Laporan Penelitian

Predict a Students Graduation Using Data Mining Techniques with Classification Method C4.5 Algorithm

Rudi Setiawan

2019

Abstract

The development of the technology world, especially the field of information technology that is increasingly sophisticated, has produced a lot of data stacks either. With data mining, it can be applied to dig up the value of this data set with the output of the knowledge which is not so much known manually. There is a wide range of techniques that can be used in implementing data mining, one of the data mining techniques is classification using C 4.5 classification technique. This time, the example taken to implement this C 4.5 classification is to predict a student graduation. Hopefully, this research can help in implementing data mining using C 4.5 classification technique for all types of fields, one of them in determining and estimate the graduation of the student.

1. Introduction

The process in terms of predicting a graduation from a student is important enough to analyze the graduation rate of a generation. The number that still use the manual way, is certainly very small possibility of accuracy. There are several ways that have not relied on the manual way anymore, one of them with data mining methods.

According to Turban DKK 2005, data mining is a process that uses statistical techniques, mathematics, artificial intelligence, machine learning to extract useful information and related knowledge from various large databases. Data Mining's ability to find valuable business information from a huge database, can be analogue to the precious metals mining of its source land.

1.1. DataMining

Data mining is a method used to extract predictive information hidden in a database, is a very potential technology for companies in empowering Data Warehouse (Sulianta, 2010).

Data mining is a method used to extract predictive information hidden in a database, is a very potential technology for companies in empowering Data Warehouse. Broadly, data mining can be grouped into two main categories, consisting of descriptive mining, which is the process of discovering important characteristics of data in one database. Data mining techniques that include descriptive mining are clustering, association, and da Sequential mining. Predictive, that is, the process to find the pattern of the data by using several other variables in the future. One of the techniques found in predictive.

1.2. Algorithm C 4.5

The C 4.5 algorithm is a group of Decision Tree algorithms. This algorithm has input in the form of training samples and samples. Training samples are sample data that will be used to construct a tree that has been tested in its true order. While the samples are the data fields that will be used as a parameter in the classification of data (Sunjana, 2010).

Data classification is a process that finds the same properties in a set of objects in a database and classifies them into different classes according to the classification model set. The purpose of the classification is to find the model of the training set that distinguishes the attribute into the

appropriate category or class, the model is then used to classify the attributes that the class has not known before. The classification technique is divided into techniques such as the decision tree.

The decision tree is a very powerful and well-known method of classification and prediction. The method of the decision tree transforms a very large fact into a decision tree representing a rule. The decision tree is also useful for exploring data, finding the hidden relationship between a number of candidate input variables with a target variable.

As the decision tree blends between data exploration and modelling, the decision tree is very good as a first step in the modelling process even when it is used as the final model of several other techniques. A decision tree is a structure that can be used to divide a large set of data into smaller sets of records by implementing a set of decision rules. With each series of divisions, the members of the resulting set are similar to one another (Berry and Linoff, 2004).mining is

Classification. Simple data mining can be said as a filter process or "mining" knowledge of a large amount of data. Another term for data mining is Knowledge Discovery in Database (KDD). Although data mining itself is part of the process phase of KDD.

1.3. Gain

To select the attribute as root, it is based on the highest gain value of the existing attributes. To calculate the gain used formulas as stated in the equation [C45-01] (Quinlan 1993, Craw 2005).

Gain (S, A) = Entropy (S) − ∑ ni = 1 ❙ Si ❙ ❙ S ❙ ∗ Entropy (Si) .. [C45-01]

Description:

S: Case Set

A: Attributes

N: Number of attribute partitions A

| The | : Number of cases on the to-I partition | S | : Number of cases in S

1.4. ConstructionofReferences

While the count of entropy values can be seen in the following equation [C45-02] (Quinlan 1993, Craw 2005).

Entropy (S) = ∑ ni = 1 (− pi) ∗ log2 (PI)

.. [C45-02]

Description:

S: Case Set

N: Number of S attribute partitions Pi: Proportion of Si against S.

2. Illustrations

All figures should be numbered with Arabic numerals (1,2,3,....). Every figure should have a caption. All photographs, schemas, graphs, and diagrams are to be referred to as figures. Line drawings should be good quality scans or through electronic output. Scans must be of high quality for acceptability purposes. For more guidelines and information to help you submit high-quality artwork, please visit http://www.elsevier.com/wps/find/authorsview.authors/authorartworkinstructions. Figures must be embedded into the text and not supplied separately. In MS word input the figures must be appropriately coded. Lettering and symbols should be clearly defined either in the caption or a legend provided as part of the figure. Figures should be placed at the top or bottom of a page wherever possible, as close as possible to the first reference to them in the paper.

The figure number and caption should be typed below the illustration in 8 pt and centre justified. The artwork has no text along the side of it in the main body of the text. However, if two images fit next to each other, these may be placed next to each other to save space.

From the training table is known that the number of cases 46 cases, students who belong to the graduated class amounting to 22 records, and students who belong to the class does not pass 24 records so that the Entropy obtained as follows:

Entropy (S) in 1 pi. LOG2 Pi

Entropy (S) = (-22/46. LOG2 (22/46)) + (-24/46. LOG2 (24/46)) Entropy (S) = 0.99864

Calculating the Gain value of each attribute then the highest Gain will be used as the root of the decision tree to be created, as an example on the gender attribute obtained Gain value of 0.058451942

Gain value of 0.058451942

3. Results and Conclusions

In this study the author conducted a research model of C 4.5 algorithm which is in the analysis of data from students of one university in Indonesia, where from the results of the analysis turned out when viewed from gender turned female gender is greater percentage compared to male gender for analysis results of this graduation prediction, then the gain gained from the age of students who work more to students with the category of age 1, and the gain gained from the age of students who do not work more to students with the category of age 3.

References

David Hand, Heikki Mannila and Padhraic Smyth, "Principles of Data Mining", A Bradford book The MIT Press, Cambridge, Massachusetts London England, 2001.

Budi Santosa, "Data Mining: Data utilization techniques for business purposes", Graha Ilmu, Yogyakarta, 2007.

Mabrur, A. G. & Lubis, R. 2012. Data Mining application to predict the criteria of credit customer. Int J Cancer.

Dian Oktafia and D.I. Crispina Pardede. (2010) Performance comparison algorithm of Decision Tree and Naive Bayes in the Resurrection Prediction, Jakarta

Ian H. Witten and Eibe Frank, "Data Mining: Practical machine learning tools and techniques", Morgan Kaufmann Publishers, San Francisco, 2005.

Howard Podeswa, "UML for the IT Business Analyst: A Practical Guide to Object-Oriented Requirements Gathering", Thomson Course Technology PTR, 2005.

Gennady L. Andrienko and Natalia V. Andrienko. (1999). Data Mining with C 4.5 and Interactive Cartographic Visualization, Germany.