# Optimization of C4.5 Algorithm Using K-Means Algorithm and Particle Swarm Optimization Feature Selection on Breast Cancer Diagnosis

Anita Ayu Septiantina [1*], Endang Sugiharti [1]

[1]Department Of Computer Science, Faculty of Mathematics and Natural Sciences, Universitas Negeri Semarang, Semarang, Indonesia
*Corresponding author: anita0509@students.unnes.ac.id

ARTICLE INFO

**Article history**
Received 11 February 2020
Revised 10 March 2020
Accepted 2 April 2020

**Keywords**
C4.5 Algorithm Optimization
K-Means
PSO
Breast Cancer

ABSTRACT

Large data requires methods to explore information so that it can provide solutions to problem solving. The method is the data mining process. In the medical world, data mining is useful in diagnosing a disease such as breast cancer. Data mining has several techniques in exploring hidden data, one of which is a classification with the C4.5 algorithm. The C4.5 algorithm has proven better results than other decision tree algorithms. In the classification process, the results of the accuracy obtained are very important. So, optimization is needed to improve classification accuracy. The C4.5 algorithm optimization is done using the K-Means algorithm for clustering processes in continuous data and the Particle Swarm Optimization feature selection process. This research aims to determine the workings of accuracy optimization in the C4.5 algorithm and the results of accuracy obtained in breast cancer diagnosis. This research uses a dataset of the Wisconsin Diagnostic Breast Cancer (WDBC) UCI Machine Learning Repository. From the results of the research, the proposed method provides an average accuracy is 97,894%. So that provides better accuracy when compared with the C4.5 algorithm, which is 94.152%. Experiments based on the proposed method proved to be able to increase the classification accuracy by 3,742%.



## 1 Introduction

Information technology in various fields already has a large volume of data storage. The collected data requires a method to explore knowledge and information so that it can provide good decisions in solving problems. Data mining is the right solution to find useful information on large data (Ahmed & Elaraby, 2014).

Data mining is useful in predicting a disease (Muslim *et al*., 2017). Breast cancer is one of the main causes of a woman's death (Sheikhpour, Sarram, & Sheikhpour, 2016). For example, women in the United States, there are 230,480 breast cancer cases, and in 2011, 39,520 breast cancer patients died (Wang, Makond, Chen, & Wang, 2014). So, it is necessary to do early detection to reduce mortality rates in breast cancer patients by applying data mining in breast cancer diagnosis (Muslim, Rukmana, Sugiharti, Prasetiyo, & Alimah, 2018).

One of the techniques in data mining is a classification with the C4.5 algorithm (Vijayarani, Janari, & Sharmila, 2015). C4.5 algorithm is one of the decision tree algorithms that provides good accuracy for solving short-term problems (Nino, Gonzalez, Palacious, Domingo, & Hernandez, 2013). C4.5 algorithm has a weakness in continuous data in building a decision tree. Research conducted by Rajeshinigo, K-Means algorithm can be used to overcome continuous data so it can improve classification accuracy (Rajeshinigo & Jebalmalar, 2017).

K-Means algorithm is one of the unsupervised learning algorithms and follows the partition method for grouping, which is most commonly used because of its simplicity and efficiency (Sugiharti & Muslim, 2016). The K-Means algorithm works by partitioning a set of data objects from one set to the right class or cluster (Devi, Sugiharti, & Arifudin, 2018).

Apart from weaknesses in continuous data, empty branches, insignificant branches, and overfitting are caused by data noise (Muslim *et al*., 2018). Then feature selection is needed to select relevant features to improve the classification process's accuracy (Liu *et al*., 2011). The feature selection algorithm is an algorithm used to find important features of high dimensional data. Features that are not considered as important (irrelevant, excessive) features will be removed (Ismi, Panchoo, & Murinto, 2016).

Particle Swarm Optimization (PSO) algorithm is one of the metaheuristic optimization algorithms for the feature selection process (Muslim *et al*., 2018). In addition, the PSO algorithm is a computational evolution technology, which is part of the swarm intelligence algorithm category. In the PSO algorithm, particle movement follows the optimal position to provide optimal solutions that it has been widely used in algorithm optimization (Huo & Ma, 2015).

Based on the background above, the researcher intends to conduct research on the optimization of the C4.5 algorithm using the K-Means algorithm and PSO feature selection so it can provide good accuracy in the diagnosis of breast cancer. The formulation of the problem in this research is (1) how the K-Means algorithm and PSO feature selection for optimization algorithm C4.5 and (2) how the accuracy of the proposed method in the diagnosis of breast cancer.

This research aims to determine the working of the K-Means algorithm and PSO feature selection to optimize the C4.5 algorithm and find out the results of accuracy obtained in the optimization of the C4.5 algorithm in the diagnosis of breast cancer.

## 2 Methods

This research will incorporate the proposed algorithm for optimization algorithm C4.5 and compare results obtained an accuracy of the C4.5 algorithm and C4.5 algorithm with the K-Means clustering process. This research's preprocessing data is the clustering process with the K-Means algorithm and the feature selection process with Particle Swarm Optimization. In building a web-based system using the Python programming language to implement the proposed method and as an application for research trials.

### 2.1 Data Collection

The data used in this research, the Wisconsin Diagnostic Breast Cancer (WDBC), was obtained from the UCI Machine Learning Repository. This dataset consists of 569 instances with 30 continuous attributes and 2 category attributes (ID Patient and class).

### 2.2 K-Means Algorithm

The clustering process uses the K-Means algorithm on each continuous attribute of the dataset, and the process is used to transform continuous value attributes into a categorical value in the form of clusters (Rajeshinigo *et al*., 2017). The K-Means algorithm consists of two phases; namely, the first phase calculates k centroid, and the second phase takes each point to the cluster that has the closest center of adjacent data points (Sahu, Parvathi, & Krishna, 2017). The K-Means algorithm takes input parameters, k as the number of clusters, and divides the dataset from n objects into k clusters. In K-Means Algorithm begins with the object k chosen randomly, representing the initial k cluster center. Then each object is assigned to one cluster based on the proximity of the object to the center of the cluster with the distance of each existing data to each centroid using the Euclidian formula in Equation 1 until the closest distance is found from each data with the intended centroid (Purwar & Singh, 2015). Where D is the distance of data, $x_1$ is centroid, and $x_2$ is data object data.

$$D(x_2, x_1) = \sqrt{\sum_{i=1}^{n}(x_2 - x_1)^2} \quad (1)$$

### 2.3 Particle Swarm Optimization

After the clustering results are obtained, then feature selection is performed using the Particle Swarm Optimization algorithm to select the relevant attributes. PSO is an iterative algorithm to find the best solution based on a population consisting of many particles. As an optimization tool, PSO provides population-based search procedures where individuals are called particles (Sumathi & Surekha, 2010). Particles that represent candidate solutions can move to the optimal position by updating their position and velocity (Arifin, 2017). Update velocity ($v_{id}^{new}$) with Equation 2 (Wang *et al*., 2014).

$$v_{id}^{new} = w \,.\, v_{id}^{old} + \; c_1.r_1\left(pb_{id}^{old} - x_{id}^{old}\right) + c_2.r_2\left(gb_d^{old} - x_{id}^{old}\right) \tag{2}$$

Where $x_{id}^{old}$ representing the position of the particle i. pBest and gBest are positions with the best objective values found by particles i and the entire population. $w$ is used to control the PSO convergence behavior while $r_1$ and $r_2$ are random parameters ranging from 0 and 1. $c_1$ and $c_2$ indicate particle control iterative movements.

In applying the PSO algorithm used to solve the feature selection problem, it can use binary digits to show features. By updating the position ($x_{id}^{new}$) with the Sigmoid ($S$) formula in Equation 3 of the velocity that has been updated above with Equation 2. The unselected feature is denoted by 0, while the selected feature is represented by 1 (Wang *et al*., 2014).

$$S = \frac{1}{1+ e^{-v_{id}^{new}}} \tag{3}$$

$$x_{id}^{new} = \begin{cases} 1, \; \textit{if } S > \textit{rand}\,(0,1) \\ 0, \; \textit{otherwise} \end{cases}$$

In this research, using particle population initialization randomly is *rand* (0,1). Repetition stops when it reaches the maximum iteration. The selected attribute is obtained by taking the highest fitness value is the classification accuracy of the C4.5 algorithm with Equation 4.

$$\textit{Fitness} = \textit{Accuracy} = \frac{\text{number of instances classified correctly}}{\text{number of instances}} \tag{4}$$

### 2.4 Data Split

After obtaining selected features, the data is distributed with the proportion of training data: testing data = 70%: 30%. Training data is used to create a model from the proposed method and testing data for model validation.

### 2.5 C4.5 Algorithm

The feature selection process results will be classified using the C4.5 algorithm, and the model will be generated and calculate the accuracy obtained. To choose attributes as roots based on the highest gain ratio value. To calculate the gain ratio's value, calculate the information gain value of each attribute used formula Equation 5 and calculate split information Equation 7 below. Then calculate the gain ratio as in Equation 8 (Muslim *et al*., 2018).

$$\textit{Gain(S,A)} = \textit{Entropy(S)} - \sum_{i=1}^{n} \frac{|Si|}{|S|} * \textit{Entropy(Si)} \tag{5}$$

With S: Set of cases, A: Attribute, n: Number of partitions attribute A, | Si |: number of cases on the i, and | S | partitions: number of cases in S.

While the calculation of the entropy value can be seen in the formula Equation 6 below.

$$\textit{Entropy(S)} = - \sum_{i=1}^{n} Pi * log_2 Pi \tag{6}$$

$$Split\ Information(A) = -\sum_{i=1}^{n} \frac{|Si|}{|S|} * log_2\left(\frac{|Si|}{S}\right) \quad (7)$$

$$Gain\ Ratio(A) = \frac{Gain(S,A)}{Split\ Information(A)} \quad (8)$$

### 2.6 Performances Measures

The model will be evaluated using a confusion matrix to determine the accuracy obtained. Evaluation of the confusion matrix model can be seen in Table 1, and the formula for calculating accuracy is shown in Equation 9 (Wang *et al*., 2014).

$$Accuracy = \frac{TN+TP}{TP+TN+FP+FN} \quad (9)$$

**Table 1.** Confusion matrix

| *Classification* | | *Predicted* | |
|---|---|---|---|
| | | *Negative* | *Positive* |
| *Actual* | *Negative* | TN | FP |
| | *Positive* | FN | TP |

Where:
*True Positive* (TP) : a number of positive cases classified as positive
*False Positive* (FP) : a number of negative cases classified as positive
*True Negative* (TN) : a number of negative cases classified as negative
*False Negative* (FN) : a number of positive cases classified as negative

## 3 Results and Discussion

### 3.1 Research Result

The algorithm used in the C4.5 algorithm optimization process is K-Means Algorithm and Particle Swarm Optimization (PSO) algorithm. The algorithm used as a comparison is the C4.5 algorithm as the algorithm of classification and optimization of the C4.5 algorithm with the clustering process using the K-Means algorithm. The results of the research that have been done are:

#### *3.1.1 The Result of Clustering*

The first stage is the clustering process using the K-Means algorithm. This process is used to transform continuous values on each dataset attribute, amounting to 30, into a categorical value in the form of clusters. In determining the number of k-clusters, researchers tested several numbers of clusters, namely k = 2, 3, 4, and 5. Each experiment will then be assessed for the quality of the number of clusters, using the C4.5 classification and selected, which has the highest accuracy. The results of accuracy for each k-cluster can be seen in Table 2. The sample clustering results in the dataset can be seen in Table 3.

**Table 2.** Accuracy results for each k-cluster

| Number of k-*cluster* | Accuracy |
|---|---|
| 2 | 89.473% |
| **3** | **95.321%** |
| 4 | 90.058% |
| 5 | 92.982% |

**Table 3.** Sample clustering result

| No. | Attribute | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | a1 | b0 | c1 | d0 | e0 | f1 | g1 | h0 | … | ad1 |
| 2 | a1 | b1 | c1 | d1 | e2 | f0 | g2 | h2 | … | ad2 |
| 3 | a1 | b1 | c1 | d1 | e0 | f2 | g1 | h0 | … | ad2 |
| 4 | a2 | b1 | c2 | d2 | e0 | f1 | g1 | h2 | … | ad1 |
| 5 | a1 | b0 | c1 | d1 | e1 | f2 | g1 | h2 | … | ad0 |

In the results of this clustering, researchers use the initial attributes (a-ad) to distinguish the results of clustering each attribute. For k = 3 (0,1,2) is the cluster result on each attribute.

### *3.1.2 Feature Selection Process*

In this research, the parameter is the number of particles was 50 particles because it was able to obtain the optimal solution. For parameters such as the number of iterations, the weight of inertia, as well as cognitive and social learning factors, followed Chiu C. Y's research. This parameter value is chosen because it provides good convergence, including:

1. Cognitive learning factors (c1) and social learning factors (c2) are set at 1.49 for each.
2. Inertia weight (w) = 0,72
3. Number of iteration = 100

There are 50 evaluation values of 50 particles spread in the search area in each iteration, where each particle is based on the cost value (the difference from evaluation of the fitness function with minimal error tolerance). Each iteration will get the best position with the lowest cost value called the best cost. After all the iterations have been done, the cost value for each iteration will be compared, and the final cost will be obtained with the lowest cost value (best cost). The best cost value for each 1st iteration to 100th iteration (with multiples of 10) can be seen in Table 4.

**Table 4.** The best cost value for each n iteration

| Iteration | Cost value |
|---|---|
| 1 | 0.0351 |
| 11 | 0.0292 |
| 21 | 0.0234 |
| 31 | 0.0234 |
| 41 | 0.0175 |
| 51 | 0.0117 |
| 61 | 0.0117 |
| 71 | 0.0117 |
| 81 | 0.0117 |
| 91 | 0.0117 |
| **Final Cost** | **0.0117** |

The final cost value is obtained the best position for selected feature recommendations, which are considered to have the highest fitness value of particles and are calculated based on classification accuracy.

### *3.1.3 Classification Process*

The results of feature selection in the form of selected attributes will be used for the classification process using the C4.5 algorithm. To choose the selected attribute as the root, it is based on the highest gain value of the existing attributes. After obtaining a model from the proposed method, then through accuracy testing, using a confusion matrix.

### *3.1.4 Performances Measures*

After constructing a model from the proposed method, the model will be evaluated to determine the results of the accuracy based on the calculation of testing data using the Confusion matrix method. Accuracy formula in Equation 9. From the results of the execution, the accuracy was 96.491%. The calculation of the confusion matrix results from the example execution can be seen in Table 5.

**Table 5**. Confusion matrix execution results

| *Classification* | | *Predicted* | |
|---|---|---|---|
| | | *Negative* | *Positive* |
| *Actual* | *Negative* | 56 | 3 |
| | *Positive* | 3 | 109 |

$$Accuracy = \frac{56+109}{171} = \frac{165}{171} = 0{,}96491*100 = 96.491\%$$

In this research, the execution was carried out ten times to find out the average number of attributes and accuracy obtained. The following are the results of 10 attempts to find the average number of selected features, and the results of each execution's accuracy can be seen in Table 6.

**Table 6.** Results of 10 experiment accuracy

| Execution | Feature | Accuracy |
|---|---|---|
| 1 | 17 | 97.660% |
| 2 | 14 | 98.245% |
| 3 | 13 | 98.245% |
| 4 | 17 | 98.245% |
| 5 | 16 | 97.660% |
| 6 | 15 | 98.830% |
| 7 | 16 | 96.491 % |
| 8 | 14 | 97.660% |
| 9 | 15 | 98.245% |
| 10 | 13 | 97.660% |
| **Mean** | **15** | **97.894 %** |

### 3.2 Discussion

In this research, an algorithm was implemented using the C4.5 algorithm, the C4.5 algorithm with the clustering process using the K-Means algorithm, and the proposed method. Each algorithm will be compared based on the results of the accuracy obtained. With the highest accuracy results obtained, the algorithm is considered to be getting better.

The first step in the optimization of the C4.5 algorithm is the clustering process using the K-Means algorithm. The clustering process is used to transform continuous values on each dataset attribute, which amounts to 30, into a categorical value in the form of clusters. Selection of the number of k-cluster effects on the accuracy of the results obtained, the quality of the number of k-cluster will be assessed using the C4.5 classification and selected with the highest accuracy. From the experiments that have been done, the number of clusters = 3 can provide better accuracy equal to 1.169% compared with the C4.5 algorithm.

The next step is the feature selection process with the Particle Swarm Optimization (PSO) algorithm. Feature selection is an algorithm used to search for important features. Features that are not considered as important (irrelevant, excessive) features will be removed, so it will provide good accuracy in the classification process.

The results of feature selection in the form of selected attributes will be used for the classification process using the C4.5 algorithm. The selected attribute as root is based on the highest gain value of the existing attributes. After obtaining a model from the proposed method, the accuracy testing phase uses a confusion matrix.

The proposed method will be executed ten times to find the selected feature average, and the accuracy obtained. The average accuracy obtained from 10 executions carried out was 97.894%. When compared with the results of the accuracy of the C4.5 algorithm, the proposed method produces an accuracy of 3,742% more accurately. This is because the feature selection process with the PSO algorithm in the classification of the C4.5 algorithm performs an optimal solution based on the swarm intelligence concept, where each particle in the search area represents a classification process.

The proposed method will be compared with the C4.5 algorithm and optimization of the C4.5 algorithm with a clustering process to determine whether the proposed method has proven to be the highest in producing accuracy. The results of the comparison of accuracy obtained can be seen in Table 7.

**Table 7.** Results of accuracy comparison

| No. | Algorithm | Accuracy |
|---|---|---|
| 1 | C4.5 Algorithm | 94,152% |
| 2 | C4.5 Algorithm + *K-Means* | 95,321% |
| 3 | *Proposed Method* | 97,894% |

## 4 Conclusion

Based on the results of the research, conclusions obtained based on the formulation of the problem is the way of K-Means algorithm works in optimizing the accuracy of the C4.5 algorithm classification process is by clustering each continuous attribute on the WDBC dataset and selecting the number of k-clusters can affect the results of the accuracy obtained. Then the feature selection process is performed using Particle Swarm Optimization. The feature selection process results will be classified using the C4.5 algorithm, and a model is produced. The model from the proposed method will be tested to determine the accuracy obtained. The process of classification C4.5 algorithm on the WDBC

dataset produces an accuracy of 94,152%. The experiment was then carried out with the K-Means algorithm to optimize the C4.5 algorithm, so the accuracy becomes 95,321%. While the experiment based on the optimization of the C4.5 algorithm using the K-Means algorithm and PSO feature selection can produce an average accuracy of 97.894%. It can then be concluded that the proposed method, proven to optimize and improve the accuracy of the C4.5 algorithm classification process in breast cancer diagnosis.

## References

Ahmed, A. B. E. D. & Elaraby, I. S. (2014). Data Mining: A prediction for Student's Performance Using Classification Method. *World Journal of Computer Application and Technology, 2*(2), 43-47. doi:10.1088/1757-899X/215/1/012036

Arifin, T. (2017). Implementasi Algoritma PSO Dan Teknik Bagging Untuk Klasifikasi Sel Pap Smear [Implementation of PSO Algorithm and Bagging Techniques for Classification of Pap Smear Cells]. *Jurnal Informatika*, *4*(2), 155-162. doi:10.31294/ji.v4i2.2129

Chiu, C. Y., Feng, C. Y., Ting, K. I., Chun, K. H. (2009). An Intelligent Market Segmentation System Using K-Means and Particle Swarm Optimization. *Expert Systems with Applications*, *36*, 4558-4565. doi:10.1016/j.eswa.2008.05.029

Devi, F. R., Sugiharti, E., & Arifudin, R. (2018). The Comparison Combination of Naïve Bayes Classification Algorithm with Fuzzy C-Means and K-Means for Determining Beef Cattle Quality in Semarang Regency. *Scientific Journal of Informatics*, *5*(2), 194-204. doi: 10.15294/sji.v5i2.15452

Huo, J. & Ma, X. (2015). Quantum Particle Swarm Optimization Algorithm Based on Dynamic Adaptive Search Strategy. *TELKOMNIKA Journal*, *13*(1), 321-330. doi:10.12928/telkomnika.v13i1.1266

Ismi, D. P., Panchoo, S., & Murinto. (2016). K-Means Clustering Based Filter Feature Selection on High Dimensional Data. *International Journal of Advances in Intelligent Informatics*, *2*(1), 38-45. doi:10.26555/ijain.v2i1.54

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

Muslim, M. A., Sugiharti, E., Prasetiyo, B., & Alimah, S. (2017). Penerapan Dizcretization dan Teknik Bagging Untuk Meningkatkan Akurasi Klasifikasi Berbasis Ensemble pada Algoritma C4.5 dalam Mendiagnosa Diabetes [Application of Dizcretization and Bagging Techniques to Improve Accuracy of Ensemble-Based Classification on the C4.5 Algorithm in Diagnosing Diabetes]. *Lontar Komputer: Jurnal Ilmiah Teknologi Informasi*, *8*(2), 135-143. doi:10.24843/LKJITI.2017.v08.i02.p07

Nino, J. T., Gonzalez, A. J., Palacious, R. C., Domingo, E. J., & Hernandez, G. A. (2013). Improving Accuracy of Decision Trees Using Clustering Techniques. *Journal of Universal Computer Science*, *19*(4), 484-501. doi:10.3217/jucs-019-04-0483

Purwar, A. & Singh, S. K. (2015). Hybrid Prediction Model with Missing Value Imputation for Medical Data. *Expert Systems with Applications*, *42*(13), 5621-5631. doi:10.1016/j.eswa.2015.02.050

Rajeshinigo, D. & Jebamalar, J. P. A. (2017). Accuracy Improvement of C4.5 using K-Means Clustering. *International Journal of Science and Research (IJSR), 6*(6), 2755-2758. Retrieved from https://www.ijsr.net/search_index_results_paperid.php?id=ART20174834.

Sahu, M., Parvathi, K., & Krishna, M. V. (2017). Parametric Comparison of K-means and Adaptive K-means Clustering Performance on Different Images. *International Journal of Electrical and Computer Engineering (IJECE)*, *7*(2), 810-817. doi:10.11591/ijece.v7i2.pp810-817

Sheikhpour, R., Sarram, M. A., & Sheikhpour, R. (2016). Particle Swarm Optimization for Bandwidth Determination and Feature Selection of Kernel Density Estimation Based Classifiers in Diagnosis of Breast Cancer. *Applied Soft Computing*, *40*, 113-131. doi:10.1016/j.asoc.2015.10.005

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. Retrieved from https://lib.unnes.ac.id/33054/

Sumathi, S. & Surekha, P. (2010). *Computational Intelligence Paradigms: Theory and Application Using Matlab.* Boca Raton: CRC Press. doi:10.1201/9781439809037

Vijayarani, S., Janani, R., & Sharmila, S. (2015). Data Mining Classification Algorithms for Hepatitis and Thyroid Data Set Analysis. *International Journal of Data Mining Techniques and Applications*, *4*(1), 43-47. Retrieved from http://www.hindex.org/2015/article.php?page=1139

Wang, K. J., Makond, B., Chen, K. H., & Wang, K. M. (2014). A Hybrid Classifier Combining SMOTE with PSO to Estimate 5-Year Survivability of Breast Cancer Patients. *Applied Soft* Computing, *20*, 15-24. doi:10.1016/j.asoc.2013.09.014