The Clustering Validity with Silhouette and Sum of Squared Errors

Tippaya Thinsungnoen[a*], Nuntawut Kaoungku[b], Pongsakorn Durongdumronchai[b],
Kittisak Kerdprasop[b], Nittaya Kerdprasop[b]
[a]Informatics Program, Faculty of Science and Technology, Nakhon Ratchasima Rajabhat University, Thailand
[b]Data Engineering Research Unit, School of Computer Engineering, Institute of Engineering,
Suranaree University of Technology, Thailand
*Corresponding Author: tippayasot@hotmail.com

Abstract

The data clustering with automatic program such as k-means has been a popular technique widely used in many general applications. Two interesting sub-activity of clustering process are studied in this paper, selection the number of clusters and analysis the result of data clustering. This research aims at studying the clustering validation to find appropriate number of clusters for k-means method. The characteristics of experimental data have 3 shapes and each shape have 4 datasets (100 items), which diffusion is achieved by applying a Gaussian distributed (normal distribution). This research used two techniques for clustering validation: Silhouette and Sum of Squared Errors (SSE). The research shows comparative results on data clustering configuration k from 2 to 10. The results of both Silhouette and SSE are consistent in the sense that Silhouette and SSE present appropriate number of clusters at the same k-value (Silhouette value: maximum average, SSE-value: knee point).

Keywords: Clustering Validity, Silhouette Measure, Sum of Squared Errors, k-means Algorithm.

1. Introduction

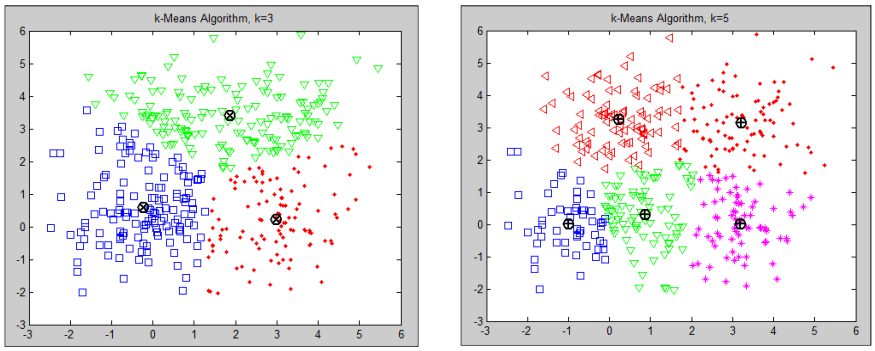
A clustering is to group data. Although the clustering is similar to the data classification in terms of data input, the clustering is learning without target class. The clustering algorithm forms groups based on object similarities[(1)]. The clustering was applied to many fields such as bioinformatics, genetics, image processing, speech recognition, market research, document classification, and weather classification[(2)]. In addition, the clustering was applied to document data analysis that was one of big data learning[(3-7)].

There are various algorithms for the data clustering. But the most popular one is k-means algorithm. The k-means algorithm is very simple in operation and suitable for unraveling compact clusters and a fast iterative algorithm[(8)]. The principle of k-means algorithm has divide n objects from dataset for k clusters that used center-based clustering methods[(2)]. In addition, each cluster has represented by the means of objects[(8)]. Although k-means is a popular technique, k-means is not known the correct number of clusters a priori. Consequently, the main challenge for these clustering methods is in determining the number of clusters[(2)]. In general, the number of clusters has been set by users or archives from knowledge of research[(1, 9-11)].

Fig. 1 shows the distribution of each cluster when k=3, and k=4. The researcher found that the determination of suitable k value is not clear as shown in fig. 1a and fig. 1b. As mentioned above about problem of clustering, there are various research for selecting an appropriate number of clusters[(10-13)]. Each of the proposed technique is suitable for each of data distribution such as Gaussianity and non-Gaussianity[(14)]. Therefore, finding the correct k-value for clustering is still a fundamental problem of clustering methods[(15-16)].

In this research, we study the clustering validity techniques to quantify the appropriate number of clusters for k-means algorithm. These techniques are Silhouette and Sum of Squared Errors. The rest of this paper is organized as follows. Section 2 discusses related research. Section 3 contains a description of methodology. Section 4 presents the results of experiments. The last section contains conclusions.

DOI: 10.12792/iciae2015.012



(a) k=3 (b) k=5

Fig. 1. Data clustering when k=3 and k=5

2. Related Research

Rousseeuw[12] have proposed the concept of the monitoring cluster. In the research proposed for Silhouette technique, which is based on the comparison of objects tightness and separation. The silhouette can reflect the data is grouped that objects are organized into groups that match it. This is a tool to assess the validity of the clustering to be used for selecting the optimal k in the cluster.

Kwedlo[17] have proposed the concept of problem solving in order to know the number of cluster using the Sum of Squared Errors (SSE). The research for developed a new method, called DE-KM (Differential Evolution Algorithm: DE) technique is the combination of algorithms, k-means clustering by tuning in DE and sort data to see the evolution. Experimental results show that the highest k values appropriate to the clustering, and DE-KM SSE values than the other methods they tested.

Shahbaba and Beheshti[2] have proposed the concept of dealing with the problem of determining the correct number of cluster to be grouped. The method used to estimate the probability of error point average (ACE), which is the difference between the actual point and estimate point. The idea is to explore how to use the k-means clustering with ACE k-means low (MACE-Means) is used with the UCI data and the other synthetic. They found that a correct number of cluster that can be spent on items that are sturdy little overlap and time less.

Jun et al.[9] have proposed the concept of clustering the documents based on the concept of reducing the dimension of the data, combined with the clustering k-means based on clustering with support vector and silhouette measure. They have experimented with the patent, documents from UCI to analyze separately each group of documents to clustering for technology forecasting.

3. Proposed Methodology

3.1 A Framework of Data Clustering and Validation Approach

The clustering is a data mining at an unsupervised learning technique[2, 17-20]. The principle of data clustering that objects in the same cluster will have to look very similar, while objects in other similar less[17]. There are various algorithms of clustering technique, for example, Basic Sequential Algorithms Scheme (BSAS), Partitioning Around Medoids Algorithm (PAM), Fuzzy c-Means Algorithm (FCM), k-means Algorithm[8] and so on.

The main steps in the work of the clustering has 5 steps[21]. There are (a) set a number of cluster for clustering (k) and cluster feature, (b) set a function for objects similarity measurement, (c) run clustering algorithm, (d) set Visualization to display cluster and (e) clustering validity analysis, as fig. 2.

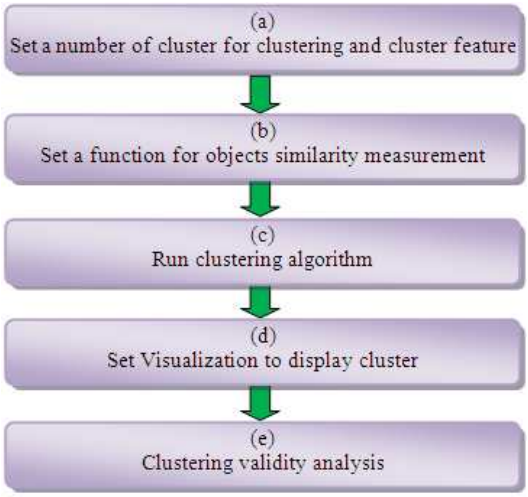


Fig. 2. Show 5 steps in clustering process.

3.2 k-Means Clustering

The k-means clustering is a technique that relies on the center of cluster. This is often represented by the average (Means) of cluster. The clustering measure the similarity of the group by iterating the measurement distance between each object and the center of each cluster[2] using Euclidean distance measuring.

The k-means algorithm is an iterative algorithm which can be described by the following steps.

Algorithm: k-means Clustering[17].

(a) Choose initial centroids $\{m_1,\ldots,m_k\}$ of the clusters $\{C_1,\ldots,C_k\}$.

(b) Calculate new cluster membership. A feature vector x_j is assigned to the cluster C_i if and only if

$$i = \arg \min_{k=1,\ldots,k} ||x_j - m_k||^2 \quad (1)$$

(c) Recalculate centroids for the cluster according.

$$m_i = \frac{1}{|C_i|} \sum_{x_j \in C_i} x_j. \quad (2)$$

(d) If none of the cluster centroids have changed, finish the algorithm. Otherwise go to Step (b).

3.3 Clustering Validity Methods

3.3.1 Silhouette Measure

The concept of Rousseeuw[12] is described as follows: the Silhouette is a tool used to assess the validity of clustering. The silhouette constructed to select the optimal number of cluster with a ratio scale data (as in the case of Euclidean distances) that suitable for clearly separated cluster. The clustering are considered average proximities as the two are dissimilarities and similarities, which work best in a situation with roughly spherical clusters.

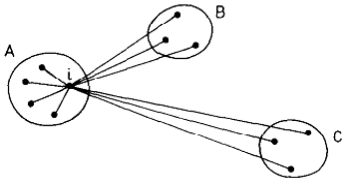


Fig. 3. Show computation *s(i)* for each object, where object *i* belong to cluster A[12].

Case #1 considered dissimilarities[12].

From fig. 3 described for take the object *i* in the data set, and assigned to cluster *A*, then define as follows:

s(i) = in case of dissimilarities.

i = object *i* belong to cluster *A*.

a(i) = average dissimilarity of *i* to all other objects of ***A***.

d(i, C) = average dissimilarity of i to all objects of *C*.

b(i) = minimum *d(i, C)*, where $C \neq A$.

B = the cluster *B* for which minimum is attained the neighbor of object *i*

The cluster *B* is like the second-best choice for object *i*: if it could not be accommodated into cluster *A*, which cluster B would be the closest competitor In Fig. 3. The number *s(i)* write this in formula:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}. \quad (3)$$

The number *s(i)* is obtained by combining *a(i)* and *b(i)* as follows:

$$s(i) = \begin{cases} 1 - a(i)/b(i) & if\ a(i) < b(i), \\ 0 & if\ a(i) = b(i), \\ b(i)/a(i) - 1 & if\ a(i) > b(i), \end{cases}$$

s(i) can will be $-1 \leq s(i) \leq 1$

Case #2 considered similarities[12].

In this case consideration similarities and define *a'(i)*, *d'(i, C)*, and put *b'(i)* = maximum *d'(i, C)*, where $C \neq A$. The numbers *s(i)* is obtained by

$$s(i) = \begin{cases} 1 - b'(i)/a'(i) & if\ a'(i) > b'(i), \\ 0 & if\ a'(i) = b'(i), \\ a'(i)/b'(i) - 1 & if\ a'(i) < b'(i), \end{cases}$$

For Example, fig. 4 shows the results silhouette of clustering, when fig. 4 (a) present clustering on k = 2 and fig. 4 (b) clustering on k = 3. The Figure shows the comparison of result: density and separation, Neighbors, the average Silhouette of each cluster. Which silhouette is used to support the evaluation clustering with the maximum of silhouette.

3.3.2 Sum of Squared Errors

The k-means clustering techniques defines the target object (x_i) to each group (C_i), which relies on the Euclidean distance measurement (m_i) is the reference point to check the quality of clustering. The Sum of Squared Errors: SSE is another technique for clustering validity. SSE is defined as follows[17].

$$SSE(X, \Pi) = \sum_{i=1}^{K} \sum_{x_j \in C_i} ||x_j - m_i||^2 \quad (4)$$

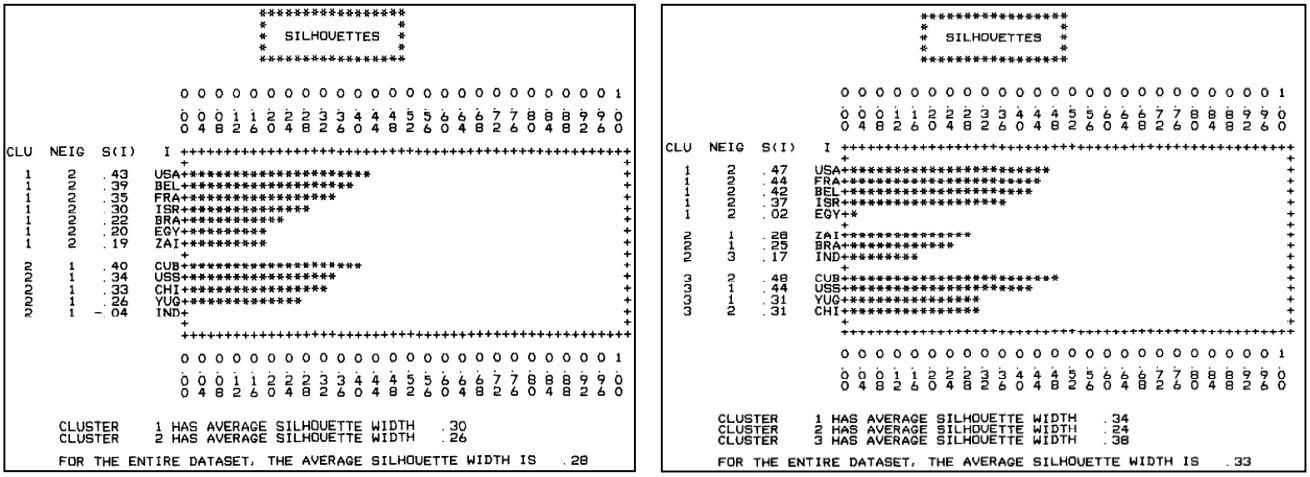
where

N = Feature vectors

$X = \{x_1, \ldots, x_i, \ldots x_N\}$, $x_i \in \Re^M$

$\Pi = \{C_1, C_2, \ldots C_K\}$ $\quad ,\forall_{i \neq j} C_i \cap C_j = \emptyset, \cup_{i=1}^{K} C_i = X, \forall_i C_i \neq \emptyset.$

$||\cdot||$ = Euclidean distance and m_i is centroid of cluster C_i which can computed as Eq.(2).



(a) k=2. (b) k=3.

Fig. 4. Shown Silhouette was present clustering when k=2 and k=3[12].

Conditions of applied the SSE for clustering, is to determine k ≥ 2[22]. When the SSE is applied in graph that generated from the relationship between the SSE and k value at knee point (Significant "knee"), which is positioned to indicate the appropriate number of cluster in the k-means clustering [8] as shown in fig. 5.

3.4 Selection an Appropriate Number of Cluster

The principle of the monitoring tool for clustering, can support the selection of correct k values for the k-means clustering, consider the following.

Fig. 4, Silhouette is used to assist in cluster monitoring. This analysis is compared between Fig. 4 (a) and (b) it is found that the average silhouette of clustering when k = 3, the value 33 will be greater than k = 2, the value 28.

Fig. 5 the SSE is used in the inspection cluster. This analysis was shows the appropriate number at the knee clearly was 5(a) k = 3, 5(b) k = 4 and 5(c) k = 5, which the appropriate number of cluster.

4. Experimentation and Results

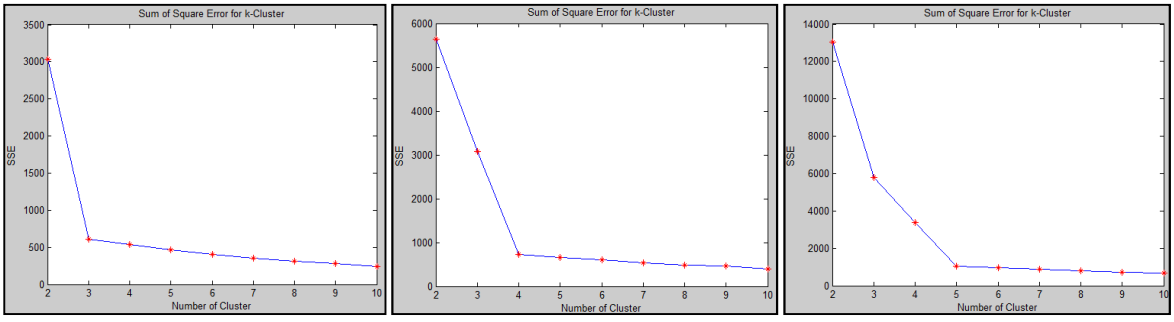
4.1 Experimental Data

The research uses data synthesized with 3 shapes and each shape have 4 datasets (100 items), which is applying a Gaussian distribution (normal distribution). Fig. 6 is the distribution of a spherical around the center of dataset, fig. 7 the distribution is non-spherical lying on the x-axis, and fig. 8 the distribution is spherical, but each group will have some overlap.

4.2 Results of k-Means Clustering Method

In experiments, the researchers repeated the k-means clustering algorithm with datasets by changing the value of k, set k = 2 to k = 10, which illustrate the specific clustering when k = 2, 4, and 6 shown in fig. 6, fig. 7 and fig. 8.

The next step is to investigate the cluster. This relies on the analysis of both Silhouette and SSE of above mentioned, are as follows.



(a) k=3 (b) k=4 (c) k=5

Fig. 5. Number of cluster consideration from the relationship between SSE and the k value.

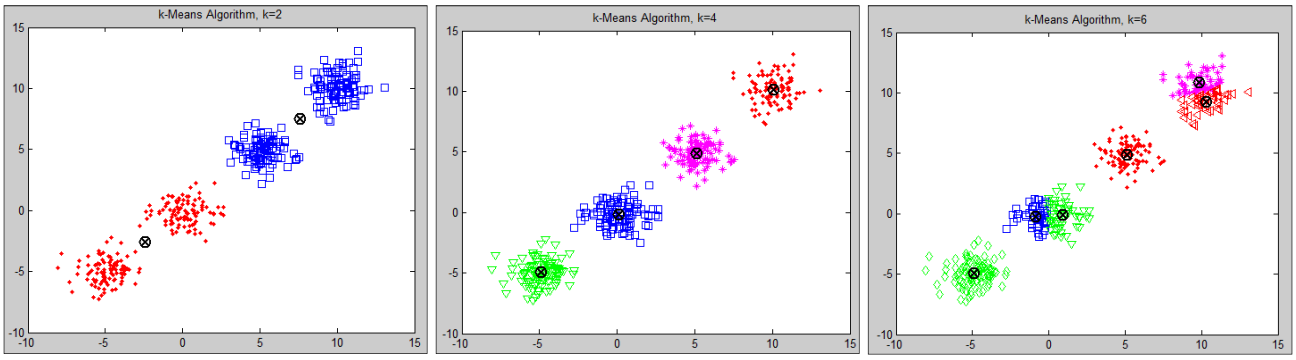


Fig. 6. Clustering with spherical data shape where k=2 (left), k=4 (middle), and k=6 (right)

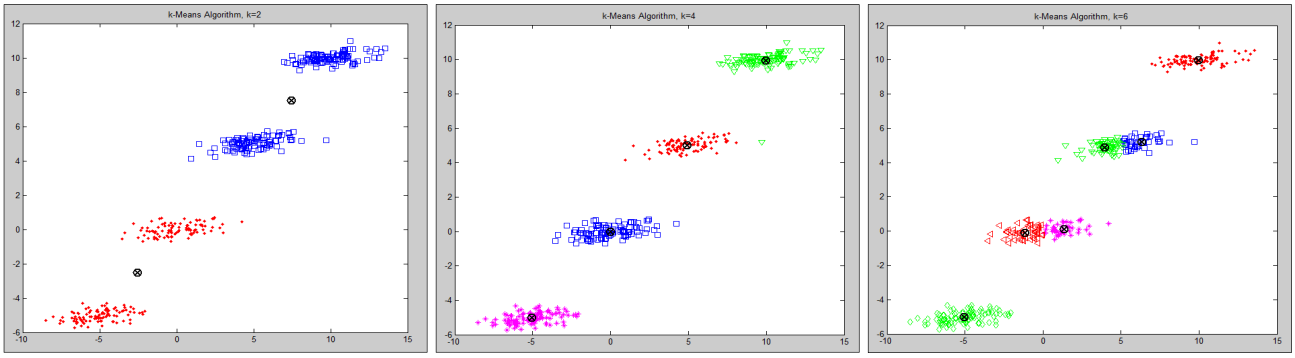


Fig. 7. Clustering with non-spherical data shape where k=2 (left), k=4 (middle), and k=6 (right)

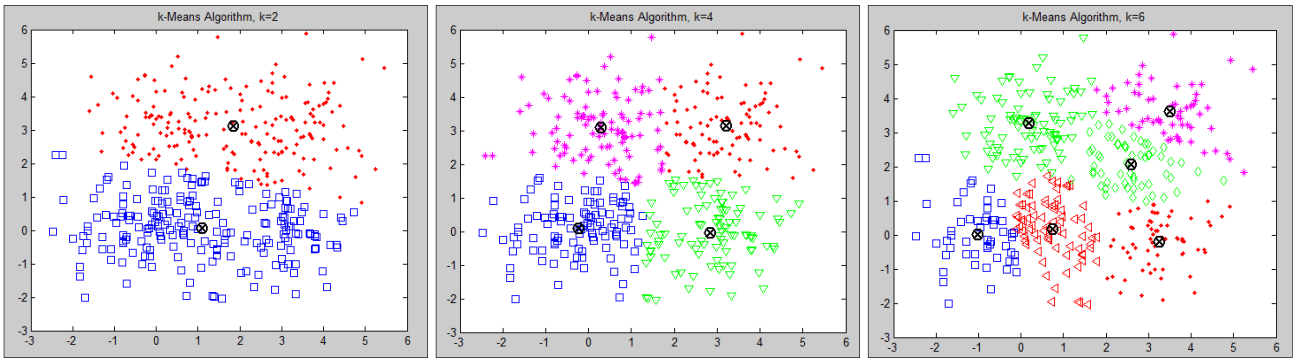


Fig. 8. Clustering with overlap data where k=2 (left), k=4 (middle), and k=6 (right)

4.3 Clustering Validity with Silhouette Measure

Consider Fig. 9 is an illustration Silhouette of a clustering technique to the k-means repeating the grouping by changing the value of k from 2 to 10, which shows a comparison of the density and separation of each cluster. Which found that the density of the k values of k = 2 and k = 4 show the density and separation is optimal.

Using the silhouette to assess the quality of clustering not silhouette diagrams only in addition need to consider the average of silhouette. It was found that the average of all silhouette values when k=4 the highest shown in table 1.

4.4 Clustering Validity with SSE

The Result of SSE for inspection the cluster is shown in Table 2. The table 2 shows the SSE value and rate of change of the SSE when k = 2 to 10, found that when k = 4 SSE is the maximum rate of change. The rate of change (%Change) defined as follows.

$$\%Change = \frac{(SSEofK_{i-1} - SSEofK_i) * 100}{SSEofK_i}. \qquad (5)$$

i can will be $i \geq 2$

where

$SSEofK_{i-1}$ = SSE values of k_{i-1}

$SSEofK_i$ = SSE values of k_i

Table 1. Show comparison of the average of the Silhouette of a k-means clustering when k = 2 to 10.

Number of Cluster	Average of Silhouette		
	Spherical	Non-Spherical	Spherical & Overlap
K=2	0.8305	0.8318	0.5262
K=3	0.7715	0.7553	0.5603
K=4	**0.9117**	**0.9018**	**0.6150**
K=5	0.8182	0.8558	0.5720
K=6	0.7109	0.7982	0.5407
K=7	0.5639	0.7466	0.5177
K=8	0.6198	0.7333	0.5104
K=9	0.5072	0.6945	0.5217
K=10	0.5162	0.6850	0.5177

As the Silhouette to assess the quality of clustering not the data in table 2 that should set correct k value only. In order to investigate the effect is therefore necessary to consider a graph showing the relationship between k and the SSE values at the knee point are shown in fig. 10.

Table 2. Show SSE values and %change from k-means algorithm when k=2 to 10.

Number of Cluster	Sum of Squared Errors					
	Spherical		Non-Spherical		Spherical & Overlap	
	SSE	%Change	SSE	%Change	SSE	%Change
K=2	5,972.97	-	6,042.24	-	1505.40	-
K=3	3,179.66	87.85	3,265.32	85.04	958.94	56.99
K=4	**771.76**	**312.00**	**834.01**	**291.52**	**608.08**	**57.70**
K=5	682.02	13.16	679.50	22.74	518.62	17.25
K=6	612.04	11.43	552.51	22.98	441.48	17.47
K=7	544.41	12.42	430.75	28.27	387.16	14.03
K=8	512.90	6.14	403.10	6.86	337.27	14.79
K=9	436.02	17.63	364.19	10.68	302.21	11.60
K=10	403.84	7.97	245.55	48.32	276.06	9.47

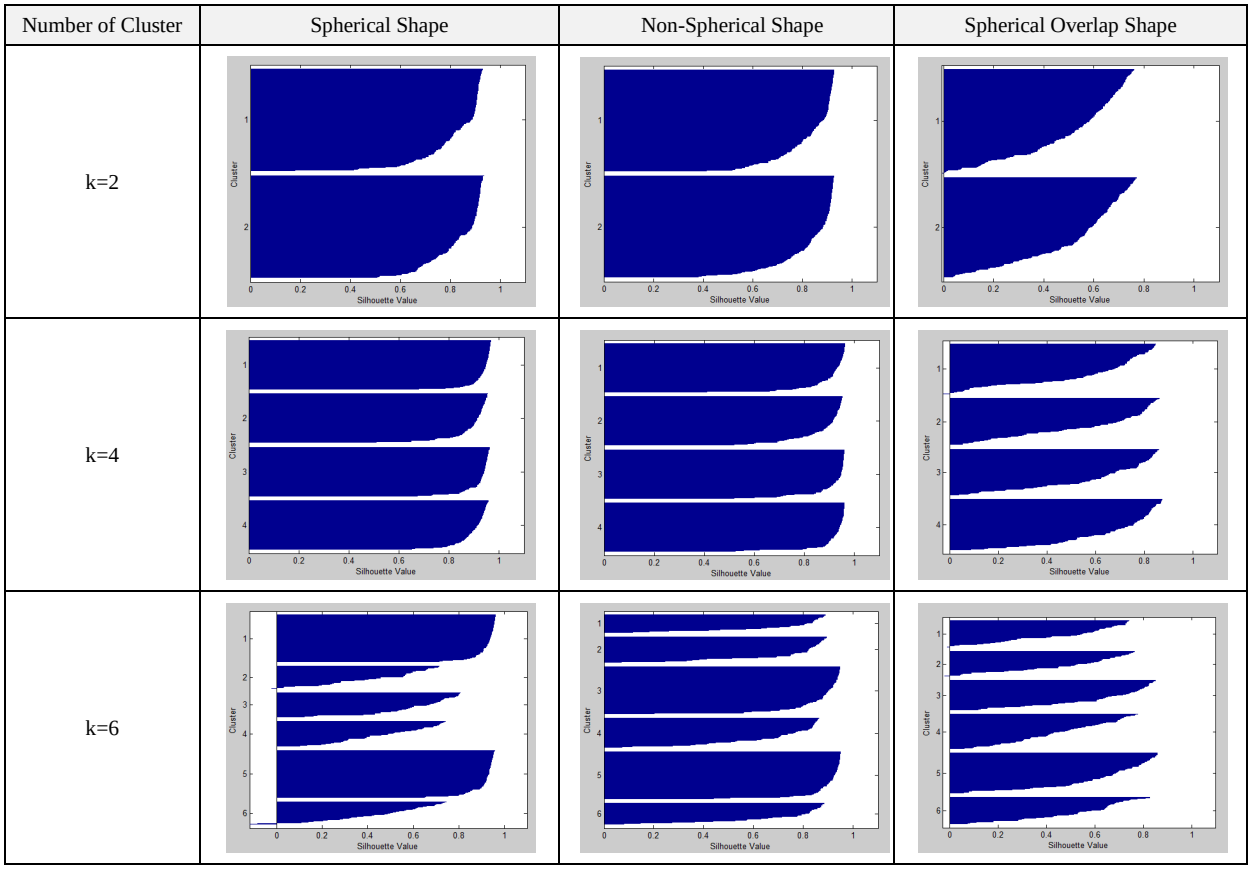
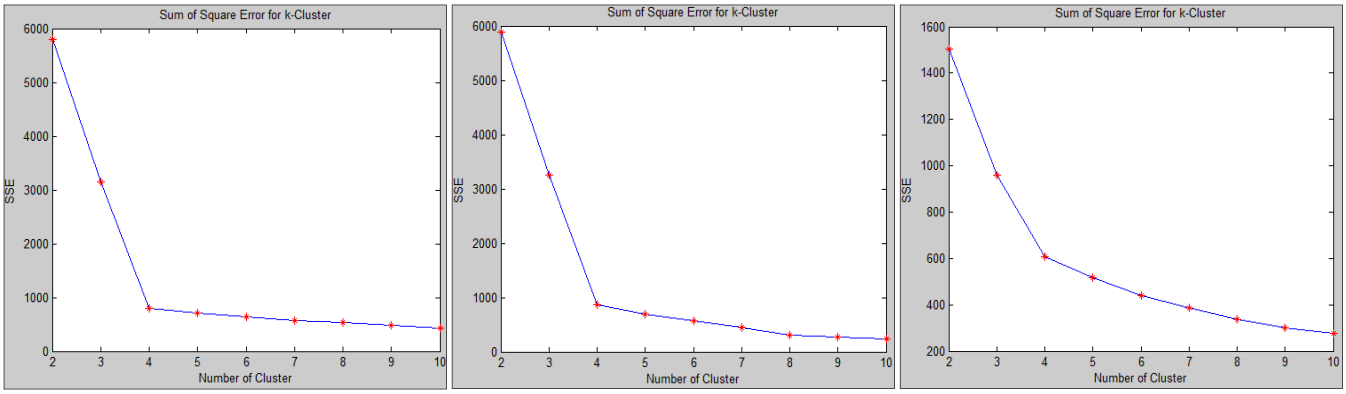


Fig. 9. Silhouette of a clustering technique when k=2, k=4, and k=6



(a) Spherical (b) Non-Spherical (c) Spherical & Overlap

Fig. 10. The graph showing the relationship between k and the SSE values of different data shapes

5. Conclusions

The results of research above when examining the Silhouette clustering analysis is to determine the k = 4 was the highest average Silhouette with all the data sets. When examining the clustering of the graph that shows the relationship between the SSE and k value with k = 4 the result was the knee point. That means the examination of both Silhouette and SSE are result inconsistent. Is that the number of cluster as the same number that k = 4.

However, a comparison of SSE and Silhouette have to attention is if the data does not overlap. Assessment the number of cluster is appropriate both SSE and Silhouette. However, when the data begin to overlap SSE will provide an assessment that is more close to the true value.

Acknowledgment

The first author has been supported by grant from the Informatics Program, Faculty of Science and Technology, Rajabaht Nakhon Ratchasrima University (NRRU). Data Engineering Research Unit has been funded by Suranaree University of Technology.

References

(1) Han, J., and Kamber, M. (2006). Data mining concepts and techniques (2nd ed.). United States of America: Morgan Kaufman Publishers.

(2) Shahbaba, M., Beheshti, S. (2014). MACE-means clustering. Singnal Processing. Vol.105, pp.216-225.

(3) Aliguliyev, R. M. (2009). Clustering of document collection–A weighting approach. Expert Systems with Applications, Vol.36, pp. 7904-7916.

(4) Isa, D., Kallimani, V. P., and Lee, L. H. (2009). Using the Self Organizing map for Clustering of Text Documents. Expert Systems with Applications, Vol.36, pp.9584–9591.

(5) Maziere, P. A. D., and Hulle, M. M. V. (2011). A clustering study of a 7000 EU document inventory using MDS and SOM. Expert Systems with Applications, Vol.38, pp. 8835–8849.

(6) Saracoglu, R., Tutuncu, K., and Allahverdi, N. (2007). A fuzzy clustering approach for finding similar documents using a novel similarity measure. Expert Systems with Applications, Vol.33, pp.600–605.

(7) Tseng, Y. H. (2010). Generic title labeling for clustered documents. Expert Systems with Applications, Vol.37, pp.2247–2254.

(8) Theodoridis, S., Pikrakis, A., Koutroumbas, K., Cavouras, D. (2010). An Introduction to Pattern Reccognition : A MATLAB Approach. Academic Press, USA.

(9) Jun, S., Park, S., and Jang, D. (2014). Document clustering method using dimension reduction and support vector clustering to overcome sparseness. Expert Systems with Applications. Vol.41, pp.3204–3212.

(10) Everitt, B. S., Landau, S., and Leese, M. (2001). Cluster analysis (4th ed.). Apnold.

(11) Jun, S., and Uhm, D. (2010). Patent and statistics, What's the connection? Communications of the Korean Statistical Society, Vol.17(2), pp.205–222.

(12) Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics. Vol.20, pp.53-65.
(13) Wang, L., Leckie, C., Ramamohanarao, K., and Bezdek, J. (2009). Automatically determining the number of clusters in unlabeled data sets. IEEE Transactions on Knowledge and Data Engineering, Vol.21(3), pp.335–350.
(14) McNicholas, P. D., Subedi, S. (2012). Clustering gene expression time course data using mixtures of multivariate t-distributions, J. Stat. Plan. Inference 142 (May (5)). p.1114–1127.
(15) Jain, A. K. (2010). Data clustering: 50 years beyond k-means, Pattern Recogn. Lett. Vol.(8) 31, pp.651–666, http://dx.doi.org//10.1016j.patrec..2009.09.011
(16) Aggarwal, C. C., Reddy, C. K. (2013). Data Clustering: Algorithms and Applications, Vol.31, CRC Press, Hoboken, New Jersey, p.648. (Chapman & Hall/CRC Data Mining and Knowledge Discovery Series, ISBN: 1466558210).
(17) Kwedlo, W. (2011). A clustering method combining differential evolution with the k-means algorithm. Pattern Recognition Letters. Vol.32, pp.1613–1621.
(18) Roiger, R. J., Geatz, M. W. (2003). Data Mining A Tutorial – Based Primer. Pearson Education, Inc. Addison Wesley. pp. 11-12.
(19) Jain, A., Murty, M. N., Flynn, P. J. (1999). Data clustering: a review. ACM Comput. Surv. Vol.31(3), pp.264–323.
(20) Kaufman, L., Rousseeuw, P. J. (1990). Finding Groups in Data: An Introduction to Cluster Analysis. Wiley.
(21) Kerdprasop, K. (2006). Density Biased Sampling for Incremental Data Clustering. Research Final Report. School of Computer Engineering, Suranaree University of Technology.
(22) Aloise, D., Deshpande, A., Hansen, P., Popat, P. (2009). NP-hardness of Euclidean sum-of-squares clustering. Mach. Learn Vol.75 (2), pp.245–248.