MAXIMUM LIKELIHOOD APPROACH TO PARAMETER SELECTION IN THE SPECTRAL CLUSTERING ALGORITHM

Cinzia Di Nuzzo[1], Salvatore Ingrassia[1]

[1] Department of Economics and Business, University of Catania, (e-mail: cinzia.dinuzzo@unict.it, salvatore.ingrassia@unict.it)

ABSTRACT: Automatic selection of the parameter in the spectral clustering algorithm through the mixture model approach has been considered. Specifically, a maximum likelihood approach using the Gaussian mixture model to select the proximity parameter in the self-tuning kernel function has been introduced.

KEYWORDS: Spectral clustering, parameters selection, gaussian mixture model.

1 Introduction

Spectral clustering methods are based on graph theory, where data are represented by the vertices of an undirected graph and the edges are weighted by the similarities between pairs of units, see von Luxburg, 2007, Shi & Malik, 2000, Ng *et al.* , 2001. Specifically, the spectral approach is based on the properties of the pairwise similarity matrix coming from a suitable kernel function. Then the clustering problem is reformulated as a graph partition problem.

Let $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\} \subseteq \mathbb{R}^p$ be a set of units. In order to cluster $\mathbf{X}$ in K clusters, the first step of the spectral clustering algorithm concerns the definition of a symmetric and continuous function $\kappa : \mathbf{X} \times \mathbf{X} \to [0, \infty)$ called kernel function. Afterwards, a similarity matrix $\mathbf{W} = (w_{ij})$ can be assigned by setting $w_{ij} = \kappa(\mathbf{x}_i, \mathbf{x}_j) \geq 0$, for $\mathbf{x}_i, \mathbf{x}_j \in \mathbf{X}$. Specifically, here, we consider the following *self-tuning* kernel function (see Zelnik-Manor & Perona, 2004)

$$\kappa(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{\varepsilon_i \varepsilon_j}\right), \qquad i, j = 1, \ldots, n, \tag{1}$$

with $\varepsilon_i = \|\mathbf{x}_i - \mathbf{x}_h\|$, where $\mathbf{x}_h$ is the h-th neighbour of point $\mathbf{x}_i$ (similarly for ε_j). Afterward, the normalized graph Laplacian is introduced as the $n \times n$ matrix $\mathbf{L}_{\text{sym}} = \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2}$, where $\mathbf{D} = \text{diag}(d_1, d_2, \ldots, d_n)$ is the *degree matrix*; d_i is the *degree* of the vertex $\mathbf{x}_i$ defined by $d_i = \sum_{j=1}^{n} w_{ij}$ and $\mathbf{I}$ denotes the $n \times n$ identity matrix. The spectral clustering algorithm works on the embedded space. Given K, let $\{\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_K\}$ be the eigenvectors corresponding to the K

smallest eigenvalues of $\mathbf{L}_{\text{sym}}$. Then the normalized Laplacian embedding is defined as the map $\Phi_{\boldsymbol{\Gamma}} : \{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \to \mathbb{R}^K$ given by $\Phi_{\boldsymbol{\Gamma}}(\mathbf{x}_i) = (\gamma_{1i}, \ldots, \gamma_{Ki})$, for $i = 1, \ldots, n$. Let $\mathbf{Y} = (\mathbf{y}'_1, \ldots, \mathbf{y}'_n)$ be the $n \times K$ matrix of the embedded data, where $\mathbf{y}_i = \Phi_{\boldsymbol{\Gamma}}(\mathbf{x}_i)$ for $i = 1, \ldots, n$. Finally, the embedded data $\mathbf{Y}$ are clustered according to some clustering procedure. Usually, this latter step is performed using the k-means algorithm, here, mixture models have been taken into account, since they are more robust approaches with respect to the choice of parameter of the spectral clustering algorithm, see Di Nuzzo & Ingrassia, 2022b for details.

As a matter of fact, in the spectral clustering algorithm, there are two free parameters to be tuned: the local scale parameter h in the kernel function (1) and the number of clusters K. Specifically, the kernel function plays an important role in the spectral clustering context because it affects the entire structure of the data. For this reason, the goal of many authors has been to find an automatic or heuristic way to select the kernel function with the corresponding scale parameter.

In this framework, given the number of clusters K, a proposal of an automatic method for parameter selection in the kernel function (1) via the Gaussian mixture model according to the maximum likelihood approach is introduced.

The rest is organized as follows: in Section 2 a maximum likelihood approach to select the parameter h in (1) is introduced; in order to confirm the validity of methodology, in Section 3 some numerical examples are shown.

2 Maximum likelihood approach to parameter selection

In this section, an automatic criterion to select the parameter h in the self-tuning kernel function (1) is introduced. Note that for the sake of simplicity, we introduce this approach by using the self-tuning kernel function (1), but it can be extended to other kernel functions proposed in the spectral clustering context, see e.g Zhang & Yu, 2011, John C.R., 2020, Park S., 2021.

The parameter h in (1) has a key role in pre-processing data because it affects the geometrical structure of the graph in terms of weight associated with any pairs of vertices in the graph. Specifically, in Di Nuzzo & Ingrassia, 2022a a graphical approach to select the parameters of the spectral clustering algorithm has been considered. The results in Di Nuzzo & Ingrassia, 2022a show that by analysing the graphic features of the embedded space and the number of the diagonal blocks of the similarity matrix $\mathbf{W}$, an optimal number of groups K can be easily selected. However, the choice of the parameter h isn't always easy to select. Therefore, without a criterion to address this problem, different values of h can be considered optimal choices.

More precisely, as h varies, we have different configurations of the data in the embedded space, so we select h such that the embedded data are fitted by a Gaussian mixture model as much as possible. Therefore, we don't apply the Gaussian mixture model for fitting a given data set, but we look for the parameter h such that the corresponding data set is fitted by the Gaussian mixture model as much as possible.

For this purpose, we analyse the maximum log-likelihood parameter estimates deriving from the Gaussian mixture model using the EM algorithm and set h according to the maximum log-likelihood. In other words, we fit a Gaussian mixture model (with a fixed number K of components), according to the maximum likelihood approach, to different data sets corresponding to different $h \in \mathcal{H}$, where $\mathcal{H} \subseteq \{1, \dots, n-1\}$ is the collection of possible parameters h considered in the numerical experiments. Then we get a set of maximum likelihood values $l_1, \dots, l_{|\mathcal{H}|}$ for each data set, and select h^* leading to the overall maximum likelihood value, i.e. $h^* = \text{argmax}_h l_h$. Our proposal is summarized in Algorithm 1.

Algorithm 1 Parameter selection h in (1)

1. $\forall h \in \mathcal{H}$, compute the spectral clustering algorithm where the last step is executed with Gaussian mixture model.
2. $\forall h \in \mathcal{H}$, compute the log-likelihood value using EM algorithm obtaining the log-likelihood set $\mathcal{L} = \left\{ l_1, \dots, l_{|\mathcal{H}|} \right\}$.
3. Select h according to the maximum log-likelihood value, i.e. h^* corresponds to $l^* = \max \mathcal{L}$.

3 Numerical examples

Numerical examples according to the proposed approach (Algorithm 1) are here presented.

Table 1. Toy data.

h	**Acc**	**ARI**	**Lik**
1	1	1	**3961.853**
2	1	1	2658.617
10	1	1	2463.996
20	0.9866	0.9444	2424.739

Table 2. Flame data.

h	**Acc**	**ARI**	**Lik**
2	0.9875	0.9501	**344.7159**
5	0.9125	0.6789	238.3863
10	0.9042	0.6517	307.1519
48	0.8583	0.5116	244.413

Toy. Toy data (`http://cs.joensuu.fi/sipu/datasets/`) consists of $n = 373$ units, $p = 2$ variables and $K = 2$ clusters. In Table 1 we list, for some parameters, the accuracy, ARI, and the log-likelihood values, the optimal choice according to Algorithm 1 corresponds to $h = 1$.

Flame. The Flame data (`http://cs.joensuu.fi/sipu/datasets/`) consists of $n = 240$ units, $p = 2$ variables and $K = 3$ clusters. In Table 2 we list ARI and log-likelihood values for some h parameters. Also in this case, the maximum log-likelihood corresponds to the maximum value for accuracy and this confirms our proposal.

Acknowledgement

Acknowledgement of financial support from PNRR MUR project PE0000013-FAIR.

References

Di Nuzzo, C., & Ingrassia, S. 2022a. A graphical approach for the selection of the number of clusters in the spectral clustering algorithm. *Pages 31–44 of:* Salvati, Nicola, Perna, Cira, Marchetti, Stefano, & Chambers, Raymond (eds), *Studies in theoretical and applied statistics*. Cham: Springer International Publishing.

Di Nuzzo, C., & Ingrassia, S. 2022b. A mixture model approach to spectral clustering and application to textual data. *Statistical methods & applications*, **31**(4), 1071–1097.

John C.R., Watson D., Barnes M.R. Pitzalis C. Lewis M.J. 2020. Spectrum: fast density-aware spectral clustering for single and multi-omic data. *Bioinformatics.*, **36**(4).

Ng, A. Y., Jordan, M. I., & Weiss, Y. 2001. On spectral clustering: Analysis and an algorithm. *Page 849–856 of: Proceedings of the 14th international conference on neural information processing systems: Natural and synthetic*. NIPS'01. Cambridge, MA, USA: MIT Press.

Park S., Xu H., Zhao H. 2021. Integrating multidimensional data for clustering analysis with applications to cancer patient data. *Journal of the american statistical association*, **116**(533), 14–26.

Shi, J., & Malik, J. 2000. Normalized cuts and image segmentation. *Ieee transactions on pattern analysis and machine intelligence*, **22**(8), 888–905.

von Luxburg, U. 2007. A tutorial on spectral clustering.tutorial on spectral clustering. *Statistics and computing*, **17**.

Zelnik-Manor, L., & Perona, P. 2004. Self-tuning spectral clustering. *In:* Saul, L., Weiss, Y., & Bottou, L. (eds), *Advances in neural information processing systems*, vol. 17. MIT Press.

Zhang, X., Li J., & Yu, H. 2011. Local density adaptive similarity measurement for spectral clustering. *Pattern recognition letters*, **32**(2), 352 – 358.