Global Journal of
Computer Sciences:
Theory and Research

Volume 07, Issue 2, (2017) 102-111

www.gjcs.eu

Kuwil method for spectral clustering algorithm

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

Suggested Citation:
Kuwil, F. H. A. (2017). Kuwil method for spectral clustering algorithm. *Global Journal of Computer Sciences: Theory and Research. 7*(2), 102-111.

Received April 24, 2017; revised June 21, 2017; accepted August 29, 2017.
Selection and peer review under responsibility of Prof. Dr. Dogan Ibrahim, Near East University, North Cyprus.

Abstract

The open issues and challenges that exist while using the spectral clustering algorithm (SCA) have led to its limited spread in practical life. This paper proposes to find an easier, faster and more accurate method to implement SCA that will lead to its wide use by statisticians, researchers, institutions and others. I suggest a new method called 'Kuwil method' for SCA on any dataset points without needing estimation or evaluation of any parameters or the use of linear algebra, not even the k_mean algorithm. The main aim is to apply an algorithm that relies on distance laws among points only. The algorithm by the Kuwil method has been applied a number of times on real data from the warehouse European Economic Association (http://ec.europa.eu/eurostat/data/database) and on unreal data. The results were highly efficient in terms of time, effort and simplification. It eliminates the problem of parameters and increases the effectiveness to give static results obtained from the first execution. No errors were seen from functions in the MATLAB language such as eigenvalues, eigenvector and *k*_mean.

Keywords: Spectral clustering, 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

Despite the importance of algorithms in many aspects of computer science such as data mining, artificial intelligence and machine learning, it is still not prevalent among financiers, researchers and statistical centres; also, most of the applications, except for image segmentation, are limited to unreal data. A lot of research and studies have been done on spectral clustering, in regard to examining the number of clustering, which shows that the multiplicity of eigenvalue one is equal to the number of clusters (this was followed to some extent by Polito and Perona in [1]). In [6], it is shown that if some conditions apply, then spectral clustering minimizes the multi-method normalized cut, a generalisation of the two-way normalized cut criterion [4], random walks [3], graph cuts and normalized cuts [4] and matrix disorder theory [2], simplifying the difficulties to make them easier to understand concurrently with the development of algorithms [2–5, 8]. Significant theoretical progress has also been done. Yu and Shi [7] proposed to swap normalized eigenvectors to get optimal segmentation. In [9], random projection tree is used. Dhillon *et al.* [10] used *k*-nearest-neighbour classifiers. I spent a long time trying to solve the problems and reducing the challenges in this algorithm, but unfortunately could not complete it. Hence, the common denominator among all previous studies and this paper is the implementation of spectral clustering algorithm (SCA) and the difference between this and the method and techniques. Therefore, it was necessary to find another simple and easy method for those who want to use this algorithm. What is needed to find the interface between the algorithm and the users by creation of a system that includes the application of optimisation and hide the complexities from users? In order to bring out the algorithms from the engineers' workshops to practical life, in this paper, we propose a general framework of the system of algorithms to be compiled in one application such as SCA, TSP and KNN.

2. Kuwil Method

According to the definition of SCA, it studies the relation among the data points themselves, then divides them into some clusters where the data in the same cluster must be connected, coherent and close to each other, while the data in different clusters are unconnected and dissimilar (Table 1). This means the relation measured by the distance among points for which we do not need any techniques or laws, except the distance law between points in coordinate or space or any dimension [11].

Input dataset $p_i = \{ p_1, p_2, ..., p_n \}$, *n* is the number of points.

Output $c_1 = \{ p_1, p_2, ..., p_{C1} \}, ..., c_k = \{ p_1, p_2, ..., p_{ck} \}$, *c*1, *c*2,..., *c*k are the clusters.

Table 1. Distance Matrix Form in SCA

0	D(p1,p2)	D(p1,p3)	→	D(p1,pn)
D(p2,p1)	0	D(p2,p3)	→	D(p2,pn)
D(p3,p1)	D(p3,p2)	0	→	D(p3,pn)
↓	↓	↓		↓
D(pn,p1)	D(pn,p2)	D(pn,p3)	→	0

3. The Algorithm

- Given data in 2D, read it in matrix **A** from external file where $p1, p2, .., pn$ are the positions of points in matrix **A**

Matrix A = $\begin{array}{ccc} p_1 & x_1 & y_1 \\ p_2 & x_2 & y_2 \\ . & . & . \\ p_n & x_n & y_n \end{array}$

- Find **Dis** matrix

$$\mathbf{Dis}_{pi,pj} = \sqrt{\left(x_i - x_j\right)^2 + \left(y_i - y_j\right)^2} .$$

- Study the correlation relation (CR) in every row between every point and all dataset points.
- We get a CR matrix(*n*,3) where it contains the positions of data points and their distances

CR matrix = $\begin{array}{ccc} p_1 & p_j & dis1 \\ p_2 & p_j & dis2 \\ . & . & . \\ p_{n-1} & p_j & dis_{n-1} \end{array}$, **j** is not sequence.

- From CR matrix, find a critical distance where $\lambda_{(min)}$ is a factor which is the required distance to create a perfect maximum number of clusters.
- Connect every two points in Dis matrix where **D(** pi, pj **)** $\le \lambda_{(min)}$**,** and put the results in a new matrix called Pos matrix

Pos = $\begin{array}{ccc} p_1 & p_j & dis1 \\ p_i & p_j & dis2 \\ . & . & . \\ p_m & p_j & dis_m \end{array}$, *m* > *n* , *i* , *j* are repeated

- Pos matrix is similar to stat transaction table in automata theory, but without conditions (stat), we use the same technique to follow and separate every cluster in Pos matrix.

4. Analysis of Methods

4.1. Three Important Coefficients

The algorithm will give the perfect maximum number of clusters with three coefficients which help the user to control the results as follows:

- $\lambda_{(\min)}$ is the critical distance to apply **SCA**
- $\lambda_{(\text{mer})}$ is the required distance to merge clusters
- $\lambda_{(\max)}$ is the required distance to prevent application of **SCA**.

4.2. Dealing with Outlier Values in Data

One of the important measures of dispersion in the statistics is the interquartile range (IQR), which is characterized by the standard deviation (SD) but not influenced by outlier values, although it is less accurate. So when you find the critical distance $\lambda_{(\min)}$, we should exclude outlier values just in case of more than two outlier distances, as explained in Section 5.3. This is because the results are negatively affected by them where the balance of distances is broken. Outlier is an extremely high or low value in our data, so we can identify an outlier in two cases as follows:

- $\boldsymbol{Q}1-1.5*\left(\boldsymbol{IQR}\right)$ lower limit for outlier
- $\boldsymbol{Q}3+1.5*(\boldsymbol{IQR})$ higher limit for outlier.

So we will ignore all points where

- $\boldsymbol{Q}3+1.5*\left(\boldsymbol{IQR}\right) <$ data points $< \boldsymbol{Q}1-1.5*\left(\boldsymbol{IQR}\right)$.

So that:
Q1 = ($25\%\boldsymbol{ile}$): IQR of the 25th percentile. Q3 = ($75\%\boldsymbol{ile}$): IQR of the 75th percentile.
$\boldsymbol{IQR}=(\boldsymbol{Q}3-\boldsymbol{Q}1)$.

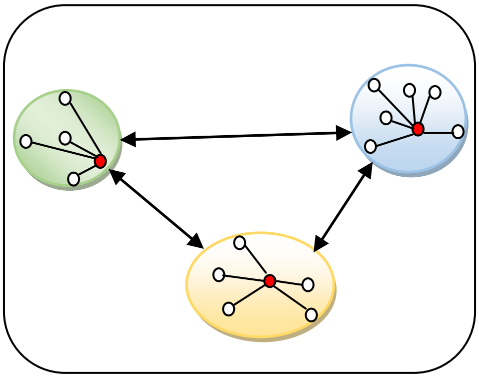

Figure 1. The Kuwil method for SCA

4.3. Characteristics of the Method

Figure 1 shows the method summary where the distance between every two points in all datasets inside every cluster is $\leq \lambda_{(\min)}$. In addition, it shows the distance required to connect any cluster with others. Also, there are some conditions for its implementation: Every cluster should contain at least three points and ignore the outlier distances, as we mentioned earlier, which represents the basic concepts for SCA. The most important advantage is that we do not need any parameters to implement it. The results are fixed from the first implementation, and we use only the distance law between the points and, therefore, do not need any functions or additional techniques. This can sometimes cause some errors in the MATLAB compiler, due to outputs of eigenvalue and eigenvector which contain complex numbers in some cases. Two or more clusters may be combined according to the coefficients obtained with the results.

5. Analysis of Methods

The method has been applied on a lot of real and unreal data.

5.1. For Real Data

Some technologies have been used to coordinate and standardize data from warehouse European Economic Association database—knowledge discovery from data (KDD), so it is necessary to implement any algorithm in data mining for executing some KDD processing to provide and prepare a dataset appropriately. The following are a list of the most important operations: data selection, data cleaning, data transformation, data integration and data mining. After preparing the data by applying the previous operations, the data will be ready to implement the SCA by the Kuwil method.

Table 2. Dataset for Experiment 1

Ser	Air pollution	Renewable energy
1	1335.99	1364.60
2	7651.31	1140.50
↓	↓	↓
270	21474.9	12080.50

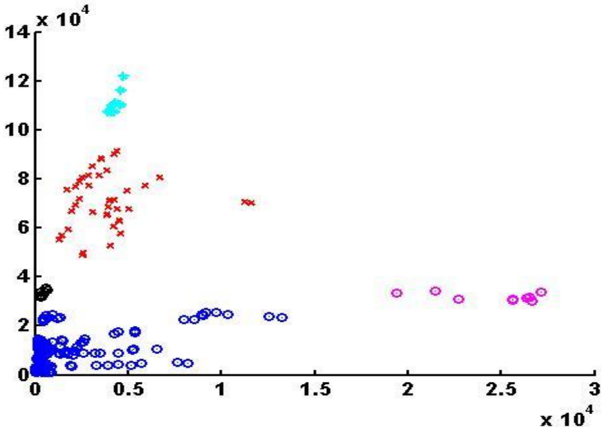

Figure 2. Result of Experiment 1

Table 2 represents the dataset that contains air pollution and renewable energies in 30 European countries in 9 years from 2006 to 2014. Figure 2 shows five clusters, each of them with a different colour. The data are close and coherent between them and far from the rest of the clusters. However, the graph does not appear to be precise, because of the scale; the distance of the vertical axis is greater than the distance in the horizontal axis. Table 3 represents the dataset for two variables. The first study classifies the people at risk of poverty or social exclusion (percentage), while the second study focuses on the population tertiary education attainment level. Both the studies were conducted (percentage) between 2008 and 2013 in 31 European countries. Figure 3 shows the results with three clusters. The nature of the data makes it impossible to increase the number of clusters where the concepts of SCA are verified.

Table 3. Dataset for Experiment 2

Ser	Risk of poverty	Education attainment
1	20.80	28.40
2	44.80	18.90
↓	↓	↓
186	16.30	33.20

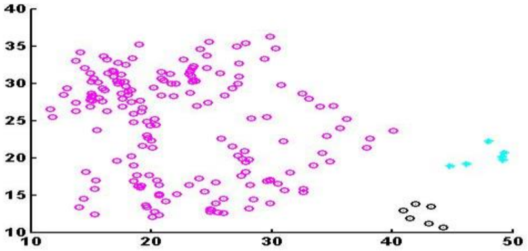

Figure 3. Result of Experiment 2

5.2. Unreal Data

The data are obtained by generation, so there are no procedures or techniques needed as in the treatment of real data. Here, we have six experiments, two in 3D and the others in 2D. The results are shown in Figure 4.

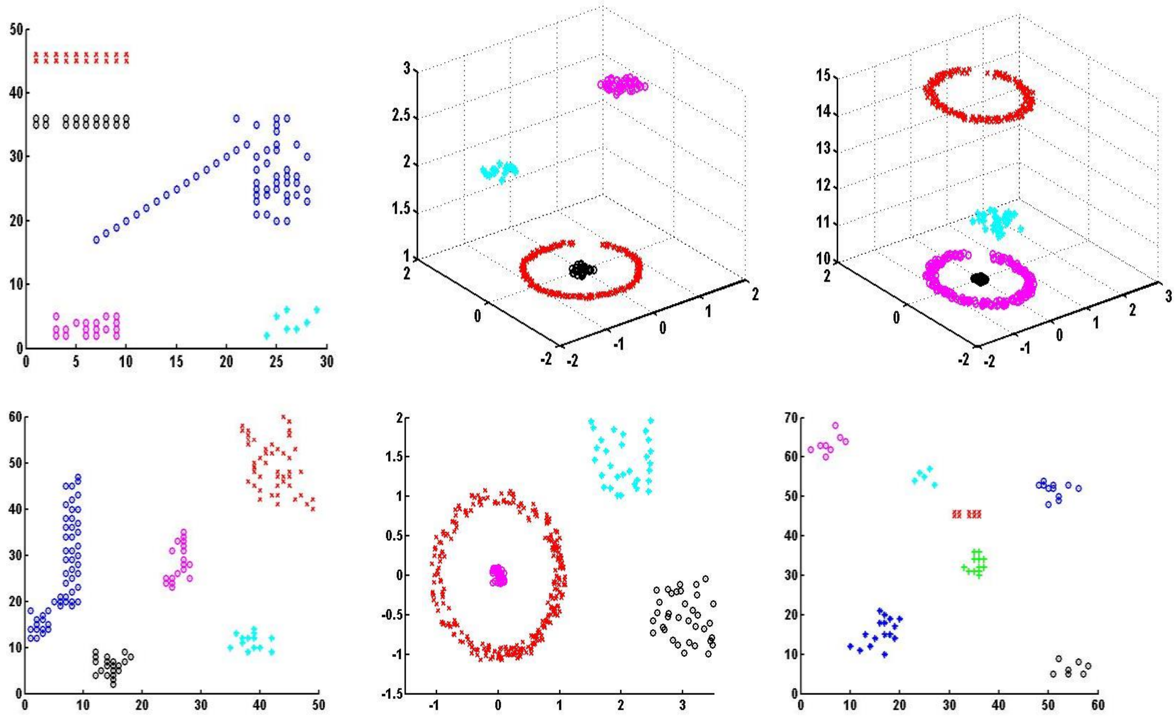

Figure 4. Results of six experiments unread data

5.3. Outlier Values

In this experiment, we will illustrate how the algorithm deals with outlier values. We will apply the algorithm to Experiment 3 after adding a few outlier points as follows:

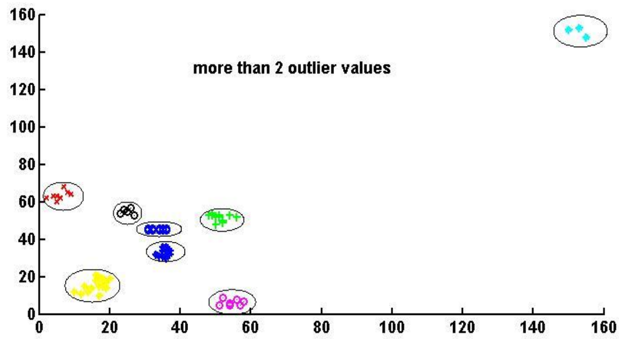

Figure 5. More than two outlier values

- With more than two outlier points (three points in this example), the algorithm deals with the situations as shown in Figure 5.
- With two outlier points and execution, and some processing, the result is as shown in Figure 6 and the outlier points always combine to the nearest cluster.
- With two outlier points, where one of the important conditions for applying SCA with the Kuwil method, the result without dealing with outlier will be all datasets in the same cluster as in Figure 7.

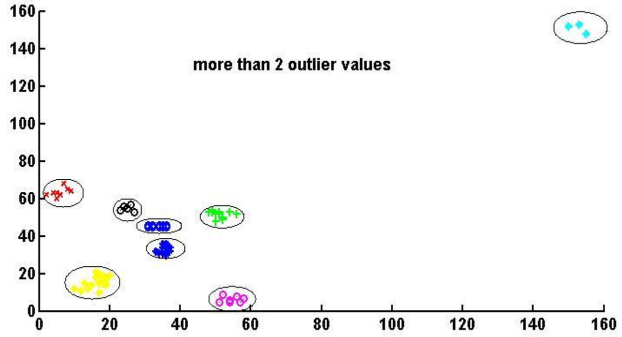

Figure 6. Outlier values with deal

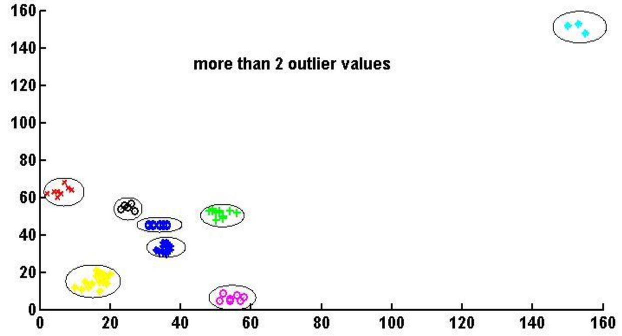

Figure 7. Outlier values without deal

6. Discussion

These were samples of the total number of experiments that the algorithm was implemented using the Kuwil method. This method provides its effectiveness with all types of real and generated data, and with two or more dimensional data. All the results were correct and compatible with the algorithm concepts. The number of clusters cannot be increased with acceptable effectiveness, but we can decrease them by merging two or more clusters by $\boldsymbol{\lambda}_{\mathbf{mer}}$. We can represent it as follows:

- $(0 < \mathrm{dis} < \lambda_{\mathrm{min}}]$ it is possible to get a number of clusters, but it does not necessarily satisfy the algorithm concepts completely.
- $[\lambda_{\mathrm{min}} \leq \mathrm{dis} < \lambda_{\mathrm{mer}}]$ permissible period for perfect applying.
- $[\lambda_{\mathrm{mer}} \leq \mathrm{dis} \ \ \lambda_{\mathrm{max}}]$ permissible period for merge clusters.
- $[\lambda_{\mathrm{max}} \leq \mathrm{dis} < \infty)$ impossible period to apply S.C.A.

7. Conclusion

In this paper, the most important characteristics of this method are based on the fact that they rely on the laws of distance between two points according to the concept of algorithm. In addition to the relative relations and solving the problems of extreme values through the science of statistics, it gives accurate coefficients and helps the user to control outputs through increasing or decreasing the number of clusters. This is done according to the nature of the data and the study, instead of requesting for finding and estimating of the implementation parameters. Using the method and its results, it is possible to say that the problems and challenges in applying the SCA, are now solved, thus opening the door for those who wants to be benefitted from this algorithm in practical life. Efficiency has increased, with less time and effort in implementation. In other words, access to the results was in a shorter and easier way. However, there is another challenge facing users: To what extent can the data be applied using SCA, as well as the measurement of the performance results. There are many cases where it has been applied, but the strength of the clusters is weak; in other cases, it was strong based on the proportion of distance differences inside and outside the clusters. In addition, it can measure the performance in the case of more than three dimensions of data.

References

[1] M. Polito and P. Perona, "Grouping and dimensionality reduction by locally linear embedding," in: *Advances in Neural Information Processing Systems*, vol. 14, 2002.

[2] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: analysis and an algorithm," in: *Advances in Neural Information Processing Systems (NIPS)*, vol. 14, 2001.

[3] M. Meila and J. Shi, "A random walks view of spectral segmentation," in: *10th International Workshop on Artificial Intelligence and Statistics* (AISTATS), 2001.

[4] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, issue 8, pp. 888–905, 2000.

[5] F. R. Bach and M. I. Jordan, "Learning spectral clustering," in: *Advances in Neural Information Processing Systems (NIPS)*, vol. 16, 2004.

[6] M. Meila, "The multicut lemma," *UW Statistics Technical Report 417*, 2001.

[7] S. X. Yu and J. Shi, "Multiclass spectral clustering," in: *International Conference on Computer Vision*, Nice, France, October 2003, pp. 11–17.

[8] M. Meila, S. Shortreed, and L. Xu, "Regularized spectral learning," *UW Statistics Technical Report 465*, 2005.

[4] F. R. Bach and M. I. Jordan, "Learning spectral clustering, with application to speech separation," *J. Mach. Learn. Res.*, vol. 7, pp. 1963–2001, 2006.

[9] S. Dasgupta and Y. Freund, "Random projection trees and low-dimensional manifolds," in: *40th ACM Symposium on Theory of Computing* (STOC), 2008.

[10] I. Dhillon, Y. Guan, and B. Kulis, “Weighted graph cuts without eigenvectors: a multilevel approach,” *IEEE Trans. PAMI*, vol. 29, issue 11, pp. 1944–1957, 2007.

[11] Delen, D., *Real-World data mining*. New Jersey: Pearson Education, Inc., 2015.