Data Mining Classification Comparison (Naïve Bayes and C4.5 Algorithms)

Leni Marlina[1], Muslim[2], Andysah Putera Utama Siahaan[3]
Faculty of Computer Science
Universitas Pembangunan Panca Budi
Jl. Jend. Gatot Subroto Km. 4,5 Sei Sikambing, 20122, Medan, Sumatera Utara, Indonesia

Abstract - *The development of data miningis inseparable from the recent developments in information technology that enables the accumulation of large amounts of data. For example, a shopping mall that records every sales transaction of goods using various POS (point of sales). Database data from these sales could reach a large storage capacity, even more being added each day, especially when the shopping center will develop into a nationwide network. The development of the internet at the moment also has a share large enough in the accumulation of data occurs. But the rapid growth of data accumulation it has created conditions that are often referred to as "data rich but information poor" because the data collected can not be used optimally for useful applications. Not infrequently the data set was left just seemed to be a "grave data". There are several techniques used in data mining which includes association, classification, and clustering. In this paper, the author will do a comparison between the performance of the technical classification methods naïve Bayes and C4.5 algorithms.*

Keywords - *Data mining, Classification , Naïve bayes, C4.5*

I. Introduction

The word "mining" means of a large number of base materials which have a long process and the literature of other disciplines such as artificial intelligence, statistics and database [1][4]. Some techniques that are often mentioned in the literature data mining are clustering, classification, association rule mining, neural networks, genetic algorithms and others. Of all the existing techniques, the distinguishes perceptions of data mining is the development of data mining techniques are applied to the database application on a large scale that turns on the application of large-scale data that can provide a lot of new challenges that could ultimately bring methodologies new [2][3]. Before data mining is becoming popular as it is today, data mining techniques can only be applied to data with small-scale only. The commencement of the application of data mining in the business world today, come to make data mining is also applied to other fields that require large-scale data analysis such as bioinformatics and defense. There are several techniques used in data mining which includes association, classification, and clustering. Association rule mining is a data mining techniques to discover the rules of associative between a combination of items. Classification is the process of finding a model or function which explain or differentiate the concept or class of data, with the aim to be able to estimate the class of an object that the label is not known. Clustering performs a set of known-based data without specific data classes. In this paper, the author will do a comparison between the performance of the classification techniques between Naïve Bayes and C4.5

II. Theories

A. *Data Mining.*

In a simple data mining is mining or the discovery of new information by looking for patterns or specific rules on very large amounts of data. Data mining is also referred to as a series of processes for adding additional value in the form of knowledge that had been unknown manually from a data set. Data mining, often referred to as knowledge discovery in databases [6]. KDD is an activity that includes the collection, use of data, historical to find regularities, patterns or relationships in large data sets.

There are some factors that define data mining:

1. Data mining is the process of digging an added value to the data collected in the past.
2. The object of data mining is that large amounts of data or complex.
3. The purpose of data mining is to find connections or patterns that may provide a useful indication

Data mining is not an entirely new field. One of the difficulties in defining the data inherited many aspects and techniques from the fields of science already established that existing first.

B. *Data Mining Technique*

Association

Association also called the market basket analysis. A typical business the problem is to analyze the sales transaction table and identify products that are often purchased together by the consumer [5][6]. Association function is often used to find a relation or correlation between the set of items. Association rule mining often calls the role in the context of the purposes of marketing strategy, catalog design, and business decision-making process. Association rule mining also has a description that is not much different that mining techniques to find an associative rule that exists between a combination of items. Traditionally, the association rules are used to find business trends by analyzing consumer transactions [9]. Important or not associative rules can be determined by two parameters, namely the support which is a percentage of a combination of items in the databaseand the confidence that the strong relationship between items in the rules associative

Classification

Classification is an act to give the group in every circumstance. Each state contains a bunch of attributes, one of which is a class attribute. This method needs to find a model that can explain the class attribute as a function of the input attribute. A decision tree is one of the most popular methods of classification because it is easy to be interpreted by humans. Here, each branching stating the conditions that must be met and the tips of trees declared class data. Decision tree algorithm C4.5 is the most famous, but the algorithm is not able to handle the data that has a large scale. The classification process is divided into two stages: learning and test [7][10]. At this stage of learning, some of the data that has been known will be fed to build the model estimates, which later in the test phase, the model that has been formed will be tested by using most other data to determine the accuracy of the model. If the accuracy is limited, then the model can be used to predict the unknown data class.

Clustering

Clustering also referred to as segmentation. This method is used to identify a natural group of a case based on an attribute group, and group data that have similar attributes. Clustering grouping based on the data without specific data classes. Even in classes that data is not yet known because, the clustering is often classified as unsupervised learning methods [8]. The principle of clustering is to maximize the similarity between members of the class and minimize similarities between classes / clusters. Clustering can be performed on the data that has several attributes that have been mapped as a multidimensional space. Most clustering algorithm to build a model through a series of repetition and stop when the models are converging or assembled.

III. Evaluation

A. *Bayes Theorema*

In probability theory and statistics, Bayes theorem is a theorem with two different interpretations. In the interpretation of Bayes theorem states how much the degree of subjective belief must rationally change when there is a new lead. In frequentist interpretation of this theory describes the representation of the inverse probability of two events. This theorem is the basis of statistical Bayes and has applications in science, engineering, economics, game theory, medicine, and law. Application of Bayes' theorem to update the trust is called Bayesian inferences. Bayes Theorem, named Rev. Thomas Bayes, describes the relationship between the conditional probability of two events A and B as follows:

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

atau

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B|A)\,P(A) + P(B|\bar{A})\,P(\bar{A})}$$

Naive Bayes algorithm is one of the algorithms contained on classification techniques. Naive Bayes classification method is a probability and statistics raised by the British scientist Thomas Bayes, which predict future opportunities based on the experience of earlier and became known as Bayes' Theorem. The naive theorem is combined with an attribute condition where it is assumed to be independent. Naive Bayes classification is assumed that there is or is not a specific characteristic of a class has nothing to do with the characteristics of other classes. Equation of Bayes' theorem is:

$$P(H|X) = \frac{P(X|A).P(H)}{P(X)}$$

Remarks :

X	= Unknown data class
H	= Hipothesis data
P(H|X)	= Hipothesisprobability
P(H)	= Prior probability
P(X|H)	= Condition probabilty
P(X)	= Probability

To explain Naive Bayes theorem, note that the classification process requires some clues to determine what classes are suitable for the samples analyzed. Therefore, the Bayes theorem above adjusted as follows:

$$P(C|F_1 \ldots F_n) = \frac{P(C)\,P(F_1 \ldots F_n|C}{P(F_1 \ldots F_n)}$$

C represents the class, while the variable F1 ... Fn represents the characteristics of instructions needed to perform the classification. The formula is explained that the chances entry of samples of certain characteristics in the class C (Posterior). It is the chance appearance of class C (before the entry of the sample, often called priors), multiplied by the likelihood of the characteristics of the sample characteristics to the class C (also called likelihood), divided by the likelihood of global characteristics of the sample characteristics (also called evidence). Therefore, the above formula can also be written simply as follows:

$$Posterior = \frac{Prior\ .\ likelihood}{evidence}$$

Evidence value is always fixed for each class on a single sample. The value of the posterior will be compared with the values of other class posteriors to determine to what class a sample would be classified. Further elaboration Bayes formula is done by describing using the product rule as follows:

$$P(C|F_1 \ldots F_n) = P(C)\,P(C|F_1 \ldots F_n|C)$$
$$= P(C)\,P(F_1|C)\,P(F_2,\ldots\ldots,F_n|C,F_1)$$
$$= P(C)\,P(F_1|C)\,P(F_2|C,F_1)\,P(F_3,\ldots\ldots,F_n|C,F_1,F_2)$$
$$= P(C)\,P(F_1|C)\,P(F_2|C,F_1)\,P(F_3|C,F_1,F_2).P(F_4,\ldots\ldots,F_n|C,F_1,F_2,F_3$$
$$= P(C)\,P(F_1|C)\,P(F_2|C,F_1)\,P(F_3|C,F_1,F_2)\ldots P(F_n|C,F_1,F_2,F_3,\ldots,F_{n-1}$$

It can be seen that the translation of these causes more and more complex factors that affect the value of the terms of probability, which is almost impossible to be analyzed one by one. As a result, the calculation becomes difficult to do. Here is used the assumption of independence is very high (naive), that each user is independent (independent) from each other. With these assumptions, then apply a similarity as follows:

$$P(P_i \mid F_j) = \frac{P(F_i \cap F_j)}{P(F_j)} = \frac{P(F_i)P(F_j)}{P(F_j)} = P(F_i)$$

Untuk $i \neq j$, sehingga :

$$P(F_i \mid C, F_j) = P(F_i \mid C)$$

From the equation above it can be concluded that the naive independence assumption makes requirements the opportunity to be simple, so the calculation becomes possible to do. Furthermore, the translation (|, ...,) 1 n P C F F can be simplified into:

$$P(C \mid F_1,\ldots,F_n) = P(C)P(F_1 \mid C)P(F_2 \mid C)P(F_3 \mid C)\ldots$$
$$= P(C)\Pi_i^n=_1 P(F_i \mid C)$$

The above equation is a model of Naive Bayes theorem which would then be used in the classification process. For classification with continuous data Density Gauss used the formula:

$$P(X_i = x_i \mid Y = y_i) = \frac{1}{\sqrt{2\pi\sigma ij}}\, e^{\frac{(x_i - \mu ij)^2}{2\sigma^2 ij}}$$

Remarks :

P	= Chance
i X	= Atribut i
i x	= Attribut value
Y	= Class
j y	= Subclass
∝	= Mean
ʃ	= Standard Deviation

The flow of Naive Bayes method is as follows:

1. Read the training data
2. Calculate the amount and probability, but if the numerical data:
   a. Find the mean and standard deviation of each parameter is numeric data.
   b. Find probabilistic value by counting the number of the corresponding data from the same category divided by the amount of data in the category.
3. Getting the values in the table mean, standard deviation and probability.

B. Algoritma C4.5

An algorithm C4.5 decision tree algorithm group. This algorithm has input in the form of training samples and samples. Training samples in the form of sample data that will be used to build a tree that has been substantiated. While samples are data field that will be used as a parameter within the classification data. C4.5 algorithms are algorithms result of the development of the algorithm ID3. Improvements from ID3 algorithm C4.5 algorithms performed in the case (Santosa, 2003):

1. Can handle with missing value
2. Can solve with continuous data
3. Pruning
4. There are rules

The measures undertaken by the C4.5 algorithm in the form of a decision tree is as follows:

1. At the beginning of the establishment of the tree will begin to create a node that symbolizes the training sample.
2. If the samples have the same class, then the node is used as a leaf node with the class label.

3. If the samples do not have the same class, then the algorithm will seek gain the highest ratio of the available attributes, as a way to select the attributes that most influence on the training sample provided. Later this attribute will be attributed to the examiner or the decision on that node. The thing to note is that when those attributes are worth continue, then attributes must be discrete first.
4. Branch for each node will be established based on the known values of attribute testing.
5. This algorithm will continue to do the same process recursively to form a decision tree for each sample in each of its parts.
6. The recursive process will stop, when one of the following conditions are met, namely:
   a. All samples were given to the node is derived from the same class.
   b. No other attributes that can be used to partition the sample further.
   c. No samples that meet test attribute.Dalam this case, a leaf is created and labeled with the class that has the largest sample (majority voting).

At this stage of the learning algorithm C4.5 has two working principles, such as:

1. Making the decision tree. The purpose of the algorithm is to construct a decision tree inducers tree data structure that can be used to predict the class of a case or a new sample that do not have class. C4.5 decision tree is doing construction with the divide and conquer method. At first only created the root node by applying a divide and conquer algorithm. It chooses to solve cases that best by calculating and comparing the gain ratio; then nodes formed at the next level, divide, and conquered algorithm will be applied again to form the leaves.
2. Making the rules (rule set). Rules are rules that form of decision trees will form a condition in the form if-then. These rules are obtained by tracking the decision tree from the root to the leaves. Each node and branching requirements will form a condition or an if, while the values contained in the leaves will form an outcome or an then.

IV. Conclusion

Each of these techniques and methods has their way. Each algorithm has advantages and disadvantages. C4.5 algorithm works by grouping several training sample data that will result in a decision tree based on the facts on the training data. While on Bayes, decision obtained based on existing experience at previous events. Bayes counts events that occur in the data into samples to determine the decision on the problems faced.

References

[1] M. J. Berry, G. Linoff, Data Mining Techniques: For Marketing, Sales, and Customer Support, New York: John Wiley & Sons, Inc, 1997.

[2] D. T. Larose, Data Mining Methods and Models, Canada: A John Wiley & Sons, Inc, 2006.

[3] A. Kumar, O. Singh, V. Rishiwal, R. K. Dwivedi, R. Kumar, “Association Rule Mining On Web Logs For Extracting Interesting Patterns Through Weka Tool,” International Journal of Advanced Technology In Engineering And Science, vol. 3, no. 1, pp. 134-140, 2015.

[4] C. D., Discovering Knowledge in Data: An Introduction to Data Mining, Canada: John Wiley & Sons, 2014.

[5] T. Krishna, D. Vasumathi, “A Study of Mining Software Engineering Data and Software Testing,” Journal of Emerging Trends in Computing and Information Sciences, vol. 2, no. 11, 2011.

[6] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. H. Witten, “The WEKA Data Mining Software: An Update,” SIGKDD Explorations, vol. 11, no. 1, pp. 10-18, 2015.

[7] D. Tomar, S. Agarwal, “A survey on Data Mining approaches for Healthcare,” International Journal of Bio-Science and Bio-Technology, vol. 5, no. 5, pp. 241-266, 2013.

[8] T. Silwattananusarn, A. D. KulthidaTuamsuk, “Data Mining and Its Applications for Knowledge Management : A Literature Review from 2007 to 2012,” International Journal of Data Mining & Knowledge Management Process, vol. 2, no. 5, 2012.

[9] S. Rajagopal, “Customer Data Clustering Using Data Mining Technique,” International Journal of Database Management Systems, vol. 3, no. 4, pp. 1-11, 2011.

[10] W. Fitriani and A. P. U. Siahaan, "Comparison Between WEKA and Salford Systemin Data Mining Software," International Journal of Mobile Computing and Application, vol. 3, no. 4, pp. 1-4, 2016.