Particle Swarm Optimization – Based on Decision Tree of C4.5 Algorithm for Upper Respiratory Tract Infections (URTI) Prediction

Dwi Meylitasari Br. Tarigan¹, Dian Palupi Rini², Sukemi²
¹Student, Faculty of Computer Science, Sriwijaya University, Palembang, Indonesia.
²Lecturer, Faculty of Computer Science, Sriwijaya University, Palembang, Indonesia.

Email : dwimeylitasaritarigan@gmail.com, dprini@unsri.ac.id, sukemiku66@gmail.com

Abstract. Data mining is related to searching data to find patterns or knowledge from the whole data. It turns out that a large data set can produce a data whose results can provide new knowledge information. Data mining is an important step in the process of finding knowledge. In this study will be discussed about data mining design using C4.5 algorithm to predict acute or non-acute URTI in children by selecting the candidate criteria used in this study so that it can contribute to the medical team in the health environment to know and follow up patients who affected by URTI. The C4.5 algorithm is used to obtain information by selecting or separating characteristics. Giving attribute weight to the C4.5 algorithm using Particle Swarm Optimization can improve the accuracy of the C4.5 Algorithm performance and can also be influenced by the selection of the right attributes, the more attributes used will result in a long time and costs that will reduce the accuracy and performance slower.

1. Introduction
At present technology can provide fast and accurate information, especially in the field of health both for health teams, doctors, nurses and even for patients themselves so that it is easier to control their health. Various fields in health produce large amounts of data each year. Most of the data cannot provide information quickly and directly so that the information obtained takes a long time. Need to be effectively carried out to provide fast information, one of which is the process of data mining from hidden information with a very large amount of data. Data mining is related to searching data to find patterns or knowledge from the whole data. It turns out that a large data set can produce a data whose results can provide new knowledge information. Data mining is an important step in the process of finding knowledge and information [1]. In data mining predictions and classifications are widely used to analyze data that can describe data classes or to predict future data. Research using classification in predicting has been widely applied to diagnose a disease. In some classification problems, datasets that have a large feature cause problems to be "miss classification", they have a large number of irrelevant features that cause constraints in classification [2].

Particle Swarm Optimization (PSO) has parameters such as position, maximum speed, constant acceleration and weight of inertia. In PSO techniques there are several ways to optimize, such as increasing the attribute weight (weight attribute) of all attributes or variables used, selecting attributes (attribute selection) and selecting features [3]. In selecting features or minimize the number of features often used in preprocessing to determine the relevant attributes that have been used before and eliminate irrelevant features that are not related to the classification. [4]
Giving attribute weight to the C4.5 algorithm using Particle Swarm Optimization can improve the accuracy of the C4.5 Algorithm performance and can also be influenced by the selection of the right attributes, the more attributes used will result in a long time and costs that will reduce the accuracy and performance slower.

Basically the level of accuracy cannot be stated before the execution of the process is complete. Then calculate based on precision where this need is to process numeric data with various precision widths so that the numerical calculation process can be optimal. [5] [6]

In this study Optimization of Attribute Weight in the C4.5 Algorithm will be discussed using Particle Swarm Optimization to predict acute or non-acute ARI in children by choosing the candidate criteria to be used in this study so that it can contribute to the medical team in the health environment to know and follow up on patients affected by URTI.

2. Literature Review

2.1. Related Works

In this related research are some results from the previous research as a reference basis for this research. From the study, researchers found several studies that support to raise topics related to the author’s research. Therefore, there will be a step in the study of several studies regarding the algorithms used and those that will be raised by the authors in this report. Prediction of Premature Birth of Babies Using C4.5 Algorithm Based on Particle Swarm Optimization [7]. This study uses a data record of 250 records and produces an accuracy rate of 93.60% with the C4.5 algorithm, while the PSO-based C4.5 produces 96.00% accuracy. Optimization of C4.5 Classification Algorithm based on Particle Swarm Optimization for Prediction of Heart Disease [8]. In his research, the data used used a public data set obtained from UCI consisting of 75 attributes which were reduced to 13 attributes. The results of the accuracy value for the C4.5 classification algorithm model is 81.25% and the PSO based C4.5 algorithm accuracy value is 93.75%.

2.2. Data Mining

Data mining is a process that employs one or more computer learning techniques to analyze and extract knowledge automatically. Data mining contains the search for trends or desired patterns in large databases to help take decisions in the future. These patterns are recognized by certain devices that can provide a useful and insightful data analysis which can then be studied more thoroughly, which may use other decision support tools [9].

2.3. C4.5 Algorithm

The C4.5 algorithm is a group of Decision tree algorithms (decision tree). This algorithm has input in the form of training sample data and samples. Training samples sample data that will be used to build a tree that has been tested for truth. While samples are data fields that will be used as parameters in data classification [10].

The C4.5 algorithm has advantages that are easy to understand, flexible, and attractive because it can be visualized in the form of an image (decision tree) [11].

In general, the C4.5 algorithm for building decision trees is as follows [12]:

1. Select an attribute as root.
2. Create a branch for each value.
3. The cases in the branch.
4. Repeat the process for each branch until all cases in the branch have the same class.

To select an attribute as a root, it is based on the highest gain value of the attributes. Before calculating the gain of an attribute, first calculate the entropy value, namely:

\[
\text{Entropy} (S) = \sum_{i=1}^{n} p_i \log_2 p_i
\]  

(1)

Where the information is:
S : set (dataset)
n : number of records
Pi : probability obtained from yes or no total case

To calculate the gain, the formula is used:

\[
Gain(A) = Entropy(S) - \sum_{i} \frac{|S_i|}{|S|} \times Entropy(S_i)
\]  

Where the information is:
S : set (dataset)
A : attribute that will be used
n : number of partition attributes A
|S_i| : number of cases on the i partition
|S| : number of cases in S

2.4. Particle Swarm Optimization (PSO)
PSO is a population-based search algorithm that is initialized with a population of random solutions and is used to solve optimization problems [13]. PSO is a global heuristic optimization method introduced by Dr. Kennedy and Eberhart in 1995 based on the research of my herd of birds and fish [14].

Each particle in the PSO is also interpreted by the speed of the particles flying through the search space at a speed that is dynamically adjusted for their historical behavior. Therefore, particles have a tendency to fly towards better and better search areas during the search process [13]. Formula for calculating displacement of position and velocity of particles, namely:

\[
V_i(t) = V_i(t-1) + c_1 r_1 [X_{pbest} - X_i(t)] + c_2 r_2 [X_{Gbest} - X_i(t)]
\]

\[
X_i(t) = X_i(t-1) + V_i(t)
\]

Information:
V_i(t) : particle velocity i during iteration t
X_i(t) : the position of the particle i during iteration t
c1 and c2 : learning rates for individual abilities (cognitive) and social influence (group)
X_{pbest} : best position for partisans i
X_{Gbest} : global best position

3. Methodology

3.1. Data Collection
At this data collection stage, techniques or methods that will be used to collect data. The data we are looking for must be in accordance with the research objectives. In data collection there are data sources, namely primary data and secondary data. Primary data is data that can only be obtained from the original source or the first, while secondary data is data that is already available so that we can simply search and collect. Secondary data can be obtained more easily and quickly because it is available, for example in libraries, companies, agencies, statistics centers, and government offices.

In this study, researchers used secondary data that already existed in the community health center in Boyolali. The attributes contained in it are the factors that cause URTI. This data have 10 Attributes 43 record data.
3.2. Modelling

There are two steps in this modelling which are tested by combining C4.5 algorithm with PSO algorithm on attribute selection. To see the results of using C4.5 algorithm and C4.5 algorithm, use PSO using the RapidMiner framework.

4. Experiment and Result

4.1. Experiment using C4.5 Algorithm

In forming the previous decision tree, changing some attributes into nominal values in accordance with the data in table 1. Then calculate Entropy and Inflation Information. Calculating total Entropy is done...
by calculating the number of "URTI" and "NO URTI" decisions in all cases. From the results of entropy calculations and information gain obtained then processed into the Decision Tree. Here are the results of the Decision Tree (Decision Tree):

![Decision Tree Using C4.5 Algorithm](image)

**Figure 2.** Decision Tree Using C4.5 Algorithm

### 4.2. Testing and Validation

Testing is done by cross validation. One type of cross validation is ten-fold cross validation. Here are the results of the ten-fold cross validation on the C4.5 algorithm.

|    | true 2 | true 1 | class precision |
|----|--------|--------|-----------------|
| pred 2 | 15  | 3 | 83.33% |
| pred 1 | 5   | 20 | 80.00% |
| class recall | 75.00% | 84.96% |

Based on testing using the Ten-Fold Cross Validation method, it produces 75.00% accuracy, 81.67% precision and a recall of 86.67%. This shows that, from the classification process carried out, it can be applied to patients with URTI.

### 4.3. Experiment with C4.5 Algorithm and PSO

In the experiment using PSO, the confusion matrix model will form matrix consisting of true positive or tuples positive and true negative or negative tuple, then enter the testing data already prepared into a confusion matrix so the results obtained in the table below.

|    | true 2 | true 1 | class precision |
|----|--------|--------|-----------------|
| pred 2 | 29  | 3 | 86.90% |
| pred 1 | 0   | 20 | 160.00% |
| class recall | 100.00% | 85.95% |

| Table 2. Confusion Matrix C4.5 Algorithm |

| Table 3. Confusion Matrix C4.5 Algorithm-PSO |
Based on testing using the Ten-Fold Cross Validation method use C4.5 Decision Tree algorithm with PSO, it produces 93.02% accuracy, 100.00% precision, and 86.96% recall. This shows that, from the classification process was carried out, it can be applied to patients with URTI.

4.4. Evaluate with ROC Curve
The test results of testing data for PSO-algorithm C4.5 against the ROC value on picture below:

![ROC Curve](image)

Fig 3. Curve of AUC PSO-Algorithm C4.5 in the ROC graph

5. Conclusion
In this paper, we discussed the theories of the algorithm C4.5, and C4.5 algorithm with Particle Swarm Optimization on the data of patients with URTI in Boyolali Health Center. From the test results by measuring the performance of both methods using confusion matrix, the ROC curve is known that C4.5 produces an accuracy value of 75%. PSO-based C4.5 method produces an accuracy of 93.02%. Thus the C4.5 method based on Particle Swarm Optimization is the best method for solving problems.

As future work, we propose to explore and experiment with various techniques in other decision tree phases, especially in determining the best attribute data to predict and improve the accuracy of results.

6. References
[1] Lakshmi, B.N., Raghunandhan, G.H. 2011 A conceptual overview of data mining (Proceedings of the National Conference on Innovations in Emerging Technology) pp 27-32
[2] Sugumaran V, Muralidharan V, Ramachandran KI. 2007. Feature Selection using Decision Tree and Classification through Proximal Support Vector Machine for fault diagnostics of roller bearing. Mechanical System and Signal Processing, pp 930-942
[3] Park, T. S., Lee, J. H., & Choi, B. 2009 Optimization for Artificial Neural Network with Adaptive Inertial Weight of Particle Swarm Optimization Cognitive Informatics (IEEE International Conference) pp 481-485
[4] O.S Qasim,Z.Y. Algamal 2018 Feature selection using particle swarm optimization-based logistic regression model (Chemometrics and Intelligent Laboratory Systems), pp 3-4.
[5] Sukemi, Hari Sudibyo, Anak Agung Putri Ratna 2012 Priority Based Computation : A Study on Paradigm Shift on Real Time Computation (IEEE International Conference on Computational Intelligence and Cybernetics) pp 129-130
[6] Moore, R.E 1962 Interval arithmetic and automatic error analysis in digital computing. Dept of mathematics of mathematics, Stanford of university, Stanford, California
[7] Ari Puspita 2016 Prediksi Kelahiran Bayi Secara Prematur dengan Menggunakan Algoritma C4.5 Berbasis Particle Swarm Optimization (Jurnal Teknik Informatika STMIK Antar Bangsa.Vol. II)
[8] Rusda Wajhillah, 2014 Optimasi Algoritma Klasifikasi C4.5 berbasis Particle Swarm Optimization untuk Prediksi Penyakit Jantung (SWABUMI, Vol.1)
[9] Daniel T. Larose. 2005 Discovering in Data Mining, An Introduction to Data Mining (Willey Interscience)

[10] Sunjana 2010 "Klasifikasi Data Nasabah Sebuah Asuransi Menggunakan Algoritma C4.5," (Seminar Nasional Aplikasi Teknologi Informasi) pp. D31-D34

[11] Hamid Abdul, Al-Ghamdi Abdullah, Naoman Amin, Madbouly Ayman 2012 A Comparative Analysis of Classification Algorithm for student college enrollment approval using data mining. pp.1-8

[12] T. H. Rahmadya and Prabowo P. W Herlawati 2013 Penerapan Data Mining Dengan Matlab. Bandung: Rekayasa Sains: Andi

[13] A. Abraham, C. Grosan and V. Ramos, Swarm Intelligence in Data Mining 2006 Verlag Berlin Heidelberg: Springer

[14] Haupt, R.L., Haupt, S.E 2004 Practical Genetic Algorithm, Second Edition. Wiley

Acknowledgments
This research was support by Dian Palupi Rini, Ph.D and Lecture of Magister Informatics Engineering Sriwijaya University for sharing their pearls of wisdom during the course of this research. We would also like to thank for family and friends for supporting.