ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8.
Available from http://www.i6doc.com/en/.

Performance assessment of quantum clustering in non-spherical data distributions

Raúl V. Casaña-Eslava[1], José D. Martín-Guerrero[2], Ian H. Jarman[1] and Paulo J. G. Lisboa[1]

1- School of Computing and Mathematical Sciences, Liverpool John Moores University, United Kingdom
2- Department of Electronic Engineering, University of Valencia, Spain

Abstract. This work deals with the performance of Quantum Clustering (QC) when applied to non-spherically distributed data sets; in particular, QC outperforms K-Means when applied to a data set that contains information of different olive oil areas. The Jaccard score can be set depending on QC parameters; this enables to find local maxima by tuning QC parameters, thus showing up the underlying data structure. In conclusion, QC appears as a promising solution to deal with non-spherical data distributions; however, some improvements are still needed, for example, in order to find out a way to detect the appropriate number of clusters for a given data set.

1 Introduction.

K-means is the most known and widely-used clustering algorithm; however, it has a number of problems, being two of the most important ones the fact that the number of clusters is not automatically selected and its difficulty to cluster properly when the dataset is not spherically distributed. Numerous works have been carried out in order to face the former problem; for instance, [1] and [2] make use of Cramérs' V statistic as stability measure to produce the Separation Concordance (SeCo) map, and then using the Area Under this Curve as metric to obtain the most consistent values of K. Nevertheless, the latter problem is difficult to be solved because of k-means design itself. In this framework, Quantum Clustering (QC) appears as a promising solution due to its ability to work well with data non-spherically distributed data..

The QC was introduced in [3] using the Schrodinger equation on probability wave function formed as a superposition of N Gaussian probability functions (1), where there are N data points of dimension d. Then, looking for solutions of the harmonic oscillator potential in ground energy eigenstate, (2 – 4), those centroids in which the potential has a local minima can be found. From the wave function in (1) the potential function $V(x)$ obtains the σ parameter; more minima appear in $V(x)$ as σ is decreased. Tuning σ can also be used for the estimation of the appropriate number of clusters.

$$\psi(x) = \sum_i^N e^{-\frac{(x-x_i)^2}{2\sigma^2}} \qquad (1)$$

ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8.
Available from http://www.i6doc.com/en/.

$$H\psi \equiv \left(-\frac{\sigma^2}{2}\nabla^2 + V(x)\right)\psi = E\psi \quad (2)$$

$$V(x) = E + \frac{\frac{\sigma^2}{2}\nabla^2\psi}{\psi} \quad (3)$$

$$E = -min\frac{\frac{\sigma^2}{2}\nabla^2\psi}{\psi} \quad (4)$$

QC has already been tested in [4]; this work made use of Single Value Decomposition (SVD) as a preprocessing step before the QC algorithm. Three known datasets of cells and genes were tested obtaining good results when dimensions were truncated to the 4th – 5th principal components before the application of QC; the corresponding Jaccard scores outperformed those achieved by k-means.

The interface called Comparative-Package-for-Clustering-Assessment (COMPACT) [5] was used to obtain the results shown in this paper. COMPACT implements several clustering algorithms and has the option of reducing the dataset´s dimensionality using SVD. The Jaccard score is used to evaluate the clusters obtained compared with the known outputs.

2 The olive oil data set

The known olive oil dataset [6] has been chosen because it presents a non-spherical distribution, and hence, it is suitable to evaluate QC performance compared to k-means. The data set presents two types of underlying structure (3 regions and 9 sub-regions) thus making the choice of the number of clusters challenging.

The olive oil dataset, consists of 572 observations with 8 characteristics, related to the fatty acid content of olive oil. This data corresponds to 3 collection regions, and 9 sub-regions.; four from Southern Italy (North and South Apulia, Calabria and Sicily), three from Umbria (Umbria, East and West Liguria) and two from Sardinia (Inland and Coastal regions).
The projected visualization of the underlying dataset is shown in Figure 1, where each data point is labelled according to the region from which it was obtained [7]; the overlapping of the data from Calabria, North and South Apulia and Sicily is remarkable.

3 QC setting-up

As previously mentioned, COMPACT was used as interface to evaluate the QC performance. There are some parameters related to preprocessing (SVD, normalization, component reduction, etc.) … and others directly related to the QC algorithm; among the latter, the most important parameter is σ, although there are more algorithm

ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8.
Available from http://www.i6doc.com/en/.

parameters that can be tuned (number of steps, rescale, QC core, % pure terms and learning rate η).

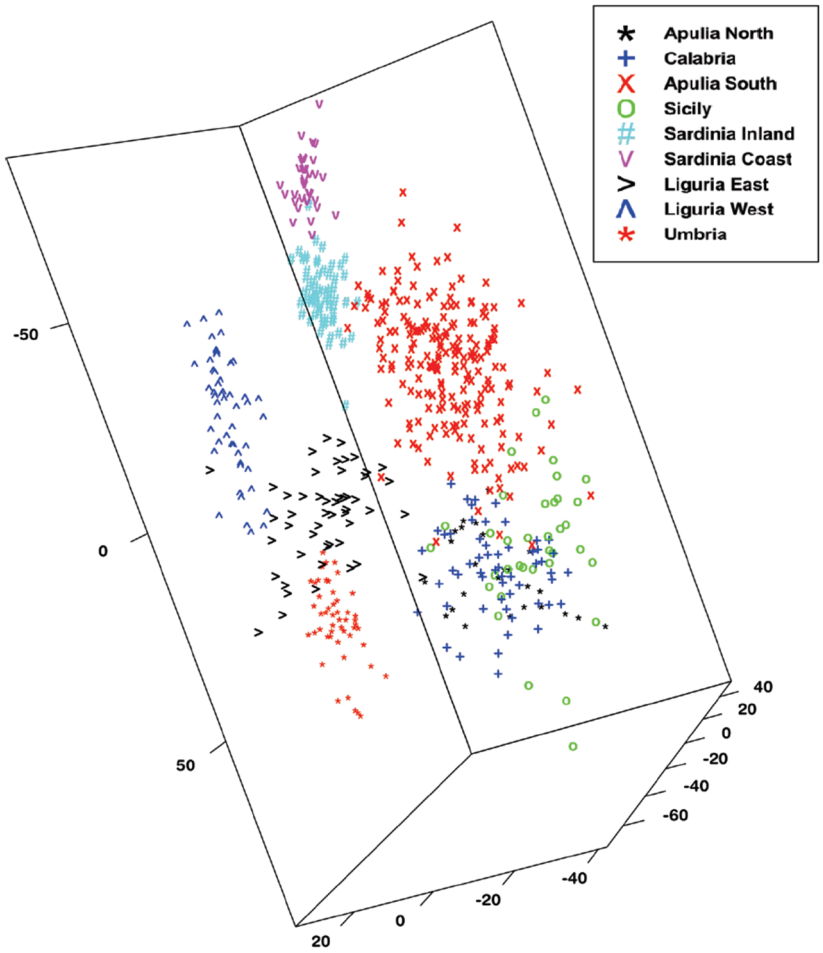

Fig. 1: Visualization of the 3 main principal components of olive oil data.

An evaluation performance based on the Jaccard score allowed to draw a number of conclusions:

- Normalization and SVD is needed in order to avoid a high number of clusters.
- When QC core is applied, the option of % of pure elements just tends to remove observations with an outlier behavior (the performance changes because the number of observations decreases).
- The optimal value of the learning rate is $\eta = 0.1$; other values of η might improve the performance but involving an unstable range of σ to assign a reasonable number of clusters.

ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8.
Available from http://www.i6doc.com/en/.

Finally, the best performance was obtained with the following combination of parameters:

- SVD pre-processing is enabled
- Normalization is applied
- QC core is not activated
- The σ belongs to [0.4, 0.6] producing 10 to 2 clusters.
- The η is 0.1
- Number of steps: 100

4 Data structure

Since the actual output of the olive oil data set is known (classes and sub-classes), it is possible to assess and analyze the performance of QC. Although one the main advantages of the QC is that the underlying data structure can be found by varying the parameter σ. Figure 2 shows that QC does not find the correct number of clusters when it obtains the best performance results, that are highlighted in black; in particular the value of σ that provides the best performance in the case of three-cluster problem leads to four clusters, while QC finds eight clusters when it achieves the best performance in the nine-cluster problem.

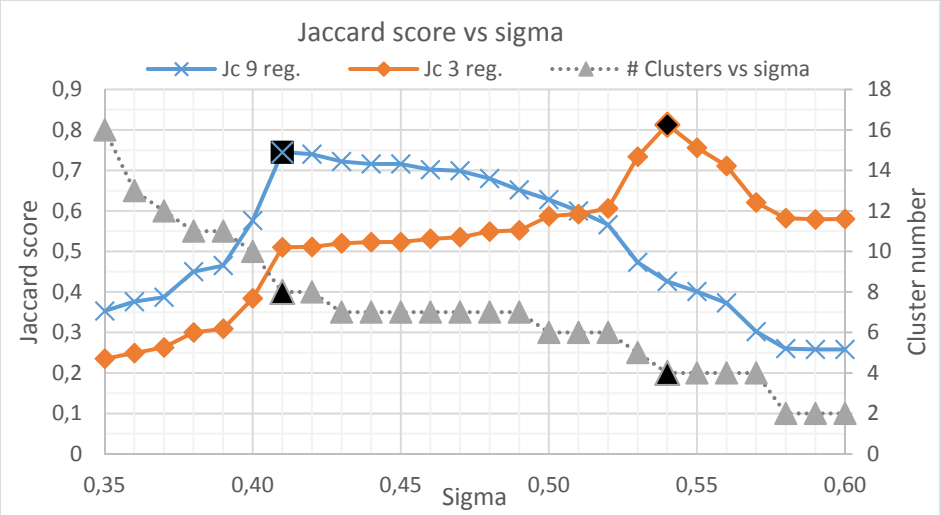

Fig. 2: Performance as σ function. Left axis shows Jaccard score and right axis shows the cluster number. Best Jaccard score results alongside the number of clusters are highlighted in black.

5 QC *vs* K-means performance

This section benchmarks QC performance versus K-Means, for the two classifications of the data set, namely, three and nine clusters. It must be emphasized that an additional

ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8.
Available from http://www.i6doc.com/en/.

advantage of QC with respect to K-Means is that QC obtains the same solution every time the algorithm is run for a particular σ value whereas K-means may provide different solutions in different runs since it is strongly depends on the initial conditions; to circumvent that bias, K-means was run 500 times to estimate the lowest SSE for each cluster number, [2, 12], then the Jaccard score has been obtained for the 3 and 9 regions, bearing in mind the lowest SSE doesn't imply the highest Jaccard score.

The results presented in Fig. 3 show that although QC nor K-Means find the correct number of clusters in either of the two problems, QC performance is considerably better than that achieved by K-Means, even taking into account that the right number of clusters is provided to K-Means. The best solutions for 3 regions are 4 clusters in both cases, with the QC Jaccard higher than K-Means. The best matching results for 9 sub-regions with K-Means is 6 clusters and QC 8 clusters, again QC scoring slightly higher (0.74 vs 0.72). The K-Means SSE decreases as K increases, as expected.

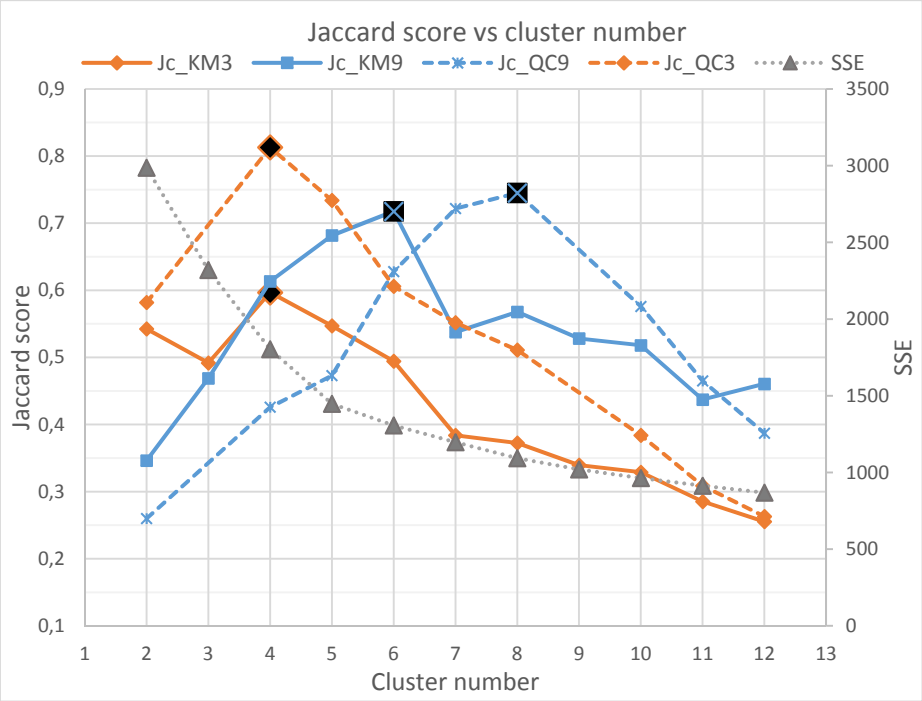

Fig. 3 Jaccard score of K-means and QC. Left axis represents Jaccard score and right axis represents SSE of K-Means. Orange lines refer to 3 cluster problem and blue ones to 9 cluster problem. Dashed line shows QC performance and the grey dot line refers to SSE of K-Means.

Table 1 depicts the best indices for both algorithms: Jaccard score, purity and efficiency (η).

Best Jaccard sc.	K-Means				QC			
Regions	Clust.	Jaccard	Purity	η	Clust.	Jaccard	Purity	η
3	4	0,597	0,619	0,943	4	0.813	0.905	0.889
9	6	0,718	0,911	0,772	8	0.745	0.794	0.924

Table 1: Jaccard score, purity and efficiency (η) are shown of clusters with the best Jaccard scores.

6 Conclusion

This work has proposed the use of QC to cluster non-spherical data distributions. QC outperforms the classical K-Means when applied to a data set containing information of different production regions of olive oil. Although, QC may not find the correct number of clusters, the performance measured in terms of Jaccard score, purity and efficiency is much better than that achieved by a K-Means that does know the number of clusters in advance.

Our ongoing and future research is related to the application of QC in more demanding environments in order to figure out its usefulness and range of application. And also, research related to the search for the correct σ for unsupervised data.

Acknowledgements:

This work has been partially supported by the Spanish Ministry of Economy and Competitiveness under project with reference number TIN2014-52033-R.

References

[1] P. J. G. Lisboa, T. A. Etchells, I. H. Jarman y S. J. Chambers, «Finding reproducible cluster partitions for the k-means algorithm,» *BMC Bioinformatics,* vol. 14, nº Suppl. 1, p. S8, 2013.

[2] S. J. Chambers, I. H. Jarman, T. A. Etchells and P. J. G. Lisboa, "Inference of number of prototypes with a framework approach to K-means clustering," *Int. J. Biomedical Engineering and Technology,* vol. 13, no. 4, pp. 323-340, 2013.

[3] D. Horn y A. Gottlieb, «Algorithm for Data Clustering in Pattern Recognition Problems Based on Quantum Mechanics,» *Physical Review Letters,* vol. 88, nº 1, p. 018702, 2002.

[4] D. Horn y I. Axel, «Novel clustering algorithm for microarray expression data in a truncated SVD space,» *Bioinformatics,* vol. 19, nº 9, pp. 1110-5, 2003.

[5] R. Varshavsky, M. Linial y D. Horn, COMPACT: A Comparative Package for Clustering Assessment, SpringerVerlag, 2005, pp. 159-167.

[6] M. Forina, C. Armanino, S. Lanteri y E. Tiscornia, «Classification of olive oils from their fatty acid composition,» de *Food Research and data Analysis, Applied Science Publishers*, London, In: Martens, M., Russwurm, H. Jr., 1983, pp. 189-214.

[7] S. J. Chambers, «A framework approach to initialisation dependent clustering.,» PhD Thesis, Liverpool John Moores University, 2015.

[8] N. Nasios y A. G. Bors, «Finding the Number of Clusters for Nonparametric Segmentation,» de *Computer Analysis of Images and Patterns*, Springer Berlin Heidelberg, 2005, pp. 213-221.