

Multiclass Spectral Clustering Based on Discriminant Analysis

Xi Li[†], Zhongfei Zhang[‡], Yanguo Wang[†], Weiming Hu[†]

[†]National Laboratory of Pattern Recognition, Institute of Automation, CAS, Beijing, China

[†]{lix, ygwang, wmhu}@nlpr.ia.ac.cn

[‡]State University of New York, Binghamton, NY 13902, USA

[‡]zhongfei@cs.binghamton.edu

Abstract

Many existing spectral clustering algorithms share a conventional graph partitioning criterion: normalized cuts (NC). However, one problem with NC is that it poorly captures the graph's local marginal information which is very important to graph-based clustering. In this paper, we present a discriminant analysis based graph partitioning criterion (DAC), which is designed to effectively capture the graph's local marginal information characterized by the intra-class compactness and the inter-class separability. DAC preserves the intrinsic topological structures of the similarity graph on data points by constructing a k -nearest neighboring subgraph for each data point. Consequently, the clustering results generated by the DAC-based clustering algorithm (DACA) are robust to the outlier disturbance. Theoretic analysis and experimental evaluations demonstrate the promise and effectiveness of DACA.

1 Introduction

In recent years, spectral clustering (SC) has been successfully applied to many domains such as circuit layout [1, 2], load balancing [3] and image segmentation [4, 5]. Based on local evidence from similarities among data points, SC finds out the best graph cuts by optimizing a particular partitioning criterion function through eigendecomposition. With effectiveness in clustering data of complex structure, SC is promising for multiclass data learning.

Shi and Malik [4] propose a normalized cut criterion for segmenting the similarity graph. Gdalyahu *et al.* [6] present a "typical cut" algorithm for graph partitioning. Ding *et al.* [7] present a min-max cut algorithm for graph partitioning and data clustering. Balanced partitions are obtained by the min-max cut algorithm. Ng *et al.* [8] present a clustering algorithm based on K-

Means after the spectral relaxation. Yu and Shi [9] propose a principled account on multiclass spectral clustering. They give a nearly global-optimal discrete clustering solution by using singular value decomposition and nonmaximum suppression in an iterative procedure. However, all the aforementioned graph-based clustering methods share a problem that the intrinsic structure of the graph is not well preserved. As a result, the clustering results are usually sensitive to outliers.

In this paper, we present a novel graph partitioning criterion called DAC, in which graph preserving based discriminant analysis is enabled to effectively capture the graph's local marginal information characterized by the inter-class separability and the intra-class compactness. By maximizing the inter-class separability and the intra-class compactness simultaneously, DAC obtains an optimal graph partitioning solution. Especially, DAC aims at preserving the intrinsic topological structures of the similarity graph on data points by constructing a k -nearest neighboring subgraph for each data point. As a result, the outlier disturbance can be reduced to a large extent.

2 Multiclass spectral clustering

An N -node weighted graph $G = (\mathbb{V}, \mathbb{E}, W)$ is used for representing the intrinsic relationships among N data points, where $\mathbb{V} = \{1, \dots, N\}$ is the node set, $\mathbb{E} \subseteq \mathbb{V} \times \mathbb{V}$ is the edge set, and $W = (w_{ij})_{N \times N}$ is an affinity matrix with the entry w_{ij} being the edge-weight between node i and node j . Clustering N data points into K classes is equivalent to partitioning \mathbb{V} into K disjoint subsets, namely, $\mathbb{V} = \bigcup_{l=1}^K \mathbb{V}_l$ s.t. $\mathbb{V}_m \cap \mathbb{V}_n = \emptyset, \forall m \neq n$. For convenience, let $\Gamma_{\mathbb{V}}^K = \{\mathbb{V}_1, \mathbb{V}_2, \dots, \mathbb{V}_K\}$.

2.1 Graph partitioning criterion

Given $\mathbb{V}_a, \mathbb{V}_b \subset \mathbb{V}$, $\text{links}(\mathbb{V}_a, \mathbb{V}_b)$ is defined as the sum of the total weighted connections between \mathbb{V}_a and

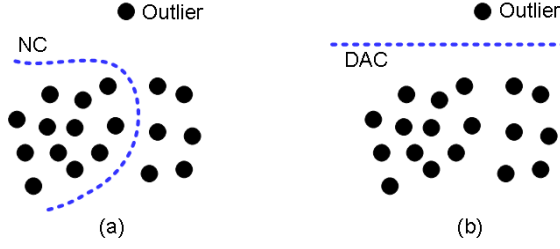


Figure 1. Graph partitioning results for NC (normalized cuts) and DAC corresponding to (a) and (b), respectively.

\mathbb{V}_b :

$$\text{links}(\mathbb{V}_a, \mathbb{V}_b) = \sum_{i \in \mathbb{V}_a} \sum_{j \in \mathbb{V}_b} w_{ij} \quad (1)$$

Moreover, the degree of \mathbb{V}_a is defined as the total links of nodes in \mathbb{V}_a to all the nodes in \mathbb{V} , i.e. $\text{degree}(\mathbb{V}_a) = \text{links}(\mathbb{V}_a, \mathbb{V})$. Subsequently, one classic objective function (i.e., K-way normalized cuts) for graph partitioning is described as follows.

K-way normalized cuts. The K-way normalized cut objective aims to minimize the cut relative to the degree of a node class. Thus, the objective is represented as:

$$\begin{aligned} \Gamma_{cut} &= \arg \min_{\Gamma_{\mathbb{V}}^K} \text{kncut}(\Gamma_{\mathbb{V}}^K) \\ &= \arg \min_{\Gamma_{\mathbb{V}}^K} \frac{1}{K} \sum_{k=1}^K \frac{\text{links}(\mathbb{V}_k, \mathbb{V} \setminus \mathbb{V}_k)}{\text{degree}(\mathbb{V}_k)} \end{aligned} \quad (2)$$

2.2 Solving K-way normalized cuts

In [9], it has been proven that solving the K-way normalized cuts is equivalent to finding the optima of an optimization program:

$$\begin{aligned} \text{maximize} \quad & f(X) = \frac{1}{K} \sum_{n=1}^K \frac{X_n^T W X_n}{X_n^T D X_n} \\ \text{subject to} \quad & X \in \{0, 1\}^{N \times K}, X_{K} = N \end{aligned} \quad (3)$$

where X is an $N \times K$ partition matrix, d denotes a $d \times 1$ vector with each element being 1, D is an $N \times N$ diagonal matrix with the m th diagonal element being the sum of the elements belonging to the m th row of W for $1 \leq m \leq N$, and X_n is the n th column of X for $1 \leq n \leq K$. However, this criterion is sensitive to the outlier disturbance owing to not effectively capturing the intrinsic topological relationships among the nodes in the graph. In order to solve this problem, we propose a discriminant analysis based criterion (i.e., DAC) to obtain an effective graph partitioning solution that is robust to the outlier disturbance. One two-class graph partitioning example is given in Fig. 1, where DAC succeeds in separating the outlier sample from the dominant sample set while NC fails.

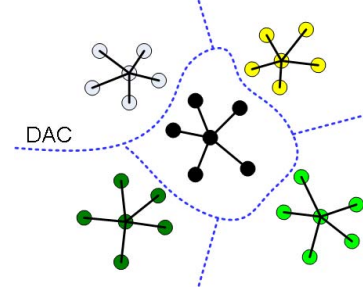


Figure 2. Illustration of the partitioning principle of DAC.

In the next section, we discuss the discriminant analysis based graph partitioning criterion (DAC) and how to find its optimal solutions.

3 Graph partitioning based on discriminant analysis

3.1 Partitioning criterion based on discriminant analysis

Before presenting the proposed graph partitioning criterion (i.e., DAC), we first give a brief introduction to the notations and symbols we use. Let $W = (w_{ij})_{N \times N}$ be the edge weight matrix, and $W^* = (w_{ij}^*)_{N \times N}$ be the k -nearest edge weight matrix reflecting the neighboring relationships among data points. w_{ij}^* is determined as:

$$w_{ij}^* = \begin{cases} 1 & \text{if } j \in N_k(i) \text{ or } i \in N_k(j) \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

where $N_k(i)$ represents the k -nearest neighborhood of data point i . Let $\widehat{W} = (\widehat{w}_{ij})_{N \times N} = (w_{ij} \cdot w_{ij}^*)_{N \times N}$, \widehat{D} be the diagonal matrix with $\widehat{d}_{ii} = \sum_j \widehat{w}_{ij}$ for $1 \leq i, j \leq N$, and $Q = \widehat{D} - \widehat{W}$. If Q is a singular matrix, it should be replaced with $Q + \epsilon I_N$, where ϵ is a small positive constant and I_N is an $N \times N$ identity matrix. Consequently, the proposed DAC is formulated as:

$$\begin{aligned} \text{maximize} \quad & g(X) = \frac{1}{K} \sum_{n=1}^K \frac{X_n^T \widehat{W} X_n}{X_n^T Q X_n} \\ & = \frac{1}{K} \sum_{n=1}^K \frac{[X_n (X_n^T X_n)^{-\frac{1}{2}}]^T \widehat{W} [X_n (X_n^T X_n)^{-\frac{1}{2}}]}{[X_n (X_n^T X_n)^{-\frac{1}{2}}]^T Q [X_n (X_n^T X_n)^{-\frac{1}{2}}]} \end{aligned} \quad (5)$$

$$\text{subject to} \quad X \in \{0, 1\}^{N \times K}, X_{K} = N$$

where X_n represents an $N \times 1$ vector formed by the n th column of X . The analytical proof of DAC is given as follows.

The intra-class compactness and the inter-class separability are respectively captured by $X_n^T \widehat{W} X_n$ and $X_n^T Q X_n$, which are formulated as:

$$X_n^T \widehat{W} X_n = \sum_{i \in \mathbb{V}_n} \sum_{j \in \mathbb{B}_i} w_{ij}; \quad X_n^T Q X_n = \sum_{i \in \mathbb{V}_n} \sum_{j \in \overline{\mathbb{B}}_i} w_{ij} \quad (6)$$

Given: A data set $Z = \{z_1, z_2, \dots, z_N\}$ and the number of classes K :

1. Create a graph $G = (V, E, W, W^*)$, where $V = \{1, \dots, N\}$ is the node set, $E \subseteq V \times V$ represents the edge set, $W^* = (w_{ij}^*)_{N \times N}$ defined in (4), and $W = (w_{ij})_{N \times N}$, i.e., $w_{ij} = \exp(-\text{dist}(z_i, z_j)/2\sigma^2)$ in which σ is a scaling factor, and $\text{dist}(\cdot)$ denotes a distance function. Typically, $\text{dist}(z_i, z_j) = \|z_i - z_j\|^2$.
2. Obtain $\widehat{W} = (\widehat{w}_{ij})_{N \times N} = (w_{ij} \cdot w_{ij}^*)_{N \times N}$ and $Q = \widehat{D} - \widehat{W}$ where \widehat{D} is the diagonal matrix with $\widehat{d}_{ii} = \sum_j \widehat{w}_{ij}$ for $1 \leq i, j \leq N$. If Q is a singular matrix, it should be replaced with $Q + \epsilon I_N$, where ϵ is a small positive constant and I_N is an $N \times N$ identity matrix.
3. Form \widetilde{P} by the normalized K largest eigenvectors of $Q^{-1}\widehat{W}$.
4. Obtain a candidate graph partitioning solution \widetilde{X} by: $\widetilde{X} = \text{Diag}(\text{diag}^{-\frac{1}{2}}(\widetilde{P}\widetilde{P}^T))\widetilde{P}$.
5. Perform the iterative refining procedure [9] on \widetilde{X} to find an optimal solution X to (5). The refining procedure is discussed in detail in steps four to eight of the algorithm in [9].

Figure 3. The flowchart of the DAC-based clustering algorithm (DACA).

where $B_i = \{u|u \in \mathbb{V}_n \text{ and } i \in N_k(u) \text{ or } u \in N_k(i)\}$, $\overline{B}_i = \{u|u \notin \mathbb{V}_n \text{ and } i \in N_k(u) \text{ or } u \in N_k(i)\}$, and \mathbb{V}_n denotes the node set corresponding to the n th class. The larger the value of $X_n^T \widehat{W} X_n$, the more compact the intra-class samples. The smaller the value of $X_n^T Q X_n$, the more separable the inter-class samples. As a result, an optimal graph partitioning solution is obtained by maximizing the $g(X)$ in (5). Clearly, *DAC* is capable of preserving the topological structures of the similarity graph by constructing a k -nearest neighboring subgraph for each node in the graph, leading to a robust graph partitioning. The partitioning principle of *DAC* can be illustrated by Fig.2, in which a 5-nearest subgraph for each node is constructed. An optimal graph partitioning boundary, denoted by a dashed curve, is obtained by *DAC* through exploring the inter-class marginal information of the graph.

3.2 Finding optimal solutions

After a sequence of simplification operations, the graph partitioning criterion (5) becomes: $g(X) = \frac{1}{K} \text{tr}\{(P^T Q P)^{-1} (P^T \widehat{W} P)\}$, where tr denotes the trace of a matrix, and $P = X(X^T X)^{-\frac{1}{2}}$ which is constrained by: $P^T P = [X(X^T X)^{-\frac{1}{2}}]^T [X(X^T X)^{-\frac{1}{2}}] = I_K$ where I_K is a $K \times K$ identity matrix, and $X^T X$ is a diagonal matrix. Thus, the criterion (5) can be rewritten as:

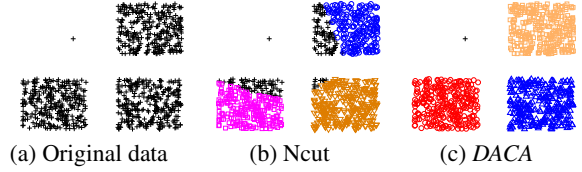


Figure 4. Clustering results of MSC and DACA. (a) shows the original data samples of dataset 1. The outlier sample in dataset 1 is plotted as “+” in the left-up side of (a). The clustering results of MSC are displayed in (b) while (c) exhibits the clustering results of DACA.

$$\begin{aligned} &\text{maximize} && h(P) = \frac{1}{K} \text{tr}\{(P^T Q P)^{-1} (P^T \widehat{W} P)\} \\ &\text{subject to} && P^T P = I_K. \end{aligned} \quad (7)$$

The optimization problem (7) has been addressed in multiclass LDA (linear discriminant analysis) learning [13]. A solution \widetilde{P} to (7) consists of the K principal eigenvectors (i.e., corresponding to the K largest eigenvalues) of the matrix $Q^{-1}\widehat{W}$. If Q is a singular matrix, $Q^{-1}\widehat{W}$ should be replaced with the matrix $(Q + \epsilon I_N)^{-1}\widehat{W}$, where ϵ is a small positive constant and I_N is an $N \times N$ identity matrix. As a result, a candidate solution \widetilde{X} to (5) is obtained by: $\widetilde{X} = \text{Diag}(\text{diag}^{-\frac{1}{2}}(\widetilde{P}\widetilde{P}^T))\widetilde{P}$, where $\text{Diag}(\cdot)$ denotes a diagonal matrix formed from its vector argument, and $\text{diag}(\cdot)$ represents a column vector formed from the diagonal elements of its matrix argument. Subsequently, the iterative refining procedure [9] may be used to find the optimal graph partitioning solution X to (5). Finally, we have the *DAC*-based clustering algorithm (*DACA*) with its specific procedure listed in Fig. 3.

4 Experiments

In order to evaluate the performance of the proposed *DAC*-based clustering algorithm (*DACA*), two datasets are used in the experiments. They are a labeled synthetic toy datasets and an unlabeled vehicle trajectory dataset collected from a real traffic scene. The first synthetic toy dataset (i.e., displayed in Fig.4(a)) is composed of 920 data samples categorized into four classes. The first class is an outlier while the other three are dominant. The last dataset is formed by a total number of 1200 vehicle motion trajectories, which is acquired from the tracker presented in [12]. k in (4) is set as 5.

Two experiments are conducted to demonstrate the claimed contributions of the proposed *DAC*-based clustering algorithm (*DACA*). They are to evaluate the clustering accuracy of *DACA* in the presence of outliers.

The first experiment is performed to evaluate the performances of two graph-based clustering techniques—*DACA* and multiclass spectral clustering (MSC) [9]

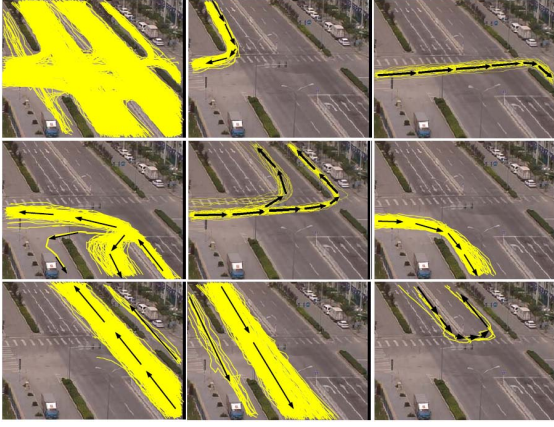


Figure 5. A trajectory clustering example. The left-up picture shows the raw trajectories while the remaining ones show the final clustering results by *DACA*.

on investigating the clustering accuracy using the first dataset. The scaling factor σ for creating a graph is set as 6. The final clustering results are shown in Fig. 4. It is clear that *DACA* succeeds in removing the outlier disturbance while MSC [9] fails. The reason is explained as follows. Since MSC does not preserve the nearest relationships among the nodes in the graph, its graph partitioning results are greatly disturbed by outliers. In contrast to MSC, *DACA* preserves the intrinsic local nearest relationships among the nodes in the graph. Consequently, *DACA* is able to remove the outlier disturbance according to the cues of the nearest relationships among the nodes in the graph.

The last experiment on the trajectory dataset is performed to showcase the performance of *DACA*. The final clustering results are shown in Fig. 5. The DFT-coefficient feature [11] is used again to represent the trajectories. Raw trajectories are clustered into nine trajectory classes. In each trajectory class, trajectories have very similar directions. We just choose to display eight dominant trajectory classes for better visualization. They are shown in Fig. 5, where the black arrows represent the directions of trajectories. The other one trajectory class is treated as outliers. From Fig. 5, it is clear that *DACA* performs well in multiclass trajectory clustering.

In summary, the proposed *DACA*, based on the novel graph partitioning criterion called *DAC*, is able to effectively reduce the outlier disturbance. Thus, it is a very promising algorithm for multiclass data clustering.

5 Conclusion

In this paper, we have proposed a discriminant analysis based graph partitioning criterion (*DAC*). In this

criterion, discriminant analysis is enabled to effectively characterize the intra-class compactness and the inter-class separability. By maximizing the inter-class separability and the intra-class compactness simultaneously, *DAC* obtains an optimal graph partitioning solution. Moreover, *DAC* preserves the intrinsic topological structures of the affinity graph by constructing a k -nearest neighboring subgraph for each node in the graph. As a result, the clustering results generated by the *DACA*-based clustering algorithm (*DACA*) is robust to the outlier disturbance.

6 Acknowledgment

This work is partly supported by NSFC (Grant No. 60520120099, 60672040 and 60705003) and the National 863 High-Tech R&D Program of China (Grant No. 2006AA01Z453). Z.Z. is supported in part by NSF (IIS-0535162). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

References

- [1] C. J. Alpert and A. B. Kahng, "Multiway partitioning via geometric embeddings, orderings and dynamic programming," *IEEE Trans. on Computer-aided Design of Integrated Circuits and Systems*, Vol.14, Iss.11, pp.1342-1358, 1995.
- [2] P. K. Chan, M. D. F. Schlag and J. Y. Zien, "Spectral k -way ratio-cut partitioning and clustering," *IEEE Trans. on Computer-aided Design of Integrated Circuits and Systems*, Vol.13, Iss.9, pp.1088-1096, 1994.
- [3] B. Hendrickson and R. Leland, "An improved spectral graph partitioning algorithm for mapping parallel computations," *SIAM J. Sci. Comput.*, Vol.16, Iss.2, pp.452-459, 1995.
- [4] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. on PAMI.*, Vol.22, Iss.8, pp.888-905, 2000.
- [5] J. Malik, S. Belongie, T. Leung, and J. Shi, "Contour and texture analysis for image segmentation," *IJCV*, 2001.
- [6] Y. Gdalyahu, D. Weinshall and M. Werman, "Self-organization in vision: stochastic clustering for image segmentation, perceptual grouping, and image database organization," *IEEE Trans. on PAMI.*, Vol. 23, Iss. 10, pp.1053-1074, Oct. 2001.
- [7] C.H.Q. Ding, X. He, H. Zha, M. Gu and H.D. Simon, "A min-max cut algorithm for graph partitioning and data clustering," in *Proc. ICDM*, pp.107-114, 2001.
- [8] A. Y. Ng, M. I. Jordan and Y. Weiss, "On spectral clustering: analysis and an algorithm," *NIPS*, MIT Press, 2001.
- [9] S.X. Yu and J. Shi, "Multiclass spectral clustering," in *Proc. ICCV*, Vol.1, pp.313-319, 2003.
- [10] W. Hu, D. Xie, T. Tan and S. Maybank, "Learning activity patterns using fuzzy self-organizing neural network," *IEEE Trans. SMC. Part B* 34(3):1618-1626,2004.
- [11] A. Naftel and S. Khalid, "Motion trajectory learning in the DFT-coefficient feature space," in *Proc. ICVS*, Jan. 2006.
- [12] D. Xie, W. Hu and T. Tan, "A multi-object tracking system for surveillance video analysis," in *Proc. ICPR*, Vol. 4, pp.767-770, 2004.
- [13] J. Ma, J.L. Sancho-Gomez and S.C. Ahalt, "Nonlinear multiclass discriminant analysis," *IEEE Signal Processing Letters*, Vol. 10, Iss. 7, pp.196-199, July 2003.