GRAPH SIGNAL PROCESSING: TOWARDS THE DIFFUSED SPECTRAL CLUSTERING

Anonymous authors
Paper under double-blind review

ABSTRACT

Graph signal processing is an emerging field of research. When the structure of signals can be represented as a graph, it allows to fully exploit their inherent structure. It has been shown that the normalized graph Laplacian matrix plays a major role in the characterization of a signal on a graph. Moreover, this matrix plays a major role in clustering large data set. In this paper, we present the diffused spectral clustering: a novel handwritten digits clustering algorithm based on the normalized graph Laplacian properties. It's a clever combination between a graph feature space transformation and the spectral clustering algorithm. Experimentally, our proposal outperforms the other algorithms of the state-of-art.

1 INTRODUCTION

During the recent years, the analysis and processing of large-scale datasets (Martelot & Hankin (2013)) using graphs has become very useful. In fact, many kinds of data domains such as social and economic networks, electric grids, neuronal networks and images databases (Martelot & Hankin (2013)) require a graph representation of their structure. Each of these structures usually carries out information that flow between different elements of the network. For example, in a neural network, a neuron is activated after receiving an electric excitation, and the activation of a neuron usually influences the nearby neurons. In the case of economic networks, we can consider the economic crisis as a flow that spreads from one bank to another. This need to represent these phenomena has lead to the development of a new field: the graph signal processing. One of the main applications of graph signal processing (Shuman et al. (2013)) today appear in the field of artificial intelligence and especially in machine learning. Partitioning a large data set into homogeneous clusters is among the most known applications in this domain. The literature gives us many clustering algorithms like the mean-shift clustering (Carreira-Perpinán (2015)), the classical k-means algorithm (Oyelade et al. (2010)) and also the well-known Gaussian Mixture Model (Neagoe & Chirila-Berbentea (2016)) but the graph signal processing framework using the spectral clustering method (Ng et al. (2002)) seems to be more efficient.

In this paper, we present a novel method of clustering handwritten digits taken from the MNIST database (Liu et al. (2003)) based on graph signal processing framework. It's a combination between a graph feature space transformation based on heat diffusion and the spectral clustering algorithm (Xiaowen (2014)).

2 GRAPH FEATURE SPACE TRANSFORMATION

2.1 HEAT DIFFUSION ON GRAPH

Given an arbitrary graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{W})$, its normalized graph Laplacian $\mathcal{L}$, a signal $\boldsymbol{x}^0$ defined on $\mathcal{G}$ such that $\boldsymbol{x}^0 = [\boldsymbol{x}_1^0, ..., \boldsymbol{x}_n^0] \in \mathbb{R}^n$ and a nonnegative constant α, we define the time-varying signal $\boldsymbol{x} \in \mathbb{R}^n$ as the solution of the following differential equation (Xiaowen (2014)):

$$\begin{cases} \boldsymbol{x}(0) = \boldsymbol{x}^0 \\ \frac{\partial \boldsymbol{x}}{\partial t} = -\alpha \mathcal{L} \boldsymbol{x} \end{cases} \quad (1)$$

This differential equation represents the heat diffusion on the graph $\mathcal{G}$ in so far as $-\mathcal{L}$ can be seen as the analogous of the well-know continuous Laplacian operator used in the classical heat diffusion description (Narasimhan (1999)). The signal $\boldsymbol{x}$ describes the temperature distribution at a given time while $\boldsymbol{x}^0$ describes the initial distribution before any diffusion. The positive constant α represents

the thermal conductivity controlling the heat diffusion rate. The greater is α the faster will be $\boldsymbol{x}$ smooth. We can easily verify that the solution of 1 takes the following form (Xiaowen (2014)):

$$\boldsymbol{x}(t) = \exp\left(-\alpha\mathcal{L}t\right)\boldsymbol{x}^0 \tag{2}$$

This formula allows us to compute the temperature distribution at any time point with only the knowledge of the initial distribution $\boldsymbol{x}^0$ and the normalized graph Laplacian $\mathcal{L}$ of the structure carrying out the signal $\boldsymbol{x}$. One can notice that $\boldsymbol{x}(t)$ tends to an uniform distribution since time grows.

2.2 DIFFUSION DISTANCE

With the same notations as above, we introduce the diffusion distance $d^{\mathcal{L}}(\boldsymbol{u}, \boldsymbol{v})$ between the signals $\boldsymbol{u}$ and $\boldsymbol{v}$ defined as:

$$d^{\mathcal{L}}(\boldsymbol{u}, \boldsymbol{v}) := \left\| \int_0^{+\infty} \exp\left(-t\right) \exp\left(-\alpha\mathcal{L}t\right) (\boldsymbol{u} - \boldsymbol{v}) dt \right\| \tag{3}$$

for a given nonnegative scalar $\alpha > 0$ which corresponds to the thermal conductivity in (1). One can notice that the diffusion distance incorporates the structure of the graph which carry out the signals $\boldsymbol{u}$ and $\boldsymbol{v}$. By this way, we can well-determine the proximity between $\boldsymbol{u}$ and $\boldsymbol{v}$ instead of using classical distances as the L_2-distance which doesn't take into account the data type. Noticing that the primitive of the matrix $\exp\left(-t\right)\exp\left(-\alpha\mathcal{L}t\right) = \exp\left(-(\boldsymbol{I}_n + \alpha\mathcal{L})t\right)$ is $-(\boldsymbol{I}_n + \alpha\mathcal{L})^{-1}\exp\left(-(\boldsymbol{I}_n + \alpha\mathcal{L})t\right)$, we can easily verify that:

$$d^{\mathcal{L}}(\boldsymbol{u}, \boldsymbol{v}) = \left\| (\boldsymbol{I}_n + \alpha\mathcal{L})^{-1} (\boldsymbol{u} - \boldsymbol{v}) \right\| \tag{4}$$

Thus, for an input vector $\boldsymbol{u} \in \mathbb{R}^n$ and a nonnegative scalar $\alpha > 0$, we define the diffusion distance transformation $\boldsymbol{u}^{\mathcal{L}}$ of $\boldsymbol{u}$ as:

$$\boldsymbol{u}^{\mathcal{L}} = (\boldsymbol{I}_n + \alpha\mathcal{L})^{-1}\boldsymbol{u} \tag{5}$$

By this way, the diffusion can be interpreted as a feature space transformation on the assumption that there exist relations between the features. These relations can easily be modeled as a graph describing the structure likely to carry out the features of any given signal.

2.3 UNDERLYING GRAPHS FOR 2D SIGNALS

For instance, let us consider the MNIST handwritten digit database (Liu et al. (2003)). Each picture of this database consists of a gray scaled image of size 28×28 pixels representing a given handwritten digit. Therefore, we can see each observation in the MNIST database as a signal $\boldsymbol{x} \in \mathbb{R}^{28 \times 28}$ where the value of each of its components corresponds to the associated pixel intensity. Nonetheless, there are relations between the 784 features imposed by the underlying structure of an image. Indeed, we build an underlying graph $\mathcal{G}$ where each pixel is linked by an unit weight edge to its four (4NN) or eight (8NN) nearest neighbors. By denoting $\mathcal{L}$ the normalized graph Laplacian of the underlying graph built, we are allowed to use the formula 5 in order to obtain the transformed versions by diffusion distance of each handwritten digit.
Figure 1d shows us an observation of the digits "1" and "2" drawn from MNIST, and their transformed versions by the 4NN and 8NN underlying graph (for $\alpha = 10$).

3 DIFFUSED SPECTRAL CLUSTERING

3.1 SPECTRAL CLUSTERING ALGORITHM

In this section, we briefly recall the principle of the spectral clustering algorithm (Xiaowen (2014)) applied in order to partition a graph into several homogeneous subgraphs (Zhao et al. (2011)). For this purpose, let us consider an undirected graph $\mathbf{G} = (\mathbb{V}, \mathbb{E}, \boldsymbol{W})$ where the vertex set $\mathbb{V} = \{\boldsymbol{v}_i\}_{i \in 1..N}$ of N vertices represents the database to cluster. We build the weight matrix $\boldsymbol{W}$ of size $N \times N$ on the model of a Gaussian kernel as follow:

$$\boldsymbol{W}_{i,j} = \exp(-\beta \parallel \boldsymbol{v}_i - \boldsymbol{v}_j \parallel_2^2) \tag{6}$$

where v_i represents the vertex at node i, $\parallel \cdot \parallel_2$ denotes the classical L_2-distance and β the inverse of the empirical variance.
Thus, we build the $N \times N$ unweighted adjacency matrix $\boldsymbol{A}$ on the model of nearest neighbors (one gives a threshold $\gamma > 0$) as follow:

Algorithm 1 Diffused Spectral Clustering

- **Input**:
  - The set $\{M_i\}_{i\in[1:N]}$ of images drawn from MNIST.
  - α: the diffusion rate. Typically, $\alpha \approx 10$ (See section 4).
  - $\mathcal{L}$: be the normalized graph Laplacian of the 4NN or 8NN underlying graph.
  - k: Number of expected clusters.
- Transform each matrix $M_i \in \mathbb{R}^{28\times 28}$ into a vector row $\boldsymbol{x}_i \in \mathbb{R}^{1\times 784}$.
- Transform each $\boldsymbol{x}_i$ into $\boldsymbol{x}_i^{\mathcal{L}} = (\boldsymbol{I}_{784} + \alpha\mathcal{L})^{-1}\,\boldsymbol{x}_i$ using the graph diffusion (4NN or 8NN).
- Let $\mathbf{G} = (\mathbb{V}, \mathbb{E}, \mathbf{W})$ be the weighted complete graph of size N carrying out the set of transformed images $\left\{\boldsymbol{v}_i = \boldsymbol{x}_i^{\mathcal{L}}\right\}$. $\boldsymbol{W}$ is computed using eq (6).
- Compute the unweighted adjacency matrix $\boldsymbol{A}$ using eq (7).
- Launch a spectral clustering using $\boldsymbol{A}$ and k as inputs to get the clusters $C_1, ..., C_k$
- **Output**: $C_1, ..., C_k$: The expected clusters.

$$\boldsymbol{A}_{i,j} = \begin{cases} 1 & \text{if the node } i \text{ is a node } j \text{ neighbor} \Leftrightarrow \boldsymbol{W}_{i,j} > \gamma \\ 0 & \text{otherwise} \end{cases} \qquad (7)$$

We are now looking at partitioning the vertices $\boldsymbol{v}_i$ of $\mathbf{G}$ into k different subsets called clusters, such that the vertices belonging to the same cluster are connected by edges of large weights which can be considered as a similarity measure. This problem has been solved using the algorithm proposed in Ng et al. (2002) which solves the following optimization problem:

$$\min_{\boldsymbol{U}\in\mathbb{R}^{N\times k}} tr(\boldsymbol{U}^T \boldsymbol{L} \boldsymbol{U}) \quad \text{such that: } \boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I}_N \text{ and } \boldsymbol{L}\text{: the normalized Laplacian of } \mathbf{G} \qquad (8)$$

Using the Rayleigh-Ritz theorem (Ng et al. (2002)), the solution of the optimization problem is the matrix containing the first k eigenvectors associated to the k smallest eigenvalues of $\boldsymbol{L}$. The clustering is then computed by applying the k-means algorithm Martelot & Hankin (2013) to the row vectors of the matrix $\boldsymbol{U}$. So, the spectral clustering reduces the dimensionality of the data before applying the k-means algorithm. This is generally done in order to avoid the curse of dimensionality which foils any partitioning algorithm based on geometric approaches like the k-means algorithm.

3.2 Diffused Spectral Clustering

We recall that the goal of this paper is to build an efficient algorithm in order to well-partition a handwritten digit database taken from MNIST Liu et al. (2003). The idea is to combine the graph feature space transformation explained in section 2 with the spectral clustering principle.

Indeed, the graph feature space transformation is used in order to reduce the geometric distance (L_2 distance) between images describing the same digits. By this way, the spectral clustering algorithm which is based on geometric approaches will be efficiently applied to group the same digits and thus to cluster the input database. The diffused spectral clustering algorithm applied to the MNIST database is summarized in Algorithm 1.

4 Experiments and result analysis

In this section, we make some discussions about the performances of the diffused spectral clustering algorithm used in order to cluster samples taken from the MNIST database. For the sake of simplicity, we only deal with 1000 samples of digits "1" and "2". The results are averaged over 100 runs. The first step of the analysis consists on choosing the best value of the diffusion rate α.

Figure 1a shows us that choosing a value of the parameter α close to 10 is a sufficient condition to reach the optimum error rate when we are dealing with the 4NN or 8NN underlying graph. Error rate means the percentage of wrong digits detected among the treated digit set.

Naturally, the lesser the error rate, the better the clustering algorithm.

By setting the parameter $\alpha = 10$, we evaluate the error rate achieved by the diffused spectral clustering in both cases of the 4NN or 8NN underlying graph.

We are then interested in comparing the diffused spectral clustering performances to those of an another clustering algorithm from the state of art like the k-means algorithm and GMM. We recall that the only difference between k-means, GMM and spectral clustering is the dimensionality reduction applied by the last one on the original data. Figure 1b shows us this comparison.

From figure 1b, we can conclude about the importance of the diffused spectral clustering. We reach a very low error rate (0.01) when we are dealing with the 4NN or 8NN underlying graph.

In order to emphasize this result, we compare it to the one obtained by computing a principal component analysis (Dong et al. (2014)) (PCA) before launching the k-means and GMM. This has been chosen to well compare the dimensionality reduction influence on the clustering result. Figure 1c shows the result of this comparison.

5 CONCLUSION AND FUTURE WORKS

The paper shows the importance of using graph signal processing framework in applications such as clustering. By this way, we have implemented a novel clustering algorithm for partitioning handwritten digits. It seems that it's very efficient and outperforms the state of art. Out algorithm is essentially based on the normalized graph Laplacian properties in terms of smoothness, and informations carried out by its eigenspace. These informations seem to be very useful to understand the implicit and explicit relationships in a database.

Our future works consist on applying the diffused spectral clustering on a database other than MNIST in order to partition data other than images. For example, we are thinking about adapting our method in order to cluster sound signals, or other complex signals carried out by complex structures like neural and social networks.
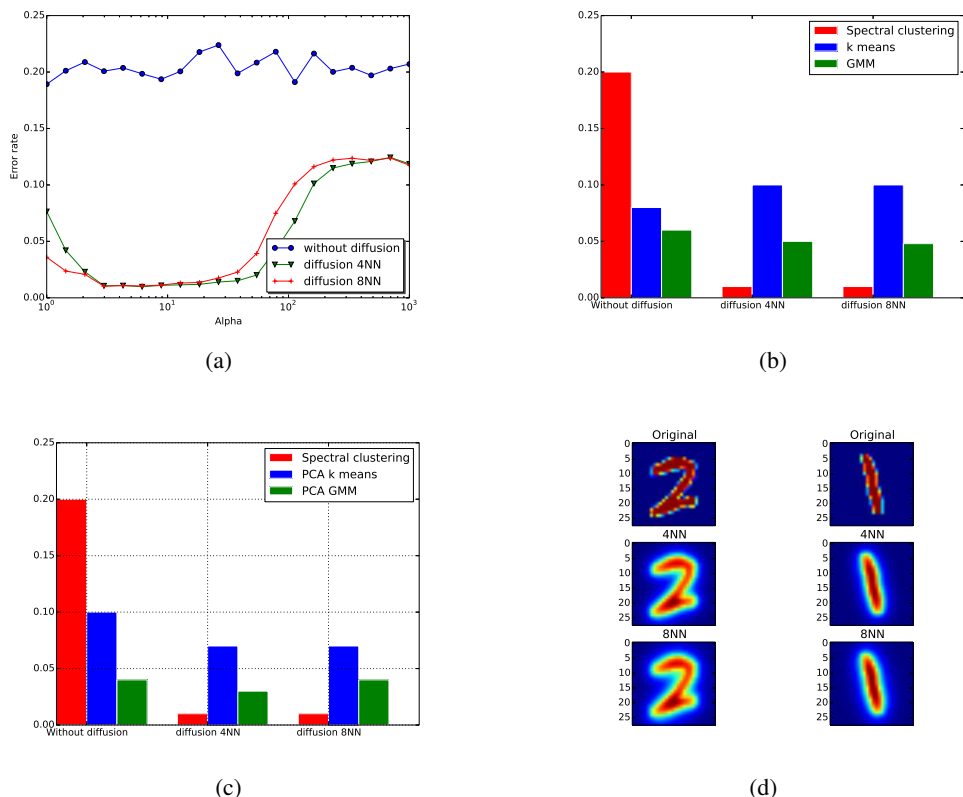

Figure 1: (a) Error rate variations for 1000 samples of digit "1" and "2" using spectral clustering with and without diffusion. (b) Clustering of 1000 samples of digit "1" and "2" using the spectral clustering vs k-means vs GMM. (c) Clustering of 1000 samples of digit "1" and "2" using both spectral clustering, k-means with PCA and GMM with PCA. (d) Observation of the digits "1" and "2", and their transformed smoothened versions using the 4NN and 8NN underlying graph.

REFERENCES

Miguel A Carreira-Perpinán. A review of mean-shift algorithms for clustering. *arXiv preprint arXiv:1503.00687*, 2015.

Xiaowen Dong, Dorina Thanou, Pascal Frossard, and Pierre Vandergheynst. Learning laplacian matrix in smooth graph signal representations. *arXiv preprint arXiv:1406.7842*, 2014.

Cheng-Lin Liu, Kazuki Nakashima, Hiroshi Sako, and Hiromichi Fujisawa. Handwritten digit recognition: benchmarking of state-of-the-art techniques. *Pattern Recognition*, 36(10):2271–2285, 2003.

Erwan Le Martelot and Chris Hankin. Fast multi-scale community detection based on local criteria within a multi-threaded algorithm. *arXiv preprint arXiv:1301.0955*, 2013.

TN Narasimhan. Fourier's heat conduction equation: History, influence, and connections. *Reviews of Geophysics*, 37(1):151–172, 1999.

Victor-Emil Neagoe and Vlad Chirila-Berbentea. Improved gaussian mixture model with expectation-maximization for clustering of remote sensing imagery. In *2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pp. 3063–3065. IEEE, 2016.

Andrew Y Ng, Michael I Jordan, Yair Weiss, et al. On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, 2:849–856, 2002.

OJ Oyelade, OO Oladipupo, and IC Obagbuwa. Application of k means clustering algorithm for prediction of students academic performance. *arXiv preprint arXiv:1002.2425*, 2010.

David Shuman, Sunil K Narang, Pascal Frossard, Antonio Ortega, Pierre Vandergheynst, et al. The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains. *Signal Processing Magazine, IEEE*, 30(3):83–98, 2013.

DONG Xiaowen. *Multi-View Signal Processing and Learning on Graphs*. PhD thesis, ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE, 2014.

Yunpeng Zhao, Elizaveta Levina, and Ji Zhu. Community extraction for social networks. *Proceedings of the National Academy of Sciences*, 108(18):7321–7326, 2011.