Grouping the community health center patients based on the disease characteristics using C4.5 decision tree

N Anwar[1], A Pranolo[2] and R Kurnaiwan[3]
Informatics Department, Universitas Ahmad Dahlan, Yogyakarta, Indonesia

nuril.anwar@tif.uad.ac.id[1], andri.pranolo@tif.uad.ac.id[2], riza.kur93@gmail.com[3]

Abstract. Community health centers (Puskesmas) is one of the important public health service facilities in Indonesia. Puskesmas serves many patients on performing the examination or treatment in every day. Accumulated medical record data is not utilized to generate new information or knowledge. One existed datamining techniques is the process of grouping an object with unknown label into a class. The C.45 algorithm is used to mine the patients diagnosis data available on 2015 2016. As a result, C4.5 algorithms can be applied for grouping disease. The first test using 85 training data has 78% accuracy level, while the second test of 115 training data reaches 88% accuracy rate.

1. Introduction

Community health centers (Puskesmas) in Indonesia is a health center owned by government. Puskesmas always improve the quality of service to the patient through the way of involving technological progress in health world. So the current government of Indonesia immediately take action such as in the form of Social Health Insurance Provider Body, called BPJS. As a service in the form of BPJS Health it can be ascertained the number of patients increases. Activities at this community health centers can generate and collect a lot of medical record data every day. Heaps of medical record data are used for operational needs. Daily medical record data is always increasing. It can be explored to be used as a source of historical data to find a new pattern and knowledge for the community health centers, community and related agencies. This way can be called as a data mining technique [1-3]. Data mining is a term used to describe the discovery of knowledge in a database. Data mining is a process that uses statistical, mathematical, artificial intelligence, and machine learning techniques to interact and identify useful information and related knowledge from large databases [1, 4, 5].

Previous studies have utilized the data mining for many purposes and techniques, such as early prediction of heart diseases [6], to predict liver diseases progress [7], gut microbiota profiles characterization in coronary artery disease patients [8], breast cancer detection [9, 10], and mind performance in Alzheimer's disease [11]. One of techniques, which used for this conducted research, is C4.5 algorithm [12]. C4.5 algorithm is based on decision tree form [12-14].

This research aims to mine valuable information from the historical patient data by using C.45 algorithms. The data will be grouped based on disease characteristics. Therefore, a system that can help the community health centers in determining the number of patients is badly needed.

2. Method

2.1. Dataset

We have collected 150 patient data from Community health centers (Puskesmas) Jetis 1 Bantul district, Indonesia. The data is an attribute owned by the patient, the data in question is the data that has at least

two columns of attributes. One column as the insert attribute column and another as the target attribute column. From each column there are values to be used for calculation, and the value of each attribute must be discrete. The application will read the input with the target attribute located in the last column of the table. Therefore, from the last column the system will recognize it as the input attribute of the system. Some components of variables were:

1) Age. This variable contains the age of each data held by the patient to be filled in the program input process. The pre-defined values for this program are 15-24 years, 25-44 years, 45-64 years, 65+ years.
2) Sex. This variable contains sex data from the patient. The existing groupings are based on the provisions made by the program: L (male) and P (female).
3) Check year. This variable contains the checking years of each patient. The predefined variables in the app based on groupings are 2015 and 2016.
4) Symptoms. This variable contains patient data based on the symptoms that occur in the patient. Symptom data obtained include atrophy, cough, shortness of breath, nausea and vomiting, fever, headache, chills. Grouping based on the provisions made by the program has 2 values that is yes and no.
5) Type of Illness. This variable is data that serves to determine the outcome of the decision. In the grouping of data has been fixed permanently to avoid errors in the calculation process of the program. Decision data has two values: "Infection" and "Degenerative"

2.2. Research design

The main purpose of design is to provide a design description to be built, as well as to understand the flow of information and processes within the system. Figure 1 determined the stages to be performed in system design. The calculation process is done by C4.5 algorithm method, to get the entropy value and gain value, which will be made a decision tree with node and node.

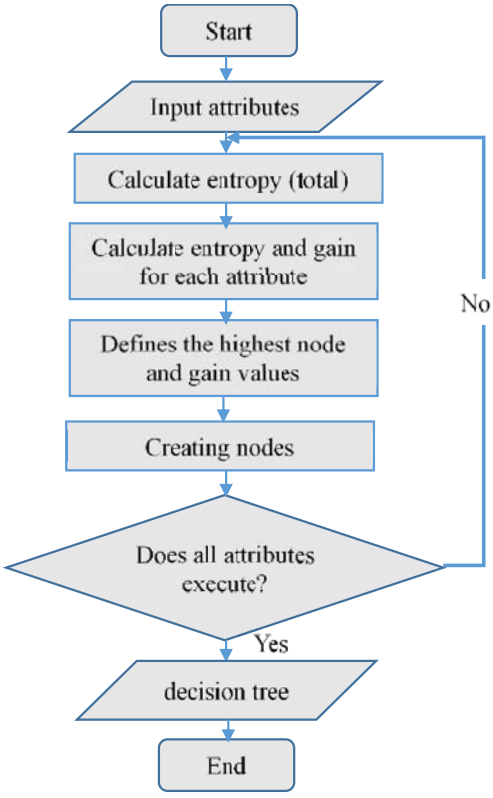

Figure 1. Flowchart algorithm C4.5.

2.3. C4.5 algorithm

In the C4.5 algorithms, decision trees are formed based on the decision-making criteria. The decision tree is a very powerful and well known method of classification and prediction. The decision tree method transforms a very large fact into a decision tree that represents the rule. Rules can be easily understood with Natural language. They can also be expressed in the form of database languages such as Structured Query Language to search for records in certain categories. In general, the C4.5 algorithm constructs a decision trees by selecting attribute as root, creating a branch for each value for the case in the branch, and the process will be repeated for each branch until the case on the branch has the same class [12, 13]. The attribute selection is based on the highest gain value, using equation (1).

$$Gain\ (S, A) = Entropy(S) - \sum_{i=0}^{n} \frac{|S_i|}{|S|} * Entropy(Si) \quad (1)$$

where S are case set, A are attributes, N is a number of attribute partition A, |Si| is the number of cases on the i-th partition, and |S| are the number of cases in S. The value of entropy is calculated before getting a gain value. Entropy is used to determine how informative an attribute is to generate an attribute. The basic formula of entropy is as in equation (2).

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

where S are case set, A are features, N is a number of partition S, p_i is a proportion from S_i to S.

3. Results and discussion

To perform data mining process, we start on data preprocessing which includes the steps of data cleaning, data integration, data selection, and data transportation. Table 1 shows the transformation result of age. In the data mining stage we calculate an $Entropy(Total)$ by using equation (2), for example $\left(-\frac{60}{150} * \ log_2\left(\frac{60}{150}\right)\right) + \left(-\frac{90}{150} * \ log_2\ \left(\frac{90}{150}\right)\right)$, then we get an entropy (total) is 0.970950. The gain (total, age) can be calculated by using equation (1), for example $0.970950 - \left(\left(\frac{20}{150} * 0\right) + \left(\frac{80}{150} * 1\right) + \left(\frac{20}{150} * 0\right) + \left(\frac{30}{150} * 0\right)\right)$, then we got $Gain(Total, Age)$ is 0.437617.

The final decision result can be seen from the result of entropy that the result is zero. If we have obtained the result of entropy zero, then next we see the result of the decision variable with the most value.

Table 1. Node 1 of C4.5.

		Number of Case (S)	Infection (S1)	Degenerative (S2)	Entropy	Gain
Total		150	60	90	0.970950	
Age						0.437617
	15-24	20	20	0	0.000000	
	25-44	80	40	40	1.000000	
	45-64	20	0	20	0.000000	
	65+	30	0	30	0.000000	
Atrophy						0.419973
	Yes	60	0	60	0.000000	
	No	90	60	30	0.918295	
Cough						0.278259
	Yes	70	50	20	0.863120	
	No	80	10	70	0.543564	
Nausea and Vomiting						0.219493
	Yes	76	50	26	0.926819	
	No	74	10	64	0.571354	
Fever						0.067272

		Number of Case (S)	Infection (S1)	Degenerative (S2)	Entropy	Gain
	Yes	115	55	60	0.998635	
	No	35	5	30	0.591672	
Headache						0.019973
	Yes	60	30	30	1.000000	
	No	90	30	60	0.918295	

Table 1 shows that the highest gain value is age variable (0.437617). Thus, age can be used as a root node. There are four attributes of age which have decisions, except the category 25-44 age that has not produced yet a decision. Furthermore, we need to process for it age category (table 2).

Table 2. Node 1.1. of C4.5.

		Number of Case (S)	Infection (S1)	Degenerative (S2)	Entropy	Gain
Age						
25-44		80	40	40	1.000000	
Atrophy						0.548795
	Yes	30	0	30	0.000000	
	No	50	40	10	0.721928	
Cough						0.323073
	Yes	45	35	10	0.824587	
	No	35	5	30	0.430827	
Nausea and Vomiting						0.454819
	Yes	56	30	26	0.558268	
	No	24	10	14	0.979869	
Fever						0.311278
	Yes	60	40	20	0.918296	
	No	20	0	20	0.000000	
Headache						0.048795
	Yes	30	20	10	0.918296	
	No	50	20	30	0.970951	

From the table 1 and table 2, we can create a knowledge representation which represented by a decision tree (figure 2). In table 2, four attributes age has each decision, 15-24 is infection and 45-64 is degenerative. Hence, no further calculation is required, but for 15-24 attribute is still needed further process. In Table 2, both Atrophy attributes have degenerative decision, so no further calculation is required. Figure 2 shows the final result of the decision tree which is containing the rules, as follows:

1) When the age is "15-14 years", then the decision is infection.
2) When the age is "45-64 years", then the decision is degenerative.
3) When the age is "65+ years", then the decision is degenerative.
4) When the age is "25-44 years", then the decision is atrophy.
5) When the atrophy is "Yes", then the decision is degenerative.
6) When the atrophy is "No", then the factor is cough.
7) When cough is "Yes", then the decision is infection.
8) When the cough is "No", then the factor is a headache.
9) When the headache is "No", then the decision is degenerative.
10) When the headache is "Yes", then factor is fever.
11) When the fever is "Yes", then the decision is infection.
12) When the fever is "No", then the factor nausea-vomiting.
13) When nausea-vomiting is "Yes", then the decision is degenerative.
14) When nausea-vomiting is "No", then the decision is infection.

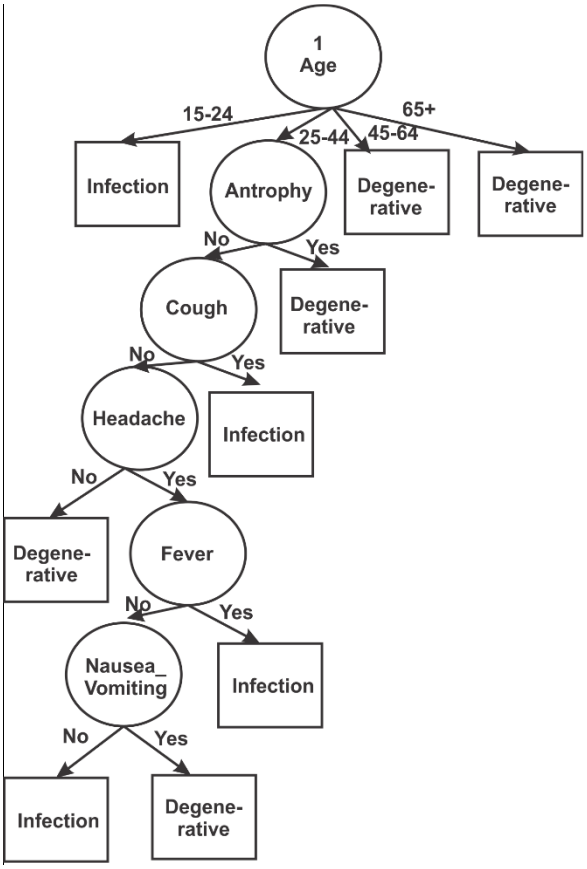

Figure 2. Final result of the decision tree.

4. Conclusion

It can be concluded that Conclusions that data mining can be used to help provide useful information in predicting the number of groupings diagnosis of patient where in this thesis use one of algorithm from data mining that is C4.5 algorithm. The result of accuracy testing with confusion matrix method, test one with amount of training data 85 and data testing 33 produce 78% accuracy with error 22%. The two tests with the amount of training data 115 and the 18 test data yielded 88% accuracy with 12% error.

References

[1] Han J and Kamber M 2006 Data Mining: Concepts and Techniques 2nd ed, The Morgan Kaufmann Series in Data Management Systems (Morgan Kaufmann)
[2] Berkhin P 2006 A survey of clustering data mining techniques *Group. Multidimens. data* vol 25 pp 25-71
[3] Cohen W W and Richman J 2002 Learning to match and cluster large high-dimensional data sets for data integration *Proc. eighth ACM SIGKDD Int. Conf. Knowl. Discov. data Min.* pp 475–80
[4] Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten I H 2009 The WEKA data mining software: an update *ACM SIGKDD Exploration Newsletter* vol 11 issue 1 pp 10-18
[5] Hand D J 2007 Principles of data mining *Drug Saf.* vol 30 no 7 pp 621–622
[6] Chaurasia V and Pal S 2017 Early prediction of heart diseases using data mining techniques

Carribean Journal of Science and Technology vol 1 pp 208-217

[7] Saad Y, Awad A, Alakel W, Doss W, Awad T and Mabrouk M 2018 Data mining of routine laboratory tests can predict liver disease progression in Egyptian diabetic patients with hepatitis C virus (G4) infection: a cohort study of 71 806 patients *Eur. J. Gastroenterol. Hepatol.* vol 30 no 2 pp 201–206

[8] Emoto T*,* Yamashita T, Kobayashi T, Sasaki N, Hirota Y, Hayashi T, So A, Kasahara K, Yodoi K, Matsumoto T, Mizoguchi T, Ogawa W and Hirata Kl 2017 Characterization of gut microbiota profiles in coronary artery disease patients using data mining analysis of terminal restriction fragment length polymorphism: gut microbiota could be a diagnostic marker of coronary artery disease *Heart Vessels* vol 32 no 1 pp 39–46

[9] Chaurasia V and Pal S 2017 A novel approach for breast cancer detection using data mining techniques *International Journal of Innovative Research in Computer and Communication Engineering* vol 2 issue 1

[10] Chaurasia V and Pal S 2017 Data mining techniques: To predict and resolve breast cancer survivability *International Journal of Computer Science and Mobile Computing IJCSMC* vol 3 issue 1 pp 10–22

[11] Ramanan S, de Souza L C, Moreau N, Sarazin M, Teixeira A L, Allen Z, Guimaraes H C, Caramelli P, Dubois B, Hornberger M and Bertoux M 2017 Determinants of theory of mind performance in Alzheimer's disease: A data-mining study *Cortex* vol 88 pp 8–18

[12] Li Y, Jiang Z L, Yao L, Wang X, Yiu S M and Huang Z 2017 Outsourced privacy-preserving C4. 5 decision tree algorithm over horizontally and vertically partitioned dataset among multiple parties *Cluster Comput.* pp 1–13

[13] Tsai C-H, Weng S-J, Chou C-A, Wu H-H and Hung W-Z 2017 Applying C4. 5 decision tree to analyze insomnia symptoms *Taiwan Gong Gong Wei Sheng Za Zhi* vol 36 no 5 pp 449–60

[14] Sharma N and Dubey S K 2014 Semantic based Web Prefetching using Decision Tree Induction *2014 5th* Int. *Conf. - Conflu. Next Gener. Inf. Technol. Summit* pp 132–7 Sep. 2014