Decision Tree Optimization in C4.5 Algorithm Using Genetic Algorithm

Irfan Sudahri Damanik¹, Agus Perdana Windarto¹, Anjar Wanto¹, Poningsih², Sundari Retno Andani², Widodo Saputra²

¹ STIKOM Tunas Bangsa, Pematangsiantar, Medan – Indonesia
² AMIK Tunas Bangsa, Pematangsiantar, Medan – Indonesia

* irfansudahri@amiktunasbangsa.ac.id

Abstract. The importance of efficiency in the space of search rules C4.5 decision tree algorithm has been the focus of a lot of researchers. Therefore, the development needs to be conducted to form a new, more efficient method but it can not be separated from the accuracy of the analysis as the results of the algorithm itself. For that purpose, by using a genetic algorithm (GA), it is expected to optimize and simplify the search rules of more complex combinations. The use of C4.5 with Hybrid genetic algorithm in search of a more effective rules requires a better understanding and a long time. But the use of the two algorithms will be mostly effective if the cases faced are very complex, having more branching condition and highly accurate.

1. Introduction

Decision tree [1] is a very strong and famous classification and prediction method. The decision tree method transforms a very large fact into a decision tree that represents the rules. Rules can be easily understood with natural language. The main benefit of using decision trees[2] is their ability to break down complex decision making processes to be simpler so that decision making will better interpret problem solutions by changing the form of data (tables) into a decision tree model, turning the decision tree model into a rule. Decision trees are also useful for exploring data, namely finding hidden relationships between a number of prospective input variables with a target variable.

In genetic algorithms there are a number of operations that support the success of genetic algorithms such as initial population generation, fitness calculations for each chromosome, chromosome selection, crossing, and chromosomal mutations. The process of crossing is the process of forming a child chromosome (offspring). Crossing aims to increase the diversity of strings in one population by crossing strings obtained from previous reproduction. There are several types of crosses including 1 point crossover, 2 points (two point crossover), and cross arithmetic[3]. The arithmetic method of crossing can be divided into 3 types, namely single arithmetic crossing, simple arithmetic crossing, and whole arithmetic crossing [4], [5]. The difference in this crossing method in producing the best chromosomes will affect the performance of genetic algorithms. The genetic algorithm will stop if a maximum number of generations is reached or the specified fitness level has been met.

2. Methodology

All processes carried out aim to build a classification process that is more efficient than before. The C4.5 algorithm [6]–[8] segmentation process predicts the data that will be the decision. The classification process in this algorithm produces branches while the leaves are classes or segments.
While the rule selection process that occurs has a large search space. This research methodology includes a genetic algorithm that determines the search optimization and reduces the search space for it, including details of the plan and the basics of the method used. As a test and training process, a dataset is obtained from the UCI Learning Machine Repository that has been donated by Candanedo, L et al (2015) about Accurate occupancy detection. This dataset contains the detection of the feasibility of an office room based on light intensity, room temperature, humidity and CO2 levels. The number of datasets consists of 8 attributes, one attribute is not included in table 1, namely the number of data. Lines on a dataset are 8143 lines. This means that the matrix in a dataset is 8 x 8143

| Table 1. Dataset UCI Learning Machine Accurate occupancy detection |
|---------------|-----------------|
| No | Label | Content |
|---|-----|--------|
| 1 | Date | time year-month-day hour:minute:second |
| 2 | Temperature | in Celsius |
| 3 | Relative Humidity | % |
| 4 | Light | Lux |
| 5 | CO2 | in ppm |
| 6 | Humidity Ratio, Derived quantity from temperature and relative humidity | kg\text{water-vapor/kg-air} |
| 7 | Occupancy | \{1=yes,0=no\} |

Source: Dataset UCI Learning Machine Accurate occupancy detection.

The dataset for Occupancy Detection from the UCI Learning Machine has a matrix of 8 x 8143 rows and columns. Sample dataset is shown in table 2 below.

| Table 2. Sampel dataset Occupancy Detection matrix 8 x 15 |
|---------|-----------------|
| No | Date | Temperature | Humidity | Light | CO2 | Humidity Ratio | Occupancy |
|---|-----|-------------|--------|------|-----|----------------|---------|
| 1 | 04/02/2015 17:51 | 23.18 | 27.272 | 426 | 721.25 | 0.004793 | 1 |
| 2 | 04/02/2015 17:51 | 23.15 | 27.245 | 419 | 701 | 0.004772 | 1 |
| 3 | 04/02/2015 17:53 | 23.15 | 27.245 | 419 | 701.5 | 0.004779 | 1 |
| 4 | 04/02/2015 17:54 | 23.15 | 27.245 | 419 | 701.5 | 0.004779 | 1 |
| 5 | 04/02/2015 17:55 | 23.1 | 27.2 | 419 | 701.6667 | 0.004757 | 1 |
| 6 | 04/02/2015 17:57 | 23.1 | 27.2 | 419 | 701 | 0.004757 | 1 |
| 7 | 04/02/2015 17:57 | 23.1 | 27.2 | 419 | 689.3333 | 0.004757 | 1 |
| 8 | 04/02/2015 17:58 | 23.1 | 27.2 | 419 | 688 | 0.004757 | 1 |
| 9 | 04/02/2015 17:58 | 23.1 | 27.2 | 419 | 688 | 0.004757 | 1 |
| 10 | 04/02/2015 18:00 | 23.075 | 27.175 | 419 | 687.5 | 0.004745 | 1 |
| 11 | 04/02/2015 18:01 | 23.075 | 27.175 | 419 | 687.5 | 0.004745 | 1 |
| 12 | 04/02/2015 18:02 | 23.1 | 27.1 | 419 | 686 | 0.004745 | 1 |
| 13 | 04/02/2015 18:03 | 23.1 | 27.166667 | 419 | 686 | 0.004745 | 1 |
| 14 | 04/02/2015 18:04 | 23.05 | 27.15 | 419 | 685 | 0.004745 | 1 |
| 15 | 04/02/2015 18:04 | 23 | 27.125 | 419 | 684 | 0.004745 | 1 |

Entropy calculation is done for all segments, then calculate Gain (Humidity), Gain (Light) and finally Gain (CO2). The results can be seen in table 3 below.

| Table 3. Entropy Calculation Analysis and Node Gain |
|-----------|-----------------|
| Node | Attribute | Nilai | Jml Nilai | Jumlah | Nilai | Jml Nilai | Jumlah |
|--------|----------|------|---------|------|--------|---------|------|
| 1      | Temperature | temp<=21.09 | 5481 | 5172 | 309 | 0.312891 |
| 2      |           | temp>21.09 | 2662 | 1242 | 1420 | 0.9967732 |
|        | Humidity | Hum<=27.93125 | 8143 | 6414 | 1729 | 0.209457 |
|        |         | Hum>27.93125 | 5719 | 4617 | 1102 | 0.7070671 |
|        | Light   | Light<200 | 8143 | 6414 | 1729 | 0.003822 |
|        |         | 200<=Light<618.53 | 2075 | 360 | 1715 | 0.6656354 |
|        |         | Light>618.53 | 18 | 4 | 14 | 0.7642045 |
|        |         | Light>927.8 | 8143 | 6414 | 1729 | 0.209457 |
Do the making of roots and leaves as before, so that the overall value of each attribute is not 0. The result is as shown in the picture below.

3. Results And Discussion

For the calculation of accuracy, availability, simplicity and maximum fitness (avg) is calculated based on the conditions so that the results are as follows.

| Node | Attribute | Nilai | Jml Nilai | Jumlah | Entropi | Gain |
|------|-----------|-------|-----------|---------|---------|------|
|      | CO2 < 405.7 | 0     | 0         | 0       | 0       | 0    |
| CO2  | 405.7 ≤ CO2 < 811.4 | 6644  | 6147      | 497     | 0.383602 | 0    |
|      | 811.4 ≤ CO2 < 1217.1 | 1030  | 190       | 840     | 0.6897502 | 0.309736 |
|      | 1217.1 ≤ CO2 < 1622.8 | 268   | 58        | 210     | 0.7535726 | 0    |
|      | CO2 ≥ 1622.8 | 201   | 19        | 182     | 0.4514037 | 0    |

Figure 1. Decision tree

The rules produced by the decision tree have been, as follows:

Table 4 Rules for Decision Tree with C4.5 Algorithm

3. Results And Discussion

For the calculation of accuracy, availability, simplicity and maximum fitness (avg) is calculated based on the conditions so that the results are as follows.

Table 5. Maximum Fitness Calculation Value
From the table shows that the maximum maximum fitness value is 0.68 obtained from calculations on the 2nd chromosome. Furthermore, if the maximum fitness is still too small then the crossing and mutation process will continue to be carried out to obtain an average accuracy of 75% or maximum average difference fitness is smaller than the specified error or $avgi-avgi-1 < \varepsilon$, with $\varepsilon = 0.05$ or iteration reaching 500. In the initial stage. The following values of the three highest rules are as follows:

**Table 6. Accuracy of Rules in Training Data**

| Kromosom ID | Rules | Training Set Accuracy | Fitness  |
|-------------|-------|-----------------------|----------|
| 2           | IF Light<200 THEN Occupancy = No | 100%      | 0.677223933 |
| 12          | IF Light<927.8 AND Light >= 618.53 AND CO2>=405.7 AND CO2<811.4 THEN Occupancy = No | 50%      | 0.26387042    |
| 10          | IF Light<618.53 AND Light >= 200 AND CO2=405.7 AND CO2<811.4 AND Humidity < 27.93235 AND Temperature < 21.09 THEN Occupancy=Yes | 50%      | 0.21387042    |

The selection process has been carried out, the results are shown in Table 4.17. The next step is to do crossing and mutation of decision rules. For rules the results of the selection process are shown in Table 6

**Table 7. Process Results Rules of Selection**

| ID Kromosom | Rules |
|-------------|-------|
| 1           | IF Light<618.53 AND Light >= 200 AND CO2>=405.7 AND CO2<811.4 AND Humidity >= 27.93235 THEN Occupancy=No |
| 2           | IF Light<618.53 AND Light >= 200 AND CO2>=405.7 AND CO2<811.4 AND Humidity >= 27.93235 AND Temperature >= 21.09 THEN Occupancy=No |
| 3           | IF Light<618.53 AND Light >= 200 AND CO2>=405.7 AND CO2<811.4 AND Humidity >= 27.93235 AND Temperature >= 21.09 THEN Occupancy=No |
| 4           | IF Light<618.53 AND Light >= 200 AND CO2>=405.7 AND CO2<811.4 AND Humidity >= 27.93235 AND Temperature >= 21.09 THEN Occupancy=Yes |
| 5           | IF Light<618.53 AND Light >= 200 AND CO2>=405.7 AND CO2<811.4 AND Humidity >= 27.93235 THEN Occupancy=No |
| 6           | IF Light<618.53 AND Light >= 200 AND CO2>=405.7 AND CO2<811.4 AND Humidity >= 27.93235 AND Temperature >= 21.09 THEN Occupancy=No |
| 7           | IF Light<618.53 AND Light >= 200 AND CO2>=405.7 AND CO2<811.4 AND Humidity >= 27.93235 AND Temperature >= 21.09 THEN Occupancy=No |
| 8           | IF Light<200 THEN Occupancy = No |
| 9           | IF Light<927.8 AND Light >= 618.53 AND CO2>=405.7 AND CO2<811.4 THEN Occupancy = No |
| 10          | IF Light<200 THEN Occupancy = No |
| 11          | IF Light<927.8 AND Light >= 618.53 AND Humidity<27.931 THEN Occupancy = Yes |
| 12          | IF Light<200 THEN Occupancy = No |
| 13          | IF Light<927.8 AND Light >= 618.53 AND Temperature<21.09 THEN Occupancy = No |
| 14          | IF Light<618.53 AND Light >= 200 AND CO2>=405.7 AND CO2<811.4 AND Humidity >= 27.93235 THEN Occupancy=No |

**Table 8. Results of chromosome conversion to rules**

| ID Kromosom | Rules |
|-------------|-------|
| 1           | IF Light<618.53 AND Light>=200 AND CO2<811.4 AND CO2>=405.7 AND Hum>=27.93125 THEN No |
| 2           | IF Light<618.53 AND Light>=200 AND CO2<811.4 AND CO2>=405.7 AND Hum>=27.93125 AND temp>=21.09 THEN No |
| 3           | IF Light<618.53 AND Light>=200 AND CO2<811.4 AND CO2>=405.7 AND Hum>=27.93125 AND temp>=21.09 THEN No |
| 4           | IF Light<618.53 AND Light>=200 AND Hum>=27.93125 AND temp<21.09 THEN No |
| 5           | IF Light<618.53 AND Light>=200 AND CO2<811.4 AND CO2>=405.7 AND Hum>=27.93125 THEN No |
| 6           | IF Light<618.53 AND Light>=200 AND CO2<811.4 AND CO2>=405.7 AND Hum>=27.93125 AND temp>=21.09 THEN No |
| 7           | IF Light<618.53 AND Light>=200 AND CO2<811.4 AND CO2>=405.7 AND Hum>=27.93125 AND temp>=21.09 THEN No |
| 8           | IF Light<200 THEN No |
| 9           | IF Light<618.53 AND Light>=200 AND CO2<811.4 AND CO2>=405.7 THEN No |
| 10          | IF Light<200 THEN No |
| 11          | IF Light<618.53 AND Light>=200 AND Hum<27.93125 THEN Yes |
| 12          | IF Light<618.53 AND Light>=200 AND CO2<811.4 AND CO2>=405.7 THEN Yes |
| 13          | IF Light<618.53 AND Light>=200 THEN No |
| 14          | IF Light<200 AND 405.7<=CO2<811.4 AND Hum>=27.93125 THEN No |
The process of calculating the maximum fitness is done and accuracy by re-matching the search value to the existing chromosome such as the previous fitness maximum calculation. Then after the calculation process is done it is generated like table 9 below.

### Table 9. Maximum first generation fitness

|    | 1   | 2   | 3   | 4   | 5   | 6   | 7   | 8   | 9   | 10  | 11  | 12  | 13  | 14  |
|----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Accuracy  | 0.33 | 0.25 | 0.25 | 0.67 | 0.33 | 0.25 | 0.5  | 1   | 0.5 | 0.33 | 0.5 | 0.25 | 0   | 0.67 |
| Number of Attributes | 3    | 4    | 4    | 3    | 4    | 4    | 1    | 2   | 3   | 2   | 4   | 1   | 3   |
| Availability    | 0.33 | 0.25 | 0.25 | 0.67 | 0.33 | 0.25 | 0.5  | 1   | 0.5 | 0.33 | 0.5 | 0.25 | 0   | 0.67 |
| Simplicity      | 0.25 | 0    | 0.25 | 0.25 | 0    | 0    | 0.75 | 0.5 | 0.25 | 0.5 | 0    | 0.75 | 0.25 |
| Gain            | 0.08 | 0.08 | 0.08 | 0.08 | 0.08 | 0.08 | 0.57 | 0.16 | 0.57 | 0.08 | 0.16 | 0.08 | 0.57 |
| Max Fitnes      | 0.15 | 0.1  | 0.21 | 0.15 | 0.1  | 0.15 | 0.68 | 0.26 | 0.49 | 0.2 | 0.16 | 0.13 | 0.56 |

From the calculation of table 9 maximum fitness, it can be seen that in the first generation the maximum value calculation is still the same as before by 0.68 but for the total maximum fitness has increased from the previous 2,191 to 3.47, meaning that the maximum fitness value for the other chromosomes is getting better than some chromosomes have better accuracy values. Do the regeneration process to average-avgi-1 <ε or up to 500 generations or more than 75% accuracy. After 20th regeneration, the maximum fitness value is obtained, as follows:

### Table 10. Generation 20 fitness maximum results

|    | 1   | 2   | 3   | 4   | 5   | 6   | 7   | 8   | 9   | 10  | 11  | 12  | 13  | 14  |
|----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Accuracy  | 1   | 0.25 | 0.75 | 0.75 | 0.33 | 0.75 | 0.5  | 1   | 1   | 0.67 | 0.5 | 0.25 | 1   | 0.67 |
| Number of Attributes | 3    | 4    | 4    | 3    | 4    | 4    | 2    | 2   | 3   | 2   | 4   | 1   | 3   |
| Availability    | 1    | 0.5  | 0.5  | 0.75 | 0.67 | 0.25 | 0.5  | 1   | 0.5 | 0.67 | 1   | 0.5 | 1   | 0.67 |
| Simplicity      | 0.25 | 0    | 0    | 0.25 | 0    | 0    | 0.5  | 0.5 | 0.25 | 0.5 | 0    | 0.75 | 0.25 |
| Gain            | 0.08 | 0.08 | 0.65 | 0.67 | 0.08 | 0.08 | 0.57 | 0.16 | 0.57 | 0.08 | 0.16 | 0.57 | 0.57 |
| Max Fitnes      | 0.28 | 0.13 | 0.58 | 0.62 | 0.18 | 0.15 | 0.15 | 0.65 | 0.31 | 0.56 | 0.25 | 0.19 | 0.68 | 0.56 |

After carrying out several hybrid stages the genetic algorithm in optimizing the C4.5 algorithm has produced several new rules, as follows:

Step 1. Enter the ruleset, datatesting, gain;
Step 2. Initialize \( p_c = 0.85, p_m = 0.09, \) max_generation = 500, a = 0.1, b = 0.1, c = 0.1, d = 0.7, \( \varepsilon = 0.05, \) mean_acuration = 0.75, avg0 = 0, \( i = 0, \) population = encode (ruleset);
Step 3. Perform the selection process, population = selection (population, max_fitness);
Step 4. Generate random numbers \( R_n, n = \) count (ruleset), datacross = (\( R_i < p_c \));
Step 5. Do crosses, population = crosses (populasidatacross);
Step 6. Generate a random number \( R_{nxm}, m = \) column (datatesting) x3, datamutmn = (\( R_{nxm} < p_m \));
Step 7. Perform mutations, population = mutations (populasidatamut);
Step 8. Connect datatesting with the ruleset, calculate max_fitnes, agvt, rata_akurasi, t = t + 1;
Step 9. Do Step 3, if avgt-avgt-1 > \( \varepsilon, t \neq \) max_generation or rata_akurasi \( \neq \) mean_akurasai;
Step 10. Change population to ruleset, ruleset = decode (population);
Step 11. Generate ruleset;

Based on the results of the last calculation, the following rules are obtained:

### Table 11. The last rule is the result of mutation

| Rules                               | Accuracy |
|-------------------------------------|----------|
| IF Light<200 THEN No                | 100%     |
| IF Light<618.53 AND Light>=200 AND Hum<27.93125 THEN No | 75%     |
| IF Light<618.53 AND Light >=200 AND CO2>=405.7 AND CO2<811.4 AND Humidity > 27.93235 THEN Occupancy=No | 75%     |
The results of the evaluation of the genetic algorithm and C4.5 process get the results of rules that are simpler and have higher accuracy values. The use of C4.5 with Hybrid genetic algorithms in the search for more effective rules requires more and longer understanding. But the use of these two algorithms will be very effective if the case faced is very complex with more branching conditions with high accuracy.

4. Conclusion
The conclusions that can be drawn from this research are:
   a. Genetic Algorithms are able to optimize rules search space with the number 20th generation
   b. With the genetic algorithm in the C4.5 algorithm, it is able to increase the number of chromosomes so that the final result has good accuracy above 75%
   c. The use of genetic algorithms is able to optimize rules search space and simplify combinations of rules that are more complex.

References
[1] Sudirman, A. P. Windarto, and A. Wanto, “Data mining tools | rapidminer: K-means method on clustering of rice crops by province as efforts to stabilize food crops in Indonesia,” IOP Conf. Ser. Mater. Sci. Eng., vol. 420, p. 12089, 2018.
[2] B. Supriyadi, A. P. Windarto, T. Soemartono, and Mungad, “Classification of natural disaster prone areas in Indonesia using K-means,” Int. J. Grid Distrib. Comput., vol. 11, no. 8, pp. 87–98, 2018.
[3] S. Alharbi and I. Venkat, “A Genetic Algorithm Based Approach for Solving the Minimum Dominating Set of Queens Problem,” J. Optim., vol. 2017, pp. 1–8, 2017.
[4] P. K. Arora, A. Haleem, M. K. Singh, H. Kumar, and M. Kaushik, “Design of a Production System Using Genetic Algorithm,” Procedia Technol., vol. 14, pp. 390–396, 2014.
[5] J. McCall, “Genetic algorithms for modelling and optimisation,” J. Comput. Appl. Math., vol. 184, no. 1, pp. 205–222, 2005.
[6] S. K. Chaturvedi, “Attack Detection over Network based on C45 & RF Algorithms,” vol. 15, no. 9, pp. 68–73, 2015.
[7] J. Kratochvil, J. Petřů, M. Pagáč, J. Holubják, and J. Mrazik, “Effect Of Chip Breakers On The Cutting Force During The Machining Of Steel C45,” vol. 11, no. March, pp. 173–178, 2017.
[8] T. Leppert, “Surface Layer Properties In Dry Turning Of C45 Steel.”