Constrained spectral embedding for K-way data clustering

Guillaume Wacquet, Emilie Poisson Caillault, Denis Hamad, Pierre-Alexandre Hébert

▶ **To cite this version:**

Guillaume Wacquet, Emilie Poisson Caillault, Denis Hamad, Pierre-Alexandre Hébert. Constrained spectral embedding for K-way data clustering. Pattern Recognition Letters, 2013, 34 (9), pp.1009-1017. 10.1016/j.patrec.2013.02.003 . hal-01536663

HAL Id: hal-01536663
https://hal.science/hal-01536663v1

Submitted on 15 Jun 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

PRL-Wacquet.tex
Click here to download Manuscript [Word or (La)TeX]: PRL-Wacquet.tex Click here to view linked References

Constrained Spectral Embedding for K-Way Data Clustering

G. Wacquet*, É. Poisson Caillault, D. Hamad, P.-A. Hébert

LISIC - Lab. of Computing, Signal and Image Processing in Côte d'Opale
Université Lille Nord de France, ULCO
62228 Calais, France
Tel: +33 (0)3 21 46 36 91, Fax: +33 (0)3 21 46 57 51

Abstract

Spectral clustering methods meet more and more success in machine learning community thanks to their ability to cluster data points of any complex shapes. The problem of clustering is adressed in terms of finding an embedding space in which the projected data are linearly separable by a classical clustering algorithm such as K-means algorithm. Often, spectral algorithm performances are significantly improved by incorporating prior knowledge in their design, and several techniques have been developed for this purpose. In this paper, we describe and compare some recent linear and non-linear projection algorithms integrating instance-level constraints (”must-link” and ”cannot-link”) and applied for data clustering. We outline a K-way spectral clustering algorithm able to integrate pairwise relationships between the data samples. We formulate the objective function as a combination of the original spectral clustering criterion and the penalization term based on the instance constraints. The optimization problem is solved as a standard eigensystem of a signed Laplacian matrix. The relevance of the proposed

*Corresponding author.
Email address: name@lisic.univ-littoral.fr (G. Wacquet)

algorithm is highlighted using six UCI benchmarks and two public face databases.
Keywords: Graph embedding, Spectral clustering, Pairwise constraints, Signed Laplacian.

1. Introduction

In many real-world applications, we are dealing with the problem of clustering of high dimensional databases for which we have little prior knowledge. Clustering aims to group data sharing similar properties to their respective categories. It was shown that, the introduction of domain knowledge in the clustering algorithms, may greatly improve their performances. Domain knowledge is generally provided in two forms: class labels (Chapelle et al., 2006) or instance constraints (Basu et al., 2008). Labelling data is a hard and long task for human experts while pairwise relationship between data is easier since it consists in simply indicating if two instances are similar (*must-link*) or dissimilar (*cannot-link*) (Wagstaff and Cardie, 2002).

Recently, spectral methods, based on graph concepts, have been developed for dimension reduction and data clustering (Saul et al., 2006; Shortreed and Meila, 2005; Von Luxburg, 2007). They meet more and more success in machine learning community thanks to their theoretical foundations and their practical applications. The problem of data clustering is considered in terms of finding an embedding space in which the projected data are linearly separable by a classical K-means algorithm. The data are represented in a graph where each vertex is associated with a data sample and the weighted edges encode the relationship between the underlying data. Usually, the embedding space is obtained by Laplacian Eigenmaps. This is carried out by selecting the eigenvectors of the graph Laplacian. Each eigen-

vector corresponds to non-linear projection of the data set. The performances of spectral algorithms depend on the way they integrate the data constraints in their design: (a) integration of constraints in the affinity matrix, (b) integration of constraints in the optimization criterion (De Bie et al., 2004; Basu et al., 2008; Wang and Davidson, 2010; Wang et al., 2012). Note that the graph built using instance constraints may contain negatively weighted edges associated to cannot-link constraints. In this situation, the obtained graph is called signed graph and its associated Laplacian matrix is called signed Laplacian matrix (Kunegis et al., 2010).

In this paper, we present and compare recent methods for data projection and clustering, using pairwise relationships, in terms of spectral theory. Among the spectral methods developed in the literature, some include the clustering step in their algorithms (spectral clustering) and others are used for dimension reduction (spectral embedding) (Saul et al., 2006). The latter can be easily used for data clustering by applying a classical K-means algorithm on the projected data. We briefly review the classical principal component analysis (PCA) and the locality preserving projection LPP (He and Niyogi, 2002) as well as their constrained variant. We develop a constrained spectral embedding algorithm for K-way data clustering. The embedding obtained by our approach is closer to the constrained Laplacian Eigenmaps (Chen et al., 2010). The algorithm optimizes an objective function which is a combination of standard spectral clustering criterion and the penalization term based on the instance constraints. The optimization problem is solved as a standard eigensystem of a signed Laplacian matrix. We show the

relevance of the algorithm on many UCI[1] benchmark datasets and two well-known face databases[2].

The paper is organized as follows. Section 2 presents some basic graph notations used for spectral methods. Section 3 describes two constrained linear projection methods: PCA and LPP methods. Spectral clustering approaches integrating implicitly and explicitly the pairwise constraints are presented in Section 4. Section 5 describes the proposed constrained K-way spectral clustering algorithm. Section 6 presents some performance study of the proposed algorithm using six UCI datasets and two public face databases. Finally, Section 7 shows some discussions and concluding remarks.

2. Basic notations of spectral methods

In this section, we present some basic notations used in the graph formalism (Von Luxburg, 2007).

- $\mathcal{X} = \{x_1, \ldots, x_i, \ldots, x_N\}$ is a dataset of N instances, $x_i \in \Re^P$, $i = 1, \ldots, N$;
- $G(V, E, W)$ is a weighted graph associated with $\mathcal{X}$: $V = \{v_1, \ldots, v_i, \ldots, v_N\}$ is the set of vertices corresponding to the N instances; E is the set of edges and W is the weight matrix indicating the affinity or closeness of pairwise instances x_i, x_j where $w_{ij} \geq 0$ and $w_{ij} = w_{ji}$;
- D is the degree matrix of graph G. D is a diagonal matrix where $d_{ii} = \sum_{j=1}^{N} w_{ij}$ is the degree of the vertice v_i.

[1] http://archive.ics.uci.edu/ml/

[2] http://www.cad.zju.edu.cn/home/dengcai/Data/FaceData.html/

Spectral methods consider the K-way clustering problem as a K-way graph-cut into K non-empty groups denoted by $\{V_1, \ldots, V_k, \ldots, V_K\}$ such as $\cup_{k=1}^{K} V_k = V$ and $V_k \cap V_l = \phi$, $k \neq l$:

- Volume of a set V_k: $vol(V_k) = \sum_{v_i \in V_k} d_{ii}$;
- Similarity within cluster V_k: $Cut(V_k, V_k) = \sum_{v_i \in V_k} \sum_{v_j \in V_k} w_{ij}$;
- Similarity between V_k and its complement $\overline{V}_k$: $Cut(V_k, \overline{V}_k) = \sum_{v_i \in V_k} \sum_{v_j \in \overline{V}_k} w_{ij}$.

Let $u_k = (u_{1k}, \ldots, u_{ik}, \ldots, u_{Nk})^T$ be the indicator vector of V_k:

$$u_{ik} = \begin{cases} 1 & \text{if } v_i \in V_k, \\ 0 & \text{if } v_i \notin V_k. \end{cases} \tag{1}$$

Using u_k, the above graph characteristics can be defined more consistently:

$$vol(V_k) = u_k^T D u_k, \tag{2}$$

$$Cut(V_k, V_k) = u_k^T W u_k, \tag{3}$$

$$Cut(V_k, \overline{V}_k) = u_k^T (D - W) u_k = u_k^T L u_k. \tag{4}$$

where:

$$L = D - W, \tag{5}$$

is called graph Laplacian matrix. L is symmetric and positive semi-definite.

Usually, in addition to unlabeled dataset, we have some kind of knowledge known as instance-level constraints: two instances are similar and their vertices "must be linked" or dissimilar and therefore the corresponding vertices "cannot be linked":

- the set of must-links is defined by: $\mathcal{M} = \{(x_i, x_j) | \ x_i \text{ and } x_j \text{ are similar}\}$.
- the set of cannot-links is defined by: $\mathcal{C} = \{(x_i, x_j) | \ x_i \text{ and } x_j \text{ are dissimilar}\}$.

In the context of spectral theory, must-link and cannot-link graphs are built from the pairwise constraints. Usually, the edges of must-link graph have positive weights while the edges of cannot-link graph may have negative weights. Therefore, the graph associated to the unlabelled data and pairwise constraints may contain negative weights. It is called signed graph and its associated Laplacian matrix is called signed Laplacian matrix: $\overline{L} = \overline{D} - W$ where $\overline{d}_{ii} = \sum_{j=1}^{N} |w_{ij}|$ is the degree of the vertex v_i.

In the next section, we present two existing methods for dimensionality reduction based on pairwise constraints: the constrained principal component analysis and the constrained locality preserving projection.

3. Constrained linear projection approaches

In many domains, we often deal with high dimensional datasets. However, all dimensions are not necessary and reducing the input space to a lower space will make the clustering problem not only computationally easier, but also allow to discover the data structure. The performances of projection methods are highly conditioned by the way they integrate the instance-level constraints in their design. In the sequel, we will briefly describe two existing constrained projection techniques based on the classical principal component analysis (PCA) and the more recent locality preserving projection (LPP) (He and Niyogi, 2002).

3.1. Constrained principal component analysis

PCA method performs dimensionality reduction by projecting the input data onto a lower dimensional space spanned by the largest eigenvectors of the data covariance matrix. The problem is to find a linear function between the input data space and the projected data space of the form:

$$y = a^T x, \text{ with } a^T a = 1, \tag{6}$$

which maximizes the objective criterion:

$$J_{PCA} = \frac{1}{N} \sum_{i=1}^{N} (y_i - m)^2, \tag{7}$$

where $m = \frac{1}{N} \sum_{i=1}^{N} y_i$. Equation (7) can also be written as:

$$\begin{aligned} J_{PCA} &= \frac{1}{N^2} \sum_{i,j} (y_i - y_j)^2, \\ J_{PCA} &= \frac{1}{N^2} \sum_{i,j} (a^T x_i - a^T x_j)^2. \end{aligned} \tag{8}$$

The solution a is the eigenvector associated with the largest eigenvalue of the data covariance matrix. Therefore, the PCA space is spanned by the top eigenvectors in which the data are best spread.

The constrained PCA (cPCA) takes into consideration the instance-level constraints sets $\mathcal{M}$ and $\mathcal{C}$. Indeed, the main idea is to look for a direction a such as the projected points $y_i = a^T x_i$ ($a^T a = 1$) satisfies PCA criterion in Equation (8) as well as the instance-level constraints (Zhang et al., 2007). The constrained PCA criterion is defined by:

$$J_{cPCA} = J_{PCA} + \underbrace{\frac{1}{|\mathcal{C}|}\sum_{(x_i,x_j)\in\mathcal{C}}(a^T x_i - a^T x_j)^2}_{\text{cannot-link constraints}} - \underbrace{\frac{1}{|\mathcal{M}|}\sum_{(x_i,x_j)\in\mathcal{M}}(a^T x_i - a^T x_j)^2}_{\text{must-link constraints}}, \quad (9)$$

with $|\mathcal{C}|$ and $|\mathcal{M}|$ are the cardinals of constraint sets $\mathcal{C}$ and $\mathcal{M}$ respectively.

Equation (9) can also be written as:

$$J_{cPCA} = \frac{1}{2}\sum_{i,j}(a^T x_i - a^T x_j)^2 \tilde{w}_{ij}, \quad (10)$$

with:

$$\tilde{w}_{ij} = \begin{cases} \frac{1}{N^2} + \frac{1}{|\mathcal{C}|} & \text{if } (x_i,x_j) \in \mathcal{C}, \\ \frac{1}{N^2} - \frac{1}{|\mathcal{M}|} & \text{if } (x_i,x_j) \in \mathcal{M}, \\ \frac{1}{N^2} & \text{else.} \end{cases} \quad (11)$$

The development of Equation (10) leads to:

$$\begin{aligned} J_{cPCA} &= \sum_{i,j}\left(a^T x_i \tilde{w}_{ij} x_i^T a - a^T x_i \tilde{w}_{ij} x_j^T a\right) \\ J_{cPCA} &= a^T X(\tilde{D} - \tilde{W})X^T a. \end{aligned} \quad (12)$$

where $\tilde{D} \in \mathbb{R}^{N\times N}$ is the diagonal degree matrix of $\tilde{W}$.

The constrained optimization criterion of Equation 12 is:

$$J_{cPCA} = a^T X\tilde{L}X^T a, \ \textbf{s.t.}\ a^T a = 1. \quad (13)$$

The *cPCA* projection space is obtained from the top eigenvectors of the eigen-system:

$$X\tilde{L}X^T a = \lambda a. \quad (14)$$

where $\tilde{L} = \tilde{D} - \tilde{W}$.

(Zhang et al., 2007) used a semi-supervised dimensionality reduction method (SSDR) which introduces penalty terms α and β in order to balance "cannot-link" and "must-link" contributions in the optimization criterion (Equation 9). It is easy to show that, $\alpha = \beta = 0$ lead to a classical PCA criterion. In their experiments, the authors proposed to choose $\alpha = 1$ and $\beta \geq 1$ in order to favour the "must-link" constraints. In (Davidson, 2009; Tang and Zhong, 2007), only pairwise constraints are used to guide the dimensionality reduction for clustering.

3.2. Constrained locality preserving projection

The locality preserving projection (denoted LPP) is a dimensionality reduction method recently used in the literature (He and Niyogi, 2002). LPP constructs the affinity matrix W using a Gaussian kernel:

$$w_{ij} = \begin{cases} e^{-\frac{||x_i - x_j||^2}{2\sigma^2}} & \text{if } x_i \text{ (respectively } x_j\text{) is among the} \\ & \text{nearest neighbors of } x_j \text{ (respectively } x_i\text{),} \\ 0 & \text{otherwise.} \end{cases} \quad (15)$$

with σ is a scale parameter. Note that there are several ways for setting the affinity matrix W. LPP criterion is defined by:

$$J_{LPP} = \sum_{i,j} (a^T x_i - a^T x_j)^2 w_{ij}, \text{ s.t. } a^T X D X^T a = 1. \quad (16)$$

In the objective function J_{LPP}, the penalty contribution w_{ij} is high if neighboring points x_i and x_j are projected far apart. Therefore, minimizing J_{LPP} attempts to ensure that if x_i and x_j are close then $y_i = a^T x_i$ and $y_j = a^T x_j$ are close as well.

A compact form of J_{LPP} using the Laplacian matrix is given by:

$$J_{LPP} = a^T X L X^T a, \text{ s.t. } a^T X D X^T a = 1. \quad (17)$$

The LPP projection space is obtained from the eigenvectors associated with the smallest eigenvalues of the generalized eigensystem:

$$XLX^T a = \lambda XDX^T a. \tag{18}$$

The constrained locality preserving projection method (denoted cLPP) integrates the constraints sets in the objective criterion (Cevikalp and Verbeek, 2008; Yu et al., 2010):

$$J_{cLPP} = J_{LPP} + \underbrace{\sum_{(x_i,x_j)\in\mathcal{M}} (a^T x_i - a^T x_j)^2}_{\text{must-link constraints}} - \underbrace{\sum_{(x_i,x_j)\in\mathcal{C}} (a^T x_i - a^T x_j)^2}_{\text{cannot-link constraints}}. \tag{19}$$

Equation (19) can also be written as:

$$J_{cLPP} = \sum_{i,j} (a^T x_i - a^T x_j)^2 \tilde{w}_{ij}, \tag{20}$$

with:

$$\tilde{w}_{ij} = \begin{cases} w_{ij} + 1 & \text{if } (x_i,x_j) \in \mathcal{M}, \\ w_{ij} - 1 & \text{if } (x_i,x_j) \in \mathcal{C}, \\ w_{ij} & \text{else.} \end{cases} \tag{21}$$

The constrained optimization criterion can be written using the Laplacian matrix:

$$J_{cLPP} = a^T X\tilde{L}X^T a, \text{ s.t. } a^T X\tilde{D}X^T a = 1. \tag{22}$$

where $\tilde{L} = \tilde{D} - \tilde{W}$.

The cLPP algorithm attempts to preserve the locality of data and to satisfy the space-level constraints at the same time.

In the context of clustering, cPCA and cLPP are usually followed by a K-means algorithm. In (Zheng et al., 2004), the concept of locality preservation is used for data clustering.

4. Constrained spectral clustering approaches

In the literature, a number of algorithms have been proposed in order to incorporate instance-level constraints into spectral clustering. They can be grouped into two categories:

- direct integration of pairwise constraints in the affinity matrix (Kamvar et al., 2003; Xu et al., 2005).
- integration of pairwise constraints in the optimization criterion (Wang and Davidson, 2010; Wang et al., 2012).

In the following, we briefly describe the spectral clustering approach.

4.1. Spectral clustering

Spectral clustering method (SC) is usually used in its normalized form (Meila and Shi, 2000; Ng et al., 2002; Shi and Malik, 2000; Shortreed and Meila, 2005; Von Luxburg, 2007). The goal is to use the graph-cut in order to partition the data into K clusters. The objective function of spectral clustering is to find a vector u which minimizes the following criterion:

$$J_{SC} = \sum_{i,j}(u_i - u_j)^2 w_{ij}, \tag{23}$$

which can be written as:

$$J_{SC} = u^T L u, \tag{24}$$

under the constraints: $u^T D u = 1$ and $Du \perp 1$ where u is the relaxed cluster indicator vector (i.e, the components of u can have real values). For simplicity of notation, we substitute u by $D^{-\frac{1}{2}}z$ and the Equation becomes:

$$J_{SC} = z^T \mathcal{L} z, \ \textbf{s.t.}\ z^T z = 1. \tag{25}$$

where $\mathcal{L} = D^{-\frac{1}{2}} L D^{-\frac{1}{2}}$ is the normalized Laplacian matrix.

The problem solution is given by:

$$\mathcal{L}z = \lambda z, \tag{26}$$

The spectral space associated to the K smallest positive eigenvalues is similar to the one obtained by the Laplacian eigenmap developed in (Belkin and Niyogi, 2002).

4.2. Integration of constraints in the affinity matrix

In (Kamvar et al., 2003), the authors adapt the spectral clustering algorithm to constrained classification problem: spectral learning algorithm (SL). They incorporate pairwise constraints into the affinity matrix:

$$\tilde{w}_{ij} = \begin{cases} 0 & \text{if } (x_i, x_j) \in \mathcal{C}, \\ +1 & \text{if } (x_i, x_j) \in \mathcal{M}, \\ w_{ij} & \text{otherwise.} \end{cases} \tag{27}$$

The spectral learning algorithm proceeds just as any other standard spectral clustering algorithm. However, the main weakness of this algorithm is that it implicitly encodes the constraints by modifying the graph Laplacian matrix. A more natural approach is to preserve the original graph Laplacian and to explicitly encode the constraints.

4.3. Integration of constraints in the optimization criterion

In (Wang and Davidson, 2010), the authors combine spectral clustering and pairwise constraints criteria in a flexible manner. The flexible constrained spectral clustering (FCSC) preserves the original graph Laplacian matrix and explicitly

encodes the constraints. FCSC is solved by a generalized eigenvalue system. This approach includes a user-specified parameter α which serves as a tradeoff factor between the structure defined by the graph Laplacian and that by the constraint matrix.

The FCSC algorithm is detailed with $K = 2$ (Wang and Davidson, 2010). The constraints matrix is defined by:

$$q_{ij} = \begin{cases} -1 & \text{if } (x_i, x_j) \in \mathcal{C}, \\ +1 & \text{if } (x_i, x_j) \in \mathcal{M}, \\ 0 & \text{else.} \end{cases} \tag{28}$$

In order to measure how well the constraints are satisfied by the cluster assignment, Wang and Davidson used:

$$u^T Q u = \sum_{i,j} u_i u_j q_{ij}. \tag{29}$$

where $u \in \{-1, +1\}^N$ is the cluster indicator vector.

The problem is then formulated as a constrained optimization problem, letting $z = D^{\frac{1}{2}} u$ and $Q_n = D^{-\frac{1}{2}} Q D^{-\frac{1}{2}}$:

$$\arg\min_z z^T \mathcal{L} z, \ \textbf{s.t.} \ z^T Q_n z \geq \alpha, z^T z = vol(G), z \neq D^{\frac{1}{2}} \mathbf{1}, \tag{30}$$

where $vol(G) = \sum_{i=1}^{N} d_{ii}$. Recently, the authors generalized the above bipartition method to a K-way constrained clustering by selecting not only the first, but the top-K generalized eigenvectors corresponding to positive eigenvalues (Wang et al., 2012).

5. Constrained spectral clustering

In this section, we develop our constrained spectral clustering algorithm (denoted cSC). Here, the objective function J_{cSC} consists in the combination of the classical spectral clustering criterion (J_{SC}) and a penalization term based on the instance constraints ($J_{C\mathcal{M}}$):

$$J_{cSC} = \gamma . J_{SC} + (1-\gamma) . J_{C\mathcal{M}}, \tag{31}$$

The regularization coefficient γ has to be adjusted in order to balance the contribution of J_{SC} and $J_{C\mathcal{M}}$, where:

$$J_{SC} = \sum_{i,j} (u_i - u_j)^2 w_{ij} = u^T L u, \tag{32}$$

and

$$J_{C\mathcal{M}} = - \sum_{(x_i,x_j)\in\mathcal{C}} (u_i - u_j)^2 + \sum_{(x_i,x_j)\in\mathcal{M}} (u_i - u_j)^2. \tag{33}$$

$J_{C\mathcal{M}}$ can be written as:

$$J_{C\mathcal{M}} = \sum_{i,j} (u_i - u_j)^2 q_{ij}, \tag{34}$$

with:

$$q_{ij} = \begin{cases} -1 & \text{if } (x_i,x_j) \in \mathcal{C}, \\ +1 & \text{if } (x_i,x_j) \in \mathcal{M}, \\ 0 & \text{else.} \end{cases} \tag{35}$$

We can rewrite $J_{C\mathcal{M}}$ as:

$$J_{C\mathcal{M}} = u^T L_Q u, \tag{36}$$

where L_Q is the Laplacian matrix of constraints graph:

$$L_Q = D_Q - Q, \tag{37}$$

with D_Q is the degree matrix of constraints graph.

Using Equations (32) and (36), Equation (31) becomes:

$$J_{cSC} = u^T(\gamma.L + (1-\gamma).L_Q)u = u^T L_{cSC} u. \tag{38}$$

where:

$$L_{cSC} = D_{cSC} - W_{cSC}, \tag{39}$$

with:

$$D_{cSC} = \gamma.D + (1-\gamma).D_Q, \tag{40}$$

and

$$W_{cSC} = \gamma.W + (1-\gamma).Q. \tag{41}$$

The constrained spectral space is obtained from the eigenvectors of the Laplacian matrix $L_{cSC} = \gamma L + (1-\gamma)L_Q$.

Note that the eigenvalues of L_{cSC} may have negative sign due to the negative edges weights of cannot-links constraints. In order to overcome this limitation, we use the signed Laplacian matrix defined by (Kunegis et al., 2010):

$$\overline{L}_{cSC} = \overline{D}_{cSC} - W_{cSC}, \tag{42}$$

where $\overline{D}_{cSC}$ is the signed degree matrix given by:

$$\overline{d}_{cSC}(i,i) = \sum_{j=1}^{N} |w_{cSC}(i,j)|, \tag{43}$$

Note that the signed Laplacian matrix is semi-definite positive. By substituting u by $\overline{D}_{cSC}^{\frac{-1}{2}} z$ to relax the discreteness condition, Equation (38) becomes:

$$J_{cSC} = z^T \overline{L}_{cSC} z, \tag{44}$$

where $\overline{\mathcal{L}}_{cSC}$ is the normalized signed Laplacian matrix defined as $\overline{\mathcal{L}}_{cSC} = \overline{D}_{cSC}^{\frac{-1}{2}} \overline{L}_{cSC} \overline{D}_{cSC}^{\frac{-1}{2}}$. The constrained spectral space is obtained from the K lowest eigenvectors of $\overline{\mathcal{L}}_{cSC}$.

It is interesting to know that, in case $K = 2$, the retained solution is the second smallest eigenvector. Indeed, the first vector (z_1) is constant and represents a trivial solution for $\lambda = 0$. The final partition is then obtained by partitioning the projected data thanks to the sign of values in z_2.

In case $K > 2$, we maintain the usage of K eigenvectors, considering that the constant vector $z_1 = (1, \ldots, 1)^T$ has no impact on the obtained spectral subspace. These K eigenvectors are then used in order to cluster the data thanks to the K-means algorithm. Input instances are assigned to their corresponding clusters obtained in the constrained spectral space.

In (De Bie et al., 2004), the authors proposed a softly constrained spectral clustering using a regularization term similar to the one of equation (31). In this work, the constraint matrix is used in order to constrain the projection of data according to their labels. However, the constraints matrix used in multiclass learning do not include the cannot link constraints.

In the context of kernel machines, Alzate and Suykens (Alzate and Suykens, 2009, 2010, 2012) revised the spectral clustering in terms of weighted kernel PCA using the least square support vector machines developed by (Suykens and Vandewalle, 1999). Indeed, Alzate et al. present interesting weighted kernel PCA approaches to deal with the framework of spectral clustering. Furthermore, in (Alzate and Suykens, 2009) the authors formulate weighted kernel PCA with pairwise constraints which leads to a constrained spectral algorithm. Their contribution is of upmost importance since it integrates, using the kernel concept, the out-of-sample extension in a natural way allowing model selection and general-

ization capabilities. However, the regularization term of cannot-link constraints is chosen much smaller than the regularization term of must-link constraints in order to avoid negative entries in the equivalent kernel matrix due to rank-1 downdates. The constrained spectral clustering approach, presented in our paper, does not require positive eigenvalues of the graph Laplacian and can deal with negative eigenvalues, which may occur due to cannot-link constraints. Moreover, the regularization coefficient in equation (31), weights the contribution of spectral clustering and must-link and cannot-link constraints. It can take any value in its interval. Thus, our approach can be applied using only pairwise constraints (regularization coefficient = 0) without the use of unlabelled data. Note that, for a regularization term equal to 1, we obtain the classical spectral clustering. It should be noted that, our approach is closer to Constrained Laplacian Eigenmap (Chen et al., 2010) while that of (Alzate and Suykens, 2009) is derived from kernel PCA and therefore, the out of sample is obtained naturally. We stress the fact that the objective of our presented work is the use constrained spectral clustering without seeking a solution for the out-of-sample problem.

6. Experiments

We propose to compare our constrained spectral clustering algorithms with cPCA (Zhang et al., 2007), cLPP (Cevikalp and Verbeek, 2008), SL (Kamvar et al., 2003) and FCSC (Wang and Davidson, 2010) algorithms described in Sections 3 and 4.

6.1. UCI databases

In this section, we evaluate the performances of the proposed constrained spectral clustering (cSC) algorithm and compare it with the presented algorithms on

six UCI databases (”Hepatitis”, ”Ionospehere”, ”Wine”, ”Dermatology”, ”Glass” and ”Ecoli”). These databases were chosen because they represent data of different sizes and different densities. Table 1 summarizes the characteristics of each database.

In our experiments, we evaluate the performances of the several clustering algorithms described in the Sections 3, 4 and 5. For linear methods (cPCA and cLPP) we used 95% of the total variance which made the maximum dimension for each dataset as follows: Hepatitis (3), Ionosphere (24), Wine (2), Dermatology (3), Glass (5) and Ecoli (5). For non linear methods, we adopted the classical strategy by keeping the first K eigenvectors as projection coordinates. In our experiments, cPCA and cLPP are followed by K-means algorithm for data clustering in the projection space. For cLPP, we set the number of neighbors to 10. For FCSC, the parameter measuring the constraints satisfaction and denoted by α is set to $0.5 \times \lambda_{\max} vol(G)$ (Wang and Davidson, 2010). Moreover, the projected data are normalized to have unit-length. For our cSC, we simply set the balancing parameter γ to 0.5.

The affinity matrix terms are of Gaussian form with a scale parameter σ^2 equal to the average of the variances of database features. The generation procedure of pairwise constraints is as follows: we randomly select pairs of instances from the dataset and create ”must-link” or ”cannot-link” constraints depending on whether the two instances belong or not to the same class. We iterate this scheme and enrich the generated constraint sets from 0% constraints (unlabelled data) to 20% constraints. The performances of algorithms are averaged over 10 repetitions of the constraints generation process.

Each clustering algorithm generates cluster label for each data instance and the

Table 1: UCI datasets used for experiments.

	Nb. Objects	Nb. Features	Nb. Classes (K)
Hepatitis	80	19	2
Ionosphere	351	34	2
Wine	178	13	3
Dermatology	366	34	6
Glass	214	9	6
Ecoli	336	7	8

clustering performance of the algorithm is evaluated by comparing the generated class label and the ground-truth label. Given a data point x_i, let $\hat{k}_i$ be the obtained cluster label and k_i the ground-truth label respectively. The agreement between the algorithm decision and the ground-truth is measured by $\delta(k_i, map(\hat{k}_i))$ (Carpaneto and Toth, 1980):

$$\delta(k_i, map(\hat{k}_i)) = \begin{cases} 1 & \text{if } k_i = map(\hat{k}_i), \\ 0 & \text{otherwise.} \end{cases} \tag{45}$$

In order to evaluate the performance of the presented spectral algorithms, we propose to use two criteria: the Accuracy and the Rand Index (Rand, 1971).

The Accuracy criterion is defined by:

$$\text{Accuracy} = \sum_{i=1}^{N} \frac{\delta(k_i, map(\hat{k}_i))}{N}, \tag{46}$$

The Rand Index is given by:

$$\text{Rand Index} = \frac{\text{number of correct decisions}}{\text{number of total decisions}}. \tag{47}$$

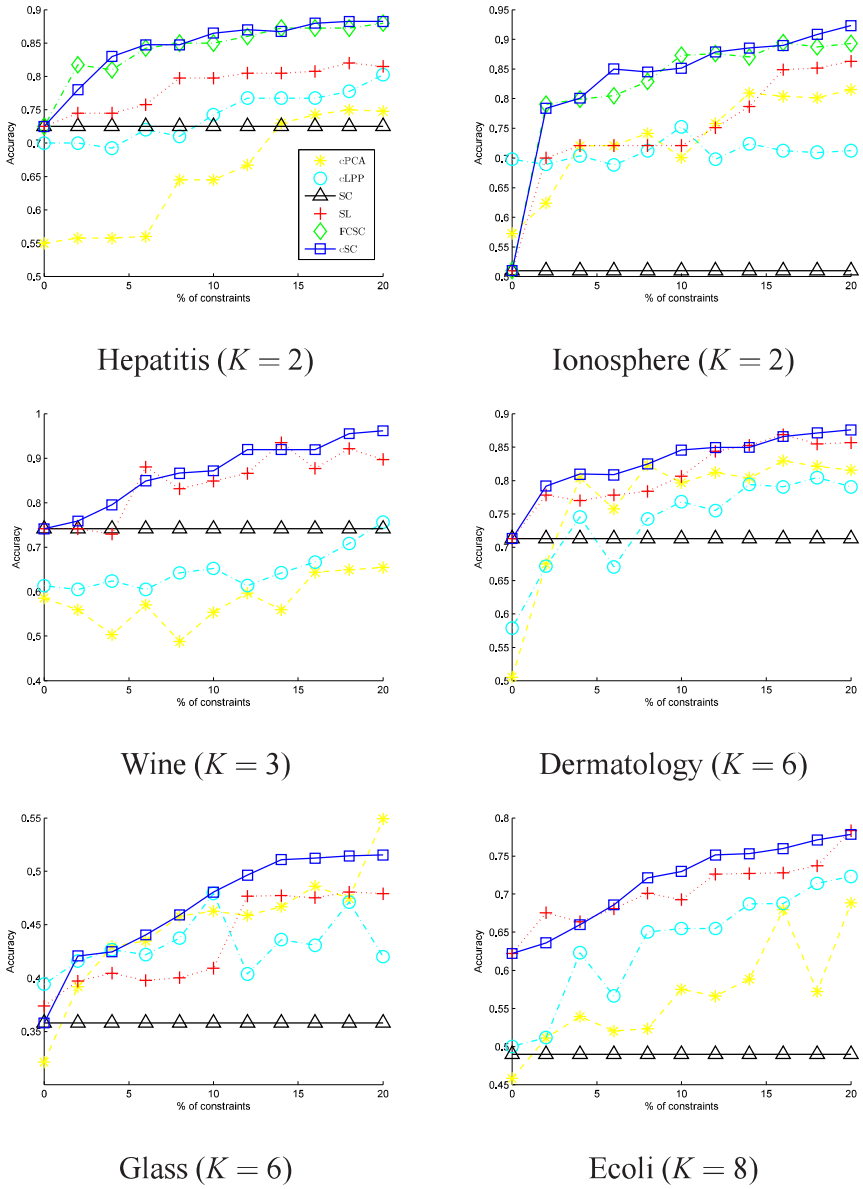
A decision is considered correct if the proposed clustering obtained by an algorithm agrees with the target clustering. More specifically, a decision is considered as correct if: two instances are in the same cluster and the algorithm partitioned them into a same cluster, or they are in different clusters and partitioned into different clusters. Rand Index values ranges from 0 to 1.

Figure 1 shows the Accuracy of the spectral algorithms depending on the constraints rates applied on the six UCI datasets. In multiclass problems, FCSC is not represented because it does not allow to obtain a feasible solution with the fixed value of θ. We can see that cSC algorithm outperforms all algorithms followed by SL algorithm. Moreover, for all databases ($K = 2$ and $K > 2$), their accuracies increase as the constraints rates increase. Globally, cPCA and cLPP produce the worst performances, specifically for the "Wine" database.

Figure 2 summarizes the Rand Index of the six spectral algorithms according to the constraints rates applied on the six databases. It is easy to see that cSC algorithm outperforms all other methods, followed by SL algorithm. Moreover, their Rand Indices increase as the constraints rates increase. The Rand Index of SC has a constant value since it does not depend on the constraints rates. cSC obtains the better Rand Indices which can be explained by the fact that, unlike SL algorithm, cSC takes into account the spatial position of data (w_{ij}) and their constraints $(-1, +1)$. Finally, Rand Indices of cPCA and cLPP which are linear projection algorithms are lower than the Rand Indices of constrained non-linear projection algorithms.

6.2. ORL and Yale Face databases

We compare the performances of the proposed algorithm with that of the presented algorithms on two well-known databases used in face recognition domain:



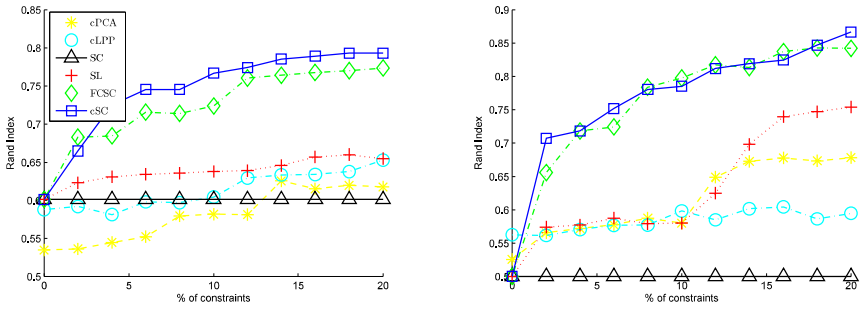
Hepatitis ($K = 2$) Ionosphere ($K = 2$)

Wine ($K = 3$) Dermatology ($K = 6$)

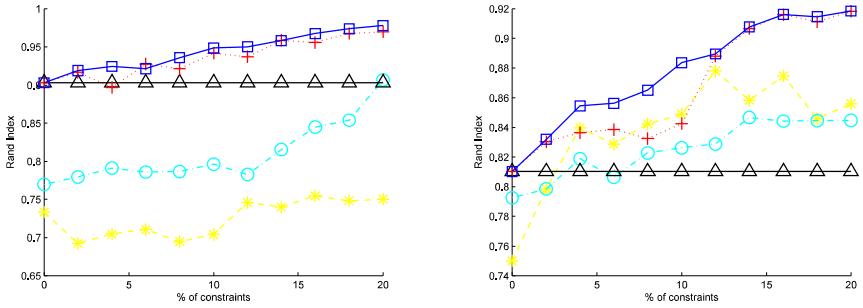
Glass ($K = 6$) Ecoli ($K = 8$)

Figure 1: Average Accuracy according to the constraints rates, on UCI datasets.

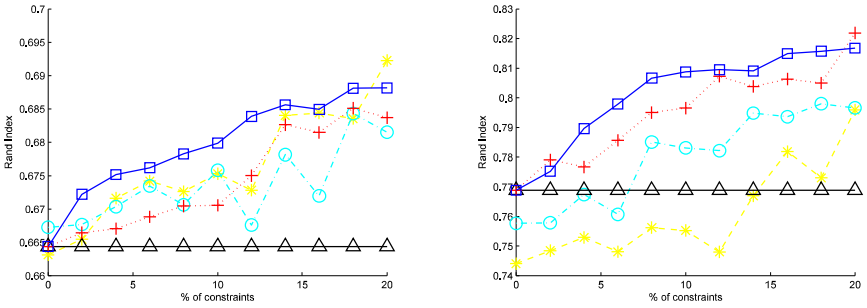
- ”Yale Face database” contains 165 grayscale images of 15 individuals. There are 11 images per subject, one per different facial expression or configura-



Hepatitis ($K = 2$) Ionosphere ($K = 2$)



Wine ($K = 3$) Dermatology ($K = 6$)



Glass ($K = 6$) Ecoli ($K = 8$)

Figure 2: Average Rand Index according to the constraints rates, on UCI datasets.

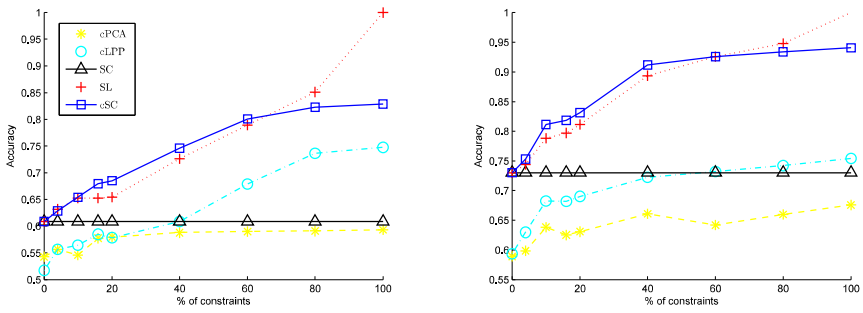
345 tion;

- ”ORL database” contains 10 different images of each of 40 distinct subjects. For some subjects, the images were taken at different times, varying the lighting, facial expressions and facial details.

Table 2 summarizes the characteristics of each database. The affinity matrix terms are of Gaussian form with a scale parameter σ^2 computed in a local way (Zelnik-Manor and Perona, 2004). The experimental protocol used is the same as in Section 6.1. For linear methods (cPCA and cLPP) we used 95% of the total variance which made the maximum dimension for each dataset as follows: Yale Face (71) and ORL (115). For non linear methods, we adopted the classical strategy by keeping the first K eigenvectors as projection coordinates.

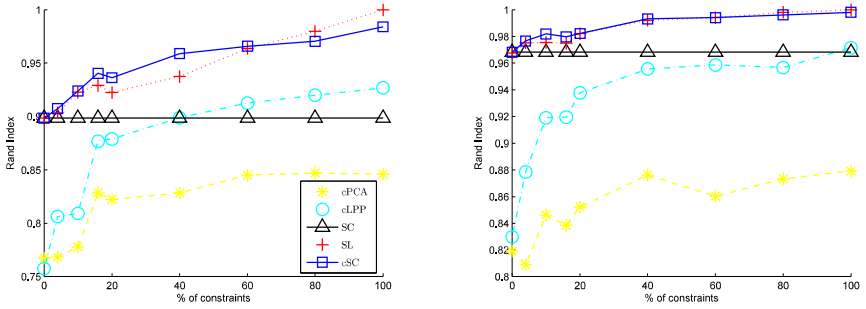
Table 2: Other datasets used for experiments.

	Nb. Objects	Nb. Features	Nb. Classes (K)
Yale Face	165	1024	15
ORL	400	1024	40



Yale Face ($K = 15$) ORL ($K = 40$)

Figure 3: Average Accuracy according to the constraints rates, on other datasets.



Yale Face ($K = 15$) ORL ($K = 40$)

Figure 4: Average Rand Index according to the constraints rates, on other datasets.

Figure 3 and 4 show respectively the Accuracy and the Rand Index of the spectral algorithms according to the constraints rates, applied on the two databases. As for UCI datasets, when the number of constraints is low, it is easy to see that cSC algorithm outperforms all algorithms, followed by SL algorithm. Moreover, for this two examples, Rand Indices and Accuracies of the linear projection algorithms (cPCA and cLPP) are lower than the performance scores obtained by the classical spectral clustering even with 20% of constraints. When the percentage of constraints is high, SL algorithm obtains the best scores. One possible explanation is that, in our experiments, the chosen γ was arbitrarily set to 0.5, so that label information was not fully exploited in the global criterion.

7. Discussions and conclusions

In this work, we presented recent linear and non-linear projection algorithms integrating constraints and applied for data clustering. PCA and LPP are linear dimension reduction methods. They find linear relationship between input data space and output projection space. PCA approach is sensitive to global data dis-

371 persion, while LPP is sensitive to local data variation. Constrained PCA (cPCA)
372 and constrained LPP (cLPP) inherit the same properties of PCA and LPP meth-
373 ods. However, contrary to cPCA, the cLPP encourages nearby data input to be
374 projected to nearby outputs, and penalizes neighboring points if they are projected
375 far apart.

376 It would be important to note that, the main step in spectral clustering method
377 is to project the data in a non linear manner and therefore the resulting spectral
378 space is similar to the one obtained by Laplacian eigenmap method (Belkin and
379 Niyogi, 2002). Indeed, spectral clustering and LPP methods optimize the same
380 objective function. However, unlike SC, the LPP method finds a linear relation-
381 ship between input data space and output projected data space. Thus, the proposed
382 method and cLPP methods inherit the same properties of SC and LPP methods.

383 Spectral clustering approaches encode instance-level constraints implicitly in
384 the affinity matrix (Kamvar et al., 2003) or explicitly in the optimization criterion
385 (Wang and Davidson, 2010). The main weakness of the spectral learning is that
386 it implicitly encodes ”must-link” and ”cannot-link” constraints by modifying the
387 Laplacian matrix. A more natural approach is to preserve the original Laplacian
388 matrix and to explicitly encode the constraints in the optimization criterion.

389 (Wang et al., 2012) proposed a smart approach which combines spectral clus-
390 tering and pairwise constraints in a flexible manner. Their FCSC algorithm allows
391 the violation of some constraints by introducing a lower-bound of satisfied con-
392 straints in the optimization criterion. However, the associated Laplacian matrix
393 may contain negative eigenvalues and lead sometimes to no solution.

394 In this work, we proposed an efficient constrained spectral clustering algo-
395 rithm which balances the unlabelled data contribution with the pairwise relation-

ships and compared its performance to the recent constrained clustering algorithms. Many UCI benchmark databases and face recognition databases (ORL and Yale Face datasets) have been used to demonstrate the relevance of the proposed algorithm compared to the most known algorithms.

In our experiments, the constraints have been randomly generated based on labelled data. However, it may happen that some generated constraints are redundant or inconsistent which may deteriorate the performance of classification algorithms (Davidson et al., 2006). Therefore, the constraints should be generated in an intelligent manner, and guided by a human expert.

Acknowledgement. The authors are very grateful to the editors and reviewers for their valuable comments and suggestions.

References

Alzate C., Suykens J.A.K., 2009. A Regularized Formulation for Spectral Clustering with Pairwise Constraints. Proceedings of the 2009 International Joint Conference on Neural Networks (IJCNN'09), Atlanta, U.S.A, pp. 141-148.

Alzate C., Suykens J.A.K., 2010. Multiway Spectral Clustering with Out-of-Sample Extensions through Weighted Kernel PCA. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 32, pp. 335-347.

Alzate C., Suykens J.A.K., 2012. A Semi-Supervised Formulation to Binary Kernel Spectral Clustering. Proceedings of the 2012 IEEE World Congress on Computational Intelligence (IEEE WCCI/IJCNN 2012), Brisbane, Australia.

Basu S., Davidson I., Wagstaff K., 2008. Constrained Clustering: Advances in

Algorithms, Theory and Applications. Chapman and Hall, CRC Press, part of the Data Mining and Knowledge Discovery Series.

Belkin M., Niyogi P., 2002. Laplacian Eigenmaps for Dimensionality Reduction and Data Representation. Neural Computation (vol. 15), pp. 1373-1396.

Cai D., He X., Han J., 2005. Document Clustering Using Preserving Indexing. IEEE Transactions on Knowledge and Data Engineering (vol. 17), pp. 1624-1637.

Carpaneto G., Toth P., 1980. Algorithm 548: solution of assignment problem. ACM Transactions on Mathematical Software.

Cevikalp H., Verbeek J., 2008. Semi-supervised dimensionality reduction using pairwise equivalence constraints. International Conference on Computer Vision Theory and Applications, pp. 489-496.

Chapelle O., Scholkopf B., Zien A., 2006. Semi-Supervised Learning. MIT Press, Cambridge MA.

Chen C., Zhang L., Bu J., Wang C., Chen W., 2010. Constrained laplacian eigenmap for dimensionality reduction. Neurocomputing Journal, pp. 951-958.

Davidson I., Wagstaff K., Basu S., 2006. Measuring constraint set utility for partitional clustering algorithms. Proceedings of the European Conference on Principles and Practice of Knowledge Discovery in Databases, pp. 115-126.

Davidson I., 2009. Knowledge Driven Dimension Reduction for Clustering. International Joint Conference on Artificial Intelligence, pp. 1034-1039.

De Bie T., Suykens J.A.K., De Moor B., 2004. Learning from general label constraints. Proceedings of the joint IAPR international workshops on Syntactical and Structural Pattern Recognition and Statistical Pattern Recognition (SSSPR 2004), Lisbon, Portugal, vol. 3138.

Han J., Kamber M., Pei J., 2011. Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers.

He X., Niyogi P., 2002. Locality preserving projections. Computer and Information Science, pp. 153-160.

He X., Yan S., Hu Y., Niyogi P., Zhang H., 2005. Face Recognition Using Laplacianfaces. IEEE Transactions on Pattern Analysis and Machine Inetlligence (vol. 27), pp. 328-340.

Kamvar S., Klein D., Manning C., 2003. Spectral Learning. International Joint Conference on Artificial Intelligence, pp. 561-566.

Kunegis J., Schmidt S., Lommatzsch A., Lerner J., De Luca E., Albayrak S., 2010. Spectral analysis of Signed Graphs for Clustering, Prediction and Visualization. SIAM, pp. 559-570.

Meila M., Shi J., 2000. Learning segmentation by random walks. Neural Information Processing Systems NIPS12, pp. 873-879.

Ng A., Jordan M., Weiss Y., 2002. On spectral clustering: Analysis and an algorithm. Neural Information Processing Systems NIPS14, pp. 849-856.

Rand W., 1971. Objective criteria for the evaluation of clustering methods. Journal of the American Statistical Association, pp. 846-850.

462 Saul L., Weinberger K., Sha F., Ham J., Lee D., 2006. Spectral Methods for Di-
463 mensionality Reduction. Semi-Supervised Learning book, MIT Press, pp. 267-
464 282.

465 Shi J., Malik J., 2000. Normalized cuts and image segmentation. IEEE Transac-
466 tions on Pattern Analysis and Machine Intelligence, pp. 888-905.

467 Shortreed S., Meila M., 2005. Unsupervised spectral learning. Proceedings of the
468 Twenty-First Conference Annual on Uncertainty in Artificial Intelligence, pp.
469 543-416.

470 Suykens J.A.K., Vandewalle J., 1999. Least squares support vector machine clas-
471 sifiers. Neural Processing Letters, vol. 9(3), pp. 293-300.

472 Tang W., Zhong S., 2007. Pairwise Constraints-Guided Dimensionality Reduc-
473 tion. Computational Methods of Feature Selection, Chapman and Hall, CRC
474 2007, pp. 295-312.

475 Von Luxburg U., 2007. A Tutorial on Spectral Clustering. Statistics and Comput-
476 ing (Vol.17, Issue 4), pp. 395-416.

477 Wagstaff K., Cardie C., 2002. Clustering with Instance-level Constraints. Interna-
478 tional Conference on Machine Learning, pp. 1103-1110.

479 Wang X., Davidson I., 2010. Flexible Constrained Spectral Clustering. Interna-
480 tional Conference on Knowledge Discovery and Data Mining, pp. 563-572.

481 Wang X., Qian B., Davidson I., 2012. On Constrained Spectral Clustering and Its
482 Applications. arXiv: 1201.5338.

Xu Q., DesJardins M., Wagsatff K., 2005. Constrained spectral clustering under a local proximity structure assumption. FLAIRS Conference, pp. 866-867.

Yu G., Peng H., Wei J., Ma Q., 2010. Robust locality preserving projections with pairwise constraints. Joint Symposium on Information Systems, pp. 1631-1636.

Zelnik-Manor L., Perona P., 2004. Self tuning spectral clustering. Advances in Neural Information Processing Systems, pp. 1601-1608.

Zhang D., Zhou Z.-H., Chen S., 2007. Semi-supervised dimensionality reduction. Seventh International Conference on Data Mining, pp. 629-634.

Zheng X., Cai D., He X., Ma W.-Y., Lin X. 2004. Locality preserving clustering for image database. Proceedings of the ACM Conference on Multimedia. ACM Press, pp 885-891.

View publication stats