Implementation of Data Mining to Analyze Drug Cases Using C4.5 Decision Tree

Sri Wahyuni
Lecturer of Computer Engineering in University of Pembangunan Panca Budi

Abstract. Data mining was the process of finding useful information from a large set of databases. One of the existing techniques in data mining was classification. The method used was decision tree method and algorithm used was C4.5 algorithm. The decision tree method was a method that transformed a very large fact into a decision tree which was presenting the rules. Decision tree method was useful for exploring data, as well as finding a hidden relationship between a number of potential input variables with a target variable. The decision tree of the C4.5 algorithm was constructed with several stages including the selection of attributes as roots, created a branch for each value and divided the case into the branch. These stages would be repeated for each branch until all the cases on the branch had the same class. From the solution of the decision tree there would be some rules of a case. In this case the researcher classified the data of prisoners at Labuhan Deli prison to know the factors of detainees committing criminal acts of drugs. By applying this C4.5 algorithm, then the knowledge was obtained as information to minimize the criminal acts of drugs. From the findings of the research, it was found that the most influential factor of the detainee committed the criminal act of drugs was from the address variable.

1. Introduction
Data Mining is one of the fastest growing fields due to the huge demands for added value of large-scale databases that are in line with the growth of information technology. Data mining can extract large data sets into new knowledge. Classification is a process of finding models that explain or distinguish concepts or classes of data, in order to be able to predict the class of an object whose class is unknown [1]. In the Classification there are several methods namely decision tree, neural network, rought set theory, Bayesian theory and fuzzy logic [2]. And each function has its own algorithm. C4.5 is a well-known and used for data classification that has numerical and categorical attributes. In previous research, Angga and Riani (2012) used data mining to predict the criteria of customers credit using the classification method used on a decision tree which later result in an application that allows the bank's funds to obtain credit marketing targets [3]. Muhammad Syahril (2011) used data mining to convert hypertension disease training data [4]. Sri wahyuni (2017) also uses data mining to analyze dropped- out students data. This research uses Decision Tree C4.5 algorithm that belong in Classification function. The decision tree method turns a very large fact into a decision tree that represents rules. The purpose of this study is to analyze the data of prisoners in Labuhan Deli prison extracted with data mining decision tree
algorithm C4.5, then it will generate new knowledge to learn about the factors of detainees doing drug crimes.

2. Literature Review

1. Understanding Data Mining

   In a book titled "Data Mining", data mining contains a searching process of trend or pattern in a large database for future decision making. It is expected that data mining devices are able to recognize these patterns of data with minimum input. These patterns are recognized by certain devices that can provide a useful and insightful data analysis that can later be studied more closely, which may use other decision-taking support tools [5]. Data mining deals with a sub-area of statistics called exploratory data analysis, which has the same objective and relies on statistical measures. Data mining is also closely related to the artificial intelligent sub-area called knowledge discovery and machine learning. An important characteristic of data mining is that the data volume is huge even though the ideas from the related study area can be applied to data mining problems, the scalability associated with the size of the data becomes an important new criterion. According to Witten et al in Anik Andriani, data mining is defined as the process of finding patterns in a sets of data. This process must be in automatic or semi-automatic behavior. The resulting pattern should mean that the pattern gives some advantages. The patterns are to be identified, validated, and later used to make a prediction. From the above notions it can be concluded that Data mining is a process of extracting stored data in large databases [16].

2. Classification

   In classification, there are target categories of variables. For example, the classification of income can be separated into three categories: high income, medium income and low income. Then to determine the income of an employee, the classification method is used in data mining.

   According to Han et al and Quinlan in their publication paper "Pattern Extraction of On-Time Graduation of STMIK AMIKOM Yogyakarta Student using Data Mining Algorithm C4.5" (Muchamad Widiarto, 2011). Classification techniques in data mining are grouped into decision tree techniques, Bayesian (Naive Bayesian and Bayesian Belief Network), neural networks (Backpropagation), a technique that is conceptually based on mining association rules and other techniques (K-Nearest Neighbor, genetic algorithm, techniques with rough and fuzzy set approach). Each technique has its own advantages and disadvantages. Data with a particular profile may be most optimal if it is classified by a particular technique, or in other words, a particular data profile might be supported by the advantages of this technique.

3. Decision Tree

   Decision tree is one of the most popular classification methods because it is easy to be interpreted by humans. The basic concept of the Decision Tree algorithm is to convert data into decision trees and rules (Defiyanti and Crispina Pardede). The concept of data in the decision tree (Muhammad Syahril, 2011) is represented as in the picture below:
   a. Data is expressed in tabular form with attributes and records.
   b. The attribute states a parameter created as a criterion in the formation of tree. One of the attributes is an attribute that states the solution data per item called the target attribute.
   c. The attribute has values named with instance. For example, the body weight attribute have instances of overweight, average and underweight.
Decision tree uses a hierarchical structure for supervised learning. The decision tree composition consists of several sections called nodes. The process of the decision tree starts from the root node until the leaf node, which is done recursively, where every branches state a condition that must be met and at the end of a tree declares the class of a data. The decision tree consists of three parts (Ratih and Isye):

a. Root node is the node located at the top of a tree.
b. Internal node is a branching node, there is only one input and has at least two outputs.
c. Leaf node is the final node, has only one input and has no output.

According to Han in Muhammad Syahril (2011), in a decision tree, each leaf node marks the class label. A node that is not the final node consists of a root and an internal node that consists of an attribute test condition on a part of a record that has different characteristics. The root node and internal node are marked with an oval shape and the leaf node is marked with a rectangular shape.

4. **C4.5 Algorithm**

According to Sujana in the National Seminar on Information Technology Applications 2010, C4.5 Algorithm is grouped as a decision tree algorithm. This algorithm has input in the form of training samples. Training samples are sample data that will be used to build a tree that has its correctness tested. On the other hand, samples are data fields that will be used as a parameter in doing data classification. The results of the classification process are in the form of rules that later can be used to predict the value of the discrete-type attribute of the new record (Veronica Mortini, 2003). In general, the C4.5 algorithm for building decision trees is as the following (Muhammad Syahril, 2011);

1. Select an attribute as a root.
2. Create a branch for each value.
3. Divide the cases in the branch.
4. Repeat the process for each branch having the same class.

To select an attribute as a root, it is based on the highest gain value of the attributes. To get the value of gain, it must first determine the entropy value.
3. Research Methods

The stages in this study follow the general pattern of scientific research as seen in Figure 2.

![Research Framework Diagram]

3.1. Research Sites

The research is conducted in the Administrative Office of Labuhan Deli Simpang Jail (Rumah Tahanan). The registration part can be analyzed through custody data with the goal to learn the rules or knowledge, which is the relationship between input and target variables.

3.2. Variables Selection

The data of detainees at the Labuhan Deli State Jail in 2011/2013 have the following format:
1. Registration Number
2. Age
3. Education Level
4. Occupation
5. Address
6. Case

In the classification technique, the output of each targeted data or class must be an integer or a discrete number. From the above detainees data which is used as the target parameter or decision variable (class) is a case containing the value of drug and non-drug parameters. The value of the drug parameter means that the detainee does the type of drug case, while the non-drug parameter value means that the detainee commits another type of crime other than drugs such as theft, robbery, fraud, etc.
4. **Decision Tree Implementation**

Here are the steps to get the Decision Tree model using RapidMiner Decision Tree C4.5 algorithm:

1. Detainees data that contains variables of the conditions and attributes of the decision or target, to know whether the prisoners committed a criminal offense that is type of drug cases and not drugs are stored with Microsoft Excel containing the case to produce the rules information.

2. Prisoner data that has been stored in Microsoft excel can be used directly on the software application Rapid Miner 5.2. By opening the RapidMiner software, then Click New.

3. Then click Repositories, choose Import Excel Sheet. Select the test data saved in Microsoft Excel, click Next.

4. The Data Import Wizard will appear, click Next, on Annotation select Name, click Next to display the Data Import Wizard Step 4.

5. In the number column. select text and id, in the age column select the polynomial and attribute, in the education level column select the polynomial and attribute, in the occupation column select binominal and attribute, in the address column select polynomial and attribute, in the case column as a decision select binominal and label, then click Next.

![Figure 4. Data Import Wizard Step 4](image)

6. Click Next. Select NewLocalRepository, then click import data. Then click Finish. Then NewProcess RapidMiner will appear.

7. Drag the data on the process screen, then click Operator, Modeling, Classification and Regression, Tree Induction, Decision Tree. Then drag it to the Process screen.

8. Click No Pre Pruning and Pruning, on the criterion select Information_Gain. Click the Run icon to display the decision tree results.

![Figure 5. Decision Tree Result](image)
5. Conclusion

From the research that has been done by the author, there are some conclusions that can be drawn and used as knowledge. The conclusions are:

1. In the selection of detainees data variables, which are the conditions attribute and the decisions attributes, they greatly affect the generated rule or knowledge.
2. The more the data being processed, the more rules or knowledge obtained.
3. The decision tree method that is processed by Rapid Miner Software Application is more effectively used in data processing of prisoner at Labuhan Deli Jail.
4. Decision tree of C4.5 Algorithm can provide information or knowledge of a rule that is easily understood by the wider community because it is described with the decision tree.

For further research and development of this application, the author would like to give some suggestions as follows:

1. From the results of rules or knowledge obtained from the data of prisoners, the factors of why the prisoners did criminal acts of drugs can be known. It is expected that the concerning parties, such as the police and the community, can anticipate or reduce the crime of drug offenses.
2. For further development, it would be better to conduct a research combination of several classification techniques to be compared so that it can be known which classification techniques are more effective and efficient.

References

[1] Iko Pramudiono. (2003). Pengantar Data Mining : Menambang Permata Pengetahuan di Gunung Data. Kuliah Umum Ilmu Komputer.
[2] Hamidah Jantan1, abdul Razak2. (2010). Human Talent Prediction in HRM Using C4.5 Classification Algorithm. ISSN : 0975-3397.
[3] Angga Ginanjar Mabrur1, Riani Lubis2. (2012). Penerapan Data Mining Untuk Memprediksi Kriteria Nasabah Kredit. Jurnal Komputa Edisi 1 Volume1.
[4] Budianis Dwi Meilani1, Achmad Fauzi Slamat2. (2011). Klasifikasi Data Karyawan Untuk Menentukan Jadwal Kerja menggunakan Metode Decision Tree.
[5] Fajar Astuti Hermawati. (2013). Data Mining. Yogyakarta : Andi.
[6] Anik Andriani. (2012). Penerapan Algoritma C4.5 Pada Program Klasifikasi Mahasiswa Dropout. Seminar Nasional Matematika.
[7] Fernandes Edy Syahputra (2012). Kebijakan Hukum Pidana Terhadap Pengaturan Tindak Pidana Narkotika Di Indonesia. Jurnal Ilmiah.
[8] Angga Ginanjar Mabrur1, Riani Lubis2. (2012). Penerapan Data Mining Untuk Memprediksi Kriteria Nasabah Kredit. Jurnal Komputa Edisi 1 Volume1.
[9] Kusrini dan Luthfi Taufiq Emha. (2009). Algoritma Data Mining. Yogyakarta : Andi.