International Journal of Fuzzy Logic and Intelligent Systems, vol. 10, no. 1, March 2010, pp. 31−36

Empirical Comparisons of Clustering Algorithms using Silhouette Information

Sunghae Jun and Seung-Joo Lee

Department of Bioinformatics & Statistics, Cheongju University, 360-764 Chungbuk, Korea
{shjun, access}@cju.ac.kr

Abstract

Many clustering algorithms have been used in diverse fields. When we need to group given data set into clusters, many clustering algorithms based on similarity or distance measures are considered. Most clustering works have been based on hierarchical and non-hierarchical clustering algorithms. Generally, for the clustering works, researchers have used clustering algorithms case by case from these algorithms. Also they have to determine proper clustering methods subjectively by their prior knowledge. In this paper, to solve the subjective problem of clustering we make empirical comparisons of popular clustering algorithms which are hierarchical and non hierarchical techniques using Silhouette measure. We use silhouette information to evaluate the clustering results such as the number of clusters and cluster variance. We verify our comparison study by experimental results using data sets from UCI machine learning repository. Therefore we are able to use efficient and objective clustering algorithms.

Key Words : Objective clustering, Subjective clustering, Silhouette Information, Number of clusters

1. Introduction

We have used many clustering algorithms in diverse data mining fields[1]. These algorithms are based on statistical methods which are depended on distance or similarity[2]. Also machine learning algorithms such as K-means clustering and SOM(self organizing maps) have been used in clustering[3]. We have to select one from many algorithms for clustering works. In given training data, most researchers have to select proper clustering algorithms by their prior knowledge[1],[2]. This determination may be not objective but subjective. In general, the performance of the clustering result is depended on the selected algorithms. So we have to determine good algorithm for the data set. To overcome the subjective problem of the clustering, we need objective selection of the clustering algorithm[4],[5]. In this paper, using Silhouette measure, we do empirical comparisons among popular clustering algorithms by experiments using UCI machine learning repository[6]. Silhouette measure is a visualization for partitioning methods[7]. Until now, the Silhouette based clustering studies have not been used widely. In our research, we can do objective clustering by Silhouette value and width. By our experiments using data sets from UCI machine learning repository, we show comparison results of objective clustering based on the Silhouette information. We wish our study is a beginning of the wide usage of the Silhouette measure.

2. Related Works

2.1 Compared clustering algorithms

Many clustering algorithms have been used in diverse data mining works. In this paper, we consider the algorithms as two methods, hierarchical and non-hierarchical clustering[1],[8].

gi={gi1, gi2, …, gin} and gk={gk1, gk2, …, gkm} are clusters, and n is smaller than m. Then gi is a sub group of gk. Hierarchical clustering algorithm is performed by their inclusion of clusters, gi, gk, …, gr (gi ⊃ gk ⊃ … gr). In general hierarchical clustering has two clustering methods which are agglomerative and divisive. The agglomerative clustering is a grouping method from small clusters. The other way, the divisive clustering is a grouping method from large clusters. That is, the agglomerative hierarchical clustering method is a bottom-up approach. Also the divisive hierarchical clustering method is a top-down approach. The following figure shows the agglomerative and divisive hierarchical clustering methods[8].

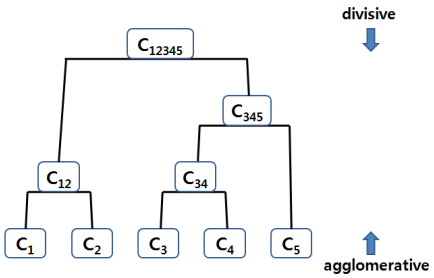

Fig. 1 Hierarchical clustering

In our experiments, we use single, complete, average, median, centroid, and ward as the agglomerative linkages of distance. Also DIANA(divisive analysis) is used for divisive hierarchical clustering[1]. Another clustering approach is non-hierarchical method in this paper. K-means clustering algorithm, PAM(partitioning around medoids), CLARA(clustering large applications), and FANNY(fuzzy analysis clustering) are popular non-hierarchical clustering

Manuscript received Aug. 30, 2009; revised Jan. 10, 2010.

methods[1].

2.2 Silhouette coefficient for optimal clustering

The silhouette measure has been used in clustering validation as a Silhouette coefficient[1]. The coefficient is based on the cohesion and separation of individual points and clusters[2],[7]. We can compute the Silhouette width as the following s(i)[9].

$$(i) = (b(i) – a(i)) / \max(a(i), b(i)) \quad (1)$$

In the above, a(i) is average distance between i and all other instances of the cluster to which i belongs. To define b(i), we let k into all other clusters. And we set d(i, k) as average distance between i with all instances of k. The smallest d(i, k) is defined as b(i). Instances with large Silhouette are good clustered and a small s(i) is represented the bad clustered of instances. In this paper, we evaluate the result by Silhouette information.

3. Empirical Comparisons of Clustering Algorithms using Silhouette Information

Cluster analysis has been used to establish maximally differentiable groups of instances from a large data set of instances[10],[11]. The instances within a group are similar across a set of variables. We standardize the data for empirical comparisons in this paper. Also we classify the clustering algorithms into the following figure for our experiments.

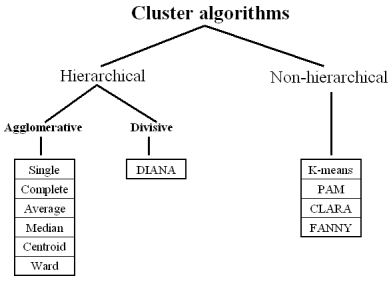

Fig. 2 Clustering algorithms

General clustering algorithms are classified into statistical methods and machine learning algorithms. The algorithms depended on statistics are consisted of hierarchical clustering and non-hierarchical clustering methods. Hierarchical clustering produces families of clusters that contain clusters that have similar instances. Also, the clustering method has several ways to determine linkage or joining clusters such as single, complete, average, median, centroid and Ward. These are divided by the distance measures. Non-hierarchical clustering produces discrete or overlapping clusters. Also, the non-hierarchical clustering does not merge clusters in a stepwise fashion instead, searches for the nearest separation into k groups among all objects. The K-means clustering algorithm is one of the simplest examples of non-hierarchical clustering. The distance measures of K-means clustering are Euclidean and Mahalanobis. PAM(portioning around methods) is another version of K-means clustering algorithm. This uses median instead of mean in K-means clustering. The PAM has been used many clustering works. So, we consider it in our empirical study. Machine learning has diverse clustering algorithms. From them, we consider fuzzy clustering and CLARA(clustering large application)[12]. These have been used as popular clustering algorithms in machine learning. Fuzzy clustering clusters data using fuzzy set which has membership functions of weights. The CLARA is a sampling-based clustering algorithm to cluster large data set. Of course, we can consider more algorithms than the clustering algorithms of this paper. But we propose not comparison results of all clustering algorithm but comparison process of the clustering algorithms by Silhouette information. So any other clustering algorithms can be compared by our approach. The following steps show our process for empirical comparison of the clustering algorithms by Silhouette information.

(Step1) Get training data.
(Step2) Select a clustering algorithm.
(Step3) Compute the Silhouette width.
For i=1 to C
(C: possible number of cluster)
(Step4) Determine #NC, the number of clusters with maximum Silhouette width.
(Step5) Compute Silhouette with among the candidate clustering algorithms.
(in this step, the number of clusters is fixed by #NC)
(Step6) Finally choose the proper clustering algorithm with maximum Silhouette width for given data.

We find appropriate number of clusters of given data by Silhouette width(Step2, Step3, and Step4). Also we can select the proper clustering algorithm for given data by the Silhouette information(Step5 and Step6). Although we make experiments about some clustering algorithms for verifying our study in next section, it is possible to extend above steps to comparison of all clustering algorithms.

4. Experiments and Results

To make experiments for our empirical comparisons, we use the data sets of UCI machine learning repository. The following table shows the information of the data.

Table 1. Numbers of instances and attributes

Data set	# of instances	# of attributes
Breast Cancer	638	9
Diabetes	768	8
Image	2310	18
Iris	150	4
Vehicle	846	18

Considering the numbers of instances and attributes, we

select diverse data set. Breast cancer is Breast Cancer Wisconsin data with 699 instances and 10 attributes originally. The data have missing values and sample code number(id number). We remove these from the data. So we get 638 instances and 9 attributes. Other data sets are used by same process with Breast cancer data. Also we find the number of classes and their labels in the following table.

Table 2. Numbers of classes and labels of each class

Data set	# of classes	# of Labels of each class
Breast Cancer	2	(444,238)
Diabetes	2	(500,268)
Image	7	(330,…,330)
Iris	3	(50,50,50)
Vehicle	4	(218,212,217,199)

In our experiments, we make an effort to verify empirical comparisons of popular clustering algorithms using Silhouette information. Also to compute the silhouette value, we use the cluster package based on R-project[9],[13]. The following figure is the Silhouette result of Iris data by cluster Package.

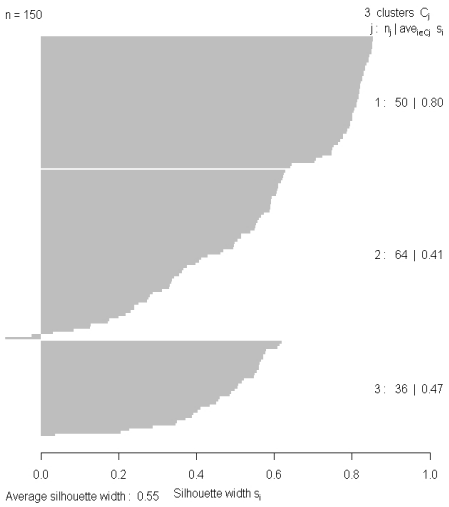

Fig. 3 Silhouette values by cluster Package

In the above figure, we know the numbers of clusters of given data is 3. All 150 instances are assigned to clusters C1, C2, and C3. C1 has 50 instances and its Silhouette value is 0.80. C2 and C3 are represented same to C1. The average Silhouette value of three clusters is 0.55. We can get the best clustering result when the Silhouette width has maximum value. First we compute average Silhouette values of Breast cancer according to traditional clustering algorithms.

Form above result, we can select the proper clustering algorithm for Breast Cancer Wisconsin as CLARA. Its Silhouette value is the largest in the competitive eleven algorithms. So we are able to expect good clustering result by CLARA in the Breast Cancer Wisconsin data. Next table shows the Silhouette result of Diabetes data.

Table 3. Average Silhouettes value of Breast Cancer Wisconsin

Clustering methods			Value
Hierarchical	Agglomerative	Single	0.4058
		Complete	0.5309
		Average	0.5885
		Median	0.2799
		Centroid	0.4058
		Ward	0.5709
	Divisive	DIANA	0.5951
Non-hierarchical (Partitioning)		K-means	0.5968
		PAM	0.5967
		CLARA	0.6160
		FANNY	0.5919

Table 4. Average Silhouettes value of Diabetes

Clustering methods			Value
Hierarchical	Agglomerative	Single	0.8061
		Complete	0.7859
		Average	0.7859
		Median	0.7300
		Centroid	0.7167
		Ward	0.5533
	Divisive	DIANA	0.5710
Non-hierarchical (Partitioning)		K-means	0.5688
		PAM	0.5053
		CLARA	0.5280
		FANNY	0.4976

The recommended clustering algorithm is determined as hierarchical-single linkage algorithm from above result. So we can select the single-linkage hierarchical clustering algorithm when we do clustering analysis about Diabetes data. We also compute the Silhouette value of Image Segmentation. The result is shown in the following.

Table 5. Average Silhouettes value of Image Segmentation

Clustering methods			Value
Hierarchical	Agglomerative	Single	0.7369
		Complete	0.4875
		Average	0.5115
		Median	0.7378
		Centroid	0.7390
		Ward	0.2997
	Divisive	DIANA	0.5801
Non-hierarchical (Partitioning)		K-means	0.2975
		PAM	0.2763
		CLARA	0.3380
		FANNY	0.0650

By the Silhouette value we can recommend the appropriate clustering algorithm for Image Segmentation data as hierarchical centroid-agglomerative clustering algorithm. In this result we find the values of centroid and median clustering methods are very close. So we conclude the median hierarchical clustering also can be recommended for Image

Segmentation. Next Iris data set is very popular in the data mining fields. We compute its Silhouette value in the following tables.

Table 6. Average Silhouettes value of Iris

Clustering methods			Value
Hierarchical	Agglomerative	Single	0.5121
		Complete	0.5136
		Average	0.5542
		Median	0.5081
		Centroid	0.5121
		Ward	0.5543
	Divisive	DIANA	0.5419
Non-hierarchical (Partitioning)		K-means	0.5528
		PAM	0.5528
		CLARA	0.5616
		FANNY	0.5354

We find the proper algorithm for Iris data clustering as CLARA. Finally we make an experiment of Vehicle data.

Table 7. Average Silhouettes value of Vehicle

Clustering methods			Value
Hierarchical	Agglomerative	Single	-0.0589
		Complete	0.3653
		Average	0.4634
		Median	0.4803
		Centroid	0.6004
		Ward	0.4256
	Divisive	DIANA	0.4994
Non-hierarchical (Partitioning)		K-means	0.4424
		PAM	0.4528
		CLARA	0.4089
		FANNY	0.4368

For the clustering Vehicle data, we can use centroid hierarchical clustering algorithm. In the table, the negative value, -0.0589(single) represents the clustering result is worse than original data. So it is insignificant.

The assessment of the clustering results cannot help but depend on the subjective knowledge of each application[14]. That is, there is not a correct answer of the clustering[14],[15]. But we need objective approaches to improve the assessment of clustering results. In this paper, we consider the proper determination of the number of clusters as one of the approaches. The objective selection of the number has been needed in the clustering[16]. We know the Silhouette value is a proper method for determining the number of clusters. So in this research, we did empirical comparisons of the method. Finally we compare the Silhouette value with popular criteria for selecting the number of clusters. AIC(Akaike's information criterion) and BIC(Bayesian information criterion) have been used for determining the number of clusters[16],[17],[18]. Generally the researchers can select the number of clusters with the minimum values of AIC or BIC measures. We make experiments by three simulation data sets. They are generated from mixture Gaussian distribution with different mean and standard deviation(s.d.). The following table shows the simulated data.

Table 8. Three Simulation data sets

Cluster	Simulation 1		Simulation 2		Simulation 3	
	mean	s.d.	mean	s.d.	mean	s.d.
1	0	0.5	0	0.5	0	0.35
2	10	0.5	2	0.5	1	0.35
3	20	0.5	4	0.5	2	0.35

In the above table, each simulation data set has 3 mixture Gaussian distributions according to different mean and standard deviation. So the number of clusters of each data set is 3. Also simulation 1, 2, and 3 data sets have 2 attributes and 500 instances respectively. We show the scatter plot of these data sets in the following figures. We compare the Silhouette value with AIC and BIC according to the distances of 3 clusters.

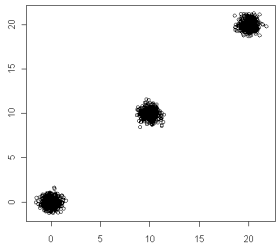

Fig. 4 Scatter plot of simulation 1 data

In the above figure, we know these 3 clusters are exactly separated. Also we are sure the number of clusters is 3. The AIC, BIC, and Silhouette values are shown the following table.

Table 9. AIC, BIC, and Silhouette values of simulation 1

# of clusters	AIC	BIC	Silhouette
2	657.16	699.66	0.63
3	**34.88**	**98.64**	**0.95**
4	41.67	126.68	0.76

The number of clusters is 3 with minimum values of AIC and BIC. Also the Silhouette has maximum value when the number is 3. So, all measures show proper results. Next we consider these measures when three clusters are closer. The following figure shows the plot of the three clusters.

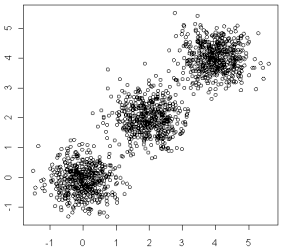

Fig. 5 Scatter plot of simulation 2 data

We compute the values of AIC, BIC, and Silhouette using simulation 2 data. The values are shown the following table.

Table 10. AIC, BIC, and Silhouette values of simulation 2

# of clusters	AIC	BIC	Silhouette
2	808.09	850.60	0.55
3	270.55	**334.31**	**0.66**
4	**252.46**	337.47	0.56

The BIC and Silhouette select the proper number of clusters. But the AIC does not find the proper number. In the experiment of simulation 2 data, the AIC has minimum value 200.66 when the number of clusters is 11. We consider some instances of each cluster are overlapping in the next simulation 3 data.

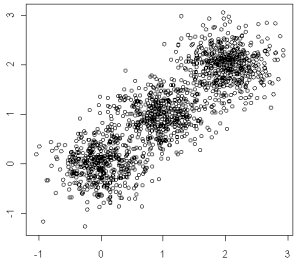

Fig. 6 Scatter plot of simulation 3 data

We also compute the values of AIC, BIC, and Silhouette like previous simulation data. The following table shows the result.

Table 11. AIC, BIC, and Silhouette values of simulation 3

# of clusters	AIC	BIC	Silhouette
2	899.92	942.43	0.58
3	453.77	517.53	**0.62**
4	**415.31**	**500.32**	0.52

In the above table, we know only Silhouette shows the result of the proper number of clusters. The AIC and BIC does not find the proper number. The minimum values of AIC and BIC are 257.50 and 473.85 when the numbers of clusters are 17 and 8 respectively.

Therefore we know the performance of Silhouette is better than AIC and BIC. When the instances of clusters are overlapping, we can expect to use the Silhouette measure for determining the number of clusters.

5. Conclusions and Future Works

In this paper, we showed comparison results of objective clustering using the Silhouette measure. Traditional clustering algorithms have needed the subjective knowledge to researchers for selecting the number of clusters. But the Silhouette algorithm does not demand this subjective information from the users. In spite of its clustering performance, the Silhouette measure has not been used extensively. In the experimental results using data sets from UCI machine learning repository, we verified empirical comparisons of popular clustering algorithms using Silhouette information. Our research was not all comparisons of all clustering algorithms. But we wish this paper is a commencement of the wide usage of the Silhouette information for objective clustering. This is the contribution of our paper. For our future works, we will apply Silhouette measure into unsupervised neural networks models like SOM for objective clustering.

References

[1] J. Han, M. Kamber, *Data Mining Concepts and Techniques*, Morgan Kaufmann, 2001.

[2] P.-N. Tan, M. Steinbach, V. Kumar, *Introduction to Data Mining*, Addison Wesley, 2006.

[3] A. S. Pandya, R. B. Macy, *Pattern Recognition with Neural Networks in C++*, IEEE Press, 1995.

[4] S. H. Jun, "An Optimal Clustering using Hybrid Self Organizing Map", *International Journal of Fuzzy Logic and Intelligent Systems*, vol. 6, no. 1, pp. 10-14, 2006.

[5] M. J. Park, S. H. Jun, K. W. Oh, "Determination of Optimal Cluster Size Using Bootstrap and Genetic Algorithm", *International Journal of Fuzzy Logic and Intelligent Systems*, vol. 13, no. 1, pp. 12-17, 2003.

[6] UCI ML Repository, http://archive.ics.uci.edu/ml/

[7] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied mathematics*, vol. 20, pp. 53-65, 1987.

[8] B. S. Everitt, S. Landau, M. Leese, *Cluster Analysis*, Arnold, 2001.

[9] M. Maechler, *Cluster Analysis Extended Rousseeuw et al.*, Package cluster, 2009.

[10] T. M. Mitchell, *Machine Learning*, McGraw-Hill, 1997.

[11] A. K. Jain, M. N. Murty, P. J. Flynn, "Data clustering: a review," *ACM Computing Surveys*, vol. 31, no. 3, pp. 264-323, 1999.

[12] D. Dumitrescu, B. Lazzerini, L. C. Jain, *Fuzzy Sets and Their Application to Clustering and Training*, CRC Press, 2000.

[13] The R Project for Statistical Computing, www.r-project.org

[14] R. Xu, D. Wunsch II, "Survey of clustering algorithms," *IEEE Transactions on Neural Networks*, vol. 16, no. 3, pp. 645-678, 2005.

[15] I. Oh, *Pattern Recognition*, Kyobo, 2008.

[16] R. C. Dubes, "How many clusters are best? – an

experiment," *Pattern Recognition*, vol. 20, no. 6, pp. 645-663, 1987.

[17] A. R. Liddle, "Information criteria for astrophysical model selection," *Monthly Notices of the Royal Astronomical Society: Letters*, vol. 377, iss. 1, pp. L74-L78, 2008.

[18] Q. Zhao, V. Hautamaki, P. Franti, "Knee Point Detection in BIC for Detecting the Number of Clusters," *Lecture Notes in Computer Science*, vol. 5259, pp. 664-673, 2008.

Sunghae Jun

He received the BS, MS, and PhD degrees in department of Statistics, Inha University, Korea, in 1993, 1996, and 2001. Also, He received PhD degree in department of Computer Science, Sogang University, Korea in 2007. He is currently Associate Professor in department of Bioinformatics & Statistics, Cheongju University, Korea. He has researched statistical learning theory and evolutionary algorithms.

Phone : +82-43-229-8205
Fax : +82-43-229-8432
E-mail : shjun@cju.ac.kr

Seung-Joo Lee

He received the BS degree in department of applied statistics from Cheongju University, Korea in 1985. Also, he received MS, and PhD degrees in department of Statistics, Dongkuk University, Korea, in 1987 and 1995. He is currently Professor in department of Bioinformatics & Statistics, Cheongju University, Korea. He has researched Bayesian statistics and multi-variate analysis.

Phone : +82-43-229-8204
Fax : +82-43-229-8432
E-mail : access@cju.ac.kr