Global Journal of
Computer Sciences:
Theory and Research

Volume 07, Issue 3, (2017) 112-122

www.gjcs.eu

Effectiveness measurement of spectral clustering algorithm

Farag Homed Ali Kuwil*, The Higher Institute for Administration and Financial Sciences, Istiklal Street, Benghazi-218, Libya.

Suggested Citation:

Kuwil, F. H. A. K. (2017). Effectiveness measurement of spectral clustering algorithm. *Global Journal of Computer Sciences: Theory and Research. 7*(3), 112-122.

Received July 23, 2017; revised October 21, 2017; accepted December 9, 2017.
Selection and peer review under responsibility of Prof. Dr. Dogan Ibrahim, Near East University, North Cyprus.

Abstract

After the Kuwil method was found for applying the spectral clustering algorithm, we need a way to make sure of the results, because in many cases the nature of data is not compatible with the algorithm; also, when the data contain more than *three* dimensions (3D) the results cannot be displayed on the monitor. So I found two techniques, first, for measuring the strength and effectiveness of S.C.A, such as some comparative relationships that measure the following: Effectiveness of algorithm applying, the strength of every cluster and the effectiveness of data correlation inside every cluster. Secondly, analysis of variance (ANOVA) for S.C.A; this depends on distance variance instead of values variance. I applied the methods above to calculate the strength and effectiveness of S.C.A, and they showed good results, so they can offer more reliability for the outputs of the algorithm. Using these relations and ANOVA for S.C.A help us to measure the data receptivity for applying the algorithm by 'Kuwil method', so the outputs will be more reliable and that will help to spread the use of this algorithm among researchers, analysts and other users.

Keywords: Spectral clustring, algorithm effectiveness, Kuwil method.

* ADDRESS FOR CORRESPONDENCE: **Farag Homed Ali Kuwil,** The Higher Institute for Administration and Financial Sciences, Istiklal Street, Benghazi, Libya. *E-mail address*: Kuwil73@gmail.com / Tel.: +90-545-442-6352

1. Introduction

Applying S.C.A may face some difficulties, based on the nature of the data under study. When data are three dimensional and more, they are impossible to be represented graphically on monitors using the current technology, while some data forms do not fit to the S.C.A application according to the algorithm definition. Therefore, I suggest testing the clustering outputs to be sure that they are reliable for decision making, by making limited use of analysis of variance (ANOVA) for S.C.A instead of graphical representation of data, using some comparative relations to test the strength of every cluster, and the effectiveness of algorithm application and data correlation inside every cluster. A lot of research and studies have been conducted on spectral clustering with regard to examining the number of clustering. Indeed, they show that the multiplicity of Eigenvalue 1 equals the number of clusters (this was followed to some extent by Polito and Perona in [1]). In [2], it is shown that if some conditions apply, then spectral clustering minimizes the multi-method normalized cut. A generalization of the two-way normalized cut criterion [3], random walks [4], graph cuts and normalized cuts [3], and matrix disorder theory [5], simplifying the difficulties to make them easier to understand concurrently with the improvement of algorithms [3–7], shows that significant theoretical progress has also been done. Yu and Shi [8] proposed to swap normalized eigenvectors to get optimal segmentation. Parallel spectral clustering [9] and spectral clustering using more details can be found in [10]. All the previous studies were about the algorithm and improvements in its application give more accurate results. The latest studies are interested in performance improvement in the speed of implementation through use of parallelism technology, but to the best of my knowledge, there is no published empirical study that attempts to test the significance of the effectiveness measurement through comparative relations or ANOVA.

2. Overview of Potential Fields

Two types of techniques were used to measure performance and to determine the possibility of applying the algorithm to multiple types of data.

2.1. Comparative Relations

According to the S.C.A concepts, the main issues we must look for about data are the distance and coherence among them to determine which data are similar and which are not. So the proposed method depends basically on the relations among the dataset points, through which we found five laws (relations), the first two of them for general measurement for implementation while the next three are for every cluster:

- **A. F**→**A**pply **F**actor indicates the receptivity of the dataset to applying **S.C.A**.
- **M. F**→**M**erge **F**actor illustrates the significance of merging two clusters or more.
- $\mathbf{ESF}_{(\mathbf{c.k-\ oth})}$ → **E**xternal **S**trength **F**actor between cluster k and other clusters.
- $\mathbf{ESF}_{(\mathbf{c.k-\ n})}$ → **E**xternal **S**trength **F**actor between cluster k and nearest cluster.
- $\mathbf{ISF}_{(\mathbf{c.k})}$ → **I**nternal **S**trength **F**actor of cluster k.

2.2. ANOVA

There are several uses of the ANOVA test, such as it is a way to find out if survey or experimental results are significant or help to figure out if there is a need to reject the null hypothesis or accept the alternate hypothesis or simply said, is it comparing groups to see if there is a difference between them [11].

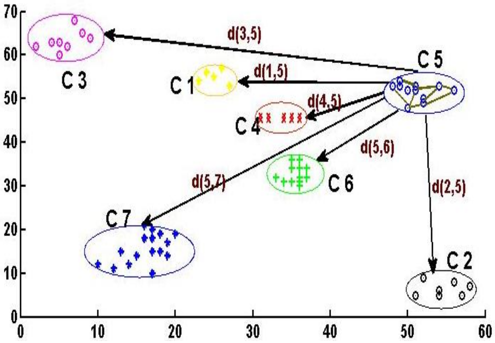

Figure 1. Distances among clusters

3. Algorithm Analysis

3.1. Comparative Relations

To clarify the proposal for comparative relations, we may look at Figure 1, which is a graph from experiment 1. The idea of measuring effectiveness is illustrated by examining the relationship among all data points inside every cluster with cluster 5; it is also used to find the relationship between this cluster and the others, that is, repeated with each cluster separately, considering that the horizontal axis scale is bigger than the vertical, which therefore causes an inaccurate perspective.

3.2. ANOVA

According to my search in statistical science, I found that ANOVA is the most important topic used to find a relation among many groups of data; therefore, I tried to adopt this law to take advantage of it for comparison among the results of the algorithm implementation, and so we use one-way ANOVA. Our use is limited to displaying the graph of each cluster, clarifying the information contained in each cluster and then comparing them in terms of distances, as is the basic idea of S.C.A. The values to be analysed for variance will be the distance among the points for each cluster. Figure 2 shows the general representation of the graphic resulting from the use of ANOVA in MATLAB, where the following can be clarified:

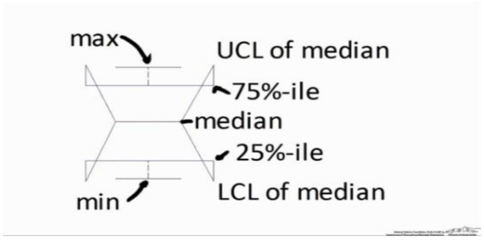

Figure 2. Graph of ANOVA in MATLAB

- **Max**—the largest distance between two points within the cluster.
- **Min**—less distance between two points within the cluster indicating at least *two* smallest coordinates in the case that it equals to 0.
- **Range**—the difference between the max and min distances.
- **Median**—it is a value which is in the middle of data.
- **25%-ile (Q1)**—interquartile range (IQR) of the 25th percentile.
- **75%-ile (Q3)**—IQR of the 75th percentile.
- **Upper Confidence Limit (UCL)**—UCL of median.
- **Lower Confidence Limit (LCL)**—LCL of median.
  We are not interested in UCL and LCL, because they are related to the hypothesis test.
- **IQR**—is a kind of dispersion measurement used to overcome the defects in the range, because it excludes the outlier values of both the sides where it depends on the calculation of the first and third quartiles: **IQR = (Q_3 – Q_1)**.

4. Experiments

In order for us to monitor the results and make comparisons, we have fixed the colours in descending order of the clusters in all the experiments, as given in Table 1. We used this method to test some data that were clustered by the Kuwil algorithm. We took *five* clustered dataset cases, *one* of them is 3D and the other *four* are in 2D, while *one* of them is real data.

Table 1. Clusters arranged by colour

Light blue	Black	pink	Red	Blue o	Green	Blue *
1	2	3	4	5	6	7

Experiment 1: Unreal Data, 2D

Table 2 shows a matrix of distances between clusters. Table 3 shows the weight of every cluster relative to the whole dataset, **$ESF_{(c.k–oth)}$**, **$ESF_{(c.k–n)}$** and **$ISF_{(c.k)}$** for every cluster, respectively. The apply factor **A.F** here is 0.158, which means that the dataset fits well the concepts of **S.C.A** definition. **M.F** = 0.555, which illustrates that the distance between the closest two clusters is not close enough to be significant for merging. Table 3 shows that **$c_{(2)}$** has the strongest **$ESF_{(c.2–oth)}$** because of its location among the other clusters. On the other hand, **$c_{(4)}$** has the weakest **$ESF_{(c.2–oth)}$**. In the graph of Figure 3, **$c_{(2)}$** has the strongest **$ESF_{(c.2–n)}$**, while we can see clearly in Figure 3 that both **$c_{(1)}$** and **$c_{(4)}$** have the weakest **$ESF_{(c.4–n)}$** , also shown in Table 3.

Table 2. Cluster distance matrix experiment 2

c	1	2	3	4	5	6	7
1	NILL	50.61	17	8.062	21	18.79	33.73
2	50.61	NILL	69.35	39.4	39.05	26.4	33.24
3	17	69.35	NILL	28.43	40.52	38.21	40.52
4	8.062	39.4	28.43	NILL	13.89	9	28.23
5	21	39.05	40.52	13.89	NILL	18.44	41.73
6	18.79	26.4	38.21	9	18.44	NILL	18.39
7	33.73	33.24	40.52	28.23	41.73	18.39	NILL

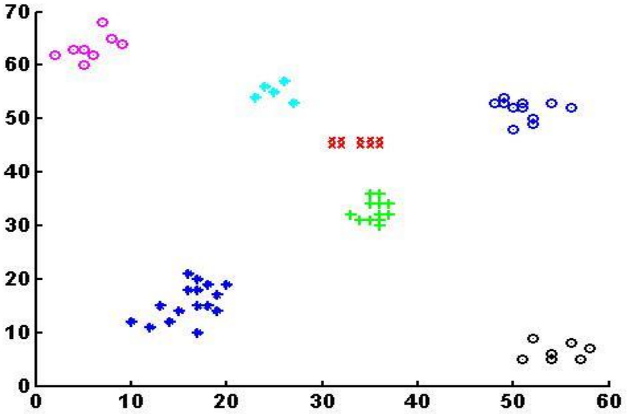

Figure 3. Result graph of experiment 1

Table 3. Clusters effectiveness for experiment 1

c	Weight	$\mathbf{ESF}_{(c.k-oth)}$	$\mathbf{ESF}_{(c.k-n)}$	$\mathbf{ISF}_{(c.k)}$
1	0.071	0.309	0.100	0.526
2	0.100	0.535	0.328	0.169
3	0.114	0.485	0.211	0.250
4	0.143	0.263	0.100	0.511
5	0.157	0.362	0.173	0.322
6	0.186	0.268	0.112	0.497
7	0.229	0.406	0.229	0.243

The strongest cluster internally is $\mathbf{c}_{(2)}$ (the lowest $\mathbf{ISF}_{(c.2)}$ = 0.169), and the weakest cluster is $\mathbf{c}_{(1)}$ ($\mathbf{ISF}_{(c.1)}$ = 0.526). Figure 4 shows the result of the MATLAB program, where *seven* clusters are represented in the graph for easy comparison and understanding.

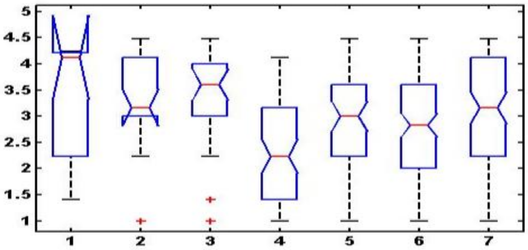

Figure 4. ANOVA result for experiment 1

Table 4 shows that $\mathbf{c}_{(1)}$ shows the highest value of IQR and thus there is more evidence of a weak relation among the points in it, unlike $\mathbf{c}_{(2)}$, where the IQR is smaller. Therefore the points are more cohesive and convergence of all clusters. Table 4 shows that IQR in all the clusters is convergent also, and the outlier values have been ignored in $\mathbf{c}_{(2)}$ & $\mathbf{c}_{(3)}$ which are coloured in red.

Table 4. ANOVA graph analysis of experiment 1

c	Range	Median	25%ile	75%ile	IQR
1	2.828	4.123	2.236	4.183	0.974
2	2.236	3.162	3.000	4.123	0.562
3	3.058	3.606	3.000	4.000	0.500
4	3.123	2.236	1.414	3.162	0.874
5	3.472	3.000	2.236	3.606	0.685
6	3.472	2.828	2.000	3.606	0.803
7	3.472	3.162	2.236	4.123	0.944

4.1. Experiment 2: Unreal Data, 2D

This case of a clustered dataset has good **A.F** (0.110), but its (**M.F** = 0.949) indicates a significant possible merging. Table 6 and Figure 5 shows that $\mathbf{c}_{(1)}$ and $\mathbf{c}_{(5)}$ are close enough to each other to be merged; they both have $\mathbf{ESF}_{(c.kn)}$ = 0.049 and $\mathbf{ISF}_{(c.k)}$ = 0.949.

Table 5. Cluster distance matrix experiment 2

c	1	2	3	4	5	6
1	NILL	42.43	54.71	27.66	3.16	37.11
2	42.43	NILL	17.12	14.42	27.29	27.86
3	54.71	17.12	NILL	16.12	40.22	9.43
4	27.66	14.42	16.12	NILL	14.76	15.03
5	3.16	27.29	40.22	14.76	NILL	29.00
6	37.11	27.86	9.43	15.03	29.00	NILL

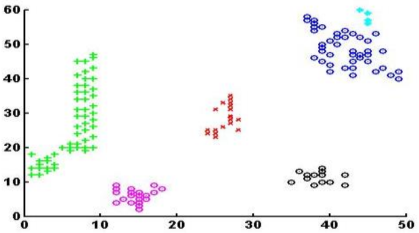

Figure 5. Result graph of experiment 2

Table 6. Clusters effectiveness for experiment 2

c	Weight	$\mathbf{ESF}_{(c.k\text{–}oth)}$	$\mathbf{ESF}_{(c.k\text{–}n)}$	$\mathbf{ISF}_{(c.k)}$
1	0.023	0.509	0.049	0.949
2	0.103	0.398	0.222	0.208
3	0.120	0.424	0.145	0.318
4	0.126	0.271	0.222	0.208
5	0.269	0.353	0.049	0.949
6	0.360	0.365	0.145	0.318

Table 7. Cluster distance matrix experiment 3

1	1	2	3	4
1	NILL	1.747	0.759	5.065
2	1.747	NILL	0.750	5.059
3	0.759	0.750	NILL	5.000
4	5.065	5.059	5.000	NILL

4.2. Experiment 3: Unreal Data, 3D

In Table 8, we see that $\mathbf{c}_{(1)}$, $\mathbf{c}_{(2)}$, and $\mathbf{c}_{(3)}$ are much weaker than $\mathbf{c}_{(4)}$, also shown in Figure 6.

Table 8. Clusters effectiveness for experiment 3

C	Weight	$\mathbf{ESF}_{(c.k-oth)}$	$\mathbf{ESF}_{(c.k-n)}$	$\mathbf{ISF}_{(c.k)}$
1	0.070	0.393	0.118	0.430
2	0.093	0.392	0.117	0.341
3	0.419	0.338	0.117	0.436
4	0.419	0.784	0.778	0.065

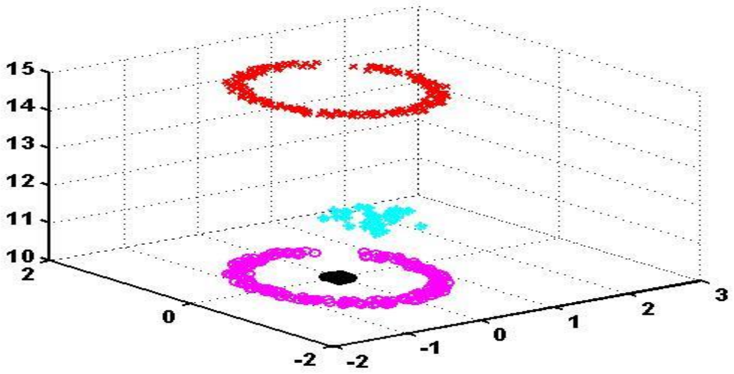

Figure 6. Result graph of experiment 3

4.3. Experiment 4: Real Data, 2D

The dataset contains two variables: Air pollution and Renewable energies in 30 European countries in 9 years from 2006 to 2014. The data were collected from the European Economic Association (http://ec.europa.eu/eurostat/data/database). We note that **A.F** in this case is good, but MF shows a significant possibility for merge. Specifically in Figure 7 and Table 10, $\mathbf{ESF}_{(c.k-n)}$ for $\mathbf{c}_{(2)}$ and $\mathbf{c}_{(5)}$ show how close the two clusters are to each other, and ($\mathbf{ISF}_{(c.5)}$ = 0.961) indicates an internal weakness relative to the distance to the nearest cluster.

Table 9. Cluster distance matrix experiment 4

C	1	2	3	4	5
1	NILL	72152	75076	16127	81989
2	72152	NILL	18819	13517	7092
3	75076	18819	NILL	22931	11679
4	16127	13517	22931	NILL	24047
5	81989	7092	11679	24047	NILL

Table 10. Clusters effectiveness for experiment 4

c	Weight	$\mathbf{ESF}_{(c.k-oth)}$	$\mathbf{ESF}_{(c.k-n)}$	$\mathbf{ISF}_{(c.k)}$
1	0.033	0.507	0.133	0.395
2	0.033	0.231	0.059	0.532
3	0.033	0.266	0.097	0.546
4	0.133	0.158	0.112	0.504
5	0.767	0.258	0.059	0.961

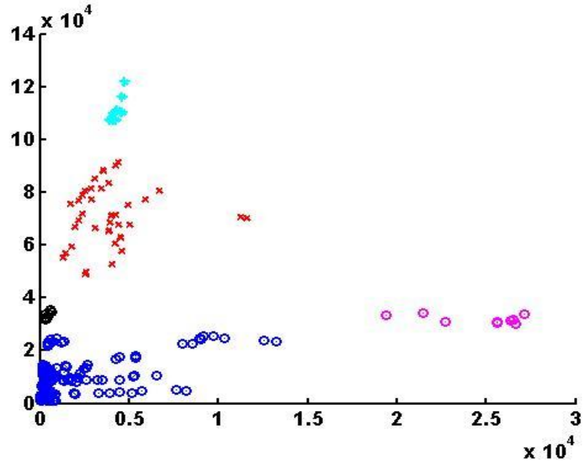

Figure 7. Result graph of experiment 4

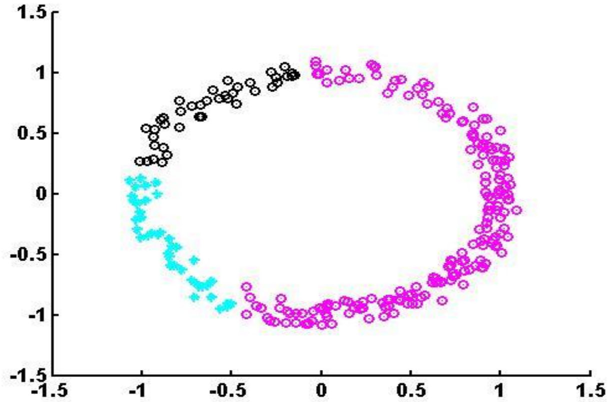

Figure 8. ANOVA result for experiment 4

Table 11 clearly shows that the second cluster is coherent and interrelated where (IQR=609), while the third one is less (IQR=2022).

Table 11. ANOVA graph analysis of experiment 4

C	Range	Median	25%ile	75%ile	IQR
1	6269	2466	1889	3768	940
2	3690	1493	1131	2349	609
3	6482	3710	1485	5528	2022
4	6458	4204	2985	5322	1169
5	6809	3113	1447	5030	1792

4.4. Experiment 5: Unreal Data, 2D

This case is quite deferent. **A.F** is relatively high (0.779) and close to *one*, which means the dataset does not fit well to the concepts of S.C.A, while **M.F** (0.943) shows a significant possibility for merging. The three other factors also illustrate the weakness of all of the clusters. So we can say that the dataset in this case cannot be clustered strongly.

Table 12. Cluster distance matrix experiment 5

C	1	2	3
1	NILL	0.145	0.12
2	0.145	NILL	0.126
3	0.12	0.126	NILL

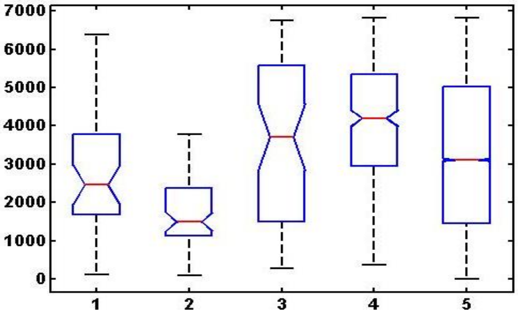

Figure 9. Result graph of experiment 5

Table 13. Cluster effectiveness for experiment 5

c	Weight	**ESF**$_{(c.k-oth)}$	**ESF**$_{(c.k-n)}$	**ISF**$_{(c.k)}$
1	0.148	0.059	0.054	0.934
2	0.148	0.061	0.056	0.901
3	0.704	0.055	0.054	0.934

5. Discussion

Before turning to our comparative empirical of two techniques of relative relations or ANOVA for distance, we first present a theoretical analysis evaluation for them to measure the effectiveness of applying S.C.A by Kuwil method. In this paper we have *five* practical experiments which are conducted on a few types of data, and dozens of experiments that cannot be mentioned in this paper.

First, we start with theoretical analysis evaluation: Adoption of the laws used to find the effectiveness of definition principles of the algorithm, with the use of the law of distance between two points and the relative relations and the use of statistical laws such as ANOVA and quartile deviation, whose results are not completely accurate and are not affected by extreme values. All the results were acceptable and logical for all the experiments with all types of data like real, unreal, 2D, 3D and different numbers of clusters. The data generated by the user are controlled in accordance with the nature of the algorithm, but the real data can accept application of the algorithm completely or partly, so we need to measure the effectiveness in the acceptance cases, which facilitates the user or researcher, whether statistical or financial, to determine that the results are acceptable according to the nature of the data under study. All this in addition to measuring the factors of merge between clusters and giving indicators to the user to decide what fits the nature of the data under study. It has become easier to apply S.C.A on more than 3D, as the evaluation of the results does not depend on the graph only but also on the mathematical and statistical measurements.

Secondly, comparative empirical: Table 14 shows the *five* cases with the most important factors **A.F** and **M.F**. Figure 10 shows that experiment 3 is the best experiment in terms of accepting the implementation of S.C.A and the less acceptable is experiment 5. The greater need to merge two clusters or more are represented in experiments 2 and 4 and the lowest is required in experiments 1 and 3.

Table 14. Comparing clusters with A.F–M.F

Experiments	No. of clusters	**A.F**	**M.F**
case 1	7	0.158	0.555
case 2	6	0.110	0.949
case 3	4	0.065	0.436
case 4	5	0.283	0.961
case 5	3	0.943	0.778

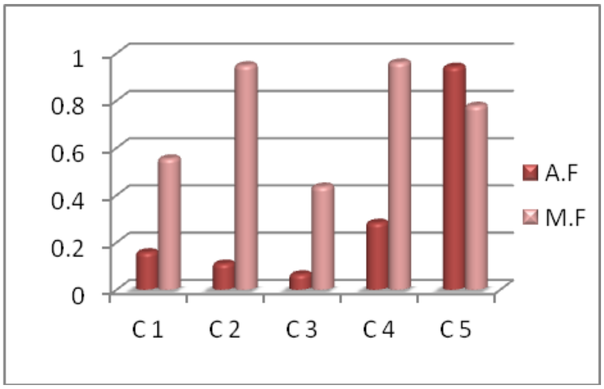

Figure 10. Comparing all results with A.F and M.F

6. Conclusion and Future Work

As we saw, the outputs of S.C.A under Kuwil method can be tested by mathematical and statistical techniques, which make the algorithm more reliable and widespread. We can also use these techniques for testing another S.C.A under another method, or even test any algorithm used in data mining and artificial intelligence after modifying the techniques, because S.C.A depends on the closest distances among the points regardless of the total distance to connect all of them, while the other algorithms depend on another factor, such as the total distance for TSP algorithm or the central points for the K-mean algorithm. So we can say that the door has been opened for further studies for evaluating the performance of the various algorithms on different types of data as well as measuring the effectiveness. Due to the focus of the method on the detailed study of the nature of data and analysis of all relationships within the dataset, the process of implementation is fast and effective in small and medium data, but huge data, such as image and sound files, need to improve by using the technique of OpenMP of parallel programming.

References

[1] Bach, F. R., & Jordan, M. I. (2004). Learning spectral clustering. In *Advances in Neural Information Processing Systems* (pp. 305-312).

[2] Chen, W. Y., Song, Y., Bai, H., Lin, C. J., & Chang, E. Y. (2011). Parallel spectral clustering in distributed systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 33*(3), 568-586.

[3] Gao, F. (2011). *Distributed approximate spectral clustering for large-scale datasets* (Unpublished Master Thesis). Applied Science: School of Computing Science, Simon Fraser University, Burnaby, BC, Canada.

[4] Judd, C. M., McClelland, G. H., & Ryan, C. S. (2017). *Data analysis: A model comparison approach to regression, ANOVA, and beyond.* US: Routledge.

[5] Meila, M. (2001). The multicut lemma. *UW Statistics Technical Report, 1,* 417-422.

[6] Meila, M. (2005). Regularized spectral learning. *UW Statistics Technical Report, 1,* 465-475.

[7] Meila, M., & Shi, J. (2001). A random walks view of spectral segmentation. In *10th International Workshop on Artificial Intelligence and Statistics* (pp. 101-110).

[8] Ng, A. Y., Jordan, M. I., & Weiss, Y. (2002). On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems* (pp. 849-856).

[9] Polito, M., & Perona, P. (2002). Grouping and dimensionality reduction by locally linear embedding. In *Advances in Neural Information Processing Systems* (pp. 1255-1262).

[10] Shi, J., & Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 22*(8), 888-905.

[11] Stella, X. Y., & Shi, J. (2003). Multiclass spectral clustering. In *International Conference on Computer Vision* (pp. 71-77).