The implementation of k-means partitioning algorithm in HOPACH clustering method

The implementation of k-means partitioning algorithm in HOPACH clustering method

K R Adzima[1], A Bustamam[2]*, D Aldila[3]

[1]Universitas Esa Unggul, Jakarta, Indonesia
[2*3]Universitas Indonesia, Depok, Indonesia

Emal:{khaola.rachma, *alhadi, aldiladipo}@sci.ui.ac.id

Abstract. Hierarchical Ordered Partitioning And Collapsing Hybrid (HOPACH) is one of the powerful clustering methods which combine the strengths of partitioning and agglomerative clustering methods. Several partition clustering methods such as PAM, K-Means, SOM, or other partitioning algorithms can be used in the partitioning process. This process is followed by the ordering steps, then continued with the agglomerative process. The number of main clusters is determined by MSS (Mean Split Silhouette) value. MSS is used to measure the heterogeneity of the clustering result. The lower the MSS value, the more homogenous each cluster members. We select the number of clusters from the clustering results with minimum MSS. In this implementation of HOPACH, we incorporate k-Means partitioning algorithm in this HOPACH clustering method, to cluster and analyze 136 DNA sequences of Ebola viruses. The clustering process is started with collecting DNA sequences of Ebola viruses from GenBank, then followed by performing features extraction of these DNA sequences using N-Mers frequency. The extraction results are compiled to be a features matrix and normalized using the *min-max* normalization with the interval [0, 1] as an input data to generate genetic distance matrix using Euclidian distance. The genetic distance matrix is used in partitioning process by the K-Means algorithm in HOPACH clustering. As the results, we obtained 8 clusters with minimum MSS (Mean Split Silhouette) 0.50266. The clustering process in this article uses the open source program R.

1. Introduction

Bioinformatics is a study of the application of computational techniques in managing and analyzing biological data [1]. The primary sources of biological data in bioinformatics are DNA, RNA, and protein. One objective of the study of DNA or other expressions of genes is to find biologically important subsections and groups of genes [7].

Clustering methods can be categorized into partitioning and hierarchical methods. Partitioning method aims to find clusters contained in the data by optimizing the function of specific objectives to improve the quality of the partition. Furthermore, hierarchical clustering is a method that has the approach to develop a binary tree-based data structure called dendrogram. The hierarchical method is divided into two different approaches, namely the so-called bottom-up as agglomerative and top-down called the divisive. The agglomerative method combines multiple groups into one large group. The divisive method divides the large group into small groups [8].

In 2001, Van der Laan and K. S. Pollard proposed a new clustering method that is Hierarchical Ordered Partitioning and Collapsing Hybrid (HOPACH). This method objective is to combine the power of both partitioning and agglomerative clustering method [7]. The HOPACH method can be applied with several partitioning algorithms, for example, K-Means algorithm, SOM, Fuzzy C- Means algorithm, and PAM [7]. A study for using the HOPACH clustering with PAM partitioning method has been successfully conducted by Muradi, et al. in 2015 [3].

One of the popular clustering methods is K-Means clustering method. K-Means method introduced by James B MacQueen in 1967 in the proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability [9]. In this method, the base of the clustering process is to place the object based on the average (mean) of the nearest cluster. K-Means clustering is included in the non-hierarchical methods that aim to categorize n objects into k clusters ($k < n$), where k is predetermined value [4].

In this research, we study the implementation of K-Means partitioning algorithm in the HOPACH clustering method for clustering DNA sequences. This study is based on the research that has been done by Bhaskar Mondal and J. Paul Choudhury which state that K-Means clustering using Euclidean distance provide better accuracy than the PAM algorithm [2].

2. Research Methodology

The stages of this study are collecting data in the form of DNA sequences, feature extraction using N-Mers frequency, normalization of the results of feature extraction using *min-max* normalization, clustering of DNA sequence using HOPACH clustering method by K-Means partitioning algorithm, and discussing the results of the clustering.

2.1. HOPACH Clustering with K-Means

HOPACH combines the power of clustering methods partitioning and agglomerative [7]. The HOPACH algorithm tries to build the tree groups of a set of elements (i.e., genes), where groups at every level are based on inequality of their medoids. The HOPACH algorithm starts from the root nodes to determine the number of branches for the nodes by doing the partitioning steps and agglomerative interchangeably. The scheme of HOPACH method can be seen in Fig.1 [6].

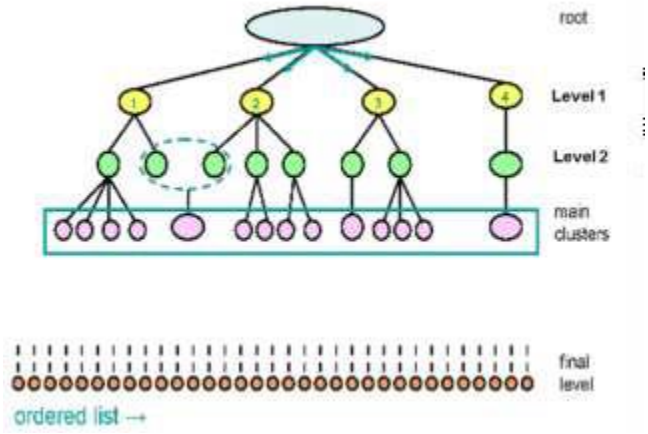

Figure 1. The scheme of HOPACH method

First Step (level 1):

1. Partitioning: the root is divided into several clusters with the partitioning algorithm.
2. Ordering: the ordering process based on the distance between the centroid of the DNA sequences.

Next Step (next level):

1. Partitioning: each clusters from the result of level 1 is divided into several clusters with the partitioning algorithm.

2. Ordering: ordering the clusters from the partitioning result.
3. Collapsing: collapsing two clusters into one cluster if it make the better MSS value.

The partitioning approach in this HOPACH method is K-Means partitioning algorithm. K-Means aim to partition the data into one or more clusters. The steps of HOPACH method using K-Means partitioning algorithm [4] can be arranged into the flowchart in Fig.2.

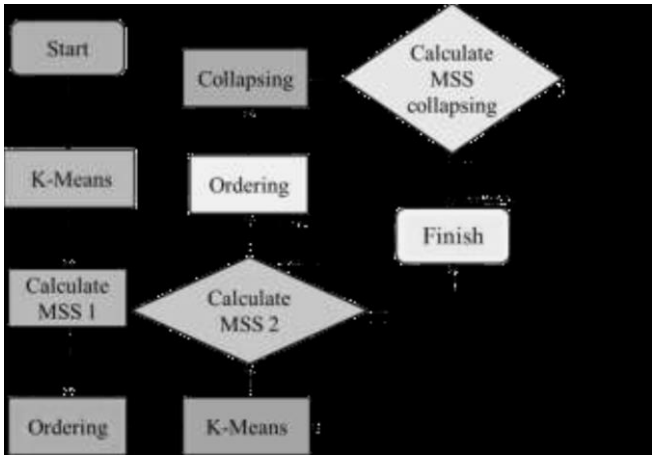

Figure 2. Flowchart of HOPACH clustering with K-Means partitioning algorithm

2.2. Data Source

This research uses 136 DNA sequences of Ebola virus which collected from GenBank in NCBI website, http://www.ncbi.nlm.nih.gov/. The selected DNA sequences are complete genome which is expected to provide more genetic information. These DNA sequences are derived from four different species, including Zaire, Bundibugyo, Tai Forest, and Sudan Ebola virus.

2.3. Feature Extraction using N-Mers Frequency

After collecting the data, feature extraction process performed on the 136 DNA sequences of Ebola virus. The method of feature extraction in this study is N-Mers frequency [10]. This method is used to determine the number of occurrences of the certain substring in a string. The intensity of the emergence of the string can be used as a cluster identifier of a string. Appearance pattern in the DNA sequences is calculated by using the four main DNA bases A, C, G, and T. Then to the power of a series of base pairs that you want to use, i.e., n. Thus, the occurrence pattern is $4n$, where $\boldsymbol{n \geq 1}$. Feature extraction used for the data set is The N-Mers frequency with $n = 3$, then the appearance pattern in DNA sequences used for feature extraction is $4^n = 4^3 = 64$.

2.4. Normalization DNA Feature Extraction

The results of the feature extraction that have been made will obtain the data with a very varied value. The values obtained from the feature extraction needs to be scaled to the specified limit value to avoid the biggest data or the smallest data because the large data range will affect the outcome of the clustering. This process is called data normalization. Data normalization in this research uses *min-max* normalization formula. *min-max* normalization maintains the relation of the original data [4].

$$v' = \frac{v - \min}{max - min}(\text{new}_{\max} - \text{new}_{\min}) + new_{min} \quad (1)$$

Where v is an element value before normalization, v' is an element value after normalization, min is a minimum element value of the matrix, max is a maximum element value of the matrix, new_{max} is the new maximum value limit, new_{min} and is the new minimum value limit.

2.5. Result and Discussion

In the first stage of the partition (level 1), 136 DNA sequences of Ebola virus are divided into 2 to 9 groups using K-Means partitioning algorithm. Then, calculate the MSS value of each clustering result using the MSS formula [9] to determine the best number of cluster.

$$MSS(k) = \frac{1}{k}\sum_{i=1}^{k} SS_i \quad (2)$$

Where k is is the number of clusters and SS_1 is the average silhouette in cluster-i.

$$S_j = \frac{b_j - a_j}{\max(a_j, b_j)} \quad (3)$$

Where S_j is the silhouette of gene j, a_j is the average distance between gene j and the other gene in the same cluster, $b_j = \min_k b_{jk}$ when b_{jk} is the average distance between gene j and other genes in the cluster k. The calculation of MSS value for $k = 2$ to $k = 9$ can be seen in Table 1.

Table 1. MSS value of partition level 1

k	2	3	4	5
MSS	0.73187	0.66838	0.66964	0.62384
k	6	7	8	9
MSS	0.54443	0.56142	0.50266	0.51641

According to the table above, the minimum MSS value is in $k = 8$, that is 0.50266. Based on the theory, the best number of the cluster has the minimum value of MSS. So, the number of the cluster in the first stage partition is 8, and the results of the classification can be seen in Fig. 3.

```
Clustering with the number of cluster =  8

The member of cluster  1 :  118 119 120
 121 122 123 124 125 126 127 128 129 130
 131 132 133 134 135 136
The member of cluster  2 :  3 9 12 15 16
 22 26 27 28 29 53 55 56 62 68 69 72 74 75
 76 86 92 94 97 99 100 101 116
The member of cluster  3 :  8 19 23 25 33
 45 63 71 73 95 98
The member of cluster  4 :  1 2 102 103
 104 105 106 107 108 109 110 111 112 113
 114 115 117
The member of cluster  5 :  4 5 7 11 14 30
 35 38 39 42 43 49 61 65 66 67 77 79 91
The member of cluster  6 :  6 10 13 31 37
 41 44 50 57 60 78 85
The member of cluster  7 :  18 21 24 32 34
 36 40 46 48 51 59 64 70 80 82 84 89 93
The member of cluster  8 :  17 20 47 52 54
 58 81 83 87 88 90 96
```

Figure 3. The result of partition level 1 based on MSS value

Having obtained the best 8 cluster and its members, each cluster has to be sorted by the centroid distance of each group. The ordering results can be seen in Table 2.

Table 2. The result of partition and ordering level 1

Cluster	Member of Cluster (DNA Ebola Virus)
1	E4, E5, E7, E11, E14, E30, E35, E38, E39, E42, E43, E49. E61. E65, E66, E67, E77, E79, E91.
2	E118, E119, E120, E121, E122, E123, E124, E125, E126, E127, E128, E129, E130, E131, E132, E133, E134, E135, E136.
3	E18, E21, E24, E32, E34, E36, E40, E46, E48, E51, E59, E64, E70, E80, E82, E84, E89, E93.
4	E17, E20, E47, E52, E54, E58, E81, E83, E87, E88, E90, E96.
5	E6, E10, E13, E31, E37, E41, E44, E50, E57, E60, E78, E85.

6	E8, E19, E23, E25, E33, E45, E63, E71, E73, E95, E98.
7	E1, E2, E102, E103, E104, E105, E106, E107, E108, E109, E110, E111, E112, E113, E114, E115, E117.
8	E3, E9, E12, E15, E16, E22, E26, E27, E28, E29, E53,E55, E56, E62, E68, E69, E72, E74, E75, E76, E86, E92, E94, E97, E99, E100, E101, E116.

In level 2, partitioning process applied to each cluster generated at level 1 using the K-Means algorithm by the same way as was done at the partition level 1. Then followed by calculating the value of MSS partition level 2. The calculation of the value of MSS partition level 2 can be seen in Table 3.

Table 3. MSS value of partition level 2

Cluster	Number of Sub-Cluster	MSS min
1	8	0.3135504
2	9	0.3372123
3	8	0.2075739
4	2	0.2459123
5	2	0.2260273
6	8	0.08065827
7	8	0.5528140
8	3	0.3541295

In the previous process, the result of the MSS value partition level 1 is 0.50266. Table 2 shows that the value of minimum MSS partition level 2 is at k = 7, that is 0.5528140, which means that the value of MSS partition level 2 is greater than that of MSS partition level 1. Because the value of MSS partition level 1 is still better than MSS value partition level 2, then the clustering process is completed. So, the clustering result of 136 DNA Ebola virus using HOPACH partition clustering algorithm K-Means is as shown in Table 3.

Dendrogram cluster 1 can be seen in Fig. 4. Cluster 1 consist of 19 DNA Ebola sequences that are derived from the same species, that is Zaire Ebola virus, and coming from the same city, Makona and the same year, 2014. The kinship between DNA Ebola sequences in cluster 1 can also be seen in Fig. 3. The genetic distance between DNA Ebola sequences can be viewed vertically. DNA sequences E11 and E12 have the minimum genetic distance, while the DNA sequences E91 and E79 have maximum genetic distance, which is 0.1039573.

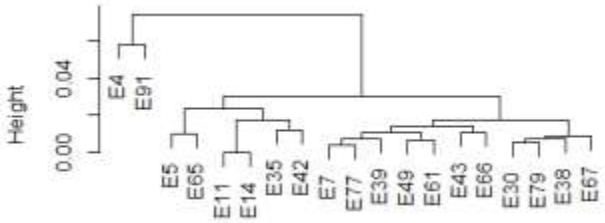

Figure 4. Dendrogram cluster 1

Dendrogram cluster 2 can be seen in Fig. 5. In cluster 2, there are 10 DNA sequences derived from Sudan Ebola virus species, i.e., E118, E119, E120, E121, E122, E123, E124, E125, E126, E127. There are 7 Ebola DNA sequences derived from Bundibugyo Ebolavirus, i.e., E128, E129, E130, E131, E132, E133, E134, E135, while derived from the Tai Forest Ebola virus species and E136 from Cote d'Ivoire Ebolavirus DNA sequence E135 and E136 have the minimum genetic distance, while DNA sequences E135 and E121, also E136 and E121 have maximum genetic distance.

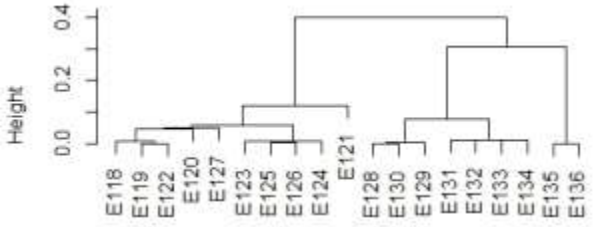

Figure 5. Dendrogram cluster 2

Dendrogram cluster 3 can be seen in Fig. 6. Cluster 3 consist of 18 DNA Ebola sequences that are derived from the same species, Zaire Ebolavirus and coming from the same city, Makona, and the same year, 2014. DNA sequences E70 and E18 have the minimum genetic distance, while the DNA sequences E32 dan E21 have maximum genetic distance.

Dendrogram cluster 4 can be seen in Fig. 7. Cluster 4 consist of 12 DNA Ebola sequences that are derived from the same species, Zaire Ebola virus and coming from the same city, Makona and the same year, 2014. DNA sequences E88 and E17, also E52 and E47 have the minimum genetic distance, while the DNA sequences E83 and E58 have maximum genetic distance.

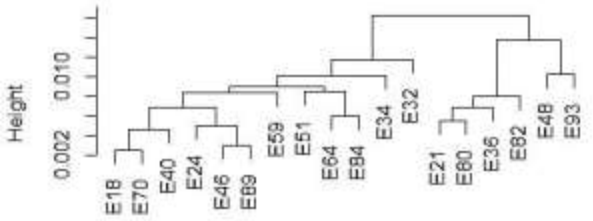

Figure 6. Dendrogram cluster 3

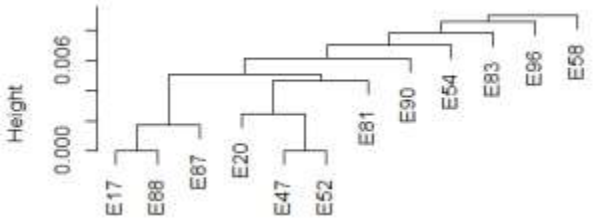

Figure 7. Dendrogram cluster 4

Dendrogram cluster 5 can be seen in Fig. 8. Cluster 5 consist of 12 DNA Ebola sequences that are derived from the same species, Zaire Ebola virus and coming from the same city, Makona and the same year, 2014. DNA sequences E44 and E31, also E57 and E60 have the minimum genetic distance, while the DNA sequences E57 and E10, also E60 and E10 have maximum genetic distance.

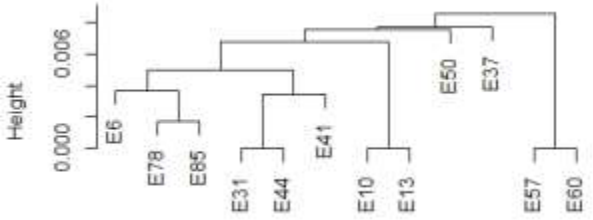

Figure 8. Dendrogram cluster 5

Dendrogram cluster 6 can be seen in Fig. 9. Cluster 6 consist of 11 DNA Ebola sequences that are derived from the same species, Zaire Ebola virus and coming from the same city, Makona and the

same year, 2014. DNA sequences E98 and E63 have the minimum genetic distance, while the DNA sequences E95 and E33 have maximum genetic distance.

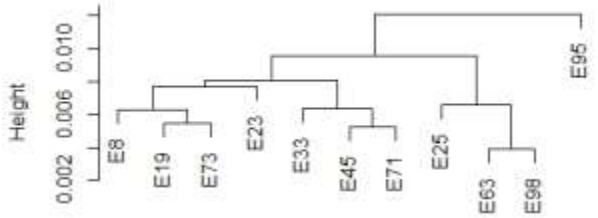

Figure 9. Dendrogram cluster 6

Dendrogram cluster 7 can be seen in Fig. 10. Cluster 7 consist of 17 DNA Ebola sequences that are derived from the same species, Zaire Ebola virus but in different years of endemic. 2 DNA Ebola sequences in this cluster are from 1976. 4 DNA sequences from 1995. 1 DNA sequences from 1994. 5 DNA sequences from 1996. 2 DNA sequences derived from 1976. DNA sequences E1 and E2, E107 and E106, also E108 and E105 have the minimum genetic distance, while the DNA sequences E112 dan E104 have maximum genetic distance.

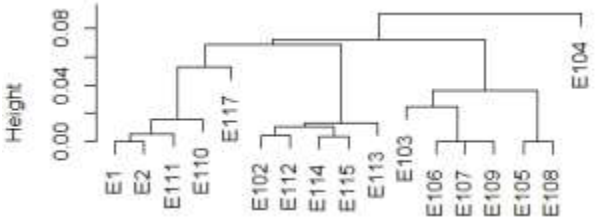

Figure 10. Dendrogram cluster 7

Dendrogram cluster 8 can be seen in Fig. 11. Cluster 8 consist of 28 DNA Ebola sequences that are derived from the same species, Zaire Ebola virus and coming from the same city, Makona and the same year, 2014. DNA sequences E12 and E9, E116 and E76, also E101 and E94 have the minimum genetic distance, while the DNA sequences E100 and E55 have maximum genetic distance.

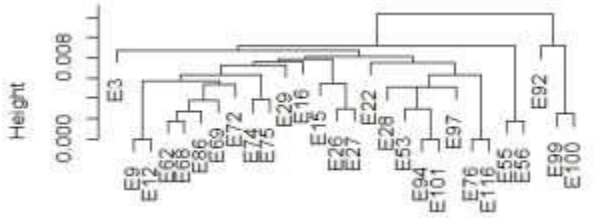

Figure 11. Dendrogram cluster 8

Based on the explanation above, each cluster of 136 DNA Ebola virus sequences from the clustering result have characteristics that are based on species and years of endemic of DNA Ebola virus sequence.

3. Conclusion and Suggestion

3.1. Conclusion

The implementation of K-Means algorithm in HOPACH method involves several steps; the first step is feature extraction process of DNA sequences of Ebola virus using N-Mers frequency. Then followed by the normalization process of feature extraction results. These steps continue to the partitioning level 1 using the K- Means algorithm. Also calculating the MSS partition level 1 to know which one is having the minimum MSS value. After that, we perform the ordering process which is

based on the distance between the centroid of the DNA sequences. Then, partitioning the result of the ordering level 1 using K-Means algorithm into several cluster and also calculating the value of MSS partition level 2. Then, compare the value of MSS partition level 1 and level 2 to determine the next process. If the value of MSS partition level 1 is still lower than MSS value partition level 2, then the clustering process is completed. If the value of MSS partition level 1 is greater than MSS value partition level 2, then continue to the ordering dan clustering process level 3. The complete clustering steps can be seen in the flowchart shown in Fig. 2.

From these clustering results, 136 DNA sequences of Ebola virus spread in 8 different clusters. The minimum MSS value from the clustering is 0.50266. Each cluster of the clustering results have the characteristics based on the species and years of endemic. The result of the clustering can be seen in Table 2.

3.2. Suggestion

The comparison of HOPACH clustering using K-Means partitioning algorithm, SOM, Fuzzy C-Means, and other partition algorithms with different case studies from this research can be done for the next study. Then, the MSS criteria for determining the best number of the cluster in every stage of HOPACH clustering can be replaced by other criteria, such as Calinski and Harabasz (CH), Hartigan (H), and Krzanowski and Lai (KL).

Acknowledgement
The research is supported by PDUPT 2018 research grant.

References

[1] Bergeron B 2002 *Bioinformatics Computing* (Cambridge: Prentice Hall PTR)
[2] Mondal B and Choudhury J P 2013 *International Journal of Computer Applications* **78** (5)
[3] Muradi H, Bustamam A, and Lestari D 2015 *International Conference on Advanced Computer Science and Information Systems (ICACSIS)* IEEE 317-323
[4] Han J and Kamber M 2000 *Data Mining: Concepts and Techniques* (San Francisco: Morgan Kaufmann Publishers)
[5] Pollard K S and Van der Laan M J 2002 *SCI2002 Proceeding* **II** 318-325
[6] Pollard K S and Van der Laan M J 2005 *Cluster Analysis of Genomic Data with Application in R* (Berkeley: University of California)
[7] Van der Laan M J and Pollard K S 2003 *Journal of Statistical planning and Inference* **177** 275-303
[8] Reddy and Aggarwal 2014 *Data Clustering: Algorithms and Applications* (New York: Taylor and Francis Group)
[9] Johnson R and D W W 2007 *Applied Multivariate Statistical Analysis* (Upper Saddle River Jersey Prentice Hall)
[10] Deorowicz S, Kokot M, Grabowski S, and Debudaj-Grabysz A 2015 *Oxford Journals of Bioinformatics* **31** Issue 10 1569-1576