Qin ZHANG, Guo-qiang ZHONG, Jun-yu DONG, 2018. An anchor-based spectral clustering method. *Frontiers of Information Technology & Electronic Engineering*, 19(11):1385-1396.
https://doi.org/10.1631/FITEE.1700262

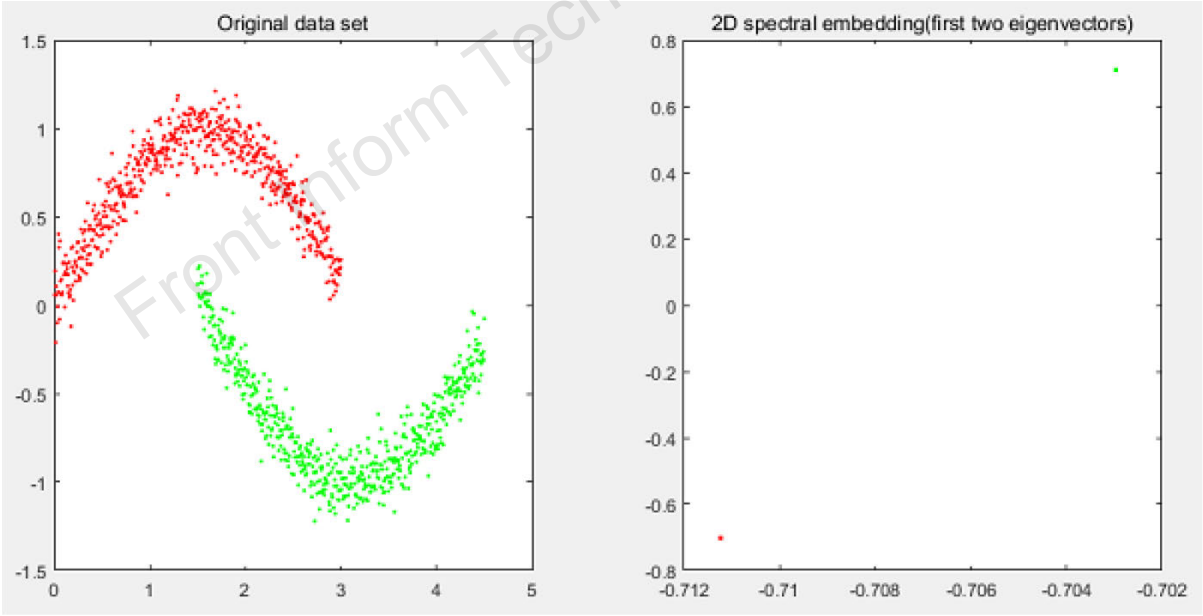
An anchor-based spectral clustering method

Key words: Clustering; Spectral clustering; Graph Laplacian; Anchors

Corresponding author: Guo-qiang ZHONG
E-mail: gqzhong@ouc.edu.cn
ORCID: Guo-qiang ZHONG, http://orcid.org/0000-0002-2952-6642

Motivation

The data points can be modeled by a graph, whose vertices and edges represent the data points and the similarity among the pair-wise data points, respectively. Then, a new representation of data can be obtained using the eigenvectors of the graph Laplacian matrix, which is low-dimensional and more separable and is often known as spectral embedding of data.



Motivation cont.

1. A small subset of the original dataset can be selected as anchors, and the rest of the data points can be represented by the linear combination of these anchor points.

2. The data-anchor mapping matrix, i.e., the linear combination coefficients matrix, ***P***, not only captures the relationship between the data and anchors, but also projects the clustering of anchor points to the original data points.

Main ideas

Use anchors to reduce the problem size to deal with the eigenvectors of the graph Laplacian:

(1) Use a probabilistic sampling method to select anchors;

(2) Solve local anchor embedding to obtain the data-anchor mapping matrix ***P***;

(3) Use data-anchor mapping matrix ***P*** to obtain the approximate clustering results of the original data set.

Flowchart of the proposed method

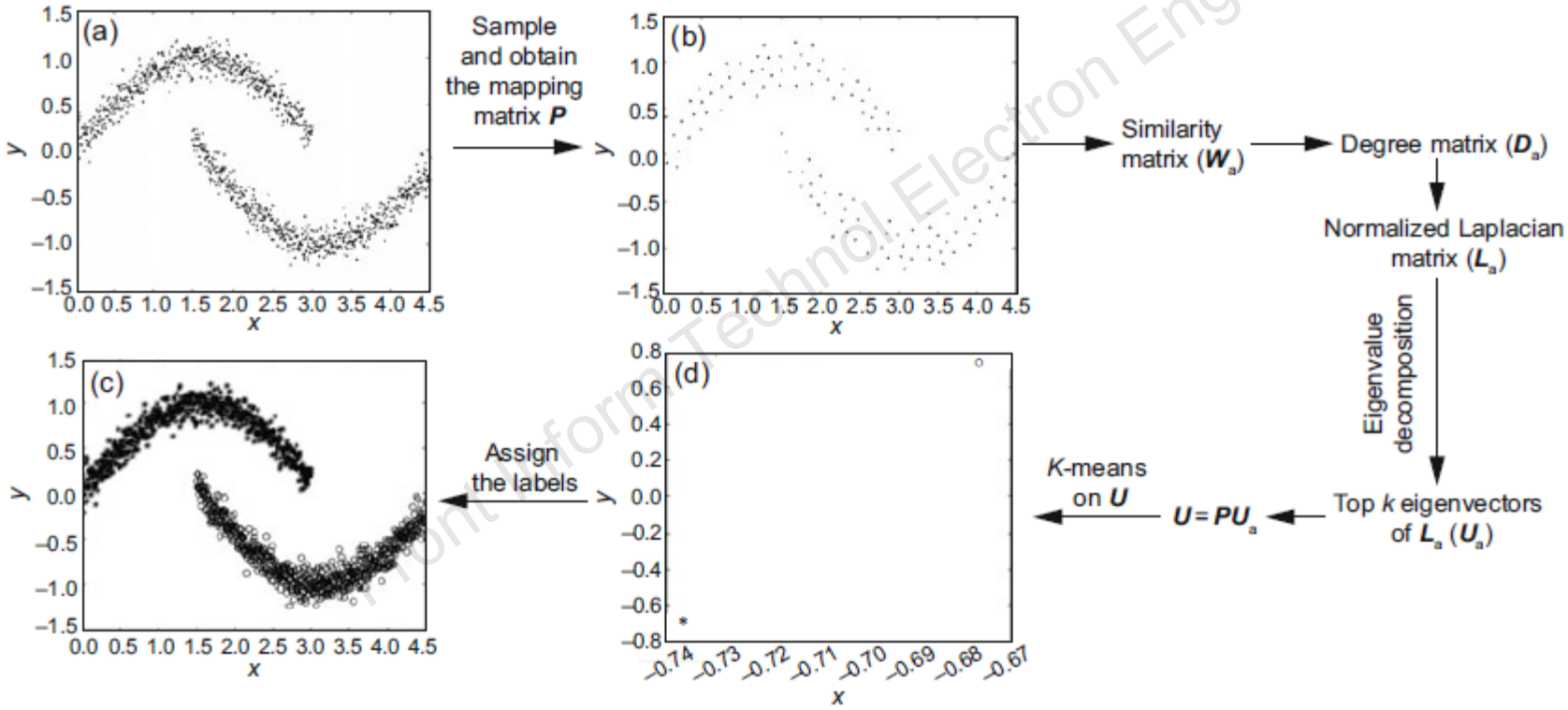


Fig. 2 Framework of the proposed anchor-based spectral clustering algorithm: (a) two-moon dataset; (b) anchor set; (c) recovered spectral clustering of the whole dataset; (d) k clusters of U

Notations

Table 1 Notations

Symbol	Description
$\boldsymbol{X}$	Data matrix with a data point in every row
$\boldsymbol{W}$	Similarity matrix of $\boldsymbol{X}$
$\boldsymbol{U}$	Eigenvector matrix of $\boldsymbol{W}$
$\boldsymbol{A}$	Anchor matrix with an anchor point in every row
$\boldsymbol{W}_{\mathrm{a}}$	Similarity matrix of $\boldsymbol{A}$
$\boldsymbol{U}_{\mathrm{a}}$	Eigenvector matrix of $\boldsymbol{W}_{\mathrm{a}}$

1. Selecting anchors

Let dist(a, $\boldsymbol{X}$) denote the shortest distance from anchor point a to dataset $\boldsymbol{X}$. The anchors can be selected as follows:

Step 1: Choose initial data point $\boldsymbol{a}$ uniformly at random from $\boldsymbol{X}$, and set $\boldsymbol{A} = \{\boldsymbol{a}\}$.

Step 2: Choose the next data point $\boldsymbol{a} \in \boldsymbol{X}$ with the probability $p(\boldsymbol{a}) = \dfrac{[\mathrm{dist}(\boldsymbol{a}, \boldsymbol{A})]^2}{\sum_{\boldsymbol{a}' \in \boldsymbol{X}} [\mathrm{dist}(\boldsymbol{a}', \boldsymbol{A})]^2}$ and add it to $\boldsymbol{A}$.

Step 3: Repeat the above two steps until a total of m anchor points have been chosen.

2. Local anchor embedding

$$\min_{\boldsymbol{P}\in\mathbb{R}^{n\times m}} \quad J(\boldsymbol{P}) = \frac{1}{2}\|\boldsymbol{X} - \boldsymbol{P}\boldsymbol{A}\|^2$$
$$\text{s.t.} \quad P_{ij} \geq 0,\ \boldsymbol{P}_i\mathbf{1} = 1,$$

3. Approximate clustering

$$K\text{means}(\boldsymbol{U}) \simeq K\text{means}(\boldsymbol{P}\boldsymbol{U}_{\text{a}})$$

Experiments: datasets

Table 3 Ten datasets used in the experiments

Dataset	Number of data points	Dimensionality	Number of classes
Protein	116	20	6
Thyroid	215	5	3
Ionosphere	351	34	2
Dermatology	366	34	6
Balance	625	4	3
Yeast	1489	8	10
Segmentation	2310	19	7
Waveform21	5000	21	3
Satimage	6435	36	6
Letter	20 000	16	26

Experiment results: quality metrics

1. External quality metric: purity

2. Internal quality metric: Davis-Bouldin index

Experiment results: average purity

Table 4 Average purity of 10 datasets

Dataset	Average purity			
	SC_NJW (baseline)	LSC	PIC	ASC
Protein	0.6466	0.4582	0.4418	**0.5095**
Thyroid	0.9721	0.8047	0.8672	**0.9419**
Ionosphere	0.7236	0.7014	0.6410	**0.7236**
Dermatology	0.8661	**0.8699**	0.7258	0.8552
Balance	0.7376	0.6545	0.5894	**0.6796**
Yeast	0.5480	**0.5343**	0.4041	0.4686
Segmentation	0.7407	**0.7016**	0.5497	0.6855
Waveform21	0.5184	0.5826	0.5620	**0.5928**
Satimage	0.7243	0.7338	0.6781	**0.7352**
Letter	–	0.3570	0.1230	**0.4126**
Mean rank	1.6667	2.4444	3.7778	2.1111

The best results are highlighted in boldface

Experiment results: average DB index

Table 5 Average Davis-Bouldin index of 10 datasets

Dataset	Average Davis-Bouldin index			
	SC_NJW (baseline)	LSC	PIC	ASC
Protein	2.4215	2.9778	3.0313	**2.9632**
Thyroid	0.9900	2.3980	**0.8295**	1.0715
Ionosphere	2.3494	1.5495	**1.3282**	2.9879
Dermatology	1.5283	1.6269	2.4315	**1.5755**
Balance	1.7529	**1.7431**	2.4860	1.8484
Yeast	1.5257	1.6693	3.4283	**1.6022**
Segmentation	1.3125	1.3534	1.8304	**1.1296**
Waveform21	1.4905	**1.4816**	2.1689	1.9026
Satimage	1.3534	1.0545	1.6821	**0.9803**
Letter	–	2.0089	7.1854	**1.8434**
Mean rank	3.1111	2.5556	1.6667	2.6667

The best results are highlighted in boldface

Experiment results: average runtime

Table 6 Average runtime of 10 datasets

Dataset	Average runtime (s)			
	SC_NJW (baseline)	LSC	PIC	ASC
Protein	0.0772	**0.0218**	0.0402	0.0291
Thyroid	0.2530	0.0365	0.0526	**0.0284**
Ionosphere	0.4657	0.0900	**0.0195**	0.0379
Dermatology	0.5783	0.0951	0.1950	**0.0580**
Balance	2.2365	0.1750	0.1364	**0.0742**
Yeast	21.7880	1.2232	0.5232	**0.3882**
Segmentation	67.4973	0.9031	**0.5142**	0.6364
Waveform21	546.9019	2.6208	**0.2073**	2.2664
Satimage	1181.9291	3.9559	5.3411	**3.4922**
Letter	–	32.4293	**9.8302**	51.3016
Mean rank	1.0000	2.5556	2.8889	3.5556

The best results are highlighted in boldface

Conclusions

1. We proposed an anchor-based spectral clustering method, called ASC, to obtain the approximate clustering. A small subset of the original dataset is selected as anchors, and the rest of the data points can be represented by the linear combination of these anchor points.

2. The data-anchor mapping matrix, i.e., the linear combination coefficients matrix, ***P***, not only captures the relationship between the data and anchors, but also projects the clustering of anchor points to the original data points.

3. The proposed method ASC is compared with two state-of-the-art methods, PIC and LSC, on 10 real-world applications using three evaluation metrics. Experimental results showed that the proposed method performed significantly better than classical spectral clustering in terms of runtime, and achieved a better or at least comparable performance (purity and runtime), compared with the state-of-the-art methods.