Decision Tree Optimization in Data Mining with Support and Confidence

Relita Buaton¹², Herman Mawengkang¹, Muhammad Zarlis¹, Syahril Effendi¹, Akim Manaor Hara Pardede¹², Yani Maulita², Achmad Fauzi³, N Novriyenni², Anton Sihombing²,Katen Lumbanbatu³

¹Program Studi Doktor Ilmu Komputer, Fakultas Ilmu Komputer dan Informasi Teknologi, Jl. Universitas No. 9A, Medan, 20155, Universitas Sumatera Utara, Indonesia
²STMIK Kaputama, Jl. Veteran No. 4A-9A, Binjai-Sumatera Utara, Indonesia

* bbcbuaton@gmail.com

Abstract. Decision Tree is a classification technique in data mining that aims to predict behaviour from database. This goal is supported by several algorithms, one of which is Iterative Dichotomiser 3 (ID3) that displays predictions in a tree structure. With the application of decision trees, warehouses or heaps of data can be processed so as to produce rules or decision trees as decision support in solving problems faced by agencies. In fact, the information or rules produced by decision trees are limited to rules using the logic of propositions. The challenge in making decisions on decision trees is how to determine algorithms with a high degree of accuracy from various algorithms in the decision tree and how to find support and confidence for each rule produced by the decision tree to add support value and confidence level of each rule produced. The resulting rule has weaknesses, namely the unavailability of support and confidence, all rules are considered equal in strength based on data before being processed, found records that vary or different amounts of data. By making support and confidence, it will be easier to make decisions based on the results obtained.

1. Introduction

The increasing transactional requirements require that database capabilities and capacities also increase. This raises the thought of knowing how to extract the database into useful information so that data mining becomes an important research material today[1]. Data mining is an analytical process of knowledge discovery in a large and complex number of data sets.Data mining is a discipline that is at the intersection of statistics, operations research, and computer science. More precisely, data mining is the result of hybridization from statistics, computer science, artificial intelligence and machine learning [2]. Generally data users and researchers need information or knowledge from existing databases by grouping, classifying, finding rules and visualizing data mining. The discovery of knowledge in the database is useful for finding trends, patterns and anomalies in a database that can be used to make accurate decisions in the future. Data mining is defined as technology to extract valuable and useful information from a set of huge amount of data and analyzing and associate between set items [3]. Data sources can be in the form of databases, data warehouses, web, other information repositories, or dynamically transmitted data. The process discovery of knowledge in the database has several steps such as data cleaning, data integration, data selection, data transformation, data mining, pattern evaluation and knowledge presentation [4]. The ID3 algorithm (Iterative Dichotomiser version 3) is one of algorithm that used for data processing. With the application of decision trees, warehouses
or heaps, data can be processed so as to produce rules or decision trees as decision support in solving
or concluding problems faced by agencies. In reality, the information or rules produced by decision
trees are limited to rules using the logic of propositions. For example if X and Y→Z (if X and Y then
Z). The resulting rule has weaknesses, namely the unavailability of support and confidence. All rules
are considered equal in strength based on data before being processed, found records that vary or
different amounts of data. By making support and confidence, it will be easier to make decisions based
on the results obtained. The challenge in making decisions on decision trees is how to process data by
determining Entropy to process data into very interesting information, how to determine algorithms
with high accuracy levels of various algorithms in the decision tree and how to find support and
confidence for each rule that the decision tree produces for increase the value of support and the level
of confidence of each rule produced.

2. Methodology

2.1. Decision Tree
Decision tree is a technique that can define or find rules automatically and can be generally applied to
data that has never been known. Decision tree is also very popular and widely used practically because
it tries to find the functions of a discrete-value that is resistant to errors (data noise). Besides decision
tree be able to learn disjunctive expressions (OR expression). There are several algorithms included in
the decision tree, namely ASISTANT, C. 45 and ID3. ID3 tries to build a decision tree top-down
(from top to bottom) starting with the determination of the attribute as root (root). To determine root
by evaluating all attributes with statistical measures, namely information gain with the aim of
measuring the effectiveness of attributes in classifying a collection of data samples. The biggest
information Gain is the attribute as root [5].
a. Entropy
To calculate information gain, first by calculating entropy as a parameter to measure the
heterogeneity(diversity) of a sample collection. If the collection of sample data is increa
singly
heterogeneous, the entropy value gets bigger. Mathematically written as follows:

\[ \text{Entropy}(S) = - \sum_{i=1}^{C} p_i \log_2 p_i \]  

Where is C: number of target attribute values (number of classification classes), pi: number of
samples for class i
b. Information Gain
The next step is measuring the effectiveness of all attributes. It aims to find the best classifier
attribute to be used as root node. The formula is as written as follows:

\[ \text{Gain}(S,A) = \text{Entropy}(S) - \sum_{v \in \text{Values}(A)} \frac{|S_v|}{|S|} \text{Entropy}(S_v) \]  

Where, A is an attribute, V is declaring a possible value for attribute A, Values (A) is the set of
possible values for attribute A, | Sv | is the number of samples for the value of v, | S | is the sum of
all data samples and Entropy (Sv) is entropy for samples that have a value of v

2.2. ID3 Algorithm
ID3 is an algorithm that performs greedy searches on all possible decision trees. The ID3 algorithm
can be implemented using recursive functions (the function that is called itself) [5]. The steps in the
decision tree construction are as follows:
a. Step 1: Determination of root node as the initial preparation of the tree.
b. Step 2: Determination of leaf node based on similarity and it is called as a class. Otherwise, information gain will be used to select the best attributes in separating sample data into individual classes.
c. Step 3: A branch will be created for each value in the attribute and the sample data will be partitioned again.
d. Step 4: This algorithm uses a recursive process to form a decision tree on each data partition. If an attribute has been used in a node, this attribute will no longer be used in the node of the child.
e. Step 5: This process stops if the condition of all samples in the node is reached in one class and no other attributes can be used to partition the sample further. In this case most votes will be applied. This means turning a node into a leaf and named it with the class in the most votes.

2.3. Support and Confidence

Association rule analysis is a data mining technique to find associative relationships of items combination. The association rule is an $A \rightarrow B$ implication statement, where $A$ and $B$ are itemsets that disjoint and meet the requirements of $A \cap B=\{\}$ [6]. Association rules want to provide information in the form of if then. This rule is calculated from probabilistic data [1]. The associative rule of items can be known with two parameters, namely support, that shows items combination percentage in database and confidence which is shown the strong relationship between items in associative rules. Association analysis aims to find all associative rules that meet the support minimum requirements (minimum support) and confidence minimum requirements (minimum confidence). There are several algorithms that have been developed regarding association rules, but there is one classic algorithm that is often used, namely the a priori algorithm. This algorithm is designed to develop frequent itemset. To develop a frequent set with two items, you can use frequent set items. In associations there are antecedent and consequent, antecedent terms to represent "if" and consequent parts to represent the "then" part. Antecedent and consequent are a group of items that do not have a shared relationship. Mathematically written as follows:

$$S = \frac{\sum (Ta + Tc)}{\sum (T)}$$  \hspace{1cm} (3)$$
Where $S$ is Support, $\sum (Ta + Tc) =$ Number of transactions containing antecedent and consequent, $\sum (T) =$ Number of transactions

$$C = \frac{\sum (Ta + Tc)}{\sum (Ta)}$$  \hspace{1cm} (4)$$
Where $C$ is Confidence, $\sum (Ta + Tc) =$ The number of transactions containing antecedent and consequent, $\sum (Ta)$ is the number of transactions containing antecedent.

2.4. Literature Review

According to Anyanwu, N.M. Shiva, G.S in 2013, the decision tree is widely applied because it is easy to implement and needs to be pruned to increase the accuracy of the results obtained. In this study also carried out a comparison algorithm ID3, C45, CART. The three algorithms have a relationship between execution time and the process of tree development with data volume [7]. According to Kakavand, S. Mokfi, T. Tarokh, J.M in 2014, this research was applied in the world of education by processing student data using a comparison of C.45, CHART and CHAID. Using the decision tree method is able to predict whether students will continue their studies or not. The accuracy of the data is 94% using the algorithm, C.45 [8]. According to Trivedi, A 2014 research was applied in the world of education by processing student data. The decision tree method is able to classify students into several groups, namely very good, medium, average and below average. With the results of this study the teacher or lecturer can more easily develop students’ abilities [9]. According to Kareem, A.I. Duaimi, G.M in 2014 that the C4.5 Algorithm performs well in building decision trees and issuing rules from continuous data sets [10]. In 2002, Adepele Olukunle presented an algorithm of data
mining association rules that were fast and suitable for handling a collection of medical image data and assessing suitability according to the proposed algorithm [11]. In 2006, Carlos Ordonez said that association rule was a technique to improve the predictive ability of heart disease, by applying search limits to reduce search. The rules found later were then evaluated with support, lift and confidence [12]. In 2010, according to Ravikumar, ACO was able to predict the risk of accidents on the road by using decision trees [13]. In 2011, Pooia Laibakhsh presented a revision of the Ant Miner algorithm using ant based rule mining by modifying the rule trimming process and introducing a dynamic pheromone evaporation approach [14]. In 2013, Divya Bhugra perfected the rules generated from the Association data mining rules using Biogeography based optimization (BBO) by sharing information between solutions that depend on the migration mechanism [15]. In 2013, Rameshkumar, using n cross validation techniques to reduce association rules that were not relevant to the collection of transactions, the approach used was a partition-based approach to support validation of the association rule [16]. Some of the above studies need to be refined because they have weaknesses in every rule produced, there is no found support and confidence that convinces each rule.

2.5. Data Mining and Knowledge Discovery

Data mining and Knowledge Discovery in Databases (KDD) are often used interchangeably to explain the process of extracting hidden information in a large database. Actually the two terms have different concepts, but are related to each other. Data mining is capable for:

a. Disclosure of pattern hidden in the database.
b. Ensure the pattern found is valid and can be used as knowledge.
c. Find new patterns that are beneficial for users.
d. Assisting organizations or users in making policy related to the knowledge obtained
e. Understood, new patterns must be understood and increase user knowledge [2]

Meanwhile, the knowledge discovery in database focuses on the stages in extracting the database, where data mining is included. In recent times, extracting databases to obtain various knowledge or information is increasingly needed. But most of the research related to this is still concentrated in the form of rules not on the truth of the rules themselves. This study is deemed necessary to convince users to take advantage of the rules or patterns that have been generated. The important task of data mining algorithms is to find hidden patterns besides being accurate and understandable [17]. To assess or rank rules that is by using the dominance factor in accordance with the rule position in a hierarchy, based on the potential dominance of interesting, technically interesting and genuinely interesting. To achieve this by designing criteria such as performance, simplicity and significance in the database [19].

3. Results And Discussion

The rules produced by the decision tree are limited to rules with the logic of propositions, which are assumed by if A and B → C (if A and B then C, or if A or B then C), are considered to have weaknesses because all rules have the same strength and need to be optimized with the following stages

a. Processing the data warehouse into a decision tree, for example the database is as follows:

| Id | GPA  | Presence | Attitude | Scholarship |
|----|------|----------|----------|-------------|
| 1  | Good | High     | Good     | Y           |
| 2  | Good | Medium   | Good     | Y           |
| 3  | Good | Medium   | Not Good | Y           |
| 4  | Good | Low      | Not Good | N           |
| 5  | Medium | High     | Good     | Y           |
| 6  | Medium | Medium   | Good     | Y           |
| 7  | Medium | Medium   | Not Good | Y           |
| 8  | Medium | Low      | Not Good | N           |
| 9  | Low   | High     | Good     | Y           |
Then process the data using the ID3 algorithm, the following decision tree is produced:

Figure 1. Decision Tree Data Acceptance of Scholarships

b. Turn the decision tree into a rule

Rule 1: if attitude = good then scholarship = Yes
Rule 2: if attitude = not good and presence = high then scholarship = No
Rule 3: if attitude = not good and presence = medium and GPA = good then scholarship = Yes
Rule 4: if attitude = not good and presence = medium and GPA = medium then scholarship = Yes
Rule 5: if attitude = not good and presence = medium and GPA = low then scholarship = No
Rule 6: if attitude = not good and presence = low then scholarship = No

c. Determine support and confidence for each rule produced by the decision tree using a priori algorithm, such as the following pseudocode.

Figure 2. Pseudo Code of Determining Rules by Apriori Algorithm

A rule with support and confidence is generated as follows:

| No | Rule | Support | Confidence |
|----|------|---------|------------|
| 1  | if attitude=good then scholarship=Yes | 55%     | 100%       |
| 2  | if attitude=not good and Presence=medium and GPA=good then scholarship=Yes | 9%      | 100%       |
### 4. Conclusion

Every rule that is produced is more optimal, has been equipped with support and confidence that can be used in making decisions.

### References

[1] L. W. Santoso, “Pembuatan Perangkat Lunak Data Mining Untuk Metode Apriori,” *J. Pendidik. Teknol. Inf. dan Komun.*, pp. 49–56, 2013.

[2] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, “From Data Mining to Knowledge Discovery in Databases,” *AI Mag.*, vol. 17, no. 3, pp. 37, 1996.

[3] N. N. Das and A. Saini, “A Study on Association Rule Mining Using ACO Algorithm for Generating Optimized ResultSet,” vol. 2, pp. 123–128, 2013.

[4] J. Han, M. Kamber, and J. Pei, *Data Mining Concepts and Techniques*, Third Edit., vol. 1. Elsevier, 2012.

[5] Suyanto, *Artificial Intelligence*. Bandung: Informatika, 2011.

[6] R. Buaton, Y. Maulita, and A. R. Febria, “Korelasi Kecerdasan Emosional dengan Prestasi Belajar Siswa Menggunakan Metode a Priori,” vol. 1, no. 1, pp. 33–43, 2017.

[7] M. N. Anyanwu, “Comparative Analysis of Serial Decision Tree Classification Algorithms,” no. 3, pp. 230–240, 2013.

[8] T. Mokfi and M. J. Tarokh, “Prediction the Loyal Student Using Decision Tree Algorithms 1,” vol. 4, no. 1, pp. 32–37, 2014.

[9] A. Trivedi, “International Journal of Advanced Research in Computer Science and Software Engineering Evaluation of Student Classification Based On Decision Tree,” vol. 4, no. 2, pp. 111–112, 2014.

[10] I. A. Kareem and M. G. Daaimi, “Improved Accuracy for Decision Tree Algorithm Based on Unsupervised Discretization,” vol. 3, no. 6, pp. 178–183, 2014.

[11] A. Oluwunkle and S. Ehiokya, “A Fast Algorithm for Mining Association Rules in Medical Image Data,” pp. 1181–1187, 2002.

[12] C. Ordonez, “Association Rule Discovery With the Train and Test Approach for Heart Disease Prediction,” vol. 10, no. 2, pp. 334–343, 2006.

[13] K. Ravikumar and A. Gnanabaskaran, “ACO based spatial data mining for traffic risk analysis,” vol. 1, no. 1, 2010.

[14] P. Lalbakhsh, M. S. K. Fasaei, and M. N. Fesharaki, “Focusing on Rule Quality and Pheromone Evaporation to Improve ACO Rule Mining,” pp. 1–5, 2011.

[15] D. Bhugra and S. Goel, “Association Rule Analysis Using Biogeography Based Optimization,” pp. 1–5, 2013.

[16] K. Rameshkumar and S. Ravi, “Relevant Association Rule Mining from Medical Dataset Using New Irrelevant Rule Elimination Technique,” 2013.

[17] K. MCGARRY, “A Survey of Interestingness Measures for Knowledge Discovery,” *Knowl. Eng. Rev.*, vol. 0, no. January, pp. 1–24, 2004.

[18] W. J. Frawley, G. Piatetsky-shapiro, and C. J. Matheus, “Knowledge Discovery in Databases : An Overview,” *AI Mag.*, vol. 13, no. 3, pp. 57–70, 1992.

[19] J. A. Major and J. J. Mangano, “Selecting among rules induced from a hurricane database,” *J. Intell. Inf. Syst.*, vol. 4, no. 1, pp. 39–52, 1995.