Optimization of the C4.5 Algorithm by Using a Genetic Algorithm for the Diagnosis of Life Expectancy for Hepatitis Patients

Margareta Ayu Riantika [1*], Riza Arifudin [1]

[1] Department of Computer Science, Faculty of Mathematics and Natural Sciences, Universitas Negeri Semarang, Semarang, Indonesia
*Corresponding author: margaretailkom@students.unnes.ac.id

ARTICLE INFO

Article history
Received 17 February 2021
Revised 18 March 2021
Accepted 12 April 2021

Keywords
Data Mining
Genetic Algorithm
Hepatitis Dataset

ABSTRACT

As technology develops rapidly, the amount of data generated experiencing rapid development, including medical data. Data can help diagnose the life expectancy of people with the disease such as hepatitis using data mining methods in the medical field. In this research, technique data mining uses a classification technique with the C4.5 algorithm and the UCI Machine Learning Repository dataset. This dataset has 19 attributes, 1 class, and 155 samples. C4.5 algorithm is optimized using the Genetic Algorithm feature selection process. This study compares the accuracy of the C4.5 algorithm before and after optimization using a Genetic Algorithm. C4.5 algorithm produces the highest accuracy of 96.23%. Meanwhile, the C4.5 algorithm, after being optimized using Genetic Algorithm, has the highest accuracy of 98.11%. The number of features selected is 15 features. Application of Genetic Algorithms in C4.5 algorithm is proven to improve the accuracy in diagnosing life expectancy of people with hepatitis as much as 1.88%.

1 Introduction

With the development of information technology today, it is beneficial for all circles of society. Information technology has become one need that is very important in everyday life. As is information technology, humans are increasingly facilitated in doing their job. In addition, information technology can be utilized in a variety of ways fields, including medicine and health (Muzakir & Wulandari, 2016). One of the branches of science that can be used to process information in the medical field is data mining. One of them is making a diagnosis to determine the patient's life expectancy hepatitis (Wibowo & Indriyawati, 2020). Hepatitis is a disease caused by fungal infections, bacteria, viruses, drugs, alcohol consumption, fat excess, or an autoimmune disease that causes inflammation of the liver and damage to liver cells (Septiani, 2014). Potential chronic hepatitis disease becomes liver cancer (liver cirrhosis). Hepatitis has been considered the fifth deadly disease globally (Oktaviani *et al.*, 2018). By properly identifying the life expectancy of people with hepatitis, doctors can do appropriate medical treatment for patients (Khomsah, 2018). Techniques in data mining that can make one of the methods used to predict a decision is technique classification. Classification technique is a data mining technique based on data attachment to sample data (Oktanisa & Supianto, 2018). In medicine and health, classification techniques can also be used to predict disease. Examples of classification techniques are trees decision (decision tree), Naïve Bayes, Neural Network, SVM, and logistic regression (Widodo & Handoyo, 2017).

The most powerful and widely used classification technique for classification and prediction is the decision tree (Perveen *et al.,* 2016). One of the algorithms developed by the decision tree is a C4.5 Algorithm. The C4.5 algorithm can predict minimum execution time and the best accuracy result (Muslim *et al.,* 2018). There are several ways to improve the accuracy results. One of them is done data preprocessing stage. Preprocessing techniques can enhance data quality and strengthen yield accuracy because data quality determines method performance prediction and usefulness of extracted knowledge (Asgarnezhad, Shekofteh, & Boroujeni, 2017).

In the data preprocessing stage, there is an attribute selection process (feature selection). The attribute selection process also plays an essential role in data mining. The attribute selection method is a crucial procedure in pattern recognition that contributes to improved classification model performance (Eid & Abraham, 2018). In medical data, the attribute selection process helps select relevant attributes so that the specified attributes can contribute to diagnosing disease and providing more accurate accuracy (Aini, Sari, & Arwan, 2018). The attribute which is not considered an important attribute (irrelevant, redundant) will be deleted. The elimination of unnecessary attributes aims to reduce high dimensional data computing workload to speed up the calculation of objective functions in the classification process. The attribute selection process is an important factor in increasing the accuracy of the classification process (Liu *et al.,* 2011). One of the algorithms that can be used for attribute selection is the Genetic Algorithm. The genetic algorithm was chosen because it can reduce data attributes. Data that initially has many attributes is reduced to several attributes with less information without reducing the data (Nugroho, Nhita, & Trantoro, 2016). This study proposes an optimization of the classification algorithm C4.5 by selecting the Genetic Algorithm method to diagnose life expectancy hepatitis disease.

2 Methods

This study uses the Genetic Algorithm method as a selection of valuable features to optimized the performance of the C4.5 algorithm. Genetic Algorithms are applied to looking for the best attributes based on the best fitness value in that generation has been determined. This study will assess the comparison of accuracy before and after applying the Genetic Algorithm to the C4.5 algorithm. Flow chart The C4.5 Algorithm with the Genetic Algorithm is shown in Figure 1.

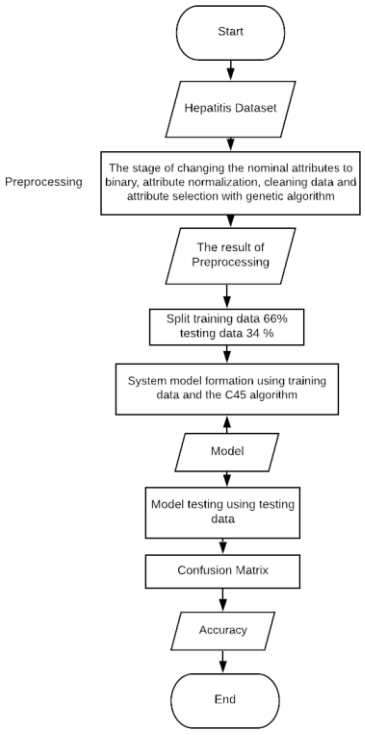

Figure 1. Flowchart C4.5 with genetic algorithm

2.1 Data Preprocessing

This study used a public dataset, namely hepatitis, available at the UCI Machine learning repository. This dataset has 20 attributes, including one attribute class and 155 instances, where six attributes are numeric, and 14 attributes are nominal. The description of the dataset can be seen in Table 1.

Table 1. Description of the hepatitis dataset

Attributes	Type	Description
Age	Numerical	Age of patient
Sex	Nominal	Sex of patient
Steroid	Nominal	Usage of steroid
Antivirals	Nominal	Usage of antiviral
Fatigue	Nominal	A feeling of constant tiredness or weakness
Malaise	Nominal	A feeling of discomfort, illness
Anorexia	Nominal	Eating disorder
Liver Big	Nominal	Enlarged liver
Liver Firm	Nominal	Hardened liver
Spleen Palpable	Nominal	Enlarged spleen
Spiders	Nominal	The symptom of spider angioma in veins
Ascites	Nominal	Abnormal buildup of fluid in the abdomen
Varices	Nominal	Abnormally dilated vessel
Bilirubin	Numerical	Concentration of Bilirubin
Alk Phosphate	Numerical	The concentration of Alkaline phosphatase
SGOT	Numerical	Concentration of SGOT
Albumin	Numerical	Concentration of Albumin
Protime	Numerical	Concentration of Prothrombin time
Histology	Nominal	Histological examination
Class	Nominal	Class of data

2.1.1 Stage of Preprocessing Nominal Attribute Data to Binary

This stage converts the attribute data with nominal types to a binary attribute with values 0 and 1.

2.1.2 Stage of Cleansing Data

In the dataset used in this current study, there are missing values. The missing values are caused by attribute data loss for various reasons, such as medical events, cost savings, anomalies, and so on. With regards to this issue, it is necessary to process the missing value data. In this study, filling in the missing values is done by replacing the missing values with the values obtained from the maximum number of frequencies in one attribute.

2.1.3 Stage of Data Normalization

Data normalization is a processing stage in which attribute data are scaled to fit within a smaller specific range, such as a range between [0–1] or [-1–0]. In this study, the normalization of intermediate xi data has used a range [0–1] where max (x) is the maximum value of attribute data, and min (x) is the minimum value of attribute data as in Equation 1.

$$\mathrm{x}i = \frac{x - min\,(x)}{max(x) - min\,(x)} \tag{1}$$

2.1.4 Stage of Data Normalization

The stages of the Genetic Algorithm in attribute selection are as follows.

1. Individual representation to know what type of data will be examined for further processing into the coding scheme. Scheme This coding will later represent each chromosome will be researched.

2. Evaluate the fitness value of each particle in the population.
3. Selection to select chromosomes as parents based on candidates the fitness value of each chromosome. Chromosome which has a value of fitness that well will be maintained.
4. Recombination. After the chromosomes are selected as parents, the recombination is often called the cross-over, aims to produce new (offspring) chromosomes with fitness values better than the previous chromosome.
5. Mutation process changes the value of genes on a chromosome.
6. Update the old chromosome value with the fitness value of the new chromosome.

If the best fitness value or maximum generation is met, stop the iteration. Otherwise, go back to step 2.

2.2 C4.5 Algorithm

C4.5 algorithm refinement of the IDE3 algorithm developed by Quinlan Ross in 1986 (Kathija, Nisha, & Sathik, 2017). In the C4.5 algorithm, attribute selection is made using Gain, Ratio, by searching Entropy value. The stages of the C4.5 algorithm in classifying the dataset are as follows:

1. Prepare training data that have been grouped into certain classes. Training data consist of 66% of the entire dataset.
2. Calculate the entropy value using Equation 3.

$$\text{entropy(S)} = \sum_{\text{i=1}}^{\text{n}} -\text{pi} * \text{log2pi} \quad (2)$$

Information:

S : Case Set
n : Number of classification classes
p_i : Proportion sample (class comparison) / opportunity for class i

Where $log_2 pi$ can be calculated by Equation 3.

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

3. Calculate Gain and splitEntropy using Equation 4 & 5.

$$gain(S, A) = Entropy(S) - \sum_{i=1}^{n} -\frac{|Si|}{|S|} * entropy(Si) \quad (4)$$

$$splitentropy(S, A) = -\sum_{i+1}^{n} -\frac{|Si|}{|S|} * log2\frac{|Si|}{|S|} \quad (5)$$

Information:

S : Case Set
A : Attribute
n : Number of classification classes
$|Si|$: The number of case on the i-partition
$|S|$: The number of case on the case set

4. Calculate *GainRatio* using Equation 6

$$gainratio(A) = \frac{gain(A)}{splitentropy(A)} \quad (6)$$

5. Repeat step 2 until all records are partitioned
6. The decision tree partitioning will stop when:
    a. There are no attributes partitioned
    b. There is no record in the empty branch

3 Results and Discussion

This study uses a hepatitis dataset, a public dataset of UCI Machine Learning Repository, where this dataset has 20 attributes with 1 class, 19 attributes, and 155 samples of data. This research was done using the PHP programming language with a framework Laravel. The results of this study are classification accuracy in diagnosing the life expectancy of people with hepatitis.

3.1 Results

This research is divided into two applications, and the first is testing classification algorithm C4.5. The C4.5 algorithm will process the hepatitis dataset based on the gain ratio value. The accuracy results of the C4.5 algorithm can be seen in Table 2.

Table 2. Accuracy results of C4.5

Execution	Accuracy
1	92.45%
2	94.34%
3	92.45%
4	96.23%
5	94.34%
6	92.45%
7	96.23%
8	94.34%
9	94.34%
10	94.34%

At the C4.5 algorithm classification stage, without the feature selection process, the highest accuracy is 96.23%. The second application is optimizing the C4.5 algorithm with Genetic Algorithms as feature selection. The parameters used as the initialization of the Genetic Algorithm process are as follows:

a. Maximum Generation : 22
b. Probability of Cross-over : 0.61
c. Probability of Mutation : 0.12
d. Number of Genes : 19
e. Total Population : 150

As for the results of that accuracy obtained are shown in Table 3.

Table 3. Result of C4.5 accuracy with Genetic Algorithm

Execution	Feature selected	Accuracy
1	14	98.11%
2	15	94.34%
3	14	98.11%
4	14	94.34%
5	17	96.23%
6	16	94.34%
7	16	96.23%
8	17	92.45%
9	14	96.23%
10	15	94.34%

The highest accuracy result of the C4.5 algorithm after applying the Genetic Algorithm is 98.11%, and the number of selected features is 14.

3.2 Discussion

This study aims to determine how it works and the results of its accuracy obtained from optimizing the C4.5 algorithm using the feature selection algorithm Genetics in the diagnosis of life expectancy

of people with hepatitis. Election features are carried out based on the best fitness value by paying attention to parameters from genetic algorithms such as maximum generation, population size, probability cross-over, and mutation probability. After going through the pre-processing stage is a classification process using the C4.5 algorithm. From that result obtained, there is an increase in accuracy results. Comparison of accuracy results Algorithm C4.5 before and after optimized using Genetic Algorithms in Figure 2.

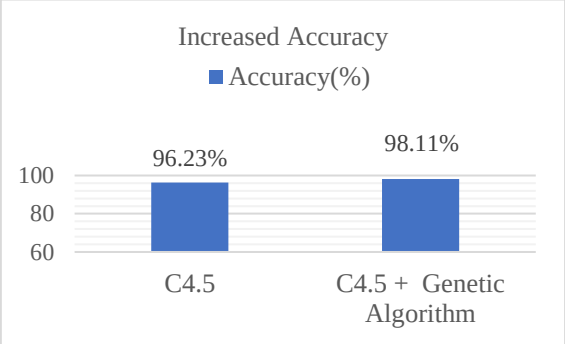

Figure 2. Increased accuracy

The composition between the training and testing data was chosen randomly, resulting in the accuracy results obtained using the C4.5 algorithm changing in each experiment. While the highest accuracy results obtained by applying the Genetic Algorithm are 98.11%, and the number of features used is 14 features. Based on the experiments conducted, there was an increase in accuracy using the Genetic Algorithm to classify the C4.5 algorithm to diagnose the life expectancy of people with hepatitis as much as 1.88%. The accuracy results in the system show the level of accuracy in determining a class consisting of die or live classes in the system. The comparison of this study with previous research can be seen in Table 4.

Table 4. The related research

Researchers	Year	Methods	Result
Septiani	2014	C4.5	77.29%.
Ramdhani	2016	C4.5 + PSO	C4.5 79,33%
			C4.5 + PSO 85,00%
Oktaviani *et al.*,	2018	Comparison of C45 and SVM	C4.5 80,6452%
			SVM 80,3279%
Buani	2018	Naïve Bayes	Naïve Bayes 83,71%
		Naïve Bayes + Genetic Algorithm	Naïve Bayes + Genetic Algorithm 96, 77%

The advantage of research through this model is that applying the Genetic Algorithm as a selection feature in the classification of the C4.5 algorithm can improve accuracy in diagnosing the life expectancy of people with hepatitis. While the drawback of this research model is that the accuracy results tend to fluctuate because it depends on the initial initialization to determine the fitness value of the Genetic Algorithm to be chosen randomly.

4 Conclusion

In this study, the application of the C4.5 algorithm is combined with the Genetic algorithm as a feature selection for the diagnosis of hepatitis dataset obtained from the UCI Machine Learning Repository. Genetic algorithms are applied to select the best features based on the best fitness values. This research resulted in the accuracy of the application of the C4.5 algorithm of 94.15%. This research produces the highest accuracy from the application of the C4.5 algorithm of 96.23%. The results of the C4.5

algorithm can be improved again by the preprocessing algorithm, namely using the Genetic algorithm as a feature selection of 98.11%, and the number of selected features is 14. In this study, it can be concluded that applying the Genetic algorithm to the C4.5 algorithm can improve the accuracy of diagnosing hepatitis by 1.88%.

References

Aini, S. H. A., Sari, Y. A., & Arwan, A. (2018). Seleksi Fitur Information Gain untuk Klasifikasi Penyakit Jantung Menggunakan Kombinasi Metode K-Nearest Neighbor dan Naïve Bayes [Selection of Information Gain Features for Cardiac Disease Classification Using a Combination of K-Nearest Neighbor and Naïve Bayes Methods]. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, 2(9), 2546-2554.

Asgarnezhad, R., Shekofteh, M., & Boroujeni, F. Z. (2017). Improving Diagnosis of Diabetes Mellitus Using Combination of Preprocessing Techniques. *Journal of Theoretical and Applied Information Technology*, 13(95), 2889- 2895.

Eid, H. F., & Abraham, A. (2018). Adaptive Feature Selection and Classification Using Modified Whale Optimization Algorithm. *International Journal of Computer Information Systems and Industrial Management Applications*, 10, 174-182.

Kathija, A., Nisha, S. S., & Sathik, M. M. (2017). Classification of Breast Cancer Data Using C4.5 Classifier Algorithm. *International Journal of Recent Engineering Research and Development* (*IJRERD*), 2(2), 13-19.

Khomsah, S. (2018). Prediksi Harapan Hidup Penderita Hepatitis Kronik Menggunakan Metode-Metode Klasifikasi [Prediction of Life Expectancy of Chronic Hepatitis Patients Using Classification Methods]. *Seminar Nasional Informatika Medis (SNIMed)*, 38-45.

Liu, Y., Wang, G., Chen, H., Dong, H., Zhu, X., & Wang, S. (2011). An Improved Particle Swarm Optimization for Feature Selection. *Journal of Bionic Engineering*, 8, 191-200. doi: 10.1016/S1672-6529(11)60020-6

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*, 983(1), 1-8. doi: 10.1088/1742-6596/983/1/012063

Muzakir, A., & Wulandari, R. A. (2016). Model Data Mining sebagai Prediksi Penyakit Hipertensi Kehamilan dengan Teknik Decision Tree [Data Mining Model as Prediction of Pregnancy Hypertension with Decision Tree Technique]. *Scientific Journal of Informatics*, 3(1), 19-26. doi: 10.15294/sji.v3i1.4610

Nugroho, D., Nhita, F., & Trantoro, D. (2016). *Prediction of Disease Using Genetic Algorithm (GA) and Naive Bayes for Data High Dimension.* eProceeding of Engineering, 3(2), 3889-3899.

Oktanisa, I., & Supianto, A. A. (2018). Perbandingan Teknik Klasifikasi dalam Data Mining untuk Bank Direct Marketing [Comparison of Classification Techniques in Data Mining for Direct Marketing Banks]. *Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK)*, 5(5), 567-576. doi: 10.25126/jtiik.201855958

Oktaviani, P. S., Ramadhani, R.D., Laksana, T. G., & Amalia, A. E. (2018). Komparasi Tingkat Akurasi Support Vector Machine (SVM) dan C4.5 dalam Mengklasifikasikan Keberlangsungan Hidup Pasien Hepatitis [Comparison of the Accuracy Level of Support Vector Machine (SVM) and C4.5 in Classifying Survival of Hepatitis Patients]. *Proceedings on Conference on Electrical Engineering, Telematics, Industrial Technology, and Creative Media*, 163-167.

Perveen, S., Shahbaza, M., Guergachib, A., & Keshavjeec, K. (2016). Performance Analysis of Data Mining Classification Techniques to Predict Diabetes. *Procedia Computer Science*, 82,115-121. doi: 10.1016/j.procs.2016.04.016

Septiani, W. D. (2014). Penerapan Algoritma C4.5 Untuk Prediksi Penyakit Hepatitis [Application of C4.5 Algorithm for Hepatitis Disease Prediction]. *Journal Techno Nusa Mandiri*, 10(1), 69-78.

Wibowo, R., & Indriyawati, H. (2020). Top-K Feature Selection untuk Deteksi Penyakit Hepatitis Menggunakan Algoritma Naïve Bayes [Top-K Selection Feature for Hepatitis Detection using Naïve Bayes Algorithm]. *Jurnal Buana Informatika*, 11(1), 1-9. doi: 10.24002/jbi.v11i1.2456

Widodo, A., & Handoyo, S. (2017). The Classification Performance Using Logistic Regression and Support Vector Machine (SVM). *Journal of Theoretical and Applied Information Technology,* 95(19), 5184-5193.