|| Volume 3 || Issue 7 || July 2018 || ISSN (Online) 2456-0774
INTERNATIONAL JOURNAL OF ADVANCE SCIENTIFIC RESEARCH
AND ENGINEERING TRENDS

Applications of Partition-Based Algorithms for Clustering Dengue Patients in Sri Lanka

Mohamed Cassim Alibuhtto[1], Nor Idayu Mahat[2]
Department of Mathematical Sciences, Faculty of Applied Sciences, South Eastern University of Sri Lanka, Sri Lanka[1]
Department of Mathematics and Statistics, School of Quantitative Sciences, Universiti Utara Malaysia, Malaysia[2]
mcabuhtto@seu.ac.lk

***Abstract*— Dengue is a mosquito-borne viral disease that has rapidly spread in different regions of Asia, Latin America, Africa, and Oceania over the past few years. Similarly, Sri Lanka is facing the same epidemic challenges for more than two decades with the spread of the disease varies according to regions in a country. This paper aims to recognize the number of clusters among 26 regions in Sri Lanka that may explained some factors that are contributing to the spread of the disease. The clustering was performed using partition-based clustering techniques namely k-means and k-medoids. The results show that the appropriate number of clusters of dengue fever data is two, and one cluster consists of two regions, and other cluster contains the rest regions. This study may help the health sector to analyses and explore the key factors associated with the dengue fever regions**

Keywords*: Clustering, dengue, data mining, k-means algorithm, k-medoids algorithm.*

I INTRODUCTION

Dengue is a viral infection caused by four types of the virus namely DENV-1, DENV-2, DENV-3, and DENV-4, belonging to Flaviviridae family. The virus spreads through bites of infected female mosquitoes Aedes aegypti and Aedes albopictus, which are fed indoors and outdoors during the daytime. These mosquitoes grow in areas with stagnant water, including ponds, water tanks, containers, and old tires. The lack of reliable sanitation and regular collection of garbage often contribute to the spread of these mosquitoes (IAMAT, 2018). Sri Lanka faces unprecedented outbreaks of dengue fever, leading to more than 100,000 cases and claiming about 300 lives in year 2017. Besides, the number of cases reported in year 2018 is three and a half times higher than the average number of cases for the same period from 2010 to 2016. The tropical and warm climate of Sri Lanka is suitable for breeding Aedes mosquitoes hence contributes to the increment number of dengue fever among Sri Lankans. Evidence has shown that dengue fever cases peaked in June 2017, equivalent to the South-West monsoon rains begins at the end of May. Approximately 43% of cases reported from the Western region. The dengue epidemic after heavy rains and floods and subsequent landslides affecting 600,000 people in 15 of the 25 districts of the country (WHO, July 11, 2017).

A study the trends in prevalence of dengue geographical region are to prevent and control the spread of dengue fever. The urban areas of southern Taiwan have the highest prevalence rate due to high population density, tropical climate and the presence of Aedes aegypti mosquitoes (Hsu, Hsieh, & Lu, 2017). A possible way to recognize the clusters of geographical regions is through cluster analysis. Cluster analysis is an unsupervised learning method that is basis of a data mining process in various fields such as finance sector, business sector and health sector and many more (Dolnicar, 2003).

Many studies were conducted to analyse dengue viruses specially focused on the early detection and control the dengue viruses (Rodrigues, Monteiro, Torres, & Zinober, 2012; Beckett, Kosasih, Faisal, & Tan, 2005). However, few studies were conducted to cluster dengue fever patients according to region wise. The notion of region clustering is to recognize the similarity among regions that explain how and why the disease spread. Therefore, this paper attempts to determine the number of optimal clusters using the partitioning clustering methods on dengue fever data. This paper is arranged as follow.

II CLUSTER ANALYSIS

The cluster analysis is a major task in data analysis and groups objects into clusters mutually such that the objects in the same cluster are similar whilst objects in different clusters are different. Apparently, different clustering algorithms have been devoted by researchers to address different conditions such as *k*-means, *k*-prototype, agglomerative clustering and many more. Therefore, the choice of clustering algorithm depends on the nature of the application (Visalakshi & Suguna, 2009). A good clustering technique will produce high superiority clusters with "high intra-class similarity" and "low inter-class similarity" (Han & Kamber, 2006).

The *k*-means clustering is a common unsupervised centroid based clustering technique that determines a specified number of non-overlapping clusters within data (Aggarwal & Reddy, 2014; Gan, Ma & Wu, 2007), and this algorithm was proposed by Macqueen in 1967. This *k*-means algorithm is used to cluster observations into groups of related observations without prior knowledge of the relationship (Kanth, Balaram, &

Rajasekhar, 2014). However, the *k*-means clustering is sensitive to outliers (Hartigan & Wong, 1979). *k*-medoids algorithm is a non-hierarchical clustering algorithm which is slightly modified from the k-means algorithm. The most common *k*-medoids clustering technique is the PAM (Partitioning Around Medoids) algorithm (Kaufman & Rousseeuw, 1990). This algorithm is more robust to noise and outliers than *k*-means algorithm (Vishwakarma, Nair, & Rao, 2017). In *k*-means algorithm, to identify the cluster *k* as representative object to minimize sum of squared Euclidean distances for the data objects, whereas *k*-medoids uses the sum of dissimilarities of data objects.

III METHODOLOGY

3.1 Data

This paper used monthly basis number of suspected dengue fever cases from January 2010 to December 2017 in twenty-six regions in Sri Lanka. This data was obtained from Epidemiology Unit, Ministry of Health of Sri Lanka. These regions are: Colombo, Gampaha, Kalutara, Kandy, Matale, Nuweraliya, Galle, Hambantota, Matara, Jaffna, Kilinochchi, Mannar, Vavuniya, Mulativu, Batticaloa, Ampara, Trincomalee, Kurunegala, Puttalam, Anurathapura, Polonnaruwa, Badulla, Moneragala, Ratnapura, Kegalle and Kalmunai.

3.2 Clustering Algorithm

Two partition based clustering algorithms i.e., *k*-mans and *k*-medoids were considered to determine the number of clusters of RDHS (Regional Director of Health Services) regions with dengue fever patients.

3.2.1 K-means Algorithm

This paper adopted the common *k*-means clustering algorithm with the setting as follow:

1. Randomly select *k* of the objects, each of which initially represents a cluster mean.
2. For each of the remaining objects, an object is assigned to the cluster to which it is the most similar, based on the distance between the object and the cluster mean.
3. Then computes the new mean for each cluster.
4. This process iterates until the criterion function converges.

The *k*-means algorithm depends on minimizing the sum of the squared error function, which is very simple and can be easily implemented.

$$J = \frac{1}{n}\sum_{i=1}^{k}\sum_{x\in C_i}\left\|x - m_i\right\|^2 \qquad (1)$$

where J is the sum of the square error for all objects in the data set, x is the point in space representing a given object; and m_i is the mean of cluster C_i and k is the given number of clusters.

3.2.2 K-Medoids Algorithm

As *k*-means can be arguable if the data may be contaminated with outliers or the data are widely scattered, then *k*-medoids algorithm is a good alternative.

The *k*-medoids algorithm works in the following way:

1. Randomly select *k* data points as initial medoids *m*.
2. Assign remaining data points to the cluster with closest medoid.
3. Determine new medoid such as for each medoid *m* and data points *x* associated to that medoid swap *m* and x and find the distance between the remaining data points and data points associated to the medoid. Select that data point with minimum distance as new medoid.
4. Repeat steps 2 and 3 until there is no change in the assignment.

3.3 Cluster Validation Measures

The evaluation of cluster results for determining the validity of clusters is the most difficult task, particularly, when no-prior information is known. There are several evaluation methods that measure the quality of clustering results, which can be categorized as internal and external indices. Besides, the internal indices allow to evaluate the clusters when the previous information is not available, whereas external indices evaluate the clusters based on existing information. In this study, Dunn index and silhouette measures are used to evaluate the clustering results. Both indices are briefly described below.

3.3.1 Dunn Index

This Dunn index was introduced by Dunn in 1974 and define as the ratio between the minimal intra cluster distance to maximal inter cluster distance. This measurement serves as a measure to find the appropriate number of clusters in a dataset, where the largest value of the index is the accurate partition based on the index. The Dunn index is as follows:

$$D_k = \min_{1\le i\le k}\left\{\min_{\substack{1\le j\le k\\ i\neq j}}\left\{\frac{d(C_i,C_j)}{\max_{1\le l\le k}\{diam(C_l)\}}\right\}\right\} \qquad (2)$$

Where $d(C_i, C_j)$ is the distance between two clusters C_i and C_j as the minimum distance between a pair of objects in the two different clusters separately and the diameter of cluster C_l, diam(C_l), as the maximum distance between two objects in the cluster.

3.3.2 Silhouette method

This index is a summation type index, and this is an indicator of how close the points of the cluster are to each other compared to the points outside the cluster. Cohesion is measured on the basis of the distance between all points in a cluster, and the separation is based on the closest neighbour distance (Kaufman & Rousseeuw, 1987; Milligan & Cooper, 1985). The silhouette index [S(i)] is defined as:

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad \text{——————} (3)$$

Where $a(i)$ be the average dissimilarity of i with all other data item within the same clusters. $b(i)$ be the lowest average dissimilarity to any other cluster.

The silhouette coefficient varies from −1 to +1, where a high value (close to one) indicates that the data is properly clustered.

IV RESULTS AND DISCUSSIONS

4.1 Descriptive of Dengue Patients from 2010 to 2017

Figure 4.1 shows the patterns of dengue patients in Sri Lanka from 2010 to 2017. The figure shows that more people are suffering from dengue fever in 2017 (40.66%) in all RDHS areas, and many of them are affected with dengue fever in Colombo (24.58%) and Gampaha (15.72%), located in the Western province. However, fewer dengue patients are reported in Kilinochi (0.23%) and Mulaitivu (0.23%), where both regions are located in Northern province of Sri Lanka.

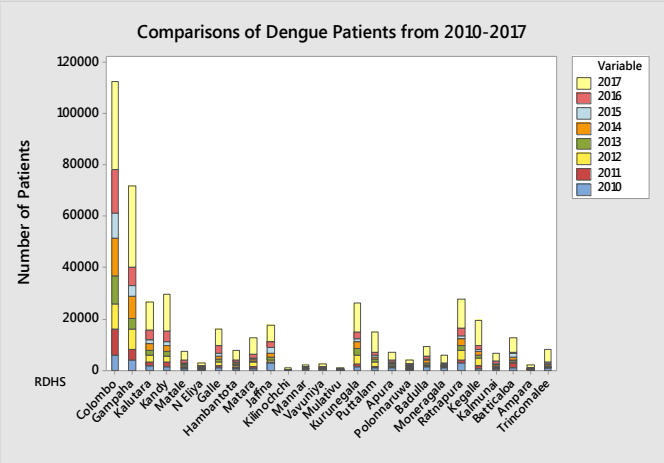

Figure 4.1. The trend of dengue patients in Sri Lanka from 2010-2017

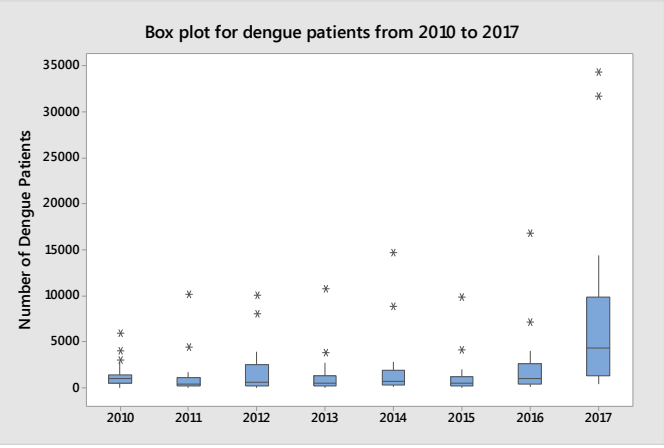

Figure 4.2 Box plot for dengue patients from 2010 to 2017.

Figure 4.2 shows the box plot of dengue fever patients from 2010 to 2017. It can be observed that, there are few outliers in each year.

4.2 Clustering Dengue Patients

The investigation of *k*-means of different values of *k* is displayed in Table 4.1. Also summarised in the table are evaluation indexes namely (full description of SS), Dunn Index, and average value of Silhouette. The highest values of Dunn index and the average silhouette is 1.2561 and 0.82 respectively, which at *k*=2. These results indicate that the potential optimal number of clusters of dengue data is two. These findings were confirmed with the silhouette plot (Figure 4.3).

TABLE 4.1 CLUSTER RESULTS OF DENGUE PATIENTS BY *K*-MEANS ALGORITHM

k	**Cluster sizes**	**clustering vector**	**SS**	**Dunn Index**	**Average Silhouette**
2	24,2	***2 2*** 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	78.3	***1.2561***	***0.82***
3	9,15,2	3 3 1 1 2 2 1 2 1 1 2 2 2 2 2 2 2 1 1 2 2 2 2 1 1 2	90.2	0.1350	0.56
4	15,9,1,1	4 3 2 2 1 1 2 1 2 2 1 1 1 1 1 1 1 2 2 1 1 1 1 2 2 1	93.1	0.2420	0.52
5	7,2,5,5,7	2 2 3 3 1 5 4 1 4 4 5 5 5 5 4 5 1 3 4 1 5 1 1 3 3 1	97.6	0.1065	0.53
6	7,1,5,5, 1,7	5 2 3 3 6 1 4 6 4 4 1 1 1 1 4 1 6 3 4 6 1 6 6 3 3 6	98.5	0.3016	0.50

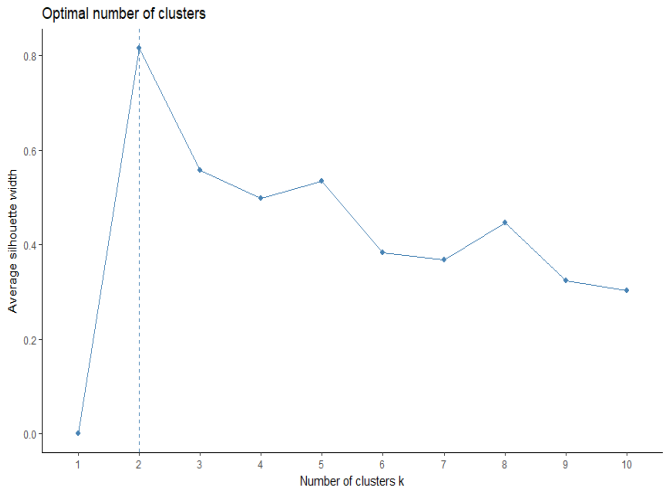

Figure 4.3 Silhouette plot for k-means algorithm

Meanwhile, the results of dengue patients clustering using *k*-medoids are Table 4.2. The results also indicated that

the clustering is optimized at k-2 as the Dunn index and average value of Silhouette are the highest at these two indices.

TABLE 4.2 CLUSTER RESULTS OF DENGUE PATIENTS BY *K*-MEDOIDS ALGORITHM

k	Cluster sizes	clustering vector	Dunn Index	Average Silhouette
2	24,2	***1 1*** 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2	***1.2561***	***0.82***
3	2,6,18	1 1 2 2 3 3 3 3 3 3 3 3 3 3 3 3 3 3 2 2 3 3 3 3 2 2 3	0.1072	0.58
4	2,5,12,7	1 1 2 2 3 4 3 3 3 3 4 4 4 4 3 4 3 2 3 3 4 3 3 2 2 3	0.1065	0.50
5	1,1,5,12,7	1 2 3 3 4 5 4 4 4 4 5 5 5 5 4 5 4 3 4 4 5 4 4 3 3 4	0.3016	0.47
6	1,1,5,7,7,5	5 2 3 3 6 1 4 6 4 4 1 1 1 1 4 1 6 3 4 6 1 6 6 3 3 6	0.3016	0.50

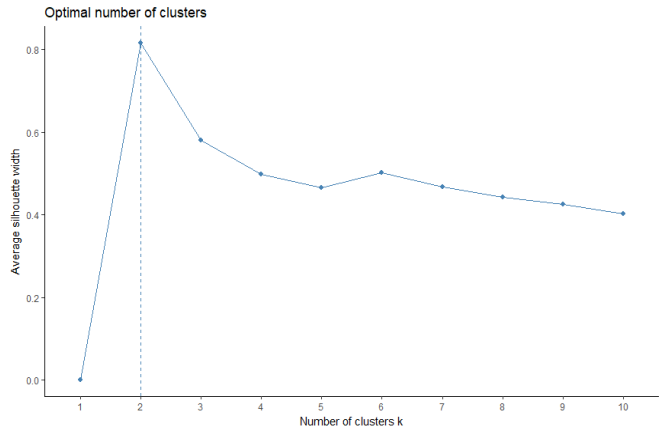

Figure 4.4 Silhouette plot for k- medoids algorithm

Based on the clustering methods, two clusters are identified, and this cluster consists of the following regions, as shown in Table 4.3.

TABLE 4.3 CLUSTERING ANALYTICAL RESULT OF DENGUE PATIENTS' STRUCTURES

Category	RDHS Areas
1	Colombo and Gampaha
2	Kalutara, Kandy, Matale, Nuweraliya, Galle, Hambantota, Matara, Jaffna, Kilinochchi, Mannar, Vavuniya, Mulativu, Batticaloa, Ampara, Trincomalee, Kurunegala, Puttalam, Anurathapura, Polonnaruwa, Badulla, Moneragala, Ratnapura, Kegalle and Kalmunai.

Table 4.3 shows the category of clusters and their objects (RDHS area). This clearly shows that Colombo and Gampaha districts are under the same group. The Western province of Sri Lanka is the most affected by dengue with cases being reported the regions of Colombo and Gampaha.

V CONCLUSIONS

Clustering is an important field in data mining techniques, and currently many researchers use cluster techniques to determine the optimal number of clusters. The aim of this paper is to identify the number of clusters in in patients with dengue fever, using partition clustering algorithm. The results of the k-means algorithm show that the number of cluster of dengue fever data is two. However, the *k*-means algorithm is not appropriate due to outlier issues. Thus, the *k*-medoids algorithm is used, which also shows the number of cluster is two. The Colombo and Gampaha regions are more affected by dengue. Therefore, since it is necessary to draw practical conclusions at the basic level of certain interventions and awareness-raising. In addition, we hope that this study will be an important platform for highlighting and discussing potential problems that determine important factors of dengue fever in certain areas.

REFERENCES

[1] Aggarwal Charu, C. and Reddy Chandan, K. (2014). Data Clustering: Algorithms and Applications. Chapman & Hall CRC Press, 88-107.

[2] Beckett, C., Kosasih, H., Faisal, I., & Tan, R. (2005). Early detection of dengue infections using cluster sampling around index cases, 72(6), 777–782.

[3] Dolnicar, S. (2003). Using cluster analysis for market segmentation - typical misconceptions, established methodological weaknesses and some recommendations for improvement. Australasian Journal of Market Research, 11(2), 5–12.

[4] Gan, G., Ma, C., Wu, J. (2007). Data Clustering: Theory, Algorithms, and Applications. Philadelphia: ASA-SIAM.

[5] Han, J. and Kamber, M. (2006). Data Mining: Concepts and Techniques. Elsevier/Morgan Kaufmann, second edition.

[6] Hartigan, J. A. and M. A. Wong. 1979. A K-means clustering algorithm. Applied Stat. 28 (1), 100–108.

[7] Hsu, J. C., Hsieh, C. L., & Lu, C. Y. (2017). Trend and geographic analysis of the prevalence of dengue in Taiwan, 2010–2015. International Journal of Infectious Diseases, 54, 43–49.

[8] Kanth, T. V. R., Balaram, V. V. S. S. S., & Rajasekhar, N. (2014). Analysis of Indian Weather Data Sets Using Data Mining Techniques. Fourth International Conference on Advances in Computing and Information Technology, 8994.

[9] Kaufman, L. & Rousseeuw, P.J., (1987). Clustering by means of medoids. In: Dodge,Y. ed., Statistical Data Analysis,

based on the L1-norm, New York: Elsevier/North-Holland, 405-416.
[10] Milligan, G. W., & Cooper, M. C. (1985). An examination procedure for determining the number of clusters in a data set. Psychometrika, 50(2), 159–179.
[11] Rodrigues, H. S., Monteiro, M. T. T., Torres, D. F. M., & Zinober, A. (2012). Dengue disease, basic reproduction number and control. International Journal of Computer Mathematics, 89(3), 334–346. [12] Visalakshi, N. K., & Suguna, J. (2009). K-means clustering using Max-min distance measure. Annual Meeting of the North American Fuzzy Information Processing Society (NAFIPS), 1–6.
[13] https://www.iamat.org/risks/dengue
[14] http://www.who.int/countries/lka/en/