Medoid-based shadow value validation and visualization

Weksi Budiaji

Abstract

One type of clustering validation results is an internal criteria validation where a silhouette index is the well-known measure and plot. Another index is a centroid-based shadow value (CSV) which depicts an opposite measure and plot to the silhouette index. A medoid-based shadow value (MSV) index is proposed in which it behaves similarly to the silhouette index and produces a network graph similar to a neighborhood graph resulted from the CSV. The network graph of the MSV has a multiplicative parameter (c) to adjust its edges visibility and is more a appropriate plot than a neighborhood graph of CSV when the data set is non-numerical.

Keywords: Cluster validation, Cluster visualization, Internal criteria, Medoid, Shadow value

1 Introduction

Cluster analysis is an unsupervised method to group objects such that homogenous objects are clustered in the same group. As an unsupervised method in which pre-defined class memberships are absent, the partitioning result from a cluster analysis has to be validated. There are three types of validation for the partitioning results; they are relative, external, and internal criteria (Webb and Copsey 2011). They differ with respect to the compactness assumption and information provided in the data.

Among the three criteria, the relative criteria do not require the compactness assumption. It based on a re-sampling scheme via either cross validation or bootstrap methods. The latter is a sampling with replacement strategy to assess the stability of clusters (Jain and Moreau 1987) and select the appropriate number of clusters (Fang 2012). The stability is then visualized in a heatmap figure (Monti et al. 2003) where a block diagonal figure depicts the most stable cluster result.

The external criteria are commonly applied in a benchmarking process with either known classes or the "gold standard" algorithm (Handl, Knowles, and Kell 2005). When a new clustering algorithm is developed, the routine process to evaluate this algorithm is by applying it in a supervised environment. Then, an evaluation measure compares this new algorithm to the existing/ gold algorithms. Two examples of external criteria applied to compare a new algorithm are the clustering accuracy rate (Ji et al. 2013) and the cluster purity (Handl, Knowles, and Kell 2005; Wu et al. 2008; Waiyama and Kangkachit 2018). Arbelaitz et al. (2013), moreover, has addressed many external criteria; e.g. Rand (Rand 1971), and adjusted Rand (Hubert and Arabie 1985).

The internal criteria, on the other hand, are applied when a real data set, which is lacking true classes, is analyzed by means of cluster analysis. Charrad et al. (2014) has cited as many as 19 internal validation indices; e.g. silhouette (Rousseeuw 1987), and gap statistic (Tibshirani, Walther, and Hastie 2001). One of the popular indices is silhouette because it offers a visualization of each cluster (Leisch 2008). It non-linearly combines the compactness and separation assumptions (Brock et al. 2008). A similar approach to silhouette has been developed namely a shadow value (Leisch 2006; Leisch 2010), in which the values are calculated based on the first and second closest centroids and can also be plotted as silhouette value.

Although the silhouette and shadow value can be plotted, they depict different figures in the same case. Well-separated clusters, for instance, are indicated by high values of silhouette (Rousseeuw 1987), while they have small indices in the case of shadow values (Leisch 2010). Because the silhouette is medoid-based, which is suitable for any type of data, and the shadow value is centroid-based, the latter gains advantage when the data consist of numerical variables such that a 2-dimensional representation of the neighborhood graph of cluster results can be produced. For any type of data, however, a neighborhood graph is not visible.

In this paper, we propose a new formula to imitate the silhouette and centroid-based shadow value characters. It is a medoid-based shadow value. The figures produced mimic both the silhouette and centroid-based shadow values where the plot is similar to the silhouette and the 2-dimensional graph is a neighborhood graph alike.

2 The Proposed Method

2.1 Shadow Value for Medoid-based Clustering

The silhouette value can be calculated via

$$si(x) = \frac{b_x - a_x}{max(a_x, b_x)},$$

where a_x and b_x are the average distance of object x to all objects within cluster and to all objects within the nearest cluster, respectively (Rousseeuw 1987). The value then has a minimum -1 and maximum 1 where the best separated clusters have a value equal to 1. The shadow value, on the other hand, is attained by

$$sh(x) = \frac{2d(x, c(x))}{d(x, c(x)) + d(x, c'(x))},$$

where $d(x, c(x))$ is the distance between object x to the first closest centroid and $d(x, c'(x))$ is the distance between object x to the second closest centroid (Leisch 2010). The poorly

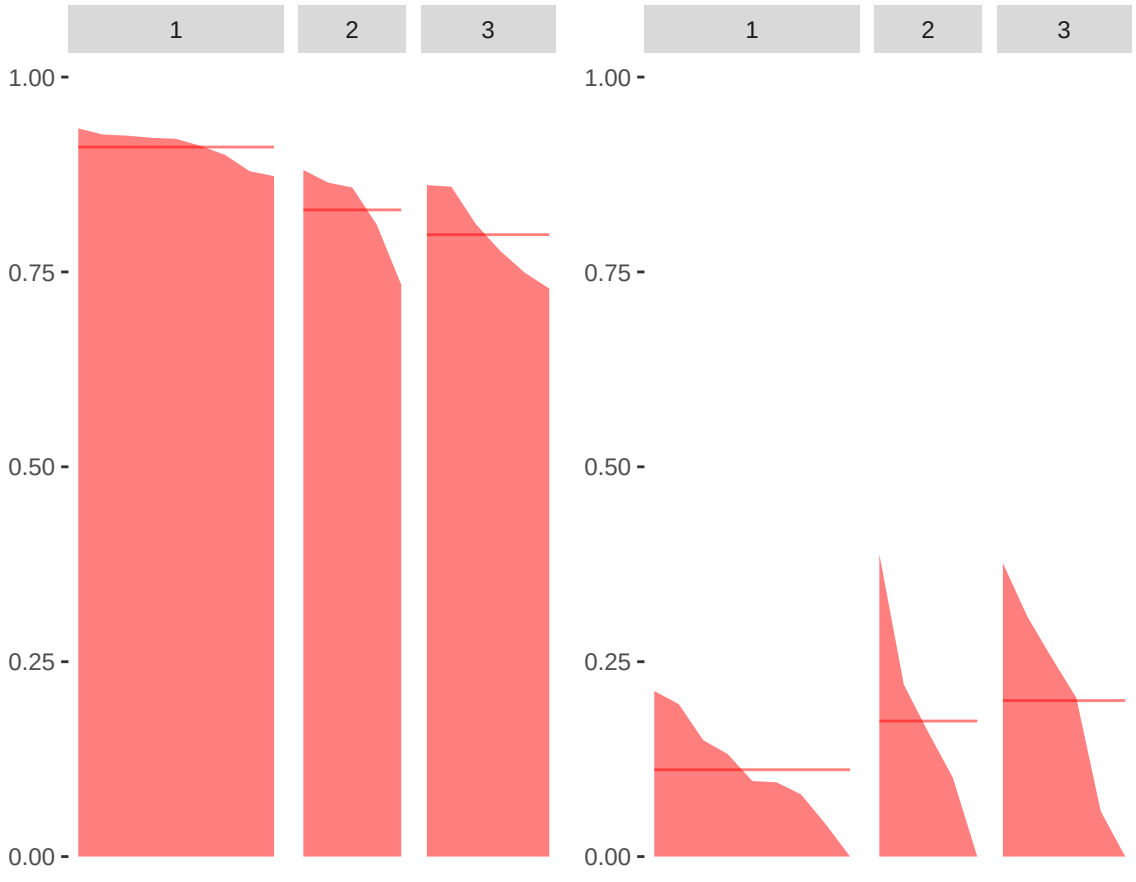

Figure 1: Silhouette (left) and shadow (right) values of well-separated clusters.

separated clusters are indicated by a shadow value of 1 meaning that the first and second closest centroids are equidistant from x.

Although the shadow value has 0 as a minimum value, which can be achieved when the object is very close to the centroid, an object that has twice the distance to the second closest centroid compared to the first second centroid achieves 0.67 as a shadow value, which is considered as a high shadow value. Figure 1 moreover, illustrates well-separated clusters via silhouette and shadow values where high peaks occur in silhouette and low peaks appear in the shadow plot. Due to the contradictory image between silhouette and shadow plots, when interpreting the plot, it requires careful consideration.

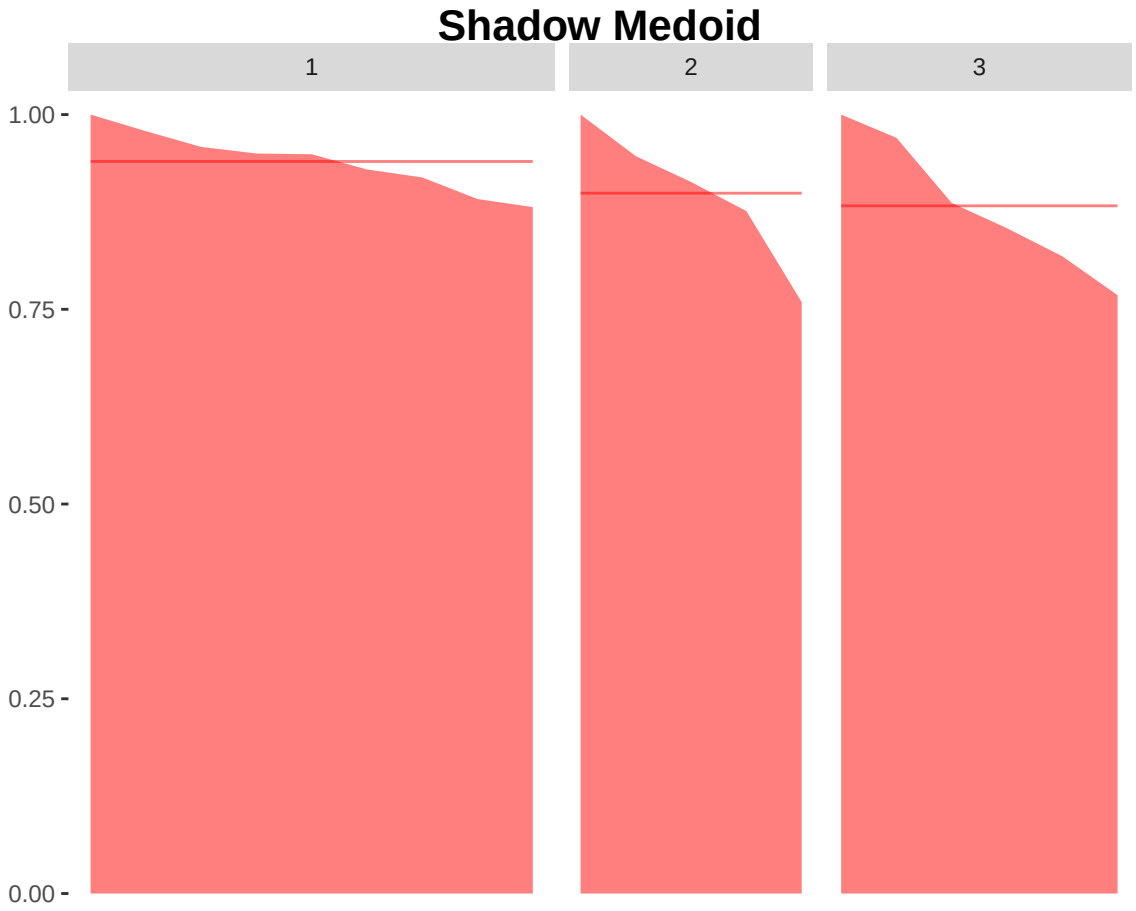

Figure 2: Medoid-based shadow value (MSV) of well-separated clusters

Table 1: Centroid-based shadow (CSV) and medoid-based shadow (MSV) values comparison

	1x	2x	3x	4x	5x	6x	7x	8x	9x	10x
CSV	1	0.67	0.50	0.40	0.33	0.29	0.25	0.22	0.20	0.18
MSV	0	0.50	0.67	0.75	0.80	0.83	0.86	0.88	0.89	0.90

A new formula is developed to calculate a new shadow value in a medoid-based clustering. To adapt the silhouette and centroid-based shadow value (CSV) characters, these following constraints are applied:

1. The lower and upper bounds of the value are 0 and 1.
2. The worst separated cluster is 0, while the best is 1.
3. The value of 0 is valid for an equidistant between the first and second closest medoids.
4. The value of 1 is achieved when the object is the medoid object.

With these constraints, the new shadow values in a medoid-based clustering are then simplified into

$$msv(x) = \frac{d(x, c'(x)) - d(x, c(x))}{d(x, c'(x))}.$$

Figure 2 illustrates the medoid-based shadow value (MSV) of well-separated clusters where it depicts a similar figure to the silhouette plot (Figure 1 left). Table 1, in addition, compares the index of CSV vs MSV in a specified distance of the second closest centroid. An object that has an equidistant between the first and second closest centroid, has CSV equal to 1 compared to 0 in the MSV.

2.2 Visualization

The CSV can be plotted in a neighborhood graph (network graph) as well. The graph has k nodes, where k is the number of clusters, and is an undirected graph with an average shadow values of the closest clusters as its edges (Leisch 2010). The cluster similarity is measured by the average shadow value within a cluster and the closest cluster. Figure 3 illustrates a neighborhood graph of well-separated clusters where all clusters have thin lines. A thick line in a neighborhood graph, on the other hand, denotes a high shadow value indicating poor-separated clusters.

The representation of either thin or thick lines in a neighborhood graph is naturally attractive where a thick line implies poorly separated clusters (close to each other). This characteristic is retained to develop a new technique of visualization. Because the neighborhood graph is a centroid-based plot, a network graph of medoid-based clustering is developed such that it fits for any type of data, i.e. numerical, binary, categorical, and mixed variables. There are two type of visualizations; they are medoids and all-object visualization.

To create a medoid-based visualization, an element of $\mathbf{M}$ ($k\ x\ k$ matrix) is calculated by

$$a_{ij} = \frac{dm_{ij} - max(\bar{d}_i, \bar{d}_j)}{dm_{ij}},$$

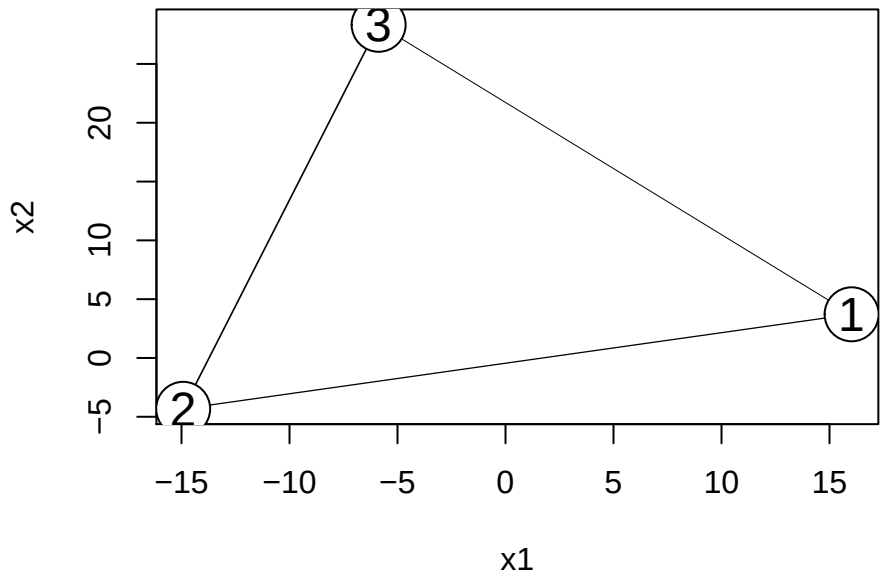

Figure 3: Neighborhood graph of well-separated clusters

where dm_{ij} is the distance between the medoid in the cluster i and j (the between cluster distance), $\bar{d}_i$ is the average distance among objects in the cluster i (the average of within cluster distance). Then, the diagonal elements of $\mathbf{M}$ are restricted to be 0. The matrix $\mathbf{M}$ is a squared matrix that is equivalent to the medoid-based shadow values among clusters.

To plot marix $\mathbf{M}$ directly in a network graph, the off-diagonal values of $\mathbf{M}$ represented as edges are converted into $1 - a_{ij}$ such that thin lines depict well-separated clusters like in a neighborhood graph. Then, the nodes and edges are laid in a 2-dimensional space via a graph layout algorithm. Battista et al. (1994) has surveyed many graph layouts, e.g. Kamada and Kawai (1989) and Fruchterman and Reingold (1991). The x and y axes are then meaningless whereas it is more relevant when the data have non-numerical variables than a neighborhood graph. Figure 4 shows a network graph of medoid-based shadow value by plotting directly matrix $\mathbf{M}$. The graph is similar to the neighborhood graph, which shows well-separated clusters

In the all-objects visualization, the information of the medoids network graph is added by all objects information. A particular medoid (node) has a/ some other nodes connected via an edge based on its cluster membership. A medoid-based shadow value matrix of all objects $\mathbf{O}$ ($n\ x\ n$ matrix) is created where all elements are "NA" except the o_{xy} elements where x is an object and y is the closest medoid. The diagonal elements of $\mathbf{O}$ are restricted to be "NA" and the o_{xy} element has a medoid-based shadow value of object x. The matrix $\mathbf{M}$ and $\mathbf{O}$ information are combined in order to create an all-object network graph. In summary, the matrix $\mathbf{M}$ indicates the separation among clusters, while the matrix $\mathbf{O}$ represents the within

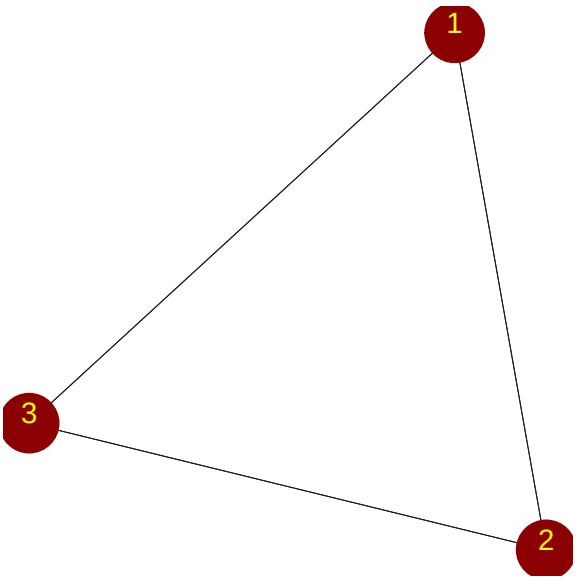

Figure 4: Medoids network graph of well-separated clusters (kamada-kawai layout)

cluster compactness.

Transforming the information of matrix **M** and **O** directly into a network topology, a medoid-based shadow value graph of all objects is produced. Figure 5 illustrates well-separated clusters in a network graph with all objects where it has a high separation and compactness. A high separation between clusters is indicated by a thin line, while a high compactness within a cluster is addressed by a thick line among objects to their cluster medoids. A constant (c) is also introduced in order to multiply the medoid-based shadow value such that the separation and compactness are more visible. The constant c is also applicable to the aforementioned medoid visualization.

3 Method

To apply the proposed MSV index and visualization, some simulated data sets are generated. Qiu and Joe (2006a) has developed an algorithm to generate numerical data set for clustering algorithm benchmarking with a pre-specified degree of separation (Qiu and Joe 2006b). The simulated data sets in this study vary in the separation degree only (well, middle, and poorly separated). The results of these three different separated clusters are compared among the existing indices, i.e. silhouette and shadow value (CSV) in the medoid setting, with the developed MSV index.

Table 2: The settings of the simulated data sets

Separation	n number of objects	p number of variables	k number of clusters
0.5	1000	2	5
0.0	1000	2	5

Separation	n number of objects	p number of variables	k number of clusters
-0.5	1000	2	5

Because this paper focusses on the different setting of degree separation, the variables of n (the number of objects), p (the number of variables), and k (the number of clusters) in the simulated data are fixed that are set as 1000, 2, and 5, respectively (Table 2). The algorithm to group the data is also fixed via partitioning around medoid (PAM) (Kaufman and Rousseeuw 1990) as a popular medoid-based algorithm. Then, each simulated data set is replicated. Although 50 replications for each simulated data set are fairly precise (Hennig 2007), the strategy to replicate the simulated data in this paper is via subsetting by choosing the number of the subset sample $m = n/2$, i.e. $1000/2 = 500$ replicates.

For real data sets, the data sets from the UCI repository (Lichman 2013), which represent well and poorly separated clusters, are also analyzed. The analyses produced in this article, moreover, are run in an Intel i3 4GB RAM using R software environment (R Core Team 2015) using the *clusterGeneration*, *cluster*, *kmed*, *ggplot2*, *geomnet*, and *flexclust* packages.

4 Results and Discussion

In this section, the MSV proposed index is applied in simulated data sets and real data sets. The simulated data sets are generated via the *clusterGeneration* package (Qiu and Joe 2015). Meanwhile, the real data sets are two data sets of UCI repository data sets (Lichman 2013) namely the well-known iris data set, and lenses data sets to represents numerical and categorical data sets partitioned by PAM via the *cluster* package (Maechler et al. 2017). The silhouette and shadow values, moreover, are obtained by the *kmed* package (Budiaji 2019). The network graph, in addition, is plotted by the *ggplot2* (Wickham 2016), *geomnet* (Tyner and H. Hofmann 2016) and *flexclust* (Leisch 2006; Leisch 2010) packages.

4.1 Simulated data

The first simulated data set (well-separated clusters) has high values in both the silhouette and MSV indices, yet it has low values in the CSV (Figure 6). The contradictory results of the CSV to the silhouette and MSV, moreover, occur in all types of simulated data set except in the middle-separated clusters where all indices produce comparable results between 0.4 and 0.6. Figure 6 also shows that the MSV has always had a higher index compared to the silhouette value. It can be explained that the span value of the MSV is shorter than the silhouette value, i.e. [0,1] compared to [-1, 1] (Rousseeuw 1987).

In summary, the proposed MSV index adapts the silhouette values well. Although it has reverse values to the CSV, it is analytically comparable to the CSV (Table 1). Thus, the MSV index is a promising index for the internal criteria evaluation of the cluster results.

For the network visualization of the simulated data, which are partitioned into 5 clusters, all objects are plotted by comparing the well, middle, and poorly separated cluster data sets.

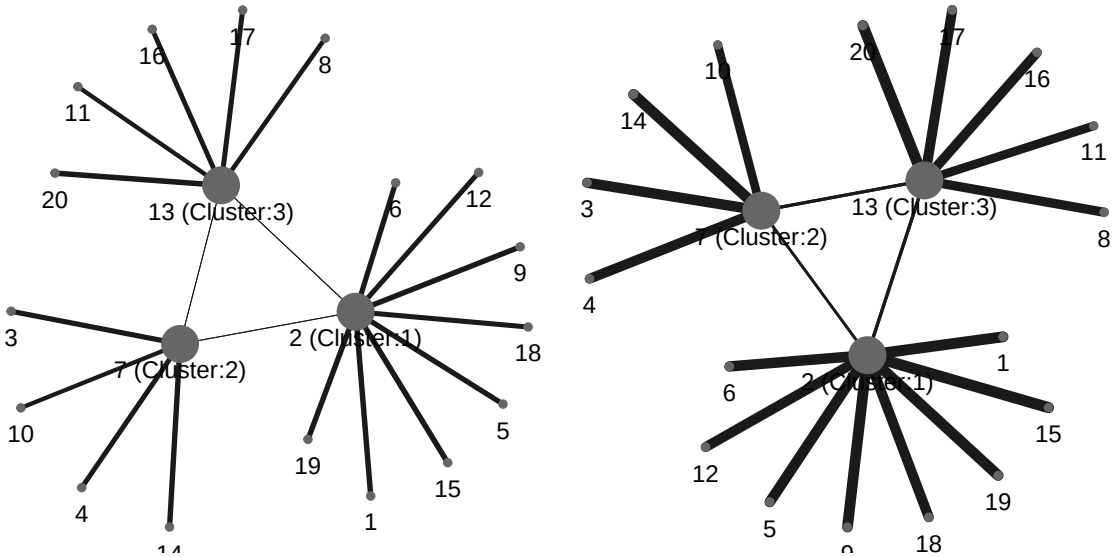

Figure 5: All-object network graph of well-separated clusters with c = 1 (left) and c = 2 (right)

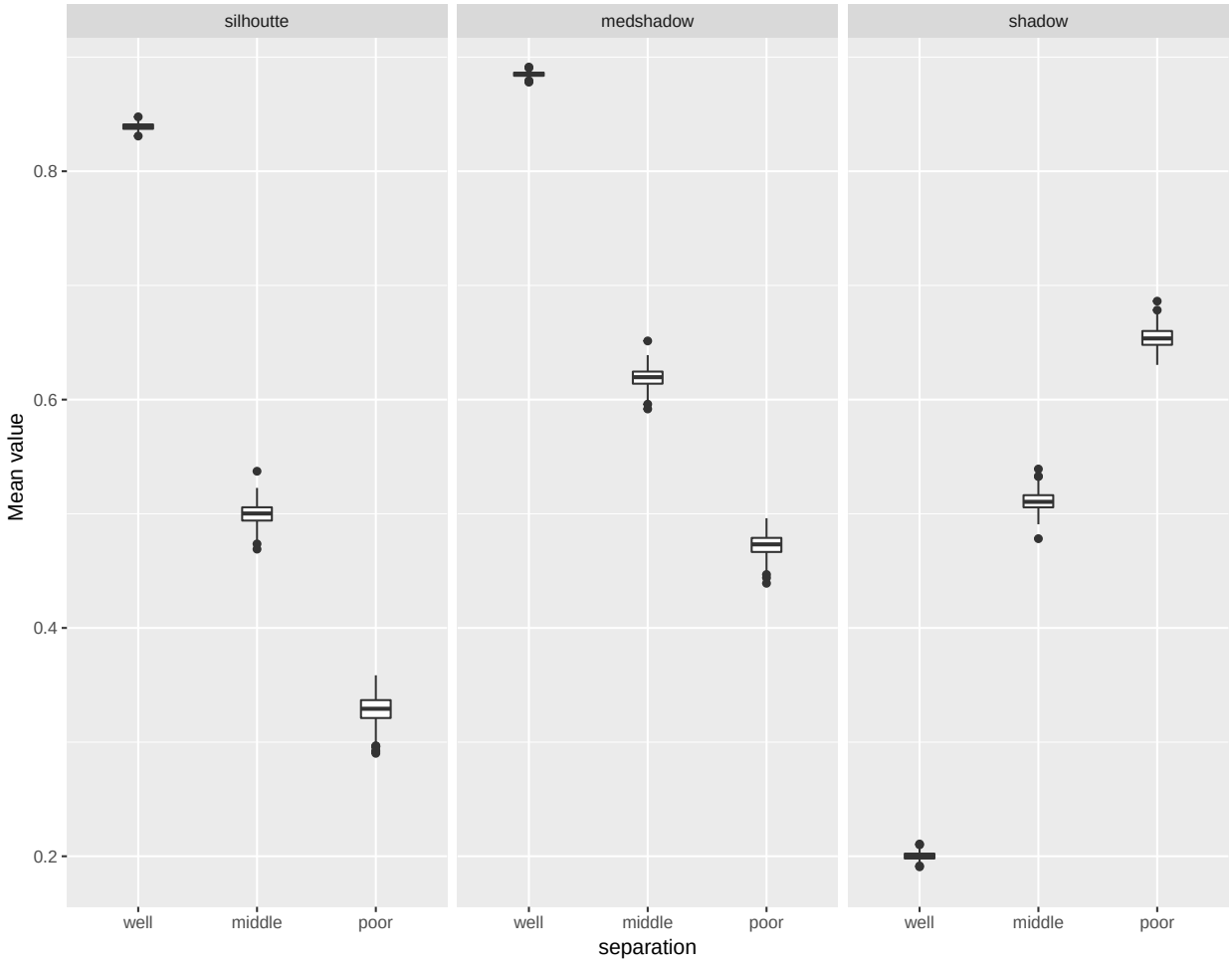

Figure 6: Boxplot of the mean value indices of the silhoutte (left), MSV (middle), and shadow (right) values

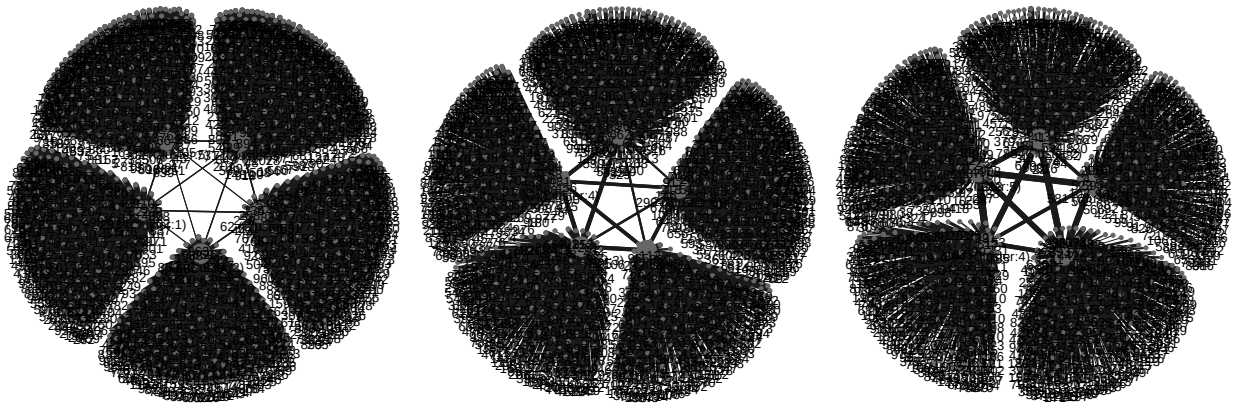

Figure 7: All-object visualizations of well (left), middle (middle), and poorly (right) separated clusters

Figure 7 shows the differences among them. The well-separated clusters (left) have thin lines among medoids and thicker lines among objects within a cluster indicating that they have high values of both separation and compactness. Meanwhile, the poorly-separated clusters (right) have opposite image where the lines among medoids are thicker than the lines among objects, which represent a low value of separation among medoids.

4.2 Real data set

4.2.1 Iris data set

Table 3: The misclassification table of the PAM algorithm in the iris data set

setosa	versicolor	virginica
50	0	0
0	48	14
0	2	36

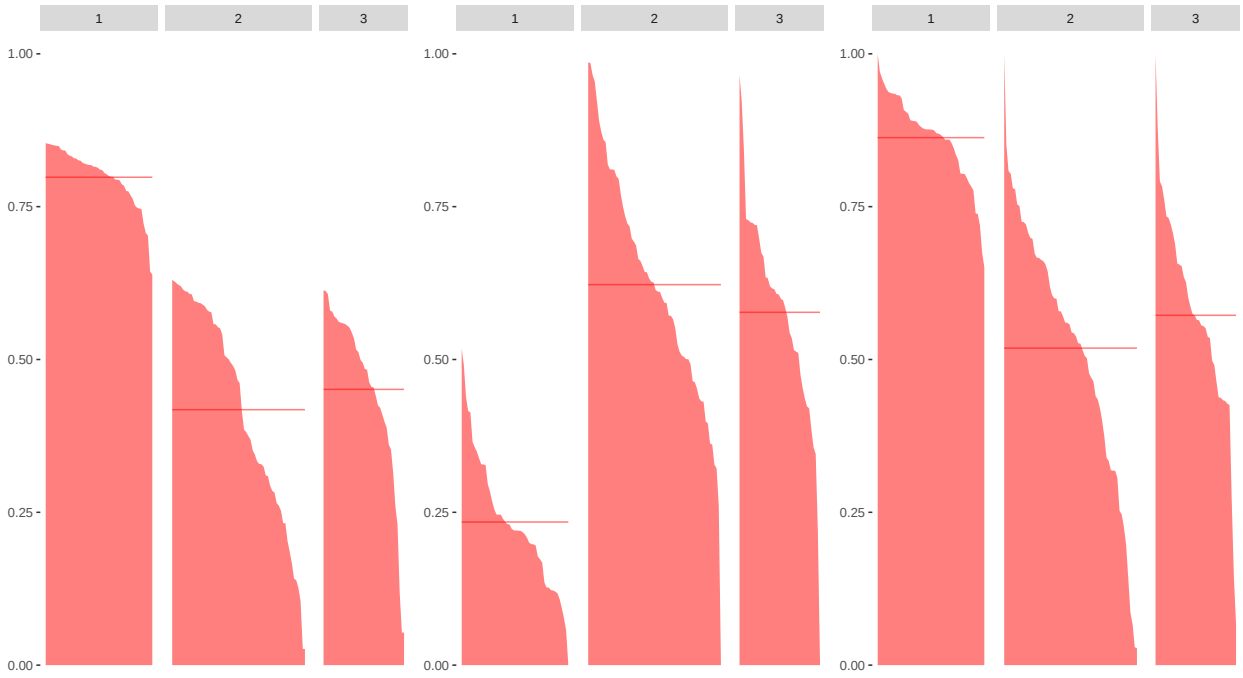

Figure 8: Silhouette (left), CSV (middle), and MSV (right) plots of the iris data set

The iris data set is a well-known data set with four numerical variables that consists of 150 objects divided into three species of iris (setosa, versicolor, and virginica). To compare the silhouette, CSV, and MSV indices, the PAM algorithm in the Euclidean distance matrix

of this data is applied with the number of clusters k equal to 3. The accuracy rate of the PAM algorithm is 86.67% (Table 3), which is 100% correct and achieved in cluster 1 (setosa class). If internal validation with silhouette, CSV, and MSV indices is plotted (Figure 8), they produce similar results and cluster 1 has the best result, i.e. high value of silhouette and MSV, and small value of CSV.

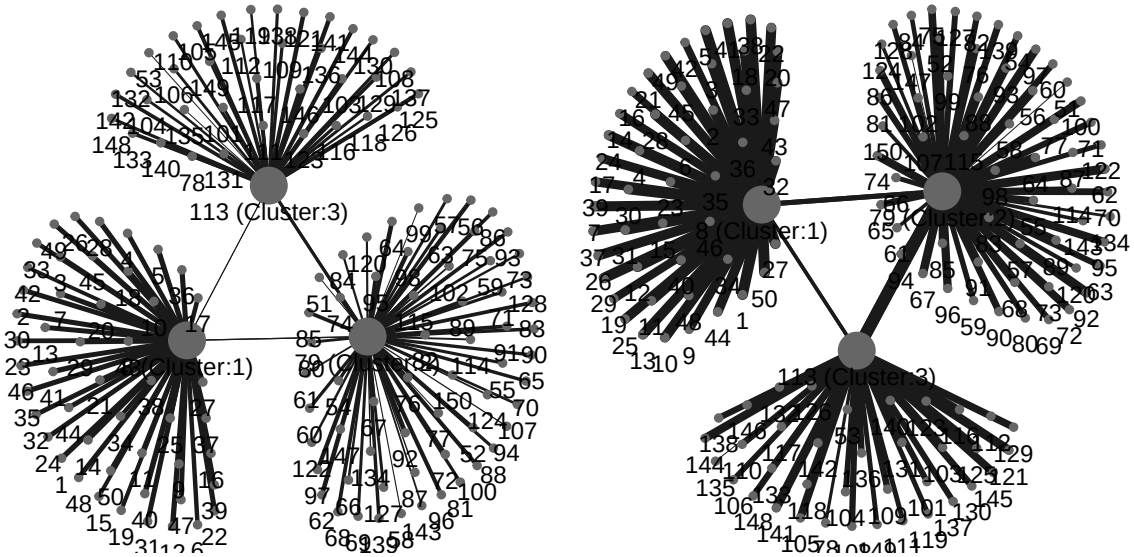

Figure 9: All-objects visualization of the iris data set with c = 1 (left) and c = 2 (right)

Figure 12 illustrates network graphs of iris data sets based on the MSV of all objects. By adjusting the value of c (multiplicative constant for MSV) into 2 (Figure 12 right), clusters 2 and 3 are discernable that they have low separation. It also shows that cluster 1 has the highest compactness among the three clusters portrayed by the thickest line within cluster 1.

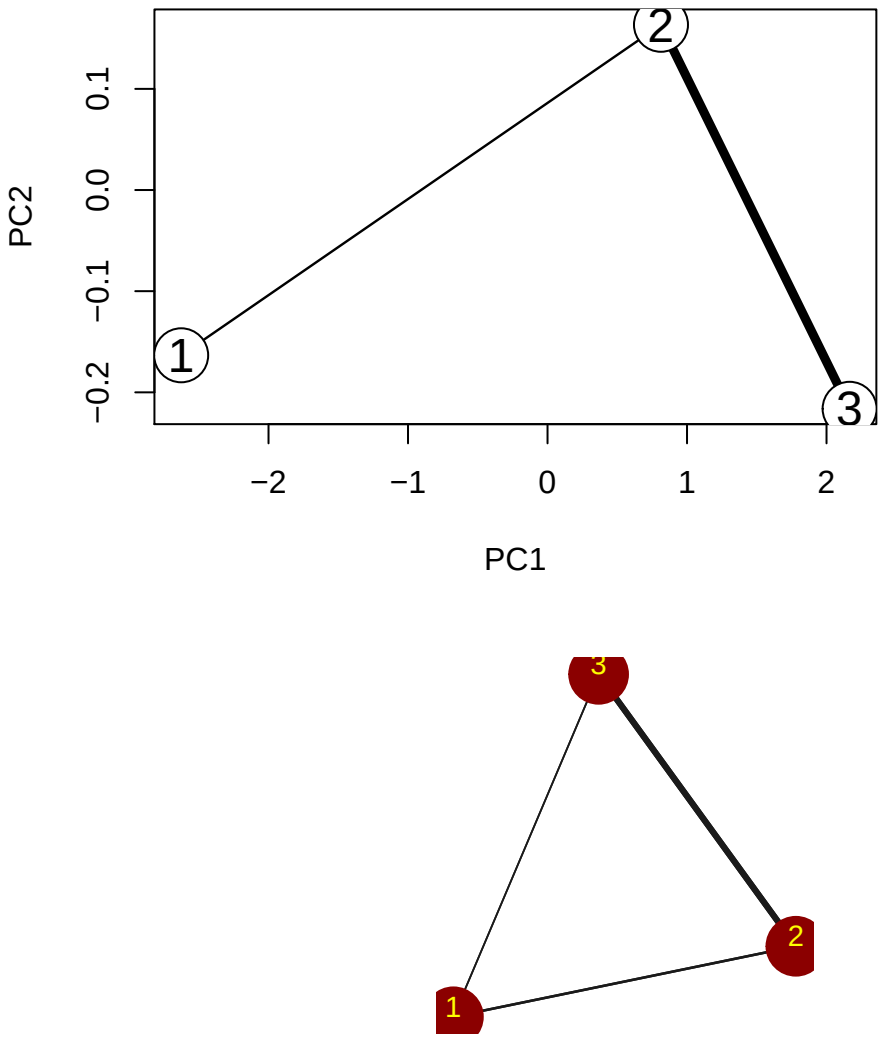

Figure 10: The neighborhood graph (top) and medoids visualization c = 2 (bottom) of the iris data set

The value of $c = 2$ is then adopted in the medoid visualization plot (Figure 10 bottom) such that it shows that cluster 1 is separable to cluster 2 and 3. If it is compared to the neighborhood graph (Figure 10 top), it depicts a similar image where clusters 2 and 3 have low separation. However, the axes in the medoid visualization are meaningless, while they can be the first and second principle components in the neighborhood graph. It also has more edges/ lines than the neighborhood graph, which draws an edge between two nodes if only at least one object has the closest and second closest to those nodes (Leisch 2006), because it is based on the squared matrix M.

4.2.2 Lenses data set

Table 4: The misclassification table of the PAM algorithm in the lenses data set

	Hard	Soft	None
Cluster 1	3	1	5
Cluster 2	1	3	4
Cluster 3	0	1	6

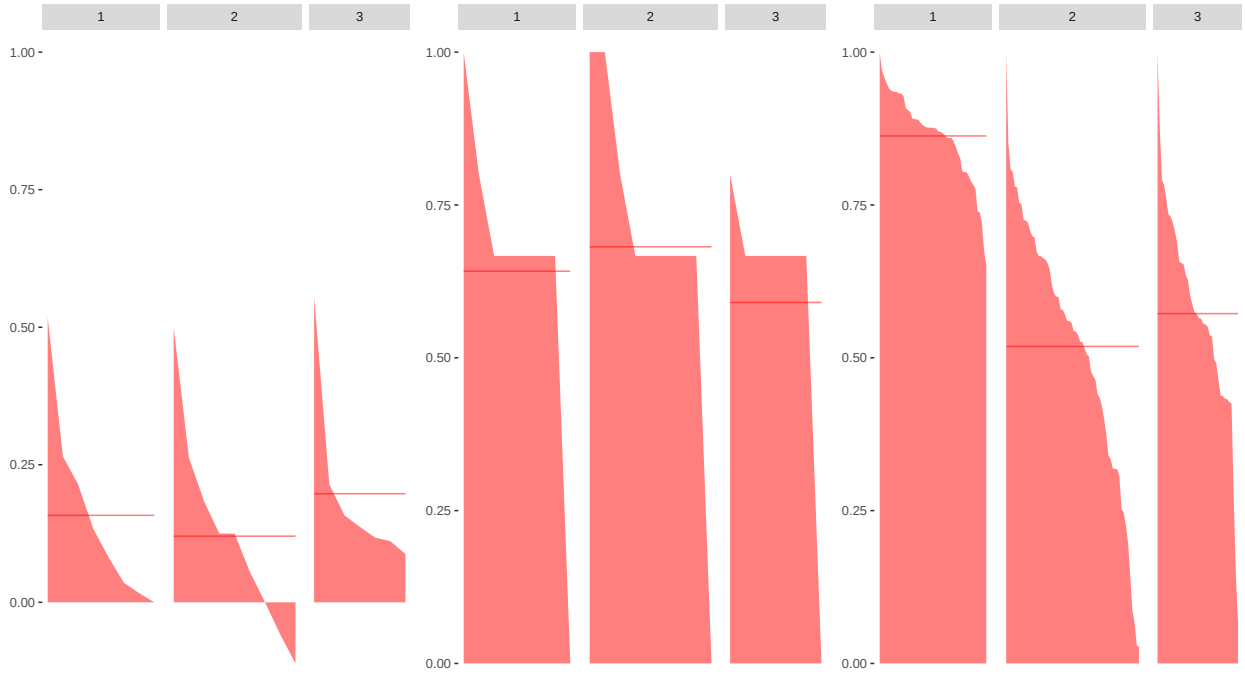

Figure 11: Silhouette (left), CSV (middle), and MSV (right) plots of the lenses data set

The lenses data set consists of 24 patients with four categorical variables. The patients are classified into three groups: hard contact lenses, soft contact lenses, and none of those two types of lenses. The PAM algorithm in the simple matching distance matrix of this data is applied with k equal to 3. The accuracy rate is low, i.e. 50% (Table 4), which indicates poor separated clusters. Figure 11 shows the three internal criteria of the clustering results indicating poorly separated clusters as well.

When the all objects are visualized in a network graph with $c = 2$, all clusters have low separation indicated by thick lines (Figure 12). The compactness within a cluster is also low representing by thin lines. Figure 13, moreover, illustrates the medoid plot with $c = 2$, adapted from the all objects network graph, in which all medoids are close to each other, i.e. poor separated. On the other hand, the neighborhood graph version of this plot is absent

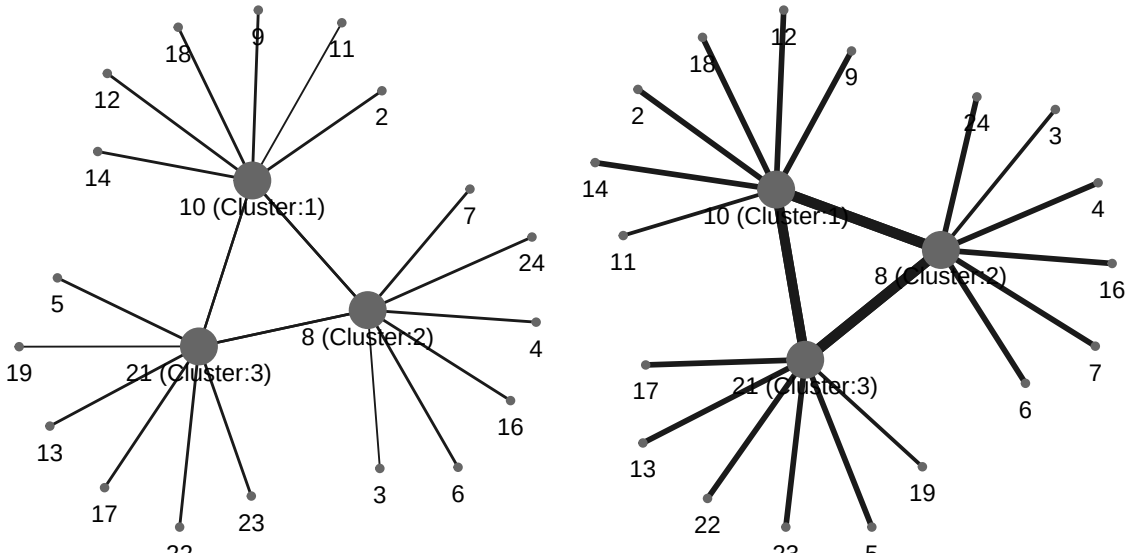

Figure 12: All-objects visualization of the lenses data set with c = 1 (left) and c = 2 (right)

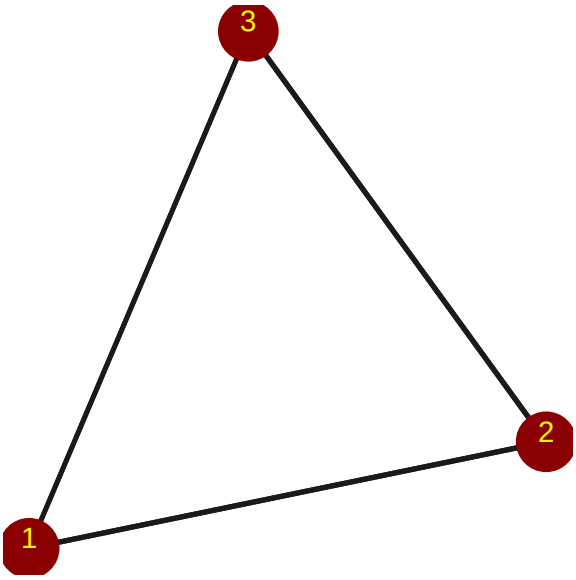

Figure 13: The medoid visualization of the lenses data set with c = 2

due to non-numerical variable data set. With this type of data set, the centroid calculation is unfeasible. Thus, a conversion of medoid-based into centroid-based, which is required to produce the neighborhood graph (Leisch 2006), is also unachievable such that a medoids visualization gains an advantage compared to a neighborhood graph.

5 Conclusion

In this paper, we proposed an internal criteria validation for cluster results, namely the medoid-based shadow value (MSV). The MSV index imitated the silhouette index behavior where the higher value of the index, the better the cluster result. While a centroid-based shadow value (CSV) could produce a neighborhood graph, the MSV was able to be visualized in all objects and medoid network graphs where the latter is a neighborhood graph alike. Both the all objects and medoid network visualizations had a parameter c to regulate the visibility of the edges. It was suggested to first apply the all-objects network graph with multiple c. Then, the c in the medoid network graph adapted the suitable c obtained from the c of all objects graph. The MSV visualization axes, in addition, were meaningless such that in non-numerical type of data set, it was preferred and more suitable than a neighborhood graph.

Acknowledgment

This research was supported by the Ministry of Research, Technology, and Higher Education of Indonesia, Kemenristekdikti, and Osterreichischer Austauschdienst, OeAD, under the Indonesia Austria Scholarship Program.

References

Arbelaitz, O., Gurrutxaga I., Muguerza J., Perez J.M., and I. Perona. 2013. "An Extensive Comparative Study of Cluster Validity Indices." *Pattern Recognition* 46: 243–56.

Battista, G. D., Eades P., Tamassia R., and I. G Tollis. 1994. "Algorithm for Drawing Graphs: An Annotated Bibliography." *Computational Geometry* 4: 235–82.

Brock, G., Pihur V., Datta Susmita, and Somnath Datta. 2008. "clValid: An R Package for Cluster Validation." *Journal of Statistical Software* 25 (4).

Budiaji, Weksi. 2019. *Kmed: Distance-Based K-Medoids*. http://CRAN.R-project.org/package=kmed.

Charrad, M., Ghazzali N., Boiteau V., and A. Niknafs. 2014. "NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set." *Journal of Statistical Software* 61 (6): 1–36.

Fang, J., Y.and Wang. 2012. "Selection of the Number of Clusters via the Bootstrap Method." *Computational Statistics and Data Analysis* 56: 468–77.

Fruchterman, T.M., and E.M. Reingold. 1991. "Graph Drawing by Force-Directed Placement." *Software-Practice and Experience* 21 (11): 1129–64.

Handl, J., Knowles J., and D.B. Kell. 2005. "Computational Cluster Validation in Post-Genomic Data Analysis." *Bioinformatics* 21 (15): 3201–12.

Hennig, C. 2007. "Cluster-Wise Assessment of Cluster Stability." *Computational Statistics and Data Analysis* 52: 258–71.

Hubert, L., and P. Arabie. 1985. "Comparing Partitions." *Journal of Classification* 2: 193–218.

Jain, A.K., and J.V. Moreau. 1987. "Bootstrap Technique in Cluster Analysis." *Pattern Recognition* 20: 547–68.

Ji, J., Bai T., Zhou C., Ma C., and Wang Z. 2013. "An Improved K-Prototypes Clustering Algorithm for Mixed Numeric and Categorical Data." *Neurocomputing* 120: 590–96.

Kamada, T., and S. Kawai. 1989. "An Algorithm for Drawing General Undirected Graphs." *Information Processing Letters* 31: 7–15.

Kaufman, L., and P.J. Rousseeuw. 1990. *Finding Groups in Data*. New York, USA: John Wiley; Sons.

Leisch, F. 2006. "A Toolbox for K-Centroids Cluster Analysis." *Computational Statistics and Data Analysis* 51: 526–44.

———. 2008. "Handbook of Data Visualization." In, edited by Chen C., Hardle W., and A. Unwin, 561–87. Springer Handbooks of Computational Statistics. Springer Verlag.

———. 2010. "Neighborhood Graphs, Stripes and Shadow Plots for Cluster Visualization." *Statistics and Computing* 20: 457–69.

Lichman, M. 2013. "UCI Machine Learning Repository."

Maechler, M., Rousseeuw P., Struyf A., Hubert M., and K. Hornik. 2017. *Cluster: Cluster Analysis Basics and Extensions*.

Monti, S., Tamayo P., Mesirov J., and T. Golub. 2003. "Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data." *Machine Learning* 52: 91–118.

Qiu, W., and H. Joe. 2006a. "Generation of Random Clusters with Specified Degree of Separation." *Journal of Classification* 23: 315–34.

———. 2006b. "Separation Index and Partial Membership for Clustering." *Computational Statistics and Data Analysis* 50: 585–603.

———. 2015. *ClusterGeneration: Random Cluster Generation (with Specified Degree of Separation)*.

R Core Team. 2015. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Rand, W. M. 1971. "Objective Criteria for the Evaluation of Clustering Methods." *Journal of the American Statistical Association* 66 (336): 846–50.

Rousseeuw, P.J. 1987. "Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis." *Journal of Computational and Applied Mathematics* 20: 53–65.

Tibshirani, R., Walther G., and T. Hastie. 2001. "Estimating the Number of Clusters in a Data Set via the Gap Statistic." *Journal of the Royal Statistical Society B* 63 (2): 411–23.

Tyner, S., and H. H. Hofmann. 2016. *Geomnet: Network Visualization in the "Ggplot2" Framework*.

Waiyama, K., and T. Kangkachit. 2018. "Constraint-Based Discriminative Dimension Selection for High-Dimensional Stream Clustering." *International Journal of Advances in Intelligent Informatics* 4 (3): 167–79.

Webb, A. R., and K. Copsey. 2011. *Statistical Pattern Recognition*. 3rd ed. West Sussex, UK: John Wiley; Sons.

Wickham, H. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. New York, US: Springer-Verlag.

Wu, X., Kumar V., Quinlan J.R., Ghosh J., Yang Q., Motoda H., McLachlan G.J., et al. 2008. "Top 10 Algorithms in Data Mining." *Knowledge and Information Systems* 14: 1–37.