,** to prevent attributes with large ranges, to speed up the learning phase by normalizing, to know how decision tree induction constructed, to study about classification and usage decision tree, to make testing using any datasets.

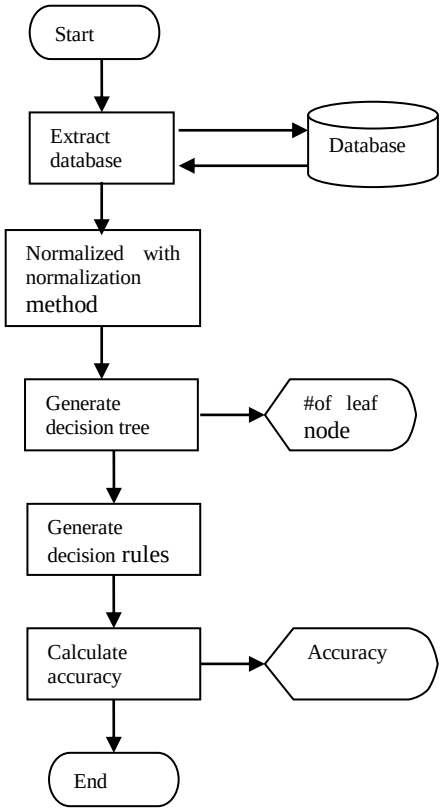

Figure 1. System Flow Diagram

3. DATA TRANSFORMATION

The data will be transformed into a form appropriate for mining. Missing data particularly for tuples with missing values for some attributes may need to be inferred. Data transformation is used for data smoothing, data aggregation, data generalization, attribute construction and data normalization**.** In normalization, data are scaled to fall within a small specified range. There are three methods in normalization. They are min-max normalization, z-score normalization and normalization by decimal scaling.

Subheadings should appear in lower case (initial word capitalized) in boldface. They should start at the left margin on a separate line.

3.1. Normalization

An attribute is normalized by scaling its values so they fall within a small-specified range, such as 0.0 to 1.0. Normalization is particularly useful for classification algorithms involving neural network, or distance measurements such as nearest neighbor classification and clustering. Trying to normalize data is an obvious solution where data are scaled so as to fall within a small specific range [4].

3.1.1. Min-max Normalization

Min-max normalization is one of the normalization methods. It is used to transform the input data as a small-specified range. The formula of min-max normalization is

$$v' = \frac{v - \min_A}{\max_A - \min_A}(new\_\max_A - new\_\min_A) + new\_\min_A$$

Where,

min-max normalization maps a value v of A to v' in the range [new_min_A, new_max_A].

min_A and max_A are the minimum and maximum values of an attribute A.

3.1.2. Z-score Normalization

Z-score normalization is one of the normalization methods. This method is used to transform the input data into a small-specified range, such as 0.0

to 1.0. The formula of z-score normalization method is

$$v' = \frac{v - \overline{A}}{\sigma_A}$$

Where,

The values for an attribute A are normalized based on the mean and standard deviation of A.

A value v of A is normalized to v'.

$\overline{A}$ and σ_A are the mean and standard deviation, respectively, of attribute A.

3.1.3. Decimal Scaling Normalization

Decimal scaling normalization is one of the normalization methods. This method is used to transform the input data into a small-specified range, such as 0.0 to 1.0. Normalization by decimal scaling normalizes by moving the decimal point of values of attribute A. The formula of decimal scaling is

$$v' = \frac{v}{10^j}$$

Where,

The number of decimal points moved depends on the maximum absolute value of attribute A.

A value v of A is normalized to v'.

j = the smallest integer such that Max ($|v'|$) < 1.

4. DECISION TREE

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 node represents class or class distributions. The top most node in a tree is the root node.

In order to classify an unknown sample, the attribute values of the sample are tested against the decision tree. A path is traced from the root to a leaf node that holds the class prediction for that sample. Decision trees can be easily converted to classification rules. Decision tree have been used in many application areas ranging from medicine to game theory and business [5].

Decision tree algorithms are quite robust to the presence of noise, especially when methods for avoiding over fitting are employed. The presence of redundant attributes does not adversely affect the accuracy of decision trees.

4.1. Decision Tree Induction

Decision tree induction is a greedy algorithm that constructs decision tree in a top-down recursive divided-and-conquer manner. Decision tree induction is relatively faster learning speed than other classification rules, can use SQL queries for accessing databases, comparable classification accuracy with other methods.

The three has types of nodes:

- A root node that has no incoming edges and zero or more outgoing edges.
- Internal nodes, each of which has exactly one incoming edge and two or more outgoing edges.
- Leaf or terminal nodes, each of which has exactly one incoming edge and no outgoing edges.

In a decision tree, each leaf node is assigned a class label. The non-terminal nodes, which include the root and other internal nodes, contain attribute test conditions to separate records that have different characteristics [4].

4.2. Attribute Selection Measure

We may think of each feature of instance as contribution a certain amount of information towards its classification. The feature's ability to classify the instance is calculates by measuring the expected information. Attribute relevance of an attribute with respect to a given class and such measures include information gain [5].

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.

Let S be a set consisting of data samples. Suppose the class label attribute has ***m*** distinct values defining ***m*** distinct class, C_i (for i=1….m). Let s_i be the number of samples of S in classes C_i

with probability s_i /s, where s is the total number of samples in set S. The expected information needed to classify a given sample is

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

An attribute ***A*** with values $\{a_1, a_2, ..., a_r\}$ can be used to partition S into the subsets $\{s_1, s_2, ..., s_r\}$, where S_j contains those samples in Z that have value a_j of A. Let S_j contain s_{ij} samples of class C_i .The expected information based on this partitioning by A is known as entropy of A. It is the weighted average:

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

The information gain obtained by this partitioning on attribute A is defined by

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

The feature with the highest information gain is considered the most discriminating feature of the given set. By computing the information gain of each feature, we obtain a ranking of features; this ranking can be used for relevance analysis to select the features to be used in classification [4].

4.3. Extracting Classification Rules from Decision Trees

The knowledge represented in a 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 [4].

5. EXPERIMENTAL RESULTS

This paper is using the three normalization methods to transform the data sets. Then the normalized datasets are evaluated by decision tree approach. This evaluation method can generate number of leaf nodes and accuracy. The system can be made testing using different datasets. In this paper, the system is made experiment using the sample dataset the name of Blood Offering table which contains four attributes and the two class that are can offer or not.

Table 1. Original Dataset

Day	Blood type	Age	Pound	Health	Class
1	1	1	1	1	No
2	1	1	1	2	No
3	2	1	1	1	Yes
4	3	2	1	1	Yes
5	3	3	2	1	Yes
6	3	3	2	2	No
7	2	3	2	1	Yes
8	1	2	2	2	No
9	1	3	2	1	Yes
10	3	2	2	1	Yes
11	1	2	1	1	Yes
12	2	2	2	2	Yes
13	3	1	1	1	No
14	1	2	2	1	Yes

The attributes the Table 1 are normalized by using min-max normalization formula. These attributes are blood type, age, pound and health. The results of min-max normalization are shown in Table 2. These results are in a small- specified range, such as 0.0 to 1.0.

Table 2. Result Table of Min_max Normalization Method

Day	Blood type	Age	Pound	Health	Class
1	1.0	0.67	0.5	1.0	No
2	1.0	0.67	0.5	0.5	No
3	0.33	0.67	0.5	1.0	Yes
4	0.67	1.0	0.5	1.0	Yes
5	0.67	0.33	1.0	1.0	Yes
6	0.67	0.33	1.0	0.5	No
7	0.33	0.33	1.0	1.0	Yes
8	1.0	1.0	1.0	0.5	No
9	1.0	0.33	1.0	1.0	Yes
10	0.67	1.0	1.0	1.0	Yes
11	1.0	1.0	0.5	1.0	Yes
12	0.33	1.0	1.0	0.5	Yes
13	0.67	0.67	0.5	1.0	No
14	1.0	1.0	1.0	1.0	Yes

Z_score normalization method is used to transform the input data. The results are within a small-specified range, such as 0.0 to 1.0. These

results are shown in Table 3. The data with small ranges are easy for classification and decision tree.

Table 3. Result Table of Z_score Normalization Method

Day	Blood type	Age	Pound	Health	Class
1	0.725	-0.15	-0.79	0.46	No
2	0.725	-0.15	-0.79	-1.16	No
3	-1.305	-0.15	-0.79	0.46	Yes
4	-0.29	0.87	-0.79	0.46	Yes
5	-0.29	-1.16	0.79	0.46	Yes
6	-0.29	-1.16	0.79	-1.16	No
7	-1.305	-1.16	0.79	0.46	Yes
8	0.725	0.87	0.79	-1.16	No
9	0.725	-1.16	0.79	0.46	Yes
10	-0.29	0.87	0.79	0.46	Yes
11	0.725	0.87	-0.79	0.46	Yes
12	-1.305	0.87	0.79	-1.16	Yes
13	-0.29	-0.15	-0.79	0.46	No
14	0.725	0.87	0.79	0.46	Yes

The results from the decimal scaling normalization method are shown in Table 4. The values of attributes in Table 4 are within a small-specified range, such as 0.0 to 1.0. They are easy for classification and decision tree.

Table 4. Result Table of Decimal Scaling Normalization Method

Day	Blood type	Age	Pound	Health	Class
1	0.02	0.01	0.00	0.01	No
2	0.02	0.01	0.00	0.00	No
3	0.00	0.01	0.00	0.01	Yes
4	0.01	0.02	0.00	0.01	Yes
5	0.01	0.00	0.01	0.01	Yes
6	0.01	0.00	0.01	0.00	No
7	0.00	0.00	0.01	0.01	Yes
8	0.02	0.02	0.01	0.00	No
9	0.02	0.00	0.01	0.01	Yes
10	0.01	0.02	0.01	0.01	Yes
11	0.02	0.02	0.00	0.01	Yes
12	0.00	0.02	0.01	0.00	Yes
13	0.01	0.01	0.00	0.01	No
14	0.02	0.02	0.01	0.01	Yes

Three different normalization methods were considered: z-score normalization, min-max normalization and decimal point normalization. This study was extended to take into account z-score and decimal scaling normalization methods. Comparisons between the three normalization methods were discussed in this paper. Data set was normalized using the three methods of normalization that were mentioned earlier. Data sets are generated from min-max, z-score and decimal scaling normalization methods respectively.

It is well known that different techniques usually generate different results when they are applied to a specific task. A data set can be redesigned in some way that helps techniques to generate better results. For example, rules generation technique could give low accuracy when it is applied to decimal scaling normalization data set, while it gives much better accuracy when it is applied to z-score or min-max normalization data sets. Designing a task in some way could help in generating better accuracy.

Decision tree methodology for data mining and knowledge discovery[6] was used to test the data sets that were designed earlier. For each data set, the accuracy, numbers of leaf nodes were computed. Table 5 summarizes how ID3 technique performs on the three normalization methods.

Table 5. Result of Three Different Approaches

Approach	Min-min	Z_score	Decimal scaling
Accuracy	83	80	80
# of leaf nodes	7	6	6

Number of leaf nodes that represents the simplicity was generated for data set. Number of leaf nodes is 7, 6 and 6 respectively. The z-score and the decimal point normalization data set were of higher simplicity than the min-max normalization data set as they give the minimum number of leaf nodes. When ID3 was applied to data set the accuracy results were as follows 94.2%, 92% and 92% respectively. The higher accuracy was noticed when the min-max normalization data set is used. The next higher accuracy was when the z-score and the decimal point normalization data sets were used.

6. CONCLUSION

The system is using three different normalization methods. When applying data mining to the real world, learning from data that fall within a large specific range is an evitable situation. Trying to normalize data is an obvious solution where data are scaled so as to fall within a small specific range. Techniques that are used to normalize data must not introduce noise. Experiments were designed to test the effect of different normalization methods on number of leaf nodes and accuracy.

REFERENCES

[1] Dunham, M.H. (2003). Data mining introductory and advanced topics,. Upper Saddle River, NJ: Pearsan Education, Inc.

[2] Benoit, Gerald. (2002). Data mining. In Cronin, B. (Ed).Annual Review of Information Science and Technology: Vol.36. Silver Spring.MD: American Society for Information Science and Technology.

[3] Breiman L., Friedman J.H, Olshen R.A., & Stone C.J.Classification and regression trees. Wadsworth.Statistics probability series. Belmont, 1984.

4] Han, J., & Kamber, M. (2001). Data mining: concepts and techniques (Morgan_Kaufman Series of Data Management Systems). San Diego: Acadernic Press.

[5] Pang-Ning, Tan Michael Steinbach and Vipin Kumar " Introduction to Data Mining".

[6]. Gora, G. and A. Wojna, (2002). RIONA: A_ new_classification_system_combining_rule_induct ion_and_instance-based_learning. Fundamenta_Informaticae,51:369-390.