Jurnal Mantik, 6 (1) (2022) 576-581

Published by:Institute of Computer Science (IOCS)

Jurnal Mantik

Journal homepage: www.iocscience.org/ejournal/index.php/mantik/index

APPLICATION OF THE C4.5 ALGORITHM FOR CLASSIFICATION OF MEDICAL RECORD DATA AT M.DJAMIL HOSPITAL BASED ON THE INTERNATIONAL DISEASE CODE

Nurul Abdillah[1], Muhammad Ihksan[2]

[12] Health Information Management, Sekolah Tinggi Ilmu Kesehatan Syedza Saintika, Jl. Prof. Dr. Hamka No.228 Padang, 25132, Indonesia

E-mail: abdillahadik15@gmail.com

ARTICLE INFO

Article history:
Received: Mar 5, 2022
Revised: Aprl 18, 2022
Accepted: May 28, 2022

Keywords:

Data Mining, Classification, Medical Records, C4.5, ICD-10

ABSTRACT

Medical record data are special patient records, often medical record data only becomes data that accumulates and is not searched to produce useful knowledge for hospitals. This study aims to determine the disease classification model based on piles of medical record data, using one of the methods in data mining. To achieve the research objectives, 4 attributes were selected according to the medical record data. The medical record data in question consists of disease diagnosis attributes based on the International Classification of Diseases-10 (ICD-10), gender, age of the patient and month of admission to the hospital. The method used in this study is the C4.5 algorithm method using the international disease code attribute as the destination label attribute for 21 international disease groups, namely: A00-B99 to Z00-Z99. In this research, the C4.5 algorithm can represent 7 attribute values for the disease code, namely A00-B99, C00-D89, I00-I99, O00-O99, P00-P96, S00-T98 and Z00-Z99. The conclusion of this study is that the C4.5 algorithm is less than optimal in producing classification of medical record data because the number of destination classes or class labels is very large and the percentage of data read is less than 50%. The resulting disease classification is only 7 classes out of 21 overall classes according to the international disease code.

1. Introduction

The development of Data Mining Science provides new innovations in terms of utilizing large data sets so that they can be useful for knowledge development, both specifically in fields related to the data and globally. Data Mining merupakan suatu rangkaian proses untuk menggali nilai tambah dari sekumpulan data berupa pengetahuan yang selama ini tidak diketahui secara manual (Handoko, 2016). Many functions that can be applied from data mining science include estimation, prediction, clustering, classification and association. To achieve these functions, various methods (algorithms) are used, such as regression for estimation, Support Vector Macsine (SVM) for prediction, Kmeans for clustering, C4.5 for classification, a priori for association (Sumanthi S. dan Sivanandam S, 2006) .

One application of data mining science, namely the problem of accumulation of medical record data in hospitals. Medical record is a file containing records and documents regarding patient identity, examination, treatment, actions and other services provided to patients. Medical records must be made in writing, complete, and clear or electronically. The administration of medical records using electronic information technology is regulated by separate regulations. Information in medical records is kept confidential by doctors, health workers and management officers and leaders of health service facilities. Medical record data continues to accumulate every day along with hospital activities. Utilization of medical records can be used as: (1) health maintenance and treatment of patients; (2) evidence in the process of law enforcement, medical and dental discipline, and medical and dental ethics enforcement; (3) educational and research needs; (4) the basis for payment of health care costs; (5) health statistical data (Kemenkes RI, 2008).

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

Data Mining sering juga disebut sebagai Knowledge Discovery in Database (KDD) yang merupakan kegiatan yang meliputi pengumpulan, pemakaian data, historis untuk menemukan keteraturan, pola atau hubungan dalam set data berukuran besar (Sari, 2016). Knowledge discovery in database (KDD) is a process to find useful information in database. The whole KDD process usually consists of steps, namely understanding the application field, creating the target data set from the raw data stored in the database, data cleaning and data preprocessing (Pupezescu, 2016).

The term knowledge discovery in database or looking for knowledge in a database or KDD for short, refers to the process of seeking knowledge in extensive data and emphasizes the application of high-level methods or certain data mining methods. This attracts the interest of researchers in conducting research developer either in the field of machine learning or machine learning(Priyadharsini, 2014).

The data mining approach is becoming very important in the healthcare industry in making decisions based on the analysis of large clinical data. Data mining plays a role in the process of extracting hidden information from large datasets and techniques such as classification, clustering, regression and association have been used by the medical field (Veenita K, 2016).

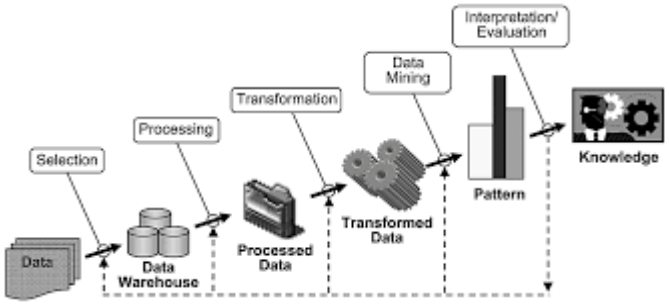

Figure 1. Knowledge discovery in database process (Susanti, 2014)

One of the classification techniques often used by researchers is the application of the C4.5 algorithm. The C4.5 algorithm is an algorithm that produces a decision tree. It has input in the form of a classification sample (Harvinder C. and AnuC, 2013). The application of the C4.5 algorithm function to produce a level of data accuracy as a dataset containing large amounts of data (Chauhan H. and Chauhan A, 2013).

2. Method

The research focuses on the process of analyzing medical record data with the C4.5 algorithm using the RapidMiner Studio 7.5 program (Tools Data Mining) to obtain classification results. There are 4 attributes used in the study, namely: (1) age grouped into the categories of toddlers, children, adolescents, adults and the elderly; (2) gender consisting of female (F) and male (L); (3) months consisting of July, August, September, October; (4) diagnosis (ICD-10) is a goal attribute consisting of a group of diseases according to the international disease code (ICD-10), namely: A00-B99, C00-D48, D50-D89, E00-E90, F00-F99, G00 - G99, H00-H59, H60-H95, I00-I99, J00-J99, K00-K93, L00-L99, M00-M99, N00-N99, O00-O99, P00-P96, Q00-Q99, R00-R99 , S00-T98, V01-Y98, Z00-Z99 (WHO, 2010).

The C4.5 algorithm starts from the process of selecting the attribute with the highest gain as the root of the tree, then creates a branch for each value, then divides the cases into branches, then repeats the process for each branch until all cases in the branch have the same class (Hitesh, 2013).

To facilitate the application of the methodology and system design, a flow chart of analysis and design is made as shown in Figure 2 below.

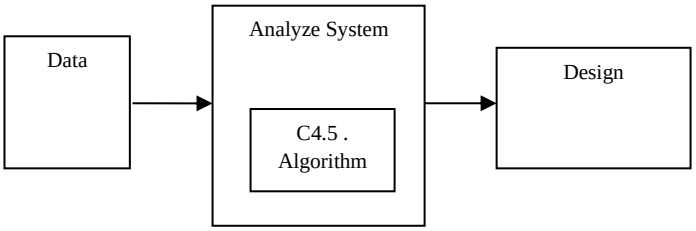

Figure 2. Analysis Flowchart

The form of a flowchart can clearly describe the process stages and steps in the classification using the C4.5 algorithm. It can be seen in Figure 3 in the form of a flowchart as follows:

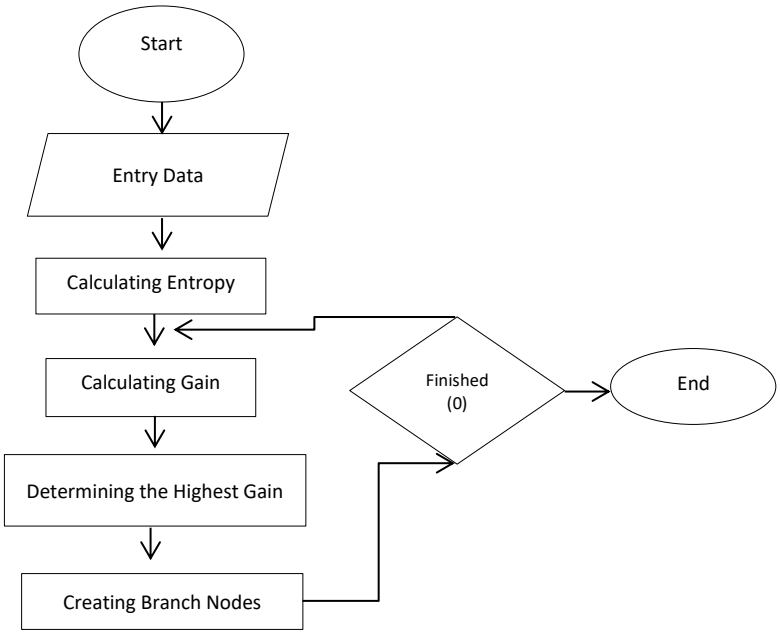

Figure 3. Process Flowchart on the C4.5Algorithm

Classification Techniques The C4.5 algorithm begins with processing and transforming data so that the raw data used for analysis is data with complete attributes and can produce a decision tree, along with the work order to determine the decision tree (Miftahul Chair, 2017).

2.1 Processing and Data Transformation

Not all attributes in the patient's medical record database were used in the study, attributes such as: date of birth, name of the doctor in charge and patient's medical record number were not needed in this study. Processing only requires four attributes, such as: Gender, Age Category, Month of Treatment and Disease Code as the objective attributes of this research, so that the data transformation process is needed to be processed, after data transformation, it will be processed using the C4.5 algorithm.

2.2 Data processing

Data processing begins with finding the total entropy of all attributes and then determining the highest gain. To get the gain value in the formation of a decision tree, it is necessary to first calculate the information value in bits from a collection of objects. The calculation form for entropy is as follows (Siska Haryati, 2015):

$$Entropy(X) = \sum_{j=1}^{k} p_j * log_2 \frac{1}{p_j} = -\sum_{j=1}^{k} p_j * log_2 p_j \quad (1)$$

where,
X : Case Collection
k : number of partitions X
pj : Proportion of Xj to X

The value of Entropy(X) indicates that X is a random attribute. The entropy value reaches a minimum value of 0, when all other pj = 0 or are in the same class. In the C4.5 tree construction, each tree node is filled with the attribute with the highest gain ratio value, with the following formula (Lusinia, 2014) :

$$Gain(a) = Entropy(X) - \sum_{j=1}^{k} \frac{|X_i|}{|X|} * Entropy(X_i) \quad (2)$$

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

a. Finding Total Entropy and Gain (Root)

The process of finding total entropy and gain is done by grouping the data correctly, then calculating the data and using the entropy and gain search formula for each data attribute. From the results of the calculations in the table above, it can be seen that the largest gain value is the "Gender" attribute of 1.346. So the "Gender" attribute becomes the root node. In the attribute "Gender" there are 2 attribute values, namely Male and Female, it is necessary to do further calculations. From this process, a temporary tree can be generated as follows:

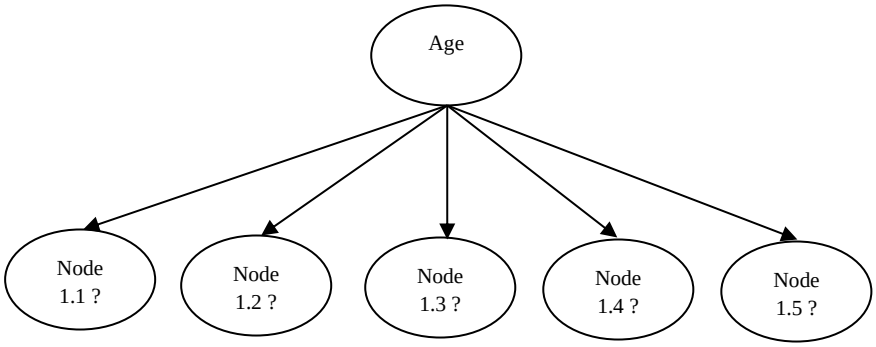

Figure 4. Temporary Decision Tree (Root)

3. Result and Discussion

The results of the analysis show that the C4.5 algorithm has succeeded in grouping diseases into 7 types of categories from 21 types of categories which are the destination labels based on the ICD (International Code Diseases) or international disease codes, so it can be said that the C4.5 algorithm is only able to define less than 50 % of existing destination label categories.

The results of testing the data that have been carried out are summarized in table 1 as follows:

Table 1. Data Test Results

AGE	TOddler	JULY	P00-P96	
		AUGUST	P00-P96	
		SEPTEMBER	P00-P96	
		OCTOBER	Z00-Z99	
	CHILDREN	JULY	MAN	Z00-Z99
			WOMAN	I00-I99
		AUGUST	MAN	C00-D89
			WOMAN	Z00-Z99
		SEPTEMBER	Z00-Z99	
		OCTOBER	MAN	C00-D89
			WOMAN	Z00-Z99
	TEENAGER	MAN	JULY	C00-D89
			AUGUST	Z00-Z99
			SEPTEMBER	C00-D89
			OCTOBER	S00-T98
		WOMAN	JULY	O00-O99
			AUGUST	C00-D89
			SEPTEMBER	C00-D89
			OCTOBER	C00-D89
	MATURE	MAN	JULY	A00-B99
			AUGUST	C00-D89
			SEPTEMBER	C00-D89
			OCTOBER	C00-D89
		WOMAN	JULY	O00-O99
			AUGUST	O00-O99
			SEPTEMBER	C00-D89
			OCTOBER	C00-D89
	ELDERLY	MAN	JULY	C00-D89

			AUGUST	I00-I99
			SEPTEMBER	I00-I99
			OCTOBER	I00-I99
		WOMAN	JULY	A00-B99
			AUGUST	Z00-Z99
			SEPTEMBER	C00-D89
			OCTOBER	C00-D89

The results of data testing that have been carried out using the Rapiminer application are as follows:

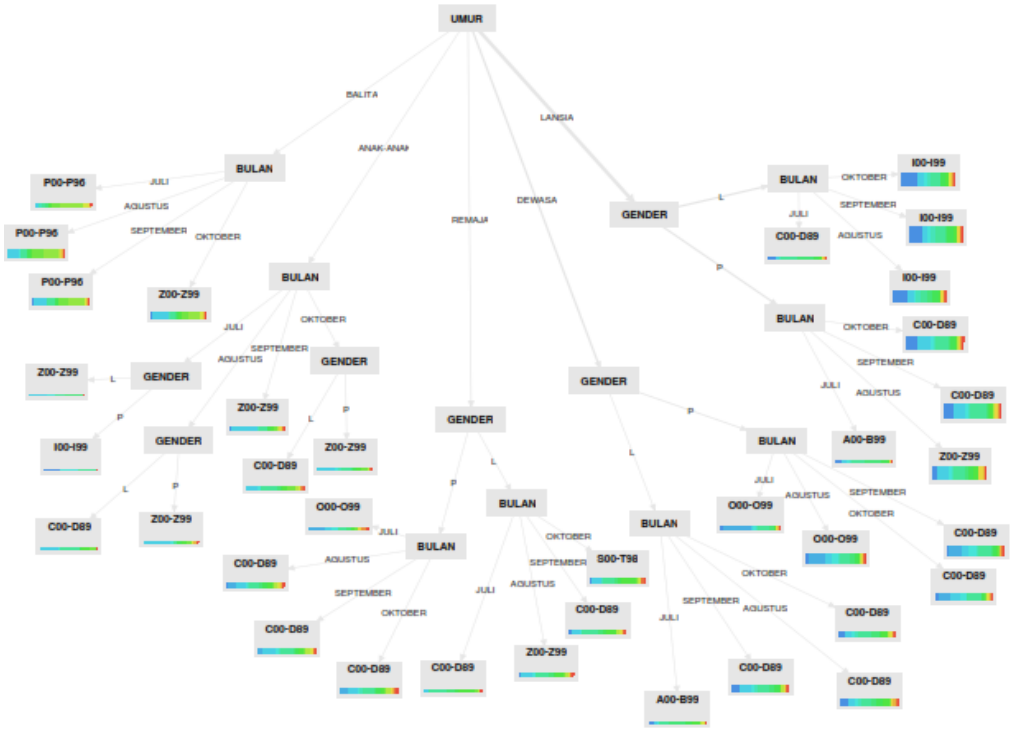

Figure 5. Decision Tree Results from Rapidminer

4. Conclusion

Based on the results of the study, it can be said that the C4.5 algorithm is less than optimal in producing medical record data classification because the destination class or class label is very large (21 class labels) and the percentage of data read is less than 50%. The resulting disease classification is only 7 classes out of 21 overall classes according to the international disease code.

References

[1] Chauhan H. and Chauhan A. (2013). Implementation of decision tree C4.5 algorithm. *International Journal of Scientific and Research Publications*, *3*(10).

[2] Handoko, K. (2016). Penerapan Data Mining Dalam Penerapan Data Mining Dalam Meningkatkan Mutu Pembelajaran Pada Instansi Perguruan Tinggi Menggunakan Metode K-Means Clustering Clustering Clustering (Studi Kasus Di Program Studi TKJ Akademi Komunitas Solok Solok Selatan). *TEKNOSI*, *2*(3), 31–40.

[3] Harvinder C. and AnuC. (2013). Implementation of decision tree C4.5 algorithm. *International Journal of Scientific and Research Publications*, *3 (10)*.

[4] Hitesh, A. G. dan G. (2013). Optimization of C4.5 Decision Tree Algorithm for Data Mining Application. *International Journal of Emerging Technology and Advanced Engineering*, *3*, 341–345.

[5] Kemenkes RI. (2008). *Permenkes RI. 2008. No. 269 / MENKES/PER/III/2008 tentang Rekam Medis*.

[6] Lusinia, S. A. (2014). *Algoritma C4.5 Dalam Menganalisa Kelayakan Kredit (Studi Kasus Di Koperasi*

Pegawai Republik Indonesia (KP-RI) Lengayang Pesisir Selatan, Painan, Sumatera Barat). *1*(2).

[7] Miftahul Chair, Y. N. N. dan N. A. R. (2017). Aplikasi Klasifikasi Algoritma C4.5 (Studi Kasus Masa Studi Mahasiswa Fakultas Matematika Dan Ilmu Pengetahuan Alam Universitas Mulawarman Angkatan 2008). *Jurnal Informatika Mulawarman*, *12*(1), 51–55.

[8] Priyadharsini, C. (2014). An Overview of Knowledge Discovery Database and Data mining Techniques. *International Journal of Innovative Research in Computer and Communication Engineering*, *2*(Special Issue 1), 1571–1572.

[9] Pupezescu, V. (2016). The Influence of Data Replication in the Knowledge Discovery in Distributed Databases Process. *ECAI 2016 International Conference Electronics (Computers and Artificial Intelligence)*, *8*, 17.

[10] Sari, C. R. (2016). Teknik Data Mining Menggunakan Classification Dalam Sistem Penunjang Keputusan Peminatan SMA Negeri 1 Polewali. *Indonesian Journal on Networking and Security. 2016*, *5*, 48–54.

[11] Siska Haryati, A. S. dan E. S. (2015). Implementasi Data Mining Untuk Memprediksi Masa Studi Mahasiswa Menggunakan Algoritma C4.5 (Studi Kasus: Universitas Dehasen Bengkulu). *Jurnal Media Infotama*, *11*, 130–138.

[12] Sumanthi S. dan Sivanandam S. (2006). Introduction to Data Mining and its Applications. *Springer*.

[13] Susanti, B. D. M. dan N. (2014). Aplikasi Data Mining Untuk Menghasilkan Pola Kelulusan Siswa Dengan Metode Naïve Bayes. *JURNAL LINK VOL 21*, *2*, 1–5.

[14] Veenita K. (2016). Chronic Kidney Disease Analysis Using Data Mining Classification Techniques. *IEEE*, 300–305.

[15] WHO. (2010). *International Statistical Classification of Diseases and Related Health Problems (ICD-10)*.