Available online at JDR Website: http://journal.unublitar.ac.id/jdr

Journal Of Development Research, 6 (2), November 2022, Pages 137-140

DOI: https://doi.org/10.28926/jdr.v6i2.205

Improving C4.5 Algorithm Accuracy With Adaptive Boosting Method For Predicting Students in Obtaining Education Funding

Mohammad Ahmad Maidanul Abrori[(1)], Abdul Syukur[(2)], Affandy[(3)], Moch Arief Soeleman[(4)]

Universitas Dian Nuswantoro, Indonesia

E-mail: [(1)]abroria96@gmail.com , [(2)]abah.syukur01@dsn.dinus.ac.id, [(3)]affandy@dsn.dinus.ac.id, [(4)]arief22208@gmail.com

Received: 12 February 2022; Revised: 12 June 2022; Accepted: 26 September 2022

Abstract

The level of accuracy in determining the prediction of the provision of educational funding assistance is very important for the education agency. The large number of data on prospective beneficiaries can be processed into information that can be used as decision support in determining eligibility for education funding assistance. The data processing is included in the field of data mining. One method that can be applied in predicting the feasibility of receiving aid funds is classification. There are several classification algorithms, one of which is a decision tree. The famous decision tree algorithm is C4.5. The C4.5 algorithm can be applied in classifying prospective recipients of educational aid funds. This study uses datasets from student data of SMK Al Fattah Kertosono. The purpose of this study is to increase the accuracy of the C4.5 algorithm by applying adaboost in classifying students who deserve education funding and not, by comparing the results before and after applying adaboost. Validation in this study uses cross validation. While the measurement of accuracy is measured by the confusion matrix. The experimental results show that there is an increase in accuracy of 7.2%. The accuracy of the application of the C4.5 algorithm reaches 91.32%. While the accuracy of the application of the C4.5 algorithm with adaboost reached 98.55%.

Keywords: Adaboost, Algorithm C4.5, Data mining

Introduction

In a previous study conducted by Aldi Nurzahputra, and Much Aziz Muslim in 2017 with the title "Peningkatan Akurasi Pada Algoritma C4.5 Menggunakan Adaboost Untuk Meminimalkan Resiko Kredit". In this study the datasets were taken from the UCI Repository of Machine Learning Datasets. The type of dataset used is the German Credit Card dataset. The data mining software used in this research is Weka 3.6.(Hardianto, 2019)

Students at SMK Al Fattah Kertosono are divided into 3 majors. Department of Computer and Network Engineering, Automotive Light Vehicle Engineering, and Institutional Financial Accounting. From the three majors, there are several students belonging to both wealthy and underprivileged families. For students belonging to underprivileged families are those who will later be included in students who are eligible for educational funding assistance. In this study, a classification system was used to predict eligible and ineligible students to receive educational funding assistance. The method used is the C4.5 Algorithm. Where students can be considered eligible for educational funding if they have met certain criteria.

According to Han J, et al. (2012) Classification is a process used to find a model (or function) by describing and distinguishing classes of data or concepts. There are several data classification algorithms, one of which is a decision tree. The C4.5 algorithm is a development of the conventional decision tree induction algorithm, namely ID3. The algorithm which is the development of ID3 can classify data using a decision tree method which has several advantages. The advantages are that it can process numeric

(continuous) and discrete data, can handle missing attribute values, produce rules that are easy to interpret, and are the fastest among algorithms that use main memory on a computer. (Quinlan, 1993). In the application of several cases of classification techniques, this algorithm is able to produce good accuracy and performance. (Degree et al., 2012)

The dataset in this study was obtained from student data at SMK Al Fattah Kertosono. The dataset is 400 students of SMK Al Fattah Kertosono. In this research, the data mining software used is Rapidminer studio. Which will later be used to calculate the level of accuracy in the C4.5 algorithm. The purpose of this study is to increase the accuracy of the C4.5 algorithm by applying the adaptive boosting method in classifying eligible and ineligible students to get educational funding assistance by comparing the results before and after the adaptive boosting method is applied. So what distinguishes this research from the previous one is the dataset and data mining software used.

Materials and Method

Decision Tree

Decision trees are very powerful and well-known classifications and predictions. The decision tree method converts very large facts into a decision tree that presents the rules. Rules can be easily understood with natural language. And they can also be expressed in the form of a database such as Structure Query Language (SQL) to search for records in certain data. A decision tree is a structure that can be used to divide a large data set into smaller sets of decision records by applying rules. In the decision tree, each node leaves a business class label. Nodes that are not final nodes consist of roots and internal nodes which consist of attribute test conditions on some records that have different characteristics. Root nodes and internal nodes are indicated by an oval shape and leaf nodes are indicated by a rectangular shape.(Muzakir & Wulandari, 2016)

Algorithm C4.5

Many algorithms can be used in the formation of decision trees, one of which is the C4.5 algorithm. The C4.5 algorithm creates a decision tree from top to bottom, where the topmost attribute is the root, and the bottommost attribute is called the leaf. The advantage of this method is that it is effective in analyzing a large number of attributes from existing data and is easily understood by the end user (Dai & Ji, 2014).

In general, the C4.5 algorithm for building a decision tree is as follows:

1. Select attribute as root
2. Create a branch for each one score
3. Divide cases in branches
4. Repeat the process for each branch until all cases in the branch have the same class (Kamagi & Hansun, 2014).

The C4.5 algorithm is one of the algorithms that has been widely used, especially in the machine learning area, which has several improvements over the previous algorithm, ID3, in terms of pruning methods. The improvements are as follows:

1. The C4.5 algorithm calculates the gain ratio for each attribute, and attribute that has a value that is highest will be selected as the node. The use of this gain ratio improves the weakness of the ID3 which using information gain.
2. Pruning can be done at the time of tree construction (tree) or during the process tree construction is complete.
3. Able to handle continues attributes.
4. Able to handle missing data.
5. Able to generate rules from a tree. (Muzakir & Wulandari, 2016)

The selection of attributes as roots is based on the highest gain value of the existing attributes. The formula used to calculate the gain is shown in equation 1.

$$\textbf{Gain (S,A)} = \textbf{\textit{Entropy}} \textbf{ (S) } - \sum_{i=I}^{n} \frac{|S_I|}{|S|} *$$

Where S is a set of cases and A is a data attribute. The value of n is the number of attribute partitions A and |Si| is the number of cases on the i-th partition. The number of cases is indicated by |S|.

Before getting the Gain value is to find the Entropy value. Entropy is used to determine how informative an attribute input is to produce an attribute. The basic formula for entropy is shown in equation 2.

$$\textbf{Entropy } (\textbf{\textit{S}}) = \sum_{i=I}^{n} - \textbf{\textit{pi}} * \log_2 \textbf{\textit{pi}}$$

Where n is the number of partitions S and pi is the proportion of Si to S.

Adaptive Boosting

Boosting is an approach in machine learning to improve accurate predictive rules by combining many relatively weak and inaccurate rules. Adaptive boosting (adaboost) is one of several variants of the boosting algorithm (Liu et al., 2015). Adaboost is an ensemble learning that is often used in boosting algorithms.

The Adaptive Boosting algorithm was the first practical reinforcement algorithm, and remains one of the most widely used and studied, with applications in various fields. Boosting can be combined with other classifier algorithms to improve classification performance. Of course intuitively, merging multiple models will help if the models are different from each other.(Freund & Schapire, 1997)

Adaboost and its variants have been successfully applied to several fields (domains) due to their strong theoretical basis, accurate predictions, and great simplicity. The steps in the adaboost algorithm are as follows.

a. Input: A collection of research samples labeled {(x_i, y_i),….,(x_N, Y_N)}, a componen-learn algorithm, the number of turns T
b. Initialize: Weight of a training sample W_i^1 = 1/N, for all i=1,….,N
c. Do for t= 1, ...,T
   1. Use the learn algorithm component to train a classification component, h_t, on the training weight sample.
   2. Calculate the training error on h_t: $\varepsilon_t = \sum_{i=1}^{N} w_i^t, y_i \neq h_t(x_i)$
   3. Set the weight for the component classifier $h_t = \alpha_t = \frac{1}{2}$ In ($\frac{1-\varepsilon_t}{\varepsilon_t}$)
d. Update training sample weights w_i^{t+1} = $\frac{w_i^t \exp\{-\alpha_t\ y_t(x_i)\}}{C_t}, i = 1, \ldots N$ C_t is a normalization constant.
e. Output = f(x) = sign ($\sum_{t=1}^{T} \alpha_t h_t(x)$).

Data Mining

Data mining is a process that uses statistical, mathematical, artificial, intelligence, and machine learning techniques to extract and identify useful information and related knowledge from large databases. (Mustafa et al., 2018)

Results and Discussion

This research uses Rapidminer Studio 7.3.001 software. RapidMiner is a data science software platform developed by the company of the same name, which provides a unified environment for machine learning, deep learning, text

Table 1. Attributes of the Al Fattah Kertosono Vocational High School Student Dataset

No	Attribute Name	Type
1	BSM Card	Qualitative
2	Parent's Income each Month	Numerik
3	KIP	Qualitative
4	Orphans	Qualitative

Table 2. Accuracy Results of C4.5 . Algorithm

accuracy: 91.32%

	true worth it	true not feasible	class precision
pred. worth it	106	22	82.81%
pred. not feasible	3	157	98.12%
class recall	97.25%	87.71%	

Table 3. Accuracy Results with the Adaboost . Method

accuracy: 98.55% +/- 1.59% (mikro: 98.54%)

	true worth it	true not feasible	class precision
pred. worth it	148	2	98.67%
pred. not feasible	4	258	98.47%
class recall	97.37%	99.23%	

mining and predictive analytics. It is used for business and commercial applications as well as for research, education, training, rapid prototyping, and application development and supports all steps of the machine learning process including data preparation, result visualization, validation and optimization. RapidMiner is developed with an open core model (Nofitri & Irawati, 2019).

The process of testing the system using the student dataset of SMK Al Fattah Kertosono which consists of 4 attributes to predict students who are considered eligible to receive educational funding assistance with a total of 400 students. The classes used are feasible and not feasible. The following attributes used in the dataset are shown in Table 1. Then after the calculation using the C4.5 algorithm with rapidminer software produces an accuracy rate of 91.32%. The results of calculations with the C4.5 algorithm can be seen in the Table 2.

After finding the level of accuracy in the C4.5 algorithm, the next step is to increase the accuracy by using the adaboost method. Based on the test results, it was found that the accuracy level using the adaboost method was 98.55%. The results can be seen in the Table 3. From the data processing that has been done using the adaptive boosting method, it is proven that it can increase the accuracy of the C4.5 algorithm. The data used can be classified properly into feasible and inappropriate classifications.

Conclusion

Model testing is done by taking a dataset from student data of SMK Al Fattah Kertosono which consists of 4 attributes and a total of 400 student data. To predict students who are eligible for educational funding assistance. Data processing using the C4.5 Algorithm produces an accuracy rate of 91.32%. While the application of the C4.5 algorithm with the addition of adaboost obtained an accuracy of 98.55%. The results of this study indicate that the application of adaboost on the C4.5 Algorithm can increase the accuracy by 7.2%.

References

Dai, W., & Ji, W. (2014). A mapreduce implementation of C4.5 decision tree algorithm. *International Journal of Database Theory and Application*, *7*(1), 49–60. https://doi.org/10.14257/ijdta.2014.7.1.05

Degree, M. M., Science, C., & Lecture, A. C. (2012). *Data Mining : Concepts and*. *05*, 703. https://scholar.google.ru/scholar?hl=ru&as_sdt=0%2C5&q=Data+Mining%3A+The+Textbook&btnG=

Freund, Y., & Schapire, R. E. (1997). A Decision -Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences*, *55*(1), 119–139. https://doi.org/10.1006/jcss.1997.1504

Hardianto, A. M. dk. (2019). Fakultas Teknik – Universitas Muria Kudus. *Prosiding SNATIF Ke-6 Tahun 2019*, *2007*, 96–101.

Kamagi, D. H., & Hansun, S. (2014). Implementasi Data Mining dengan Algoritma C4.5 untuk Memprediksi Tingkat Kelulusan Mahasiswa. *Jurnal ULTIMATICS*, *6*(1), 15–20. https://doi.org/10.31937/ti.v6i1.327

Liu, H., Tian, H. Q., Li, Y. F., & Zhang, L. (2015). Comparison of four Adaboost algorithm based artificial neural networks in wind speed predictions. *Energy Conversion and Management*, *92*, 67–81. https://doi.org/10.1016/j.enconman.2014.12.053

Mustafa, M. S., Ramadhan, M. R., & Thenata, A. P. (2018). Implementasi Data Mining untuk Evaluasi Kinerja Akademik Mahasiswa Menggunakan Algoritma Naive Bayes Classifier. *Creative Information Technology Journal*, *4*(2), 151. https://doi.org/10.24076/citec.2017v4i2.106

Muzakir, A., & Wulandari, R. A. (2016). Model Data Mining sebagai Prediksi Penyakit Hipertensi Kehamilan dengan Teknik Decision Tree. *Scientific Journal of Informatics*, *3*(1), 19–26. https://doi.org/10.15294/sji.v3i1.4610

Nofitri, R., & Irawati, N. (2019). Analisis Data Hasil Keuntungan Menggunakan Software Rapidminer. *JURTEKSI (Jurnal Teknologi Dan Sistem Informasi)*, *5*(2), 199–204. https://doi.org/10.33330/jurteksi.v5i2.365

Quinlan, J. R. (1993). J. Ross Quinlan_C4.5_ Programs for Machine Learning.pdf. In *Morgan Kaufmann* (Vol. 5, Issue 3, p. 302). http://www.springerlink.com/index/10.1007/BF00993309

