Implementation of MD Heuristic Method for Classifying Numerical Data In Data Preprocessing

Sri Wahyun1,3*, Muhammad Zarlis2, Solikhun1, Deny Jollyta1, M. Safii1, Indri Sulistianingsih3

1 Student of Faculty Computer Science in Universitas Sumatera Utara
2 Lecturer of Faculty Computer in Universitas Sumatera Utara
3 Lecturer of Sains and Technology in Universitas Pembangunan Panca Budi, Medan, Indonesia

* sriwahyuni@dosen.pancabudi.ac.id

Abstract. Implementation of Data Mining with various methods will generate new knowledge and information in the form of rules which is very useful, but in implementation of data mining experience some weakness in data processing process. One weakness is that not all variable data can be easily processed, if the data is processed has a numerical value or data in the form of numbers then the division of classification is very difficult to do. If the classification of data is so many variables then the branch on the decision tree is also very much and result in less accurate results of rules or knowledge generated. MD Heuristic method can be used to classify numerical data or numbers that have a large range of data. By implementing the MD Heuristic Method in classifying the data in numerical form it can easily classify the data to make the data classification more precise and accurate, by calculating and averaging between the upper and lower bounds of the large data set. The use of MD Heuristic can facilitate preprocessing data with decision tree algorithm C4.5. The result of this research is detainee data at Labuhan Deli Detention Center, that is the data of age of detainee can be classified with more accurate to facilitate in process of data and more shorten rule or knowledge in result of decision tree image.

1. Introduction

Data Mining is the process of mining large amounts of data stored in the database, the purpose is to find out the hidden information from the results of extracting and mining [1]. In Data Mining there are several different methods and algorithms according to their respective purposes and functions [2]. One of the methods is the Classification Method with the Decision Tree C4.5 algorithm or commonly called the Decision Tree, because the rule or knowledge generated in the form of a decision tree image can be read and easily understood by the public [3]. However, in some cases data mining data processing can’t be processed by data mining especially with decision tree C4.5 algorithm method, because the data in the form of numeric or numbers and has a huge set of data sets, for example is age data and income data that has a large number of scores, making it difficult to do classification on data preprocessing before data can be processed with Data Mining [4]. Therefore, for the data can be classified to be easier and more accurate classification results can then apply the method of MD Heuristic in classifying data. In the detainee data, especially the data of detainee age at Labuhan Deli Detention Center.
Appropriate implementation to overcome the problem is with MD Heuristic method with the aim of classifying data that is in detainee data at Labuhan Deli Detention Center. In detainee age data where the set of data sets has a large range that is difficult to classify, will result in large and irregular branching in the decision tree so that the rule or knowledge is difficult to obtain information. Before the data mining process can be implemented, it is necessary to do the cleaning process on the data that became the focus of KDD. The cleaning process includes removing data duplication, checking inconsistent data, and correcting data errors, such as typographical errors. Enrichment process is also carried out, the process of "Data Enrichment" existing data with other relevant data or information required for KDD, such as external data or information.

1.1. Grouping Data Mining
Data mining is divided into several groups based on the task that can be done, namely [5]:
- Description.
- Estimates.
- Prediction.
- Classification.
- Clustering.
- Association.

1.2. MD Heuristic Method
The heuristic searching technique is a strategy to perform a selective state space search process which guides the search process that we perform along the path of the greatest possible success, and overrides the time-consuming and non-effective. Heuristics is a technique that develops efficiency in the search process, but with the possibility of sacrificing completeness. In order to apply such heuristics well in a particular domain, a Heuristic Function is required. This heuristic function is used to evaluate the state of the individual problem situation and determine how far it can be used to get the desired solution.
Types of Heuristic Searching:
- Generate and Test.
- Hill Climbing.
- Best First Search.
- Alpha Beta Prunning.
- Means-End-Anlysis.
- Constraint Satisfaction.

2. Methodology
In general, the research phase can be seen in Figure 1.

![Figure 1. Research stages](image-url)
2.1. Data Collection

The data taken has a format in Microsoft Excel data taken as many as 30 detainee data taken randomly on Detention Center, with the number of 12 items set of variable age of detainees.

### Table 1. Sample of Detainee Data

| No | Registration Number | Age | Education | Job/Vocation | Address       | Case     |
|----|---------------------|-----|-----------|--------------|---------------|----------|
| 1  | BI 206/09           | 28  | SMP       | Jobless      | Ps. Tuan      | Drugs    |
| 2  | BI 245/10           | 27  | SD        | Jobless      | Medan Deli    | Drugs    |
| 3  | BI 436/10           | 27  | SMA       | Jobless      | H. Perak      | Drugs    |
| 4  | BI 440              | 41  | SMP       | Jobless      | Medan Labuhan| Drugs    |
| 5  | BI 453/10           | 30  | SMP       | Driver       | Medan Labuhan| Drugs    |
| 6  | BI 455/10           | 47  | SMA       | Jobless      | Medan Deli    | Drugs    |
| 7  | BI 490/10           | 32  | SMA       | Jobless      | Ps. Tuan      | Drugs    |
| 8  | BI 491/10           | 29  | SD        | Jobless      | Sunggal       | Drugs    |
| 9  | BI 508/10           | 31  | SMP       | Jobless      | Medan Deli    | Drugs    |
| 10 | BI 50/11            | 37  | SD        | Laborer      | Medan Belawan| Drugs    |
| 11 | BI 525/10           | 41  | SMP       | Jobless      | Medan Labuhan| Drugs    |
| 12 | BI 187/12           | 33  | SMA       | Employees    | Kec. Medan Deli| Not Drugs|
| 13 | BI 280/11           | 25  | SMP       | Entrepreneur | Kec. Ps. Tuan| Not Drugs|
| 14 | BI 283/11           | 35  | SMP       | Entrepreneur | Kec. Ps. Tuan| Drugs    |
| 15 | BI 289/11           | 21  | SD        | Jobless      | Kec. Medan Belawan| Not Drugs|
| 16 | BI 290/11           | 39  | SD        | Jobless      | Kec. Medan Deli| Drugs    |
| 17 | BI 291/11           | 34  | SMA       | Entrepreneur | Kec. Ps. Tuan| Not Drugs|
| 18 | BI 296/11           | 24  | SMP       | Jobless      | Kec. Ps. Tuan| Not Drugs|
| 19 | BI 306/11           | 34  | SMA       | Entrepreneur | Kec. Ps. Tuan| Drugs    |
| 20 | BI 335/11           | 21  | SD        | Jobless      | Kec. Medan Deli| Drugs    |

2.2. Process

In this study will be discussed specifically to perform pre-process data grouping especially data Age of detainees in Labuhan Deli Detention Center. Age grouping based on transformation with MD Heuristic algorithm implementation as follows:

a. Set of Cuts

\[ A = \{ 21, 24, 25, 27, 28, 29, 30, 33, 34, 37, 41, 47 \} \]

The set of cuts variable table is shown in table 2.

### Table 2. Variable

| Variable | Value       |
|----------|-------------|
| Pa1      | (21-24)     |
| Pa2      | (24-25)     |
| Pa3      | (25-27)     |
| Pa4      | (27-28)     |
| Pa5      | (28-29)     |
| Pa6      | (29-30)     |
| Pa7      | (30-33)     |
| Pa8      | (33-34)     |
| Pa10     | (37-41)     |
| Pa11     | (41-47)     |

b. Next form the Discernibility Formula as follows:

\[ \psi \left( 4, 2 \right) = a4, a5, a6, a7, a8, a9, a10 \]
\[ \psi \left( 4, 3 \right) = a4, a5, a6, a7, a8, a9, a10 \]
\[ \psi \left( 4, 5 \right) = a7, a8, a9, a10 \]
\[ \psi \left( 4, 7 \right) = a6, a7, a8, a9, a10 \]
\[ \psi \left( 4, 8 \right) = a10 \]
\[ \psi \left( 4, 9 \right) = - \]
\[ \psi \left( 4, 11 \right) = a3, a4, a5, a6, a7, a8, a9, a10 \]
\[ \psi \left( 4, 14 \right) = a9, a10 \]
The International Conference on Computer Science and Applied Mathematic

IOP Conf. Series: Journal of Physics: Conf. Series 1255 (2019) 012060   doi:10.1088/1742-6596/1255/1/012060

\[ \psi (13, 1) = a5, a6, a7, a8, a9, a10 \]  \[ \psi (13, 2) = a4, a5, a6, a7, a8 \]  \[ \psi (13, 3) = a4, a5, a6, a7, a8 \]  \[ \psi (13, 4) = a9 \]  \[ \psi (13, 5) = a7, a8 \]  \[ \psi (13, 6) = a6, a7, a8 \]  \[ \psi (13, 7) = a6, a7, a8 \]  \[ \psi (13, 8) = a9 \]  \[ \psi (13, 9) = a9, a10 \]  \[ \psi (13, 10) = a9 \]  \[ \psi (13, 11) = a3, a4, a5, a6, a7, a8 \]  \[ \psi (13, 12) = a3, a4, a5, a6, a7, a8 \]  \[ \psi (13, 13) = a3, a4, a5, a6, a7, a8 \]  \[ \psi (13, 14) = a9 \]  \[ \psi (13, 15) = a2, a3, a4, a5, a6, a7, a8 \]  \[ \psi (13, 16) = a9 \]

c. Form a Table representation of the Discernibility Formula, as shown in table 4.

|        | PA 1 | PA 2 | PA 3 | PA 4 | PA 5 | PA 6 | PA 7 | PA 8 | PA 9 | PA 10 | PA 11 |
|--------|------|------|------|------|------|------|------|------|------|-------|-------|
| 4,1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1     | 1     |
| 4,2    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1     | 1     |
| 4,3    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1     | 1     |
| 4,4    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1     | 1     |
| 4,5    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1     | 1     |
| 4,6    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1     | 1     |
| 4,7    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1     | 1     |
| 4,8    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1     | 1     |
| 4,9    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1     | 1     |
| 4,11   | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1     | 1     |
| 4,12   | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1     | 1     |
| 4,13   | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1     | 1     |
| 4,14   | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1     | 1     |
| 4,15   | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1     | 1     |
| 4,16   | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1     | 1     |
| 6,1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1     | 1     |
| 6,2    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1     | 1     |
| 6,3    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1     | 1     |
| 6,4    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1     | 1     |
| 6,5    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1     | 1     |
| 6,6    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1     | 1     |
| 6,7    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1     | 1     |
| 6,8    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1     | 1     |
| 6,9    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1     | 1     |
| 6,10   | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1     | 1     |
| 6,11   | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1     | 1     |
| 6,12   | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1     | 1     |
| 6,13   | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1     | 1     |
| 6,14   | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1     | 1     |
| 6,15   | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1     | 1     |
| 6,16   | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1     | 1     |
| 10,1   | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1     | 1     |
| 10,2   | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1     | 1     |
| 10,3   | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1     | 1     |
| 10,4   | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1     | 1     |
| 10,5   | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1     | 1     |
| 10,6   | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1     | 1     |
| 10,7   | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1     | 1     |

Table 3. Representation of Discernibility An example
d. The result from table 4 is then selected in the column with the highest number 1, then delete the column and row that has the value 1 on the deleted column, repeat the process continuously until the value 1 in the table becomes empty.

e. After deletion is done on the variable set of cuts column, after the process is sequenced at the above stages, the order of deletion is as follows: Pa1, Pa7, Pa8, Pa9, Pa10, Pa11

Based on the order of occurrence of the 1 most value then from the order of occurrence of the highest to the lowest in the input value on set Of Cut variable value in add then divided into 2 as it adds two values. Then rounding off the number 0.5 to facilitate the process of classification of the data. Then the results obtained are as follows:

\[
\begin{align*}
\text{Pa1} &= \frac{(21 + 24)}{2} = 22.5 \text{ rounding to 23} \\
\text{Pa7} &= \frac{(30 + 33)}{2} = 31.5 \text{ rounding up to 32} \\
\text{Pa8} &= \frac{(33 + 34)}{2} = 33.5, \text{ rounding to 34} \\
\text{Pa9} &= \frac{(34 + 37)}{2} = 35.5 \text{ rounding to 36} \\
\text{Pa10} &= \frac{(37 + 41)}{2} = 39 \\
\text{Pa11} &= \frac{(41 + 47)}{2} = 44
\end{align*}
\]

3. Result and Discussion

The result of the classification of detainee data on the variable age of the detainee is using MD Heuristic Method to be less than before the variables there are 12 data variables that exist prior to classification using MD Heuristic method will result in 12 classifications of data that if processed directly in Data Mining then with a variable that has a classification of 12 sets of data then the results of data mining will also have difficulty in the formation of decision trees or decision tree and the rule
or knowledge generated is very difficult to know the knowledge and dominant factors to be sought. After pre-processing with MD Heuristic method it can be seen on table 5.

| Table 4. The Final Result of Classification Ages |
|-----------------------------------------------|
| Age   | Classification |
|-------|----------------|
| *23   | 1              |
| 23 – 32 | 2            |
| 33 – 34 | 3            |
| 35 – 36 | 4            |
| 37 – 39 | 5            |
| 40 – 44 | 6            |
| 44*   | 7              |

Information:
- Classification 1 = * 23 is the age in the classification 1 under the age of 23, while the age of 23 years is in the classification 2.
- Classification 7 = 44 * is age above the age of 44 that is age 45 and so on, while age 45 is classification 6.

So there is a final conclusion there are 7 classifications of data aged detainee in Labuhan Deli Detention Centre after Preprocessing with MD Heuristic Method previously done Preprocess with MD Heuristic Method there are 12 classifications.

4. Conclusion

Based on the results obtained from this study using classification calculation method of MD Heuristic can produce smaller classification after data preprocessing that is become 7 classification which previously exist 12 classification. The classification of too many sets of items in the Data Mining process will result in Decision Tree data with difficult to read rules and knowledge and the number of branches generated by the reader. With a smaller classification will generate a decision tree that can be easily understood so that the rules and new knowledge generated can be useful and can be used as solution of the problem faced.

References

[1] Wahyuni, S. 2018 “Implementation Of Data Mining To Analyze Drug Cases Using C4.5 Decision Tree”, Journal of Physics: Conference Series. IOP Conf. Series: Journal of Physics: Conf. Series 970 (2018) 012030. http://doi:10.1088/1742-6596/970/1/012030.

[2] Wahyuni, S. 2018. “Implementasi Metode Decision Tree C4.5 Untuk Menganalisa Mahasiswa Drop Out”. ETHOS Jurnal Penelitian dan Pengabdian. Vol. 6, No.1:39-48, 2018.

[3] H. Siahaan, H. Mawengkang, S. Efendi, A. Wanto, and A. P. Windarto, “Application of Classification Method C4.5 on Selection of Exemplary Teachers,” in IOP Conference Series, 2018, pp. 1–6

[4] Kusrini, Emha Taufiq Luthfi. “Algoritma Data Mining”, Penerbit CV. Andi Offset, 2009.

[5] S. 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 Conference Series: Materials Science and Engineering, vol. 420, no. 12089, pp. 1–8, 2018

[6] Garcia, S., Luengo, J. and Herrera, F. (2016) ‘Tutorial on practical tips of the most influential data preprocessing algorithms in data mining’, Knowledge-Based Systems, 98, pp. 1–29. doi: 10.1016/j.knosys.2015.12.006.

[7] Hariyanto, Eko., Rahim, Robbi., “Arnold’s Cat Map Algorithm in Digital Image Encryption”, International Journal of Science and Research (IJSR), Vol 5 Issue 10, 2016, pp 1363-1365, https://doi.org/10.21275/ART20162488.