Login : Jurnal Teknologi Komputer
ISSN : 2302-9692 (print) | 2723-8695 (online)
Vol. 14, No. 1, June 2020, pp. 7-12
LOGIN http://login.seaninstitute.org/index.php/Login

Analysis of Patient Disease Trends Based on Medical Record Data Using the C4.5 Algorithm

Eva Darnila, Hafiz Al-Kautsar, Yaumil Iksan
Department of Informatic, Universitas Malikussaleh, Indonesia

Article Info

Article history:

Received Apr 16, 2020
Revised May 20, 2020
Accepted Jun 11, 2020

Keywords:

Data mining,
C4.5 Algorithm,
Medical Record,
Disease Trends,
Bireuen.

ABSTRACT

Data mining is the process of finding patterns from large data sets using descriptive, estimation, prediction, classification, clustering and association techniques. The C4.5 algorithm is one of the most popular algorithms of the classification method in data mining which is the development of the Iterative Dichotomizer 3 (ID3) algorithm. By using the algorithm C4.5, the authors are interested in conducting research on medical record data that is in the Regional General Hospital dr. Fauziah Bireuen. A medical record is a file containing notes and documents regarding the patient's identity, examination results, medication, actions and other services that have been provided to patients. The purpose of this study was to find trends in patient disease using the C4.5 algorithm based on 4 data variables, namely age, gender, address and diagnosis. From the research results, it was found that disease trends in children, adults and the elderly in all Bireuen regions and all genders were F00-F99, namely mental and behavioral disorders. While the disease trend for infants is A00-B99, namely certain infections and parasitic diseases. Then for adolescents in North Bireuen and West Bireuen the emerging disease trend is F00-F99, namely mental and behavioral disorders, while for adolescents in South Bireuen who are male, the emerging disease trend is I00-I99, namely diseases of the car and mastid process, while adolescents in South Bireuen are female, the trend of disease that appears is M00-M99, namely diseases of the musculoskeletal system and connective issue.

Corresponding Author:

Eva Darnila,
Department of Informatic, Universitas Malikussaleh,
Tengku Nie, Cot Rd, Reuleut Tim., Muara Batu, Kabupaten Aceh Utara, Aceh.
Email: eva.darnila@unimal.ac.id

1. INTRODUCTION

Data mining is the process of finding or extracting new patterns from large data sets involving methods from statistical data and artificial intelligence. Data Mining is an activity that includes the collection and use of historical data that determines the cohesiveness, patterns and relationships in large data sets. In data mining, there are several techniques that can be used, namely description, estimation, prediction, classification, clustering and association [1].

Classification is the process of placing certain objects into a set of categories, based on each object's property. The classification process is based on four fundamental components, namely class, predictor, training set and data set testing [2].

This technique can classify new data by manipulating the classified data by using the results to provide a number of rules. These rules are used for new data to be classified. This technique uses supervised induction, which takes advantage of a collection of tests from classified records to define additional classes.

Among the most popular classification models are Decision Trees, Bayesian Classifier, Neural Networks, Statistical Analysis, Genetic Algorithms, Rough Sets, K-Nearest Neighbor Classifier, Rule-based Methods, Memory Based Reasoning and Support Vector Machines [2].

Among the several methods that can be used for classification is the decission tree method. The decision tree method is a method that can transform very large facts into a decision tree that represents rules. Rules can be easily understood with natural language [3].

A decision tree is a structure that can be used to divide large data sets into smaller record sets by applying a set of decision rules. With each set of dividers, the members of the result set are similar to each other [3].

There are many algorithms that can be used in the formation of a decision tree, including ID3, CART, and C4.5. The C4.5 algorithm is a very powerful and well-known method of classification and prediction. The C4.5 algorithm is a development of the Iterative Dichotomizer 3 (ID3) algorithm. The C4.5 algorithm has the same basic working principles as the ID3 algorithm, it's just that the C4.5 algorithm uses an induction approach where the C4.5 algorithm divides data based on the selected criteria to make decision tree [4].

By using the C4.5 algorithm, the authors conducted an approach to research patient disease trends by using medical record data or medical record data of patients in hospital.

A medical record is a file containing notes and documents regarding the patient's identity, examination results, medication, actions and other services that have been provided to patients. This medical record data is used by doctors and paramedics as a consideration for medical services, for example drug administration and treatment methods and so on [5].

2. RESEARCH METHOD

In this study using the C4.5 algorithm. There are 4 steps in determining a decision tree using the C4.5 algorithm, [6]. In this case, we choosing an attribute as the root (root), creating a branch for each value. Dividing cases into branches and repeating the process in each branch, until all cases are in the branch have the same class.

To select the root attribute, it is based on the highest gain ratio value of the existing attributes. To calculate gain, a formula is used as shown in the following equation.

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

Where,

S : Case Set
A : Attribute
N : Jumlah partisi atribut A
|Si| : Number of partitions attribute A
|S| : Number of cases in S

Meanwhile, the calculation of the entropy value can be seen in the following equation:

$$Entropy(S) = \sum_{i=1}^{n} -pi * log_2 pi \quad (2)$$

S : case set
N : number of partitions S
pi : the proportion of Si to S

Meanwhile, the calculation of the split info value can be done using the following equation:

$$SplitInfo(S,A) = \sum_{i=1}^{n} \frac{\mathrm{Si}}{\mathrm{S}} * \log_2 \frac{\mathrm{Si}}{\mathrm{S}} \quad (3)$$

S : Case set
A : Attribute
Si : Number of samples for attribute A

Then to calculate the gain ratio, it can be done using the following equation:

$$GainRatio(S,A) = \frac{Gain(S,A)}{SplitInfo(S,A)} \quad (4)$$

S : Case Set

A : Attribute
Gain(S,A) : Information gain on attribute A
SplitInfo(S,A) : Split information on attribute A

The system scheme for data processing using algorithms is C4.5 as follows

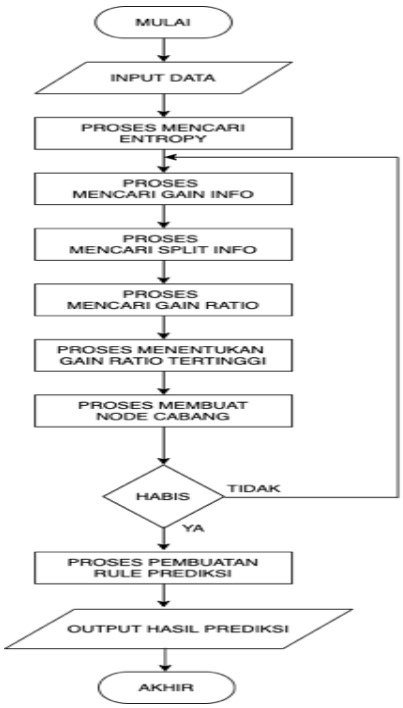

Figure 1. System Schematic

Medical record data that has been previously taken from the research site are then classified and transformed into a data format that is in accordance with what is needed by the system, namely by removing unnecessary fields and discarding incomplete data, so that the transformation results are obtained with the format as in the table 1.

Table 1. Transformation Result Data

No	Gender	Age	Adress	Diagnosis
1	Male	Adult	Bireuen Utara	O00-O99
2	Male	Children	Bireuen Utara	H00-H59
…	…	…	…	…
5959	Female	Baby	Bireuan Selatan	Z00-Z99

3. RESULTS AND DISCUSSION

From the calculation step, the search results are obtained as in table (2) tabulation of the root node search results data. From the search results, it was found that the variable that had the highest gain ratio was the age group attribute with a gain ratio of 0.045734219. So that the age group variable will become the root node with 5 branches, namely the attributes that exist in the age group variable, namely babies, children, adolescents, adults and the elderly.

Table 2. Tabulation of Root Node Search Results

			A00-B99	…	Z00-Z99	Entropi	Gain	Split Info	Gain Ratio
Total		5959	274	…	249	3.295822276			
Gender							0.01276539	0.948610512	0.013456935
	Male	3770	175	…	183	3.263954286			
	Female	2189	99	…	66	3.315956295			
Group Age							0.05623664	1.229640332	0.045734219
	Baby	19	18	…	0	0.297472249			

	Children	140	7	…	0	1.551784107			
	Youth	321	20	…	3	3.223084207			
	Adult	4140	191	…	196	3.315167372			
	Elderly	1339	38	…	50	3.228070281			
Address							0.020527226	1.293494158	0.015869593
	Bireuen Utara	3723	173	…	149	3.205308554			
	Bireuen Barat	1523	70	…	66	3.369322984			
	Bireuen Selatan	713	31	…	34	3.439888574			

From the search results for node 1.1, the variable with the highest gain ratio is the address variable with a gain ratio of 0.023388715. So that the address variable will become node 1.1 with 3 branches, namely the attributes in the address variable, namely North Bireuen, South Bireuen and West Bireuen.

Table 3. Tabulation of Node 1.1 Search Results (Adult Partitions)

			A00-B99	…	Z00-Z99	Entropi	Gain	Split Info	Gain Ratio
Total		4140	191	…	196	3.315167372			
Gender							0.017153046	0.937681944	0.018293032
	Male	2674	106	…	148	3.246416931			
	Female	1466	85	…	48	3.39212854			
Address							0.02963311	1.266983229	0.023388715
	Bireuen Utara	2655	126	…	113	3.199435894			
	Bireuen Barat	1006	42	…	55	3.39889769			
	Bireuen Selatan	479	23	…	28	3.524673222			

From the results of the search for node 1.2, it is found that the variable with the highest gain ratio is the address variable with a gain ratio of 1.685775259. So that the address variable will become node 1.2 with 3 branches, namely the attributes in the address variable, namely North Bireuen, South Bireuen and West Bireuen.

Table 4. Tabulation of Node 1.2 Search Results (Youth Partition)

			A00-B99	…	Z00-Z99	Entropi	Gain	Split Info	Gain Ratio
Total		321	20	…	3	3.223084207			
Gender							0.180076593	0.978718372	0.183992248
	Male	188	14	…	2	3.106800459			
	Female	133	6	…	1	2.952834269			
Address							2.444628588	1.450150947	1.685775259
	Bireuen Utara	153	16	…	2	3.168759931			
	Bireuen Barat	120	4	…	1	2.995759769			
	Bireuen Selatan	48	0	…	0	2.594899095			

From the results of the search for node 1.3, it is found that the variable that has the highest gain ratio is the address variable with a gain ratio of 0.2693954. So that the address variable will become node 1.3 with 3 branches, namely the attributes in the address variable, namely North Bireuen, South Bireuen and West Bireuen.

Table 5. Tabulation of Node 1.3 Search Results (Partition of Children)

			A00-B99	…	Z00-Z99	Entropi	Gain	Split Info	Gain Ratio
Total		140	7	…	0	1.551784107			
Gender							0.080062726	0.898058793	0.089150874
	Male	96	5	…	0	1.423793797			
	Female	44	2	…	0	1.576290657			
Addres							0.405958142	1.506923066	0.2693954
	Bireuen Utara	69	6	…	0	1.663407928			

	Bireuen Barat	36	0	…	0	0.030052853
	Bireuen Selatan	35	1	…	0	1.273102441

From the search results for node 1.4, it is found that the variable with the highest gain ratio is the address variable with a gain ratio of 0.21308411. So that the address variable will become node 1.4 with 3 branches, namely the attributes in the address variable, namely North Bireuen, South Bireuen and West Bireuen.

Table 6. Tabulation of Node 1.4 Search Results (Elderly Partition)

			A00-B99	…	Z00-Z99	Entropi	Gain	Split Info	Gain Ratio
Total		1339	38	…	50	3.228070281			
Gender							0.056153152	0.974091737	0.057646677
	Male	796	35	…	33	3.288290461			
	Female	543	3	…	17	3.001321969			
Address							0.124796003	0.585665461	0.21308411
	Bireuen Utara	838	17	…	34	3.131295036			
	Bireuen Barat	357	21	…	10	3.181401082			
	Bireuen Selatan	144	0	…	6	2.746519665			

From the search results for node 1.5, it is found that the variable with the highest gain ratio is the address variable with a gain ratio of 0.082826277. So that the address variable will become node 1.5 with 3 branches, namely the attributes in the address variable, namely North Bireuen, South Bireuen and West Bireuen.

Table 7. Tabulation of Node 1.5 Search Results (Baby Partitions)

			A00-B99	C00-D48	Z00-Z99	Entropi	Gain	Split Info	Gain Ratio
Total		19	18	0	0	0.297472249			
Gender							0.013438509	0.629249224	0.021356417
	Male	16	15	0	0	0.337290067			
	Female	3	3	0	0	0			
Addres							0.126676854	1.529428332	0.082826277
	Bireuen Utara	8	8	0	0	0			
	Bireuen Barat	4	3	0	0	0.811278124			
	Bireuen Selatan	7	7	0	0	0			

After obtaining the variables that will become node 1.1, node 1.2, node 1.3, node 1.4 and node 1.5, the next step is to look for the next node from each of these partitions. Since the only variable available is gender, the gender will automatically be the nodes of the branches of node 1.1, node 1.2, node 1.3, node 1.4 and node 1.5.

Because the gender variable has two attributes, namely male and female, automatically all nodes at depth 3 have male and female branches. Then by looking at the diagnosis where the amount of data is more significant, a final node is made with the label variable, namely diagnosis. So that from all the calculations above, the final decision tree is obtained as follows:

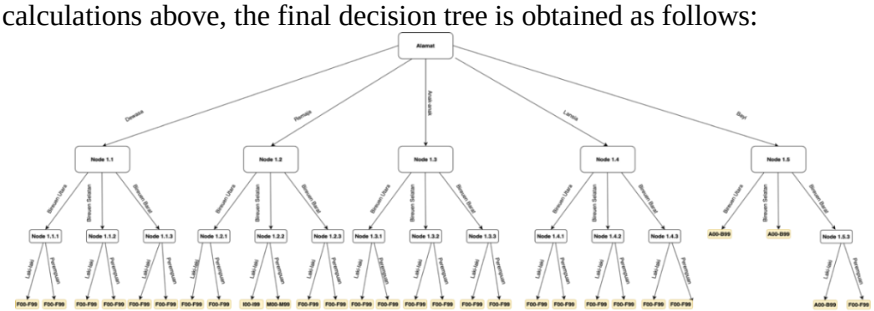

Figure 2. Final Decision Tree Results

4. CONCLUSION

Based on the results of the decision tree that were successfully formed using the C4.5 algorithm was based on medical record data obtained from dr. Fauziah Bireuen, the trend of disease in children, adults and the elderly in northern, western and southern Bireuen is F00-F99, namely mental and behavioral disorders. Meanwhile, the disease trend for infants in North Bireuen, West Bireuen and South Bireuen is A00-B99, namely certain infections and parasitic diseases. Then for adolescence, there were differences in disease trends between North Bireuen and West Bireuen and South Bireuen regions. Meanwhile, for adolescents in North Bireuen and West Bireuen the emerging disease trend is F00-F99, namely mental and behavioral disorders, while for adolescents in South Bireuen who are male, the emerging disease trend is I00-I99, namely diseases of the car and mastid process, while adolescents in South Bireuen are female, the trend of disease that appears is M00-M99, namely diseases of the musculoskeletal system and connective issue.

REFERENCES

[1] Rian R., Sarjon D., Gunadi W. N. "Analisis Rekam Medis untuk Menentukan Pola Kelompok Penyakit Menggunakan Algoritma C4.5". Jurnal Riset Sistem Information dan Teknik Informatika, halaman 391-397, 2018.

[2] Wisti D.W. "Komparasi Metode Klasifikasi Data Mining Algoritma C4.5 dan Naïve Bayes Untuk Prediksi Penyakit Hepatitis". Jurnal Pilar Nusa Mandiri Volume 13, halaman 76-84, 2017.

[3] Xiny u J., Yizheng L., Chunhui M. "Medical Record Semantic Analy sis Based on Weighted LDA". International Symposium on Computational Intelligence and Design, halaman 591-596, 2016.

[4] Adhatrao K., Gaykar A., Dhawan A., Jha R., & Honrao V. "Predicting Students Performance Using ID3 And C4.5 Classification Algorithms." International Journal of Data mining & Knowledge Management Process, halaman 39-52, 2013.

[5] Hapsara, HR.,. "Pembangunan Kesehatan di Indonesia; Prinsip Dasar, Kebijakan, Perencanaan dan Kajian Masa Depannya". Yogyakarta: Gama Press, 2004.

[6] Muhammed Z.J., & Wagner M.JR. "Data Mining and Analysis: Fundamental Concepts and Algortihms". New York: Cambridge University Press, 2014.