Implementation of K-Means Clustering on Patient Data Of National Social and Healthcare Security by Using Java

Parasian Silitonga
Computer Science Departement
St.Thomas Catholic University
Medan, Indonesia
parasianirene@gmail.com

Irene Sri Morina
Information System Section
Adam Malik Hospital
Medan, Indonesia
morina_ginting@yahoo.com

Abstract—*Currently the patient data stack is still focused on compliance with reports and charts of hospital patients, disease data and patient care costs. The existing stack of data does not yet present a custom pattern for the data. Data mining is a process of finding meaningful relationships from a set of data by examining data stored in storage media by using pattern recognition techniques such as statistical and mathematical techniques. One technique in data mining is clustering. The purpose of this study was to find a trend pattern of patient illness based on disease code in Indonesia Case Base Groups (Ina CBG's). Tests were conducted against data on the users of then National Social and Healthcare Security. This research is done by using K-Means Clustering method which is implemented with Java programming language.*

Keywords-Data Mining; Clustering; K-Means Clustering; Indonesia Case Base Groups; National Social and Healthcare Security; Java.

I. INTRODUCTION

Currently the patient data stack which is available in hospitals is generally limited to reports and charts of hospital patients, disease data and patient care costs. The existing pile of data does not present the existing pattern of disease spread. By knowing the pattern of disease penyeberan then indirectly hospital can anticipate service priority based on pattern of disease with highest tendency.

Clustering is a technique of grouping records on a database based on certain criteria. The clustering results are given to end users to give an idea of what is happening in the database [5]. Clustering performs grouping of data without specific data classes. Even clustering can be used to label the unknown data class. Therefore clustering is often classified as an unsupervised learning method [4].

One method that can be done to classify data base is K-Means Clustering method. This method divides the data into several groups and can accept input in the form of data without the class label [2]. This method partitions the data into clusters / groups so that data that have the same characteristics are grouped into the same cluster and data that have different characteristics are grouped into other groups.

National Social and Healthcare Security is a government-owned insurance agency established under Presidential Regulation No. 12 of 2013. The National Social and Healthcare Security covers all Indonesian citizens and officially operates from 1 January 2014. With the existence of the National Social and Healthcare Security program, communities receive health service rights from hospitals and community health centers designated by the government as hospitals or health centers providing services to the National Social and Healthcare Security.

Java is an object oriented programming. Java was created after C ++ and was designed so that it was small, simple, and portable [3]. Java was developed in 1991 by a group of Sun engineers led by Patrick Naughton and James. Originally this language was named Oak inspired by the name of a wooden tree named Oak, but because there was already a programming language called Oak, it was renamed Java inspired by a cup of coffee [3].

In this research, the research data is sourced from patient medical record data of Haji Adam Malik General Hospital Medan from 2014 until 2015. The result of this research is tested through the application that is produced by using Java programming language which is produced in this research.

II. DATA MINING

Data mining is a term used to describe the discovery of knowledge in a database. Data mining uses a variety of techniques such as statistics, mathematics, artificial intelligence, and machine learning that extracts and identifies useful information and assembled knowledge from large databases. Data mining is a finding of meaningful relationships, patterns, and trends by checking in a large set of data stored in storage by using pattern recognition techniques such as statistical and mathematical techniques [6].

Basically, data mining has the utility and the task to specify the patterns that must be found in the data mining process. In general, data mining tasks can be divided into two categories :

- *Predictive*

  The purpose of a predictive task is to predict the value of a particular attribute based on the value of the other attributes. Predictable attributes are commonly known as non-free targets or variables, while the attributes used to make predictions are known as independent variables.

- *Deskriptive*

  The purpose of the descriptive task is to derive patterns (correlations, trends, clusters, trajectories, and anomalies) that summarize key relationships in the data. The task of descriptive data mining is often referred to as an inquiry and often requires postprocessing techniques for validation and explanation of results.

There are 7 (seven) stages of data mining process, where the first 4 (four) stages are also called preprocessing data (consisting of cleaning data, data integration, data selection, and data transformation), which in its implementation takes about 60% of the whole process . Then data mining, pattern evaluation and presentation presentation.

III. K-MEANS CLUSTERING

K-Means Clustering is a fairly simple clustering algorithm that partitions the dataset into several clusters k. The algorithm is fairly easy to implement and run, relatively fast, easy to customize and widely used.

This method is one of the non-hierarchical data clustering methods that group data in the form of one or more clusters / groups. The data that have the same characteristics are grouped in one cluster / group and the data having different characteristics are grouped with other clusters / groups so that the data in one cluster / group has small variation level [1].

The purpose of this data clustering is to minimize the objective function set in the clustering process, which generally tries to minimize the variation within a cluster and maximize the variation between clusters.

In general, K-Means Clustering method is done with the following steps [12]:

- Select the number of clusters k.
- Initialization of cluster center k. In general, cluster centers are given initial values with random numbers.
- Allocate all data / objects to the nearest cluster. The proximity of two objects is determined by the distance of the two objects. Likewise the proximity of a data to a particular cluster is determined the distance between the data with the cluster center. To distance all data to each cluster center point can use Euclidean distance theory according to Equation 1.

  $$D(i,j) = \sqrt{(x_{1i} - x_{1j})^2 + (x_{2i} - x_{2j})^2 + \cdots + (x_{ki} - x_{kj})^2}. \quad (1)$$

  Where :
  D(i,j) = The distance of the i data to the cluster center j
  $x_{k,i}$ = The distance of the data to the cluster center j
  $x_{k,j}$ = The j-th center point of the k-data attribute
- Recalculate cluster center with current cluster membership. The cluster center is the average of all data / objects in a particular cluster.
- Check each new cluster center user object. If the cluster center does not change again then the clustering process is complete. Or, go back to step 3 until the center of the cluster does not change anymore.

IV. NATIONAL SOCIAL AND HEALTHCARE SECURITY

National Social and Healthcare Security is an institution established to organize social security program in Indonesia according to Law Number 40 Year 2004 and Act Number 24 Year 2011. In accordance with Law Number 40 Year 2004 regarding National Social Security System, National Social and Healthcare Security is a legal entity non-profit.

Based on Law Number 24 Year 2011, the National Social and Healthcare Security will replace a number of social security institutions in Indonesia. The objective of this program is to meet the appropriate health needs given to every member of the community who has paid contributions or fees paid by the government and private institutions. With the existence of National Social and Healthcare Security Implementation Program, all layers of society receive the right of health services from hospitals that have been appointed by the government as a hospital provider of National Social and Healthcare Security.

V. INDONESIA CASE BASE GROUP (INA CBG'S)

Based on Regulation of the Minister of Health of the Republic of Indonesia Number 69 Year 2013 on Health Service Tariff Standard At Health Facility. The data obtained are in the form of Indonesia Case Base Groups data comprising Indonesian Case Base Groups Code Description Indonesian Case Base Groups and Indonesia Case Base Groups Tariff.

Indonesia Case Base Groups data consists of 789 Indonesian Case Base Groups codes for tariffs applicable to Class A Hospitals. Indonesia Case Base Groups is an application used by hospitals to file claims with the government. The Indonesia Case Base Groups system is developed from the UNU-IIGH casemix system (The United Nations University-International Institute for Global Health) and is guided by the International Classification of Diseases (ICD).

VI. CLUSTERING OF PATIENT DISEASE DATA

Data selection is the first process done to use the data needed in the process of mining. The selection of data comes from Adam Malik Haji General Hospital Medan, in the form of primary data and secondary data (Silitonga, Parasian., 2017).

- *Ina-CBG's Data*

  The Ina-CBG data consists of 789 INA-CBG Codes for rates applicable to Class A Hospitals. Ina-CBG data is stored in tables with formats such as Table I.

TABLE I. INA CBG'S DATA STRUCTURE

Data Attribute	Information
InaCBG'sCode	InaCBG Data-Based Disease Code
Description	Disease Description

- Patient Data Users of the National Social and Healthcare Security
  User patient data of Social Health Insurance Administering Board, is patient data which treatment at General Hospital of Haji Adam Malik Medan. These patient data are presented in table form as in Table II.

TABLE II. PATIENT DATA STRUCTURE

Data Attribute	Information
MedicalRecordNumber	Patient's Medical Record Number
PatientName	Patient's Name
DateOfBirth	Patient's Date of Birth
DateOfEntry	Patient's Date of Entry
DateOfExit	Patient's Date of Exit
InaCBG'sCode	InaCBG Data-Based Disease Code

A. Data Transformation

Stages of data transformation is the process of converting data into the appropriate form. The initial change is done by changing the data format so that the number of diagnoses of the patient's disease every month (January to December) is known. The result of data transformation performed in accordance with the table form in Table III.

TABLE III. TRANSFORMTATION DATA STRUCTURE

Data Attribute	Information
InaCBG'sCode	InaCBG Data-Based Disease Code
Year	Year of Data
Count	Number of Patient

B. K-Means Clustering

Based on the result of data transformation that has been done, if we have the available data is like in Table IV.

TABLE IV. PATIENT DATA SAMPLE

No.	Kode INA CBG'S	2014	2015
1	A-4-10-I	4	3
2	A-4-10-II	9	2
3	A-4-10-III	3	14
4	A-4-11-I	4	30

The initial stage of K-Mean Cluster calculation is to randomly generate clusters and calculate the distance between the cluster center and the data:

a) Suppose that the number of clusters to be formed is 2 clusters with 2 iterations.
b) Suppose the fisrt cluster is derived from second data (9; 2) and second cluster is derived from third data (3; 14)
c) Calculate the cluster's center distance with the data using Eq. 1.

- Calculate the first data distance to the center of the first cluster,

  D11= $\sqrt{(4 \quad 9)^2 + (3 \quad 2)^2}$= 5,10

- Calculate the first data distance to the center of the second cluster,

  D12= $\sqrt{(4 \quad -3)^2 + (3 \quad -14)^2}$= 11,05

- Calculate the second data distance to the center of the first cluster,
  D21=$\sqrt{(9 \quad -9)^2 + (2 \quad -2)^2}$ = 0
- Calculate the second data distance to the center of the second cluster,
  D22= $\sqrt{(9 \quad -3)^2 + (2 \quad -14)^2}$= 13,42
- Calculate the third data distance to the center of the first cluster,
  D31=$\sqrt{(3 \quad -9)^2 + (14 \quad -2)^2}$ = 13,41
- Calculate the third data distance to the center of the second cluster,
  D32=$\sqrt{(3 \quad -3)^2 + (14 \quad -14)^2}$ = 0
- Calculate the fourth data distance to the center of the first cluster,
  D41=$\sqrt{(4 \quad -9)^2 + (30 \quad -2)^2}$ = 28,44
- Calculate the fourth data distance to the center of the second cluster,
  D42=$\sqrt{(4 \quad -3)^2 + (30 \quad -14)^2}$ = 16,03

TABLE V. FIRST CLUSTER ITERATION CENTER

No.	Kode	January	February	Cluster 1	Cluster 2
1	A-4-10-I	4	3	5,10	11,05
2	A-4-10-II	9	2	0	13,42
3	A-4-10-III	3	14	13,42	0
4	A-4-11-I	4	30	28,44	16,03

d) From Table IV of Iteration Cluster 1, select the smallest cluster to produce Table VI.

TABLE VI. FIRST CLUSTER ITERATION CENTER

No.	Kode	January	February	Cluster 1	Cluster 2
1	A-4-10-I	4	3	*	
2	A-4-10-II	9	2	*	
3	A-4-10-III	3	14		*
4	A-4-11-I	4	30		*

e) Calculate the cluster's center

- Cluster 1 is only the 1st and 2nd data, so it is obtained:

  C11 = (4+3)/2 = 3,5

  C12 = (9+2)/2 = 5,5

- Cluster 2 consists of the 3rd and 4th data so it is obtained:

  C21 = (3+14)/2 = 8,5

  C22 = (4+30)/2 = 17

f) In the second iteration is obtained :

- Calculate the first data distance to the center of the first cluster,

  D11= $\sqrt{(4 - 3{,}5)^2 + (3 - 5{,}5)^2}$= 2,55

- Calculate the first data distance to the center of the second cluster,

  D12= $\sqrt{(4 - 8{,}5)^2 + (3 - 17)^2}$= 14,71

- Calculate the second data distance to the center of the first cluster,
  D21=$\sqrt{(9 - 3{,}5)^2 + (2 - 5{,}5)^2}$ = 6,52
- Calculate the second data distance to the center of the second cluster,
  D22= $\sqrt{(9 - 8{,}5)^2 + (2 - 17)^2}$= 15,01
  - Calculate the third data distance to the center of the first cluster,
    D31=$\sqrt{(3 - 3{,}5)^2 + (14 - 5{,}5)^2}$ = 8,51
  - Calculate the third data distance to the center of the second cluster,
    D32=$\sqrt{(3 - 8{,}5)^2 + (14 - 17)^2}$ = 6,26
  - Calculate the fourth data distance to the center of the first cluster,
    D41=$\sqrt{(4 - 3{,}5)^2 + (30 - 5{,}5)^2}$ = 24,51
  - Calculate the fourth data distance to the center of the second cluster,
    D42=$\sqrt{(4 - 8{,}5)^2 + (30 - 17)^2}$ = 13,76

The calculation results are then entered into the table as in Table VII:

TABLE VII. SMALLEST CLUSTER CENTER SECOND ITERATION

No.	Kode	2014	2015	Cluster 1	Cluster 2
1	A-4-10-I	4	3	*	
2	A-4-10-II	9	2	*	
3	A-4-10-III	3	14		*
4	A-4-11-I	4	30		*

g) Based on the results of clustering conducted up to the second iteration, it was found that the A-4-10-I and A-4-10-II Disease Codes were Cluster-1 (C1) cluster members, while the A-4-10-III Disease Codes and A-4-11-I belongs to the Cluster-2 (C2) cluster.

VII. K-MEANS CLUSTERING USING JAVA

Clustering K-Means patient disease data in this study was conducted using programs built using the Java language and MySQL Database Management System. The resulting program corresponds to the class diagram as in Figure 1.

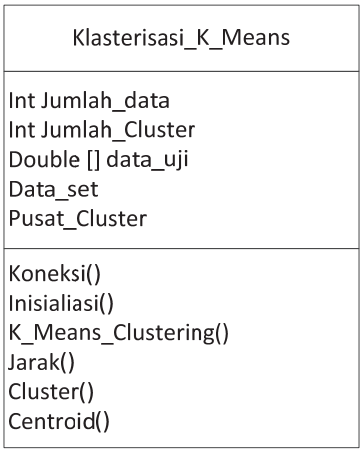

Figure 1. K-Means Clustering Class Diagram

The test was conducted with data as many as 1587 records data sourced from patient hospital data from 2014 and 2015. The output generated from the program is presented as in Figure 2.

Figure 2. Cluster's Number Input

Data clustering is done as many as 4 clusters. The clustering results using K-Means Clustering obtained the results as shown in Figure 3. The initial center of the cluster is presented as in Figure 4.

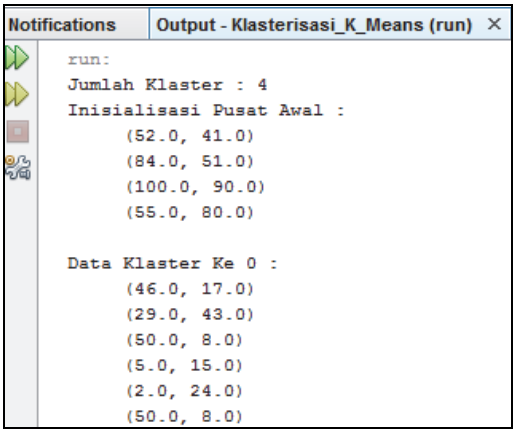

Figure 3. K-Means Clustering Clustering Result Data

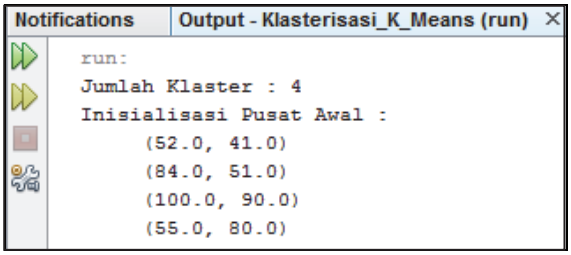

Figure 4. Cluster Start Center

ACKNOWLEDGMENT

The research was funded by the Institute for Research and Community Service to the Catholic University of St. Thomas. The authors would like to thank the Rector of the Catholic University of St. Thomas, Chairman of the Institute for Research and Community Service, the Dean of the Faculty of Computer Science and Director of Haji Adam Malik General Hospital for all the assistance provided.

REFERENCES

[1] Agusta, Yudi "K-Means-Implementation, Problems and Related Methods", Jurnal Sistem dan Informatika, Vol. 3., 2007.

[2] Berkhin Pavel. (2002) 'Survey of Clustering Data Mining Techniques', Accrue Software, Inc.

[3] Horstmann, Cay. S., Gary Cornel, "Core Java ™ Volume I – Fundamentals Eight Editon", Prentice Hall Sun Mycrosystem Press, 2008.

[4] Ian H and Eibe Frank, "Data Mining Practical Machine Learning Tools and Techniques", Morgan Kaufmann Publishers, San Francisco, 2005,

[5] Jiawei Han and Micheline Kember, "Data Mining: Concepts and Techniques Second Edition" , Morgan Kaufmann Publishers, San Francisco, 2006.

[6] Larose, Daniel, "Discovery Knowledge in Data", A Jhon Wiley & Sons, Inc Publication. Canada, 2005.

[7] Manning, Christopher D., Prabhakar Raghavan, Hinrich Schutze, "An Introduction to Information Retrieval", Cambridge: Cambridge University Presss, 2009.

[8] Mardiana T, Rudy D, "Bag-of-Word clusters Using Weka", Jurnal Edukasi dan Penelitian Informatika (JEPIN), Vol. 1, No. 1, ISSN 2460-7041, 2005.

[9] Regulation of the Minister of Health Republic of Indonesia Number 269 / Menkes / Per / III Year 2008 on the Implementation of Medical Record in Hospital.

[10] Silitonga, Parasian "Clustering of Patient Disease Data by Using K-Means Clustering", International Journal of Computer Science and Information Security (IJCSIS), Vol. 15, No. 7, pp 219-221, ISSN 1947-5500, July 2017

[11] Shofari, B, "Management of Medical Record Service System at Hospital", Jakarta : Rineka Cipta, 2002.

[12] Witten, I. H and Frank, E, "Data Mining : Practical Machine Learning Tools and Techniques Second Edition" , Morgan Kauffman, San Francisco, 2005.