Scientific Journal of Informatics
Vol. 3, No. 2, November 2016
p-ISSN 2407-7658 http://journal.unnes.ac.id/nju/index.php/sji e-ISSN 2460-0040

Identification of Tuberculosis Patient Characteristics Using K-Means Clustering

Betha Nurina Sari
Departement of Informatics, Faculty of Computer Science, University of Singaperbangsa Karawang
Email: betha.nurina@staff.unsika.ac.id

Abstract

In Indonesia, tuberculosis remains one of the major health problems unresolved. Indonesia is second ranked in the world as the country with the most tuberculosis cases. The purpose of this research is to study how K-means clustering applied to the treatment of tuberculosis patients data in order to identify the characteristics of tuberculosis patients. The results of K-means clustering validated by gene shaving and silhoutte coefficient. The experiment results indicate the optimum clusters value obtained from the K-mean clustering that has been validated by gene shaving and silhouette coefficient. K-means clustering divided four groups of tuberculosis patients based on their characteristics. There were divided at a category of disease (pulmonary TB, Extra Pulmonary TB and both), the age of the patient and the results of treatment of tuberculosis.

Keyword: characteristic, clustering, K-means, patient, tuberculosis

1. INTRODUCTION

Tuberculosis is an infectious disease caused by *Mycobacterium tuberculosis*. Each year, the WHO estimates that 8.7 million new cases and 1.4 million died of tuberculosis cases. In Indonesia, tuberculosis remains one of the major health problems unresolved. Indonesia is second ranked in the world as the country with the most tuberculosis cases, after India, Indonesia, and China [1]. According to data from Persahabatan Hospital, the number of new patients about 1500 per year. Efforts to control tuberculosis cases are implementing the DOTS strategy (Direct Observed Treatment Shortcourse) which has been implemented at the clinic or hospital within 6-9 months [2].

The application of clustering in the data patient parameter had been done to grouping the variable of Electronic Health Record (EHR), which comprises 24 value lab results clinics and 60 concept clinic of clinic records by using a single linkage agglomerative clustering [3]. Clustering is also applied to data tuberculosis patients in Ethiopia based on spatial data patient. The clustering results are used as material planning control program national tuberculosis can be more effective by identifying the group and target interventions [4]. K-means clustering method ever be applied to identify subgroups of patients based on the response to the Patient - Physician Discordance Scales (PPDS),health status and clinical visits [5].

The purpose of this research is to study how the k-means clustering applied to the treatment of tuberculosis patients data for 6-9 months. The clustering is expected to identify subgroups of patients based on patient characteristics and results of their examinations to treatment.

2. METHOD

a. Research Data

The data used is data administration reports of tuberculosis at the Persahabatan Hospital, East Jakarta. Patient data consists of a progress report on the implementation of the DOTS strategy in a period of 6-9 months. The amount of data used for this study were 235 patient data. The data will be used consists of 11 variables, the 3 variables result of examination of sputum (the beginning of treatment, the second month, and the end of treatment), 3 the results of weight measurement (the beginning of treatment, the second month, the end of treatment), sex, control taking medication, age, type of tuberculosis (pulmonary or extrapulmonary) and category of the intensive phase.

b. Research Methods

Each of tuberculosis patients came to the hospital to undergo inspection pursuant to which the DOTS strategy related recorded sputum examination results and weight. Their incomplete patient data on several parameters needed because at the time the examination is not measured or officer negligent in the treatment of patients recorded in the card. The methods used in solving this problem can be seen in Figure 1:

After a tuberculosis patient examination data entered into the system, to ensure that all parameters have value then applied to linear interpolation. It's as has been done by interpolating the data Hripcsak clinical lab test results and records the concept of patient medical records are still incomplete [6].

After data is interpolated, clustering techniques are applied using K-means clustering which consists of three steps:
1. Determine the centroid / midpoint of each cluster with random.
2. Determine the distance of each object on the coordinates of the midpoint.
The algorithm K-Means will do the repetition step by step until a stable (no object changed) [7].

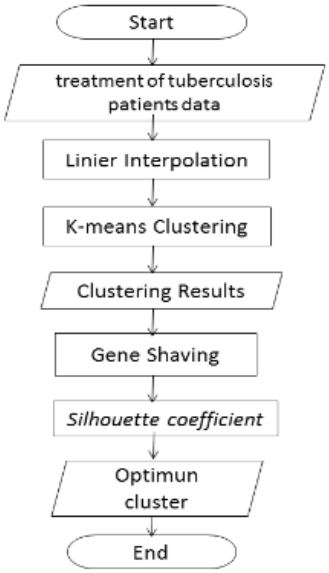

Figure 1. Research Method

Calculation of "dissimilarity" or the distance between the parameters by centroid using the Euclidean distance, A and B:

$$d(A,B) = ((A-B)^2)^{\frac{1}{2}} \quad (1)$$

where A and B are variable values will be calculated distance, d is the distance of each object on the coordinates of the midpoint.

3. Classifying the object is based on the minimum distance.

Testing the performance of algorithm K-means clustering was conducted by Gene Shaving. There are three kinds of cluster variance,ie within variance (VW), between variance (V_B), and the total variance (V_T)[8]. Counting cluster variance is as follows:

$$V_W = \frac{1}{a}\sum_{j=1}^{a}\frac{1}{n}\sum_{i\in S_n}(x_{ij}-\mu_j)^2 \quad (2)$$

where a is the number of attribute values that exist in the data (number of columns), n is the number of data (the sheer number of rows), x is the value of data and m_j is the average of the data on each attribute. V_W is used to view the results of variation of the spread of the existing data on a cluster (internal homogeneity).The smaller the value of V_W, the better cluster, the because it shows the coherent members cluster.

$$.V_B = \frac{1}{a}\sum_{j=1}^{a}(\mu_j-\mu)^2 \quad (3)$$

where a is the number of attribute values that exist in the data (number of columns), m_j is the average data at each attribute and m is the value of cluster centroid .V_B is used to view the results of variation data dissemination inter-cluster (external homogeneity).The larger the value V_B, the better the cluster is formed.

$$V_T = \frac{1}{na}\sum_{i\in S_n}\sum_{j=1}^{a}(x_{ij-}\mu)^2 = V_W + V_B \quad (4)$$

where a is the number of attribute values that exist in the data (number of columns), n is the number of data (the sheer number of rows), x is the value of data and m is the value of centroid of the cluster.

Having obtained the value of the three types of cluster variance,the can be calculated magnitude variance ratio between the variance within the variance between the following manner:

$$R^2 = 100\frac{V_W}{V_T} = \frac{\frac{V_B}{V_W}}{1+\frac{V_B}{V_W}} \quad (5)$$

The value of R^2 indicates the level of coherent members in one cluster.Rated R^2 will show better results when a large value, which is determined from the value V_B are getting bigger and the value of V_W which is getting smaller. The value of R^2 will be stored in the form of a matrix KXK sized according to a number of the clusters.The average of the value of R^2 of everything is stored in the value calculation function. Tahap next gap in the following manner:

$$\mathrm{Gap(k)} = R^2 - \overline{R}^2 \quad (6)$$

The best of cluster gap value will indicate the value of optimum k- cluster.

$$\hat{k} = \mathrm{argmax}_k\ \mathrm{Gap(k)} \quad (7)$$

In addition to the technique of shaving gene,to validate the results of clustering could use the silhouette coefficient. Silhoutte is one popular technique for determining the optimum k value of k-means clustering.formula Silhouette is as follows:

$$s(i) = \frac{b(i)-a(i)}{\max\{a(i),b(i)\}} \quad (8)$$

where a(i) is the average distance between points i and all points in cluster A and b (i) is the average distance between the point i and the points in the cluster closest to cluster A, namely cluster B [9].
Interpreting results of silhouette coefficient can be shown on the chart interpretations of the interval coefficient [10], which is in table 1.

Table 1. Interpretation of silhouette coefficient

Type	Interval silhouette coefficient	Interpretation
1	0.71 – 1.0	The strong structure has been discovered
2	0.52 – 0.70	Reasonable structure has been found
3	0.26 – 0.50	The weak structure may mock
4	< 0.25	Not found substantial structure

3. RESULT AND DISCUSSION

Experiments were carried out as many as four scenarios to input a different number of clusters (k), starting with k = 2, 3, 5, and 7 with iterations 20 times to get the K-means clustering convergent. Experiment K-means clustering using MATLAB functions K-means with parameter replicates for iteration according to the procedure and parameter 'distance' and 'sqEuclidean' which shows the use of euclidean distance to calculate the distance to the centroid point.

Plot the results of K-means clustering for k = 2, 3, 5, and 7 can be seen in Figure 4. Difference members cluster are indicated by differences in color, while the color equation indicates that the data are entered in the same cluster.

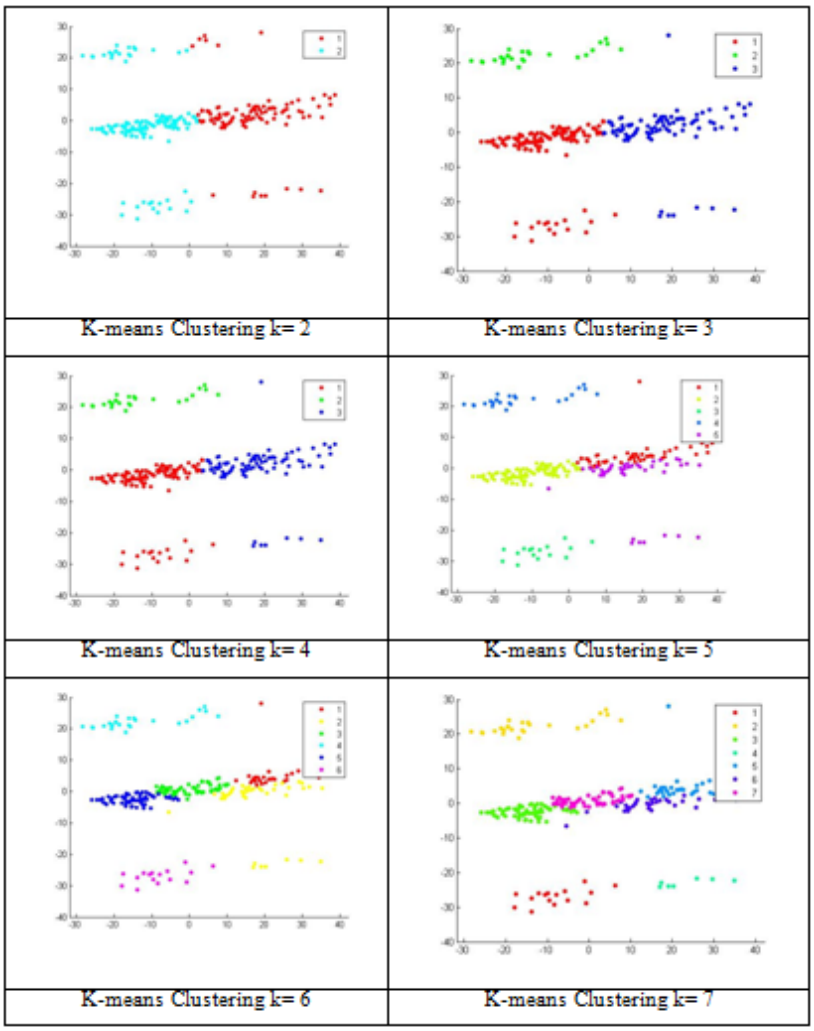

Figure 2. Plot of the Results K-means clustering

At every stage of experiments, the K-means clustering for k = 2, 3, 5, and 7 were calculated the within variance values,the variance between values and the total variance values.The three types of cluster variance were used to calculate the ratio cluster variance.The value of the variance within the K-means clustering results for k = 2, 3, 5, and 7 can be seen in Table 2. The value of variance within this indicated the variance in a single cluster,the smaller the value, the more coherent, compact and similar members in the cluster.

Table 2. Within Variance

The number of clusters	**Within Variance pada Cluster at-**						
	1	2	3	4	5	6	7
k=2	30.888	29.642					
k=3	22.800	16.675	27.508				
k=4	16.675	25.297	15.664	25.973			
k=5	18.751	14.164	17.523	16.676	21.246		
k=6	15.203	21.187	11.361	16.675	11.232	17.523	
k=7	15.364	15.939	7.028	22.340	6.6173	11.970	5.964

Besides within variance,experiment K-means clustering also scores between variance. Between variance calculated the distance between cluster one with cluster another formed at. clusterThe following Table 3 shows the variance between clusters in cluster k = 4 and Table 4 shows the total variance cluster in the clustering at k = 4.

Table 3. The Between Variance at k=4

	Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 4**
Cluster 1		228.754	59.260	130.006
Cluster 2	228.754		62.460	109.011
Cluster 3	59.260	62.460		75.456
Cluster 4	130.006	109.011	75.456	

Table 4. The Total Variance at k=4

	Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 4**
Cluster 1		254.051	74.924	155.979
Cluster 2	245.429		78.123	134.984
Cluster 3	75.936	87.757		101.429
Cluster 4	146.682	134.308	91.120	

After getting the within variance value, the between variance value and the total variance value, the next step was to compute the value of the variance ratio by applying the formula 5.The variance ratio at k = 4 can be seen in Table 5.

Table 5. The Ratio Variance at k=4

	Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 4**
Cluster 1		90.042445	79.0937151	83.3484155
Cluster 2	93.2055819		79.949818	80.7584414
Cluster 3	78.0400444	71.173435		74.3929053
Cluster 4	88.6315265	81.164779	82.8095955	

Having obtained the matrix variance ratio as shown in Table 5, it is averaged for the calculation gap cluster. At k = 4, the average ratio variance is obtained 81.88423 so that a large can be calculated gap on the results of clustering k = 4.The calculation results gap can be seen in Table 6.

Table 6. Calculation cluster gap at k=4

	Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 4**
Cluster 1		8.1582201	-2.79051009	1.46419031
Cluster 2	11.3213567		-1.9344072	-1.12578385
Cluster 3	-3.84418087	-10.710791		-7.49131997
Cluster 4	6.74730128	-0.719446	0.92537024	

In addition to using the techniques of gene shaving to validate the results of K-means clustering, experiments were also conducted with visualizing from results the clustering by Silhoutte plot. Silhoutte is to measure how well the grouping members in the cluster, namely to see how the value of the Silhoutte coefficient.The higher the Silhoutte coefficient is the better clusters. the formed

The results of the experimental implementation of the k-means algorithm to cluster number k = 2, 3, 5, and 7 on the treatment of tuberculosis patients data who have been through validation gene shaving techniques and Silhoutte results obtained maximum value cluster gap and mean of Silhoutte coefficient to determine the optimum k on clustering.The maximum cluster gap value and the mean Silhoutte of each cluster can be seen in Table 7.

Table 7. Maximum cluster gap and mean of silhoutte coefficient at k=2..7

Number of Cluster	**Gap Maximum**	**Mean of silhoutte coefficient**
2	0.4251	0.5150
3	8.9038	0.5324

4	**11.3214**	**0.5497**
5	8.9618	0.5310
6	9.1751	0.4560
7	8.2559	0.4837

From table 6 max cluster gap values and mean of silhouette coefficient can be seen that at k = 4 indicated the optimum value. On the results of k-means clustering for k = 4 maximum cluster gap worth 11.3214, which is the maximum value cluster gap of the largest than others. Similarly, in calculation results from the mean value silhouette results clustering when k = 4 at 0.5497, the value is the highest value than others. The mean of Silhoutte coefficient was classified type 2, which can be interpreted that the cluster has found a reasonable structure or be referred by producing a reasonable cluster.

Results k-means clustering for k = 4 in tuberculosis patient data can divide groups of patients based on specific characteristics. Four clusters generated in this study are shown in Table 7.

Tabel 7. Identify Characteristics from *Clustering* Results at k = 4

Cluster	Total of Membership	Characteristics
Cluster 1	105	patients with pulmonary tuberculosis categories Age of patients in the range 15-46 years (min: 15, max: 46) patients had a median age 29 years (young)
Cluster 2	20	patients with pulmonary tuberculosis categories and extra-pulmonary patients in the age range of 16-55 years (min: 16, max: 55) the average age of patients 33 years of results for the conversion treatment (negative sputum test results)
Cluster 3	85	patients with category pulmonary tuberculosis patients in the age range of 43-81 years (min: 43, max: 81) patients had a median age of 57 years (old age)
Cluster 4	25	patients with extrapulmonary tuberculosis categories age of patients in the range 15-49 years (min: 15 max: 49) Age of patients averaged 29 years of results for the conversion treatment (negative sputum test results)

The result of identification characteristics of the patients showed that the k-means clustering algorithm can be applied to the treatment of tuberculosis patients data. The cluster results may indicate that the pattern in the group of patients at some data variables, such as the category of tuberculosis (TB Pulmonary, extra-pulmonary TB, or both), age of the patient, and the results of treatment of tuberculosis.

4. CONCLUSION

Application of k-means clustering in the data treatment of tuberculosis patients produce that k = 4 is the optimum k cluster.This was validated by the technique of shaving gene and Silhoutte coefficient. The gene shaving generated cluster gap value processed from the count within wariance, between variance, total variance and variance ratio. The Silhoutte calculate the mean Silhoutte value. The value of k optimum cluster can be determined from the results of the cluster in which the value of k gap maximum and the highest mean Silhoutte value. The results of this study found that the k-means clustering algorithm can divide groups of tuberculosis patients based on their characteristics, namely by category of disease (pulmonary TB, Extra Pulmonary TB and both), the age of the patient and the results of treatment of tuberculosis. It can be used as a consideration in the decision-making related to the treatment of tuberculosis by observing the characteristics of the patient.

5. REFERENCE

[1] World Health Organization. 2015. Global Report Tuberculosis 2015. https://www.health-e.org.za/wp-content/uploads/2015/10/Global-TB-Report-2015-FINAL-2.pdf accessed at 1 September 2016

[2] Ministry of Health of the Republic of Indonesia.. 2011. Laporan situasi terkini perkembangan tuberkulosis di Indonesia Januari-Desember 2011. http://id.scribd.com/doc/ 209975729/Kementerian-Kesehatan-RI-Laporan-Situasi-Terkait-Perkembangan-Tuberkulosis-Di-Indonesia-2011. diakses 1 September 2016

[3] Hripcsak, George, dan Albers DJ. 2013.Correlating electronic health record concepts with healthcare process events. J Am Med Inform Assoc. Vol.20 : 311-18.

[4] Tadesse, Takele,et all. 2013. The Clustering of Smear-Positive Tuberculosis in Dabat, Ethiopia: A Population Based Cross Sectional Study. Plos One. Vol. 8.5: 1-6.

[5] Sewitch, Maida J, Karen Leffondre, dan Patricia L. Dobkin. 2004. Clustering patients according to health perceptions Relationships to psychosocial characteristics and medication nonadherence. Journal of Psychosomatic Research. Vol.56: 323 – 332.

[6] Hripcsak, George, Albers DJ, Perrote A. 2011. Exploiting time in electronic health records correlations. J Am Med Inform Assoc. Vol.18; 109-15

[7] Tekmono, Kardi. 2007. K-means Clustering Tutorial. http://www.croce.ggf.br/dados/ K%20mean%20Clustering1.pdf. diakses 1 September 2016.

[8] Hastie, Trevor, et al. 2000. 'Gene Shaving' as a method for identifying distinct sets of genes with similar expression patterns. Genome Biology. Vol.1(2): 1-21

[9] Al-Zoubi, Moh'd Belal dan Mohammad al Rawi. 2008. An Efficient Approach for Computing Silhoutte Coefficients. Journal of Computer Science. Vol.4(3): 252-255

[10] Struyf, Anja, Mia Hubert, dan Peter J.Rousseeuw. 1997. Clustering in an Object-Oriented Environment. Journal of Statistical Software. Vol. 1(4) : 1-30.