Proceeding of the 2nd International Seminar and Conference on Global Issues (ISCoGI) 2016: European and Asian in the Age of Globalization: Cooperation and Challenge
Wahid Hasyim University, Semarang, Indonesia. Nov. 25-26, 2016
Paper No. A.3

Merger C4.5 Algorithm and Adaboost for Determining the Department IPA Students Graduation in SMA Islam Sultan Fatah Wedung Demak

Mochamad Subchan Mauludin
Department of Informatics, University of Wahid Hasyim Semarang
Jl MenorehTengah X / 22 Sampangan, Semarang, Indonesia
aan.subhan18@gmail.com

Laily Hermawanti
Information Engineering Department, University of Sultan Fatah
Jl. Diponegoro 1A Jogoloyo, Demak, Indonesia

Abstrak-Algoritma C4.5 merupakan salah satu algoritma yang diusulkan oleh para peneliti data mining di bidang pendidikan misalnya menentukan kelulusan siswa. Sekolah adalah tempat pendidikan bagi anak. Macam-macam sekolah adalah Taman Kanak-kanak, Sekolah Dasar, Sekolah Menengah Pertama, Sekolah Menengah Atas, Perguruan Tinggi dan lain-lain. Di Sekolah Menengah Atas biasanya berhasil tidaknya Ujian Nasional dan Ujian Sekolah suatu sekolah ditentukan oleh nilai-nilai siswa di sekolah. Jika siswa di sekolah menghasilkan nilai Ujian Nasional dan Ujian Sekolah yang baik berarti siswa di sekolah tersebut banyak yang lulus. Maka dari itu, kelulusan siswa perlu ditentukan. Algoritma yang digunakan dalam penelitian ini adalah penggabungan algoritma C4.5 dan *Adaboost* untuk meningkatkan akurasi dalam menentukan kelulusan siswa. Penelitian ini menggunakan *dataset* kelulusan siswa jurusan IPA SMA Islam Sultan Fatah Wedung Demak sejumlah 230 data. Atribut-atribut penelitian ini terdiri dari nomor peserta, nama peserta, nilai Bahasa Indonesia, nilai Bahasa Inggris, nilai Matematika, nilai Fisika, nilai Kimia, nilai Biologi, dan Keterangan. Penggabungan algoritma C4.5 dan *Adaboost* menggunakan *dataset* kelulusan siswa SMA Islam Sultan Fatah jurusan IPA menghasilkan akurasi sebesar 99.57% +/- 1.30%. Akurasi penggabungan algoritma C4.5 dan *Adaboost* lebih tinggi dari pada penelitian sebelumnya dalam menentukan kelulusan siswa.

Kata kunci: algoritma C4.5, *Adaboost*, siswa, nilai

Abstract-C4.5 algorithm is one of the proposed algorithms by data mining researchers in the field of education for example, determine students' graduation. School is a place of education for children. Various school is kindergarten, elementary school, junior high, high school, universities and others. In high school are usually the success of the National Examination and the Examination Schools a school is determined by the values of students in school. If the students in the school resulted in the National Examination and the Examination Schools, which either means that many students in the school who graduate. Therefore, the students' graduation needs to be determined. The algorithms used in this study is the incorporation of C4.5 and Adaboost algorithm to improve the accuracy in determining students' graduation. This study uses the dataset graduate students majoring in science SMA Islam Sultan Fatah Wedung of Demak some 230 data. Attributes this study consisted of a number of participants, participant names, the value of Indonesian, English value, the value of Mathematics, Physics values, values Chemistry, Biology value, and description. Merger C4.5 algorithms and Adaboost using datasets graduation of students SMA Islam Sultan Fatah produces an accuracy of +/- 99.57% 1:30%. Accuracy merger C4.5 and Adaboost algorithm higher than in previous studies to determine students' graduation.

Keywords*:* Algorithm C4.5, Adaboost,, Students, Grades

1. Introduction

Education is learning knowledge, skills and habits of a group of people who passed down from one generation to the next through teaching, training and research. Although education is compulsory in the majority of the place up to some age, the form of education with present at their schools often was not carried out and a small number of parents to choose home schooling education, e-learning or similar to their children.

School is the place education for a child. Types of school is kindergarten, primary school, junior high school, senior high school, college and others.Senior high school it usually works whereabouts of national examination and the examination a school determined by values students in a school.If students in a school produce values national examination and the good schools means students in a school was widely who passed.Therefore, it is necessary to determine school graduation.

Data mining are the activities do extraction to get the information important and not previously known. Data mining divided into several groups on the basis of work will be is a prediction, classifications, clustering, estimation, and association (Larose, 2005). Algorithms used in the classification of data mining are Naïve Bayes, K-Nearest Neighbor, C4.5 and others. Classifications in research data mining be applied in the health sector, education, sales, and others (Larose, 2005).

Research conducted by Vikas Chirumamilla, Bhagya honey, Sasidhar Velpula, and Indira Sunkara, using atribut-atribut consisting of CGPA, backlogs, aptitude, technical, articulate, core skills level, and achievements. This research algorithm C4.5 and Naive Bayes An algorithm C4.5 produce accuracy of 88.89 % while algorithms Naive Bayes produce accuracy of 77.78 % (Chirumamilla et al., 2014).

To research this will use the merger algorithm the merger C4.5 and Adaboost algorithm using data set graduation science class high school Islamic Sultan Fatah of Wedung Demak expected its accuracy can increase of previous studies.

2. Research Method

This research using a process cross-standard industry-data mining (crisp-dm with stages research covering understanding business, understanding data, data processing, modeling and evaluation (Larose, 2005).

3. Results and Discussion

Stage Understanding Busines

Stage understanding the research business this take to implement the merger algorithm the merger algorithm C4.5 and Adaboost to increase the accuracy in determining school graduation .

Stage Understanding Data

Stage understanding lab data it uses dataset graduation science class high school islamic sultan fatah of wedung demak a number of 230 data. Atribut-atribut this research consisting of number participants, the list of participants, value indonesian language, value english, math scores, value physics, chemical value, value biology, and a statement.

Stage Data Processing

Processing techniques preliminary data (data pre-processing) used in this research is (Han and Kamber, 2006).

1. Data cleaning can be used for data that missing value. Because found the data have missed not unfilled (missing value) in data. Preliminary data processing done to fill value missing value to the work replace missing value done.
2. Data reduction used to produce data set the volume has smaller. One strategy data reduction used in this research is attribute a subset selection. Attribute a subset selection used to reduce the data size set by removing atribut-atribut irrelevant or redudant.

Stage Modeling

The model used for this stage using algorithms C4.5 the merger and Adaboost .

Algorithm C4.5

Algorithm C4.5 is one of the branches of an algorithm decision tree .In addition, algorithm C4.5 do classifications by dividing data to assemblage of parts smaller an examination of attribute at sample. Tree decision divide next class to be the roots of trees more specific (Wu and Kumar, 2009). To choose attribute with roots, based on value the gain the highest of atribut-tribut that is. To count the gain used formula as follows (Han and Kamber, 2006).

$$Gain(S,A) = Entropy\ (S) - \sum_{i=1}^{n} \frac{|Si|}{|S|} * Entropy\ (S) \qquad (1)$$

Information:

S : the set of cases

A : attribute

N : the sum of partition attribute A

$|Si|$: the number of cases in partition

$|S|$: the number of cases in S

So that will be gained by the gain of the value of an attribute that most highest. The gain is one of atribute selection measure that is used to select test atribute every node on tree. An attribute with information highest gain chosen as test an attribute of a node.

In the meantime, the calculation of the value of entropy can be seen on similarities (Han and Kamber, 2006):

$$Entropy\ (S) = \sum_{i=1}^{n} -pi * \log_2 pi \qquad (2)$$

Information:

S : the set of cases

A : attribute

N : the sum of partition S

Pi : the proportion of S

In general algorithm C4.5 to build tree decision is as follows (Han and Kamber, 2006).

a. select attributes as the root of

b. for branches to each value

c. to cases in the branch

d. repeat the process of each branch until all cases in branches have the same class.

To choose attributes as roots, based on value the gain the highest of atribut-atribut that is. To count the gain used formula as mentioned in the formula (Han and Kamber, 2006),

$$Gain(S,A) = Entropy\ (S) - \sum_{i=1}^{n} \frac{|Si|}{|S|} * Entropy\ (S) \qquad (3)$$

By:

S : the set of cases

A : attribute

N : the sum of partition attribute A

$|Si|$: the number of cases in partition to i

$|S|$: the number of cases in S

While the calculation of value entropy can be seen in the formula 2 following (Han and Kamber, 2006).

$$Entropy\ (S) = \sum_{i=1}^{n} -pi * \log_2 pi \qquad (4)$$

S : the set of cases
A : features
N : the sum of partition S
Pi : the proportion of S

Algorithm Adaboost

Algorithms Adaboost is to build the power of classification as a combination of linear, adaboost acronym adaptive boosting developed by freund and schapire (Wu and Kumar, 2009). Algorithms Adaboost steps (Wu and Kumar , 2009).

Data input
D = { (x_1 , y_1) , (x_2 , y_2) , ... , (X_m , Y_m) };
Learning algorithms a frail (weak learner) L;
an integer T said the number of iteration .
The process of:
Initialization heavy distribution :
D_1 (i) = , $\frac{1}{m}$ for all i = 1 , ... , m
For t = 1 , ... , T
Train learner basic/weak h_1 of D use distribution D_1
H_1 = L (D ,D_t)

Calculate a h_t: $\varepsilon_t = Pr_x \sim D_{t,y} I[h_t(X_i) \neq y_i]$
If ε_t>0.5 then break
if t & gt; 0.5 then break
Weight of h_t:$\frac{1}{2}$In($\frac{1-\varepsilon t}{\varepsilon t}$)

$$D_{t+1}(i) = \frac{D_t(i)}{Z_t} \text{x} \begin{cases} exp(-\propto_t)\ if\ h_t(x_i) = y_i \\ exp(\propto_t)\ if\ h_t(i) \neq y_i \end{cases} \qquad (5)$$

Update distribution, where, where Z_t is a factor normalization that activates D_{t+1} as distribution :

$$H(x) = sign(\sum_{t=1}^{T} \propto_t h_t(x)) \qquad (6)$$

End,output
Strong clasification

$$H(x) = sign(\sum_{t=1}^{T} \propto_t h_t(x)) \qquad (7)$$

Error measured by taking into account distribution D_t in which algorithm learner weak trained.In practice, algorithm learner weak may be a the algorithms that can use weight D_t in samples of training. Or, if this is not allow, part of sample training can di-resampling according to D_t, and the result of resampling who do unweighted can be used to train algorithm weak learner.

Stage Evaluation

Evaluation and validation to research was used in the confusion the matrix.The merger algorithm C4.5 and Adaboost use dataset students SMA Islam Sultan Fatah produce accuracy of 99.57 % + / - 1.30 %. Accuracy to research this who uses the merger algorithm C4.5 and Aadaboost higher than previous studies.

4. Conclusion

Accuracy to research this who uses the merger algorithm C4.5 and Adaboost higher than in previous studies in determining completion rates . Research keep that algorithm C4.5 and Adaboost is one of the algorithms that right in diagnosing determine completion rates.

References

Larose D.T., (2005), Discovering Knowledge in Data: An Introduction to Data Mining, United States of America: John Wiley & Sons, Inc.

Han J., and Kamber M, (2006). Data Mining Concept and Techniques, 2nd ed, United States of America: Diane Cerra.

Chirumamilla V, Bhagya ST, Velpula S, and Sunkara I., (2014). A Novel approach to predict Student Placement Chance with Decision Tree Induction, International journal of systems and technologies.

Wu X, and Kumar V., (2009). The Top Ten Algorithms in Data Mining, Taylor & Francis Group, LLC