Cluster Quality Analysis Using Silhouette Score

Ketan Rajshekhar Shahapure
Department of Computer Science
and Electrical Engineering
University of Maryland, Baltimore County
Email: ketans1@umbc.edu

Charles Nicholas
Department of Computer Science
and Electrical Engineering
University of Maryland, Baltimore County
Email: nicholas@umbc.edu

***Abstract*—Clustering is an important phase in data mining. Selecting the number of clusters in a clustering algorithm, e.g. choosing the best value of k in the various k-means algorithms [1], can be difficult. We studied the use of silhouette scores and scatter plots to suggest, and then validate, the number of clusters we specified in running the k-means clustering algorithm on two publicly available data sets. Scikit-learn's [4] silhouette score method, which is a measure of the quality of a cluster, was used to find the mean silhouette co-efficient of all the samples for different number of clusters. The highest silhouette score indicates the optimal number of clusters. We present several instances of utilizing the silhouette score to determine the best value of k for those data sets.**

1. Introduction

Determining the optimal number of clusters for a data set is an important problem in certain clustering algorithms, especially the well-known k-means and similar algorithms [1]. There is no one-size-fits-all method to determine the value of k, the optimal value for a given data set may well depend on the methods used for measuring similarities and the initial seed values used for partitioning. A solution is to inspect the dendrogram resulting from hierarchical clustering, but this remains a somewhat subjective and expensive approach, since hierarchical clustering is intrinsically slower than k-means. Hierarchical clustering could still be applied on several small subsets of the data, to find a reasonable estimate of k. We choose the more direct method of analyzing the silhouette scores [5] which measure the quality of clusters. A high average silhouette coefficient value indicates good clustering and helps in deciding the optimal value of the number of clusters k [3]. We present examples of this approach, along with 2-d and 3-d scatter plots to support if not validate the results.

We propose to investigate whether the silhouette score can be used for validation of the number of clusters obtained by running k-means clustering algorithm on each of several data sets. Dimensionality reduction is done to reduce the number of features and generate a 2D or 3D scatter plot which helps in visually analyzing the number of clusters and validating the result. The following steps are carried for the analysis

1) Obtain the data in the form of tuples.
2) Run sklearn's k-means algorithm on the data set.
3) Obtain cluster labels by fitting the data.
4) Calculate mean silhouette coefficient by passing tuples and cluster labels.
5) Repeat this for different numbers (i.e. values of k) of clusters.

2. Quality measurement

Scikit-learn's silhouette score function computes the mean silhouette coefficient of all samples. The silhouette coefficient is calculated by taking into account the mean intra-cluster distance a and the mean nearest-cluster distance b for each data point. The silhouette coefficient for a sample is $(b - a)/max(a, b)$.

- A silhouette score with a value near + 1 means the data point is in the correct cluster.
- A silhouette score with a value near 0 means the data point might belong in some other cluster.
- A silhouette score with a value near -1 means, the data point is in (a) wrong cluster.

The analysis of silhouette scores for different data sets is given below.

2.1. Iris data set

This is a classic multi-class classification data set provided by scikit-learn. The data set consists of 3 classes, 4 dimensions or features, and 150 samples. Figure 1 shows

```
For k = 2 The average silhouette_score is : 0.681046169211746
For k = 3 The average silhouette_score is : 0.5528190123564091
For k = 4 The average silhouette_score is : 0.4980505049972867
For k = 5 The average silhouette_score is : 0.4887488870931048
For k = 6 The average silhouette_score is : 0.3678464984712235
For k = 7 The average silhouette_score is : 0.3588294450965675
For k = 8 The average silhouette_score is : 0.34901133143367136
For k = 9 The average silhouette_score is : 0.3247749396939589
For k = 10 The average silhouette_score is : 0.3190032789646385
```

Figure 1. Silhouette scores for Iris data set

the silhouette scores for different number of clusters with k ranging from 2 to 10. It can be observed that the silhouette

score is the highest for k = 2. In addition, selecting k = 4 or k = 5 results in silhouette scores that are more or less equally bad. Therefore, k = 2 or k = 3 are the only two reasonable choices for this data set.

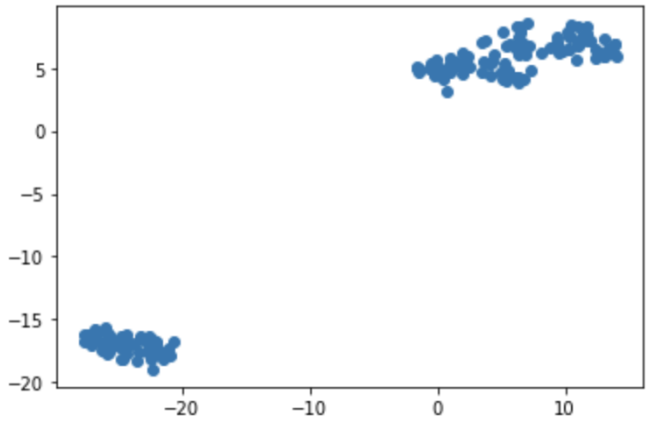

Figure 2. 2-d scatter plot for Iris data set

Inspection of Figure 2 shows that the Iris data set can be clustered into either 2 or 3 distinct clusters. In fact, we suppose that most people would say that $k = 2$ would be obvious. However, the silhouette score suggests that $k = 3$ is also a reasonable choice.

2.2. Clustering Basic Benchmark S-1 set

The S-1 data set [2] is widely used for benchmarking of clustering algorithms. The 2-D data is synthesized, consisting of N=5000 data points, and k=15 Gaussian clusters with different degrees of cluster overlap. In Figure 3 we see

```
For k = 10 The average silhouette_score is : 0.5977853892321484
For k = 11 The average silhouette_score is : 0.6044255565839454
For k = 12 The average silhouette_score is : 0.6373956265848799
For k = 13 The average silhouette_score is : 0.6488861306110127
For k = 14 The average silhouette_score is : 0.6898844415289176
For k = 15 The average silhouette_score is : 0.711278614093076
For k = 16 The average silhouette_score is : 0.6851826041302341
For k = 17 The average silhouette_score is : 0.6637768213342798
For k = 18 The average silhouette_score is : 0.6409205877720164
For k = 19 The average silhouette_score is : 0.6105366730233175
For k = 20 The average silhouette_score is : 0.5795987839208244
```

Figure 3. Silhouette score for S-1 data set

that the silhouette score for k = 15 is the highest, which is what we expect for this data set. Visual inspection of the 2-d scatter plot in Figure 4 supports the claim.

Results based on more data sets are presented in a longer companion paper [6]. In those results, some values of k resulted in silhouette scores that were very close, suggesting that in some data sets there might be two (or more?) excellent choices for k. This is a question for future work.

3. Conclusion

The silhouette score was obtained for different values of k for several data sets. The data was subjected to dimensionality reduction and plotted. We observed that the silhouette score provides a way to find a good value of k to specify in k-means clustering algorithms.

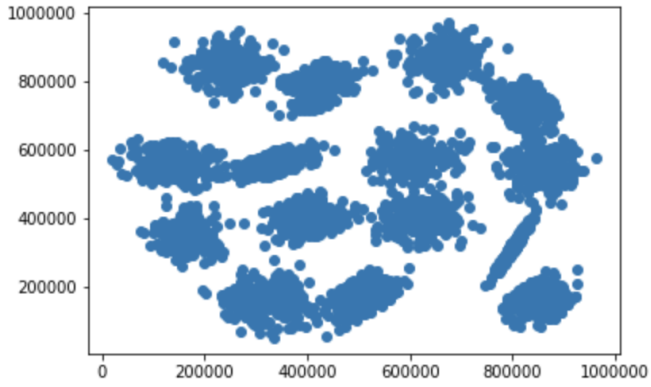

Figure 4. 2-d scatter plot for S-1 data set

4. Future Work

We want to continue this work with larger data sets, to explore the question of similar silhouette scores for different values of k. According to Volkovich et al. (2007) [7], sampling can help find the best value of k, as well as good candidates for initial seed values for the clusters. If the data set is quite large, sampling to suggest k as well as initial seed values following approaches could prove beneficial. For example, we could take a small random sample, say 1:100000, and use our results to suggest a value of k and the initial seed values. Done repeatedly, we could gain statistical confidence, so to speak, in the resulting value of k. If we keep track of the centroids produced after each of those sampled runs, that might lead to a better choice of initial centroids while running k-means on the entire data set.

References

[1] Edward W Forgy. Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *Biometrics*, 21:768–769, 1965.

[2] Pasi Fränti and Sami Sieranoja. k-means properties on six clustering benchmark datasets. *Applied Intelligence*, 48(12):4743–4759, 2018.

[3] Alboukadel Kassambara. Determining the optimal number of clusters: 3 must know methods. *Available online: https://www. datanovia. com/en/lessons/determiningthe-optimal-number-of-clusters-3-must-know-methods/.(accessed on 31 April 2018)*, 2017.

[4] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[5] Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53 – 65, 1987.

[6] Ketan Rajshekhar Shahapure. Cluster quality analysis using silhouette score, 2020. M.S. Writing Project, Department of Computer Science and Electrical Engineering, UMBC.

[7] Vladimir Volkovich, Jacob Kogan, and Charles Nicholas. Building initial partitions through sampling techniques. *European Journal of Operational Research*, 183(3):1097–1105, 2007.