Journal of Soft Computing Exploration

The Implementation of Z-Score Normalization and Boosting Techniques to Increase Accuracy of C4.5 Algorithm in Diagnosing Chronic Kidney Disease

Hestu Aji Prihanditya[1], Alamsyah[2]
[1,2]Computer Science Departement, Faculty of Mathematics and Natural Sciences,
Universitas Negeri Semarang, Semarang, Indonesia

Article Info

Article history:

Received Jul 28, 2020
Revised Aug 1, 2020
Accepted Aug 26, 2020

Keywords:

C4.5 Algorithm;
Zscore;
Boosting;
Data Mining
Chronic Kidney Disease

ABSTRACT

In the health sector, data mining can be used as a recommendation to predict a disease from the collection of patient medical record data or health data. One of the techniques can be applied is classification with the C4.5 algorithm. The increasing accuracy can be conducted in data transformation using zscore normalization method. In addition, the implementation of the ensemble method can also improve accuracy of C4.5 algorithm, namely boosting or adaboost. The purpose of this study was determinin the implementation of zscore normalization in the pre-processing and adaboost stages of the C4.5 algorithm and determing the accuracy of the C4.5 algorithm after applying zscore and adaboost normalization in diagnosing chronic kidney disease. In this study, the mining process used k-fold cross validation with the default value k = 10. The implementation of the C4.5 algorithm obtained an accuracy of 96% while the accuracy of the C4.5 algorithm with the zscore normalization method obtained an accuracy of 96.75%. The highest accuracy was obtained from the addition of the boosting method to the C4.5 algorithm and zscore normalization obtained the accuracy of 97.25%. The increasing accuracy was obtained of 1.25% which compared to the accuracy C4.5 algorithm.

Corresponding Author:

N Hestu Aji Prihanditya,
Computer Science Departement, Faculty of Mathematics and Natural Sciences,
Universitas Negeri Semarang, Semarang, Indonesia
Email: hestuuaji@gmail.com

1. INTRODUCTION

In this technological era, the development of data is growing very rapidly and large. In health sector, it saves a lot of data that can be processed and produce to information or new knowledge. The processing data that can be extracted as information and knowledge from datasets is called data mining. By the existence of data mining, it is expected that it can be able to provide the knowledge as a recommendation as decision making for experts in health sector. Data mining is the process of utilizing several mathematical methods and machines that useful for identifying information from large data [1].

In data mining, there is a pre-processing stage, one of techniques is transformation. Transformation is the process of transforming data so that it can be suitable for data mining [2]. In data transformation stage, there are several methods for conducting the data transformation process such as smoothing, generalization, normalization, aggregation and attribute construction. Normalization is a process where a numeric attribute is mapped or scaled within a certain range in the data mining process [3]. Data normalization is useful to minimize data refraction in data mining because attribute values in data usually have different ranges. There are several normalization Journal of Soft Computing Exploration , Vol. 5, No. 1, May 2018 2 techniques that are often used including normalization of zscore. Zscore normalization is the normalization of data when

the range of data is not known with certainty by calculating using the average value and standard data deviation [4].

Data mining has several techniques such as estimation, prediction, classification, clustering, and association. One of the data mining techniques used to predict a decision is classification. Classification is one technique that aims to extract the model into categorical classes [5]. One type of algorithm in data mining classification is the C4.5 algorithm. C4.5 algorithm is an algorithm developed by J. Ross Quinlan from algorithm ID3 that uses the gain ratio as the separation of criteria [6].

In the health sector, data mining can be used as a recommendation as to predict disease from a collection of patient medical record data or health data. Using the classification method, the data such as age, blood pressure, urine concentration and other attributes can be used as supporting factors to make recommendations for predicting the possibility of patients who suffering from chronic kidney disease. Kidney disease is a disease which has not normal kidney function almost as much as 90% and not characterized by certain symptoms [7]. The diagnostic study mostly used a chronic kidney disease dataset obtained from the UCI repository of machine learning datasets. In the classification algorithm, accuracy explained how precisely the algorithm can classify data. Accuracy is very discussedable because if an accuracy has little value or result then it will cause a misinterpretation of classification.

The development of machine learning using the ensemble method can improve accuracy in the way of combining several classifying components. The ensemble method that can be used to improve accuracy on a classifier is bagging and boosting. Boosting is preferred because it has a tendency to increase accuracy higher than bagging. Adaboost is a very popular boosting algorithm to improve classification accuracy. The algorithm can be used in diagnosing a disease, one of which is chronic kidney disease.

So many researchers have conducted research on the C4.5 algorithm specifically using chronic kidney disease datasets from the UCI repository of machine learning datasets and research gaps have been found from these studies. In a research conducted by Sujatha & Ezhilmaran [8], the accuracy of the C4.5 algorithm was 97% for the chronic kidney disease dataset. The preprocessing method was used to replace missing value, then for the data separation using the k-fold cross validation method with a value of k = 2,3,4,5,6. Another research was conducted by Celik et al., [9], in this study obtained in the accuracy result of 96.7% for the C4.5 algorithm.

The purpose of this research was determining the implementation of zscore normalization in the pre-processing stages and adaboost to the C4.5 algorithm and determining the accuracy of the C4.5 algorithm after implementing zscore and adaboost normalization in diagnosing chronic kidney disease.

2. METHOD

2.1 Data Mining

Data mining is a process of exploration of data that has a large number of records and has been taken a certain pattern [10]. The systematically the data mining process has 3 main steps, namely:

2.1.1 Preprocessing

The preprocessing of data consists of cleaning data, data transformation, dimension reduction, selection of feature subset and so on.

2.1.2 Build models and evaluate validity

Building a model and validation means conducting an analysis of the formed model and choosing the model that has the best performance, at this stage of research used the classification method. Classification is a method in data mining that is used to predict class labels in data [11].

2.1.3 Implementation

Implementation means applying a model to new data to form certain knowledge. The data mining process can be seen in Figure 1.

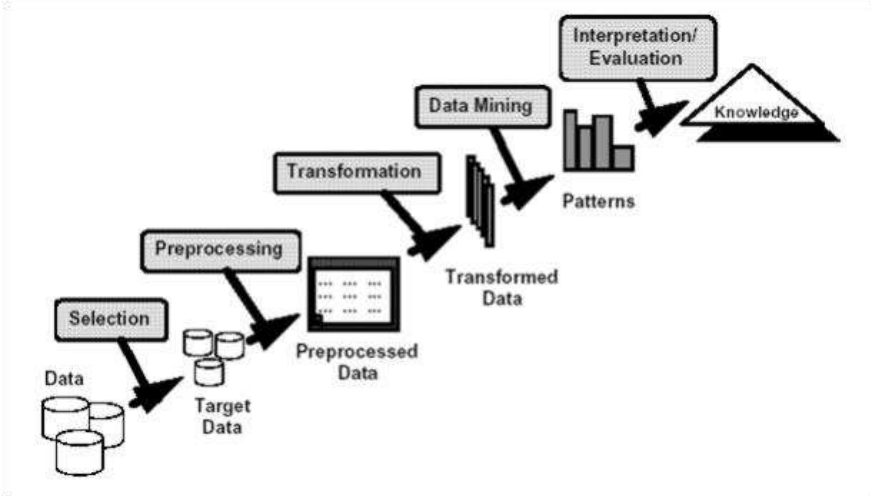

Figure 1. Data Mining Process

2.2 Z-Score Normalization

The normalization is a process in preprocessing stage by decompositing data of numeric attributes which can convert values in data into a certain range [12]. There are several methods that are usually applied in data normalization, including: min- max normalization, z-score normalization and normalization by decimal scaling. Z- score normalization maps a vi value from attribute E to v 'into a range that was previously unknown, can be seen in Equation 1.

$$v' = \frac{v_i - E_i}{std(E)} \tag{1}$$

Description:
v' = result of normalization value.
v = the value to be normalized in attribute
E_i = the mean value of attribute
std(E) = standard deviation attribute E.

2.3 C4.5 Algorithm

Decision tree is a classification method that converts data into a tree as a rule representation [13]. In the decision tree there is a very famous classification algorithm, namely C4.5 algorithm. Algorithms is a way to solve problems using certain instructions to produce the output [14]. The C4.5 algorithm is an algorithm introduced by Quinlan which is an improvement from the ID3 algorithm. In ID3, the induction decision tree can only be performed on categorical features (nominal/ ordinal), while numeric types (internal / ratio) cannot be used. The C4.5 algorithm is also defined as an algorithm that uses gain ratio as a split attribute selection [15].

The stages form a decision tree using C4.5 algorithm: Prepare the training data from existing data recap and have been grouped in certain classes. Next, determine the root of the tree by calculating the highest gain value for each attribute. For conducting that step, calculate the entropy index first using Equation 2 below

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

Description:
S = Set of Case
n = Number of Partitions S
pi = The proportion of Si to S
Where *log2pi* can be calculated using Equation 3 below.

$$log(X) = \frac{\ln(X)}{\ln(2)} \tag{3}$$

For calculating the gain can use Equation 4 below.

$$Gain(S,A) = Entropy(S) - \sum_{i=1}^{n} \frac{|\ _i|}{|S|} * Entropy(S_i) \quad (4)$$

The criteria for choosing the C4.5 feature is the gain ratio, which can be formulated by the following Equation 5.

$$GainRatio(A) = \frac{\text{Gain}(A)}{\text{SplitEntropy}(A)} \quad (5)$$

For calculating Split Entropy can used Equation 6 as follows

$$SplitEntropy_A(S) = -\sum_{i+1}^{n} \frac{|S_i|}{|S|} * log_2 \frac{|S_i|}{|S|} \quad (6)$$

Description:
S = Set of Case
A = Attributs
n = The number of A Attributr Partition
$|S_i|$ = The Case Number in i Partition
$|S|$ = The Case Number

Repeat the steps of determining root by calculating the highest gain value until all records are filled. The process of partitioning the tree will stops when: (1) There is no attribute in the partition which partitioned again. (2) There is no record in an empty branch.

The C4.5 algorithm has several weaknesses, including: (1) By a value of 0 or a value close to 0 it does not have any contribution to the classification and makes the tree size more complex. (2) Data that has noise tends to result in overfitting [16].

2.4 Adaptive Boosting (Adaboost)

Adaboost or Adaptive Boosting is a machine learning algorithm by Yoav Freud and Robert Schapire which is often used to improve the performance of certain algorithms from a set of strong or weak classifiers [17]. Adaboost can be combined with other algorithm classifiers to improve classification performance.

The method of the adaboost algorithm is as follows:

1. Initialize: weight of training sample$w_n^1 = 1/N$, which is n=1,...,N.
2. Do for t= 1, ...,T
3. Use component learn algorithm to train a classification component, h_t , to sample weight of training.
4. Training by minimizing error training or error function in $h_t: \varepsilon_t = \sum_{i=1}^{N} w_i^t , y_i \neq h_t(x_i)$
5. Update the weight sample of training $w_i^{t+1} = \frac{w_i^t \exp\{-\alpha_t y_i h_t(x_i)\}}{c_t}, i = 1, \ldots, N$ C_t is the constant normalization.

Output: $f(x) = sign\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right)$ to make prediction using the last model. The stages of work flow can be seen in Figure 2.

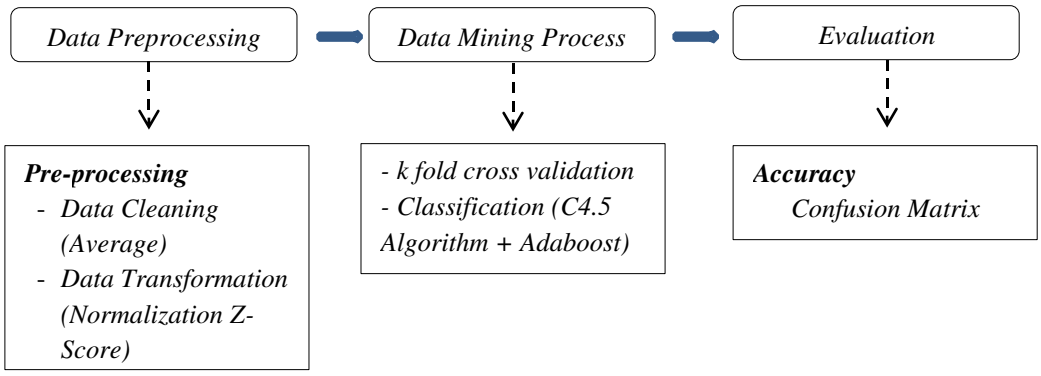

Figure 2. The stage of workflow the Implementation C4.5 using Adaboost and Z-Score Normalization

3. RESULTS AND DISCUSSIONS

This study measures the accuracy of the C4.5 algorithm with the implementing of zscore normalization and adaboost using MatLab software. The data used in this study is the chronic kidney disease dataset obtained from the UCI repository of machine learning. The chronic kidney disease dataset consists of 400 data records which is divided into 24 attributes and 1 class attribute. The attributes consist of 11 numeric attributes and 14 nominal attributes.

This dataset had .arff format the it required to rewrite in the same form stored with the extension .xlsx. Before the classification process was conducted using the C4.5 algorithm, the data must be prepared in advance so it can be ready to be processed or well known as data pre-processing.

3.1 Handling Missing Value (Cleaning Data)

Cleaning data is a process of eliminating noise and handling data that has a missing value in a record. Data which has a missing value is usually symbolized by the question mark "?" in the data record. Therefore, it needs to be given the treatment or handling of missing value, by applying the average technique. The sample data consist missing values is shown in Table 1.

Table. 1 The Data with Missing Value

Sg	Al	Su	Bgr	Bu
?	?	?	98	86
1,01	3	2	157	90
1,015	3	0	76	162
1,015	2	0	99	46
?	?	?	114	87
1,025	0	3	263	27
1,025	1	0	100	31
1,025	2	0	173	148

The average calculation to replace the missing value data using average model as follows.

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \quad (07)$$

a. The value to replace the missing value of Sg attribute:
$\bar{x}(Sg) = \frac{\sum_{1}^{353} x(Sg)}{353} = \frac{359,145}{353} = 1,01$

b. The value to replace the missing value of Al attribute:
$\bar{x}(Al) = \frac{\sum_{1}^{354} x(Al)}{354} = \frac{360}{354} = 1,01$

c. The value to replace the missing value of Su attribute:
$\bar{x}(Su) = \frac{\sum_{1}^{351} x(Su)}{351} = \frac{158}{351} = 0,45 = 0$

The chronic kidney disease datasets with the handling of missing values is presented in Table 2.

Table 2. The Data After Handling Missing Value

Sg	Al	Su	Bgr	Bu
1,01	**1,01**	**0**	98	86
1,01	3	2	157	90
1,015	3	0	76	162
1,015	2	0	99	46
1,01	**1,01**	**0**	114	87
1,025	0	3	263	27
1,015	1	0	100	31
1,015	2	0	173	148

3.2 Data Transformation Stage

The data transformation stage was conducted to normalize chronic kidney disease dataset using zscore normalization. It was transforming the numerical type data into patterns that coulb be identified to know the range values between dataset attributes so that data became simpler and had an even range of values between numeric attributes. The results of zscore normalization calculations can be seen in Table 3.

Table 3. The Implementation Result of Zscore Normalization

Sg	Al	Su	Bgr	Bu
1,01	1,01	0	**-0,7**	**0,6**
1,01	3	2	**0,1**	**0,7**
1,015	3	0	**-1**	**2,1**
1,015	2	0	**-0,7**	**-0,2**
1,01	1,01	0	**-0,5**	**0,6**
1,025	0	3	**1,5**	**-0,6**
1,015	1	0	**-0,6**	**-0,5**
1,015	2	0	**0,3**	**1,8**

3.2 Data Mining Stage

In this research the data distribution was conducted automatically by using k-fold cross validation with the default value k = 10. The testing result of the C4.5 algorithm by using the k-fold cross validation in the chronic kidney disease dataset can be seen in Table 4.

Tabel 4. The Accuracy Result of C4.5

Algorithm	Accuracy Result
C4.5	96%

a. The Implementation C4.5 Algorithm with Zscore Normalization
The implementation of classification was conducted by applying the C4.5 algorithm and zscore normalization method. The testing result of the C4.5 algorithm and zscore normalization as a pre-processing process by using the k-fold cross validation in chronic kidney disease datasets can be seen in Table 5.

Table 5. The Accuracy Result of C4.5 Using Zscore Normalization

Algorithm	Accuracy Result
C4.5 + Zscore Normalization	96,75%

b. The Implementation C4.5 Algorithm and Adaboost with Normalization Zscore in Pre-Processing
This classification conducted by implementing the C4.5 algorithm and adaboost as ensemble learning. Before the mining process was conducted, the pre-processing was processed using zscore normalization. In this adaboost, the training set used for each classifier was selected based on the performance of the previous classifier. The distribution data was conducted by using k-fold cross validation with the default value k = 10. After the data was divided into training and testing data, then the data was processed with the C4.5 then boosting. The accuracy result was obtained and can be seen in Table 6.

Table 6. The Accuracy Result of C4.5 using Zscore Normalization and Adaboost

Algorithm	Accuracy Result
C4.5 + Zscore Normalization + Adaboost	97,25%

4. CONCLUSION

The implementation of classification was conducted by applying the C4.5 algorithm which obtained an accuracy of 96%. While the accuracy results by applying the C4.5 algorithm and zscore normalization obtained an accuracy of 96.75%. Then the best accuracy of the C4.5 algorithm was obtained by applying the zscore normalization method with boosting obtained an accuracy of 97.25%. Its accuracy result was higher and occur the increased accuracy by 1.25% compared to the accuracy results of the C4.5 algorithm.

REFERENCES

[1] Sugiharti, E. & Muslim, M.A. (2016). On-line Clustering of Lecturers Performance of Computer Science Department of Semarang State University Using K-Means Algorithm. Journal of Theoretical and Applied Information Technology, 83(1): 64-71.

[2] Tamilselvi, R., Sivasakthi, B., & Kavitha, R. (2015). An Efficient Preprocessing and Postprocessing Techniques in Data Mining. International Journal of Research in Computer Applications and Robotics,3(4): 80-85.

[3] Saranya, C., & Manikandan, G. (2013). A Study on Normalization Techniques for Privacy Preserving Data Mining. International Journal of Engineering and Technology (IJET), 5(3): 2701-2704.
[4] Goyal, H., Sandeep, Venu, Pokuri, R., Kathula, S., Battula, N. (2014). Normalization of Data in Data Mining. International Journal of Software and Web Science (IJSWS). 32-33.
[5] Han, J., Kamber, M. & Pei, J. (2011). Data Mining Concepts and Techniques, 3rd ed. USA: Morgan Kaufmann Publisher.
[6] Muslim, M.A., Herowati, A.J., Sugiharti, E., & Presetiyo, B. (2018). Application of the pessimistic pruning to increase the accuracy of C4.5 algorithm in diagnosing chronic kidney disease. Journal of Physics: Conference Series, 983(1).
[7] Bala, S. & Kumar, K. (2014). A Literature Review on Kidney Disease Prediction using Data Mining Classification Techniques. International Journal of Computer Science and Mobile Computing, 3(7): 960-967.
[8] Sujatha, R. & Ezhilmaran. (2016). Performance Analysis of Data Mining Classification Techniques for Chronic Kidney Disease. International Journal of Pharmacy & Technology, 8(2): 13032-13037.
[9] Celik, E., Atalay, M., & Kondiloglu, A. (2016). The Diagnosis and Estimate of Chronic Kidney Disease Using the Machine Learning Methods. International Journal of Intelligent Systems and Applications in Enggineering, 4(1): 27-31.
[10] Chary, N., & Rama, B. (2017). A Survey on Comparative Analysis of Decision Tree Algorithms in Data Mining. International Journal of Advanced Scientific Technologies, Engineering and Management Sciences (IJASTEMS), 3(1): 91-95.
[11] Handarko, J.L. & Alamsyah. (2015). Implementasi Fuzzy Decision Tree untuk Mendiagnosa Penyakit Hepatitis. Unnes Hournal of Mathematic, 4(2): 1-9.
[12] Mishra, A.K., Choudhary, A., & Choundhary, S. (2016). Normalization and Transformation Technique Based Efficient Privacy Preservation In Data Mining. International Journal of Modern Engineering and Research Technology, 3(2): 5- 10.
[13] 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.
[14] Sampurno, G.I., Sugiharti, E., & Alamsyah, A. (2018). Comparison of Dynamic Programming Algorithm and Greedy Algorithm on Integer Knapsack Problem in Freight Transportation. Journal of Soft Computing Exploration, 5(1): 49.
[15] 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.
[16] Muslim, M.A., Rukmana, S.H., Sugiharti, E., Prasetiyo, B., & Alimah, S. (2018). Optimization of C4.5 algorithm-based particle swarm optimization for breast cancer diagnosis. Journal of Physics: Conference Series, 983(1).
[17] Korada, N.K., Kumar, N.S.P., & Deekshitulu, Y.V.N.H. (2012). Implementation of Naïve Bayesian Classifier and Ada-Boost Algorithm Using Maize Expert System. International Journal of Information Sciences and Techniques (IJIST), 2(3): 63-75.