SEMI-SUPERVISED K-WAY SPECTRAL CLUSTERING USING PAIRWISE CONSTRAINTS

Guillaume Wacquet, Pierre-Alexandre Hébert, Émilie Caillault Poisson and Denis Hamad
Université Lille Nord de France, F-59000 Lille, France
Laboratoire d'Informatique, Signal et Image de la Côte d'Opale
ULCO, 50 rue Ferdinand Buisson, B.P. 699, F-62228 Calais Cedex, France

Keywords: K-way spectral clustering, Semi-supervised classification, Pairwise constraints.

Abstract: In this paper, we propose a semi-supervised spectral clustering method able to integrate some limited supervisory information. This prior knowledge consists of pairwise constraints which indicate whether a pair of objects belongs to a same cluster (*Must-Link* constraints) or not (*Cannot-Link* constraints). The spectral clustering then aims at optimizing a cost function built as a classical *Multiple Normalized Cut* measure, modified in order to penalize the non-respect of these constraints. We show the relevance of the proposed method with an illustrative dataset and some UCI benchmarks, for which two-class and multi-class problems are dealt with. In all examples, a comparison with other semi-supervised clustering algorithms using pairwise constraints is proposed.

1 INTRODUCTION

The term ”spectral clustering” refers to a family of unsupervised clustering algorithms. It is more and more used thanks to its effectiveness and its simplicity of implementation which comes down to the eigenvectors extraction of a similarity matrix computed on the dataset (Ng et al., 2002). Similarity matrix gathers the complete information used by the method, telling for each pair of instances how close they are. Contrary to some traditional clustering algorithms such as K-means algorithm, the spectral clustering method allows to deal with ”non-globular” clusters of points.

In recent years, methods incorporating prior knowledge in their clustering process have emerged as relevant and effective in several applications, such as image segmentation (Meila and Shi, 2000), information retrieval or document analysis (Han and Kamber, 2006). The prior knowledge is generally provided in two forms: class labels, and pairwise constraints. Labelling data is a hard and long task. Pairwise constraints simply indicate if two instances must be in the same cluster (*Must-Link*) or not (*Cannot-Link*). They are easier to collect from experts than labels (Wagstaff and Cardie, 2000). However, few works take an interest in semi-supervised methods allowing to deal with multiclass problems ($K \geq 2$). Indeed, recent algorithms mainly focus on two-class problems ($K = 2$).

In this paper, we propose a new algorithm able to integrate constraints in the multiclass spectral clustering process, using a penalty term in a way similar to the constrained Principal Components Analysis (Zhang et al., 2007) used in dimension reduction. The proposed algorithm aims at minimizing the MNCut (*Multiple Normalized Cut*) criterion, while penalizing the non-respect of the given set of constraints. Moreover, a convenient weight, easily interpretable, is introduced in order to balance the MNCut and the penalty term, i.e. the impact of the original structure of the data and the contribution of the constraints. This method is compared with two recent algorithms, and some proposed variants, on an artificial sample and UCI datasets (*http://archive.ics.uci.edu/ml/*). The results are finally presented, for different proportions of known constrained pairs.

The paper is organized into three sections. The first one is theoretical and presents the spectral clustering algorithms and some semi-supervised methods dealing with pairwise constraints. The second one presents our semi-supervised K-way spectral clustering method. The last section assesses the performances of our method and some recent algorithms using synthetic dataset and public databases extracted from UCI repository.

Wacquet G., Hébert P., Caillault Poisson É. and Hamad D..
SEMI-SUPERVISED K-WAY SPECTRAL CLUSTERING USING PAIRWISE CONSTRAINTS.
DOI: 10.5220/0003682500720081
In *Proceedings of the International Conference on Neural Computation Theory and Applications* (NCTA-2011), pages 72-81
ISBN: 978-989-8425-84-3

2 STATE OF ARTS: K-WAY SPECTRAL CLUSTERING, PAIRWISE CONSTRAINTS

2.1 Graph Embedding and MNCut

Spectral clustering is generally considered as a clustering method aiming at minimizing a *Normalized Cut* criterion between $K = 2$ clusters (NCut), or a *Multiple Normalized Cut* between $K \geq 2$ clusters (MNCut) (Meila and Shi, 2000)(Ng et al., 2002)(Shi and Malik, 2000). The first measure, NCut, assesses how strongly a cluster of points (or vertices in a graph) is linked to the other points, in relation to its own cohesion. The second one deals with multiple clusters ($K \geq 2$) and is set to the average of the NCut measures over the whole clusters.

2.1.1 Notations

In order to prepare the NCut minimization problem formulations, some notations are first introduced, using an usual graph formalism.

- Let $\mathcal{X} = \{x_1, \ldots, x_i, \ldots, x_N\}$ be a set of N objects, to be clustered;
- this set $\mathcal{X}$ is described by a weighted graph $G(V,E,S)$: V is the set of nodes corresponding to the objects; E is the set of edges between the nodes and S is a weights matrix whose elements $S_{ij} = S_{ji} \geq 0$ tell how strongly related (or close) objects x_i and x_j are;
- let D be the degree matrix of graph G, i.e. a diagonal matrix whose components are equal to the degrees of the nodes: $D_{ii} = \sum_{j=1}^{N} S_{ij}$;
- let $C = \{C_1, \ldots, C_K\}$ be a partitioning of $\mathcal{X}$ into non-empty disjoint K subsets;
- each group C_k is described by its volume $Vol(C_k) = \sum_{x_i \in C_k} D_{ii}$ and its "cohesion" degree $Cut(C_k, C_k) = \sum_{x_i \in C_k} \sum_{x_j \in C_k} S_{ij}$;
- the Cut between two groups is defined by $Cut(C_k, C_{k'}) = \sum_{x_i \in C_k} \sum_{x_j \in C_{k'}} S_{ij}$.

2.1.2 MNCut Minimisation as Eigenproblem

In a two-class problem, the *Normalized Cut* between subsets C_1 and C_2 is defined as:

$$NCut(C_1, C_2) = Cut(C_1, C_2)\left(\frac{1}{Vol(C_1)} + \frac{1}{Vol(C_2)}\right). \tag{1}$$

In a K-way clustering problem, NCut criterion is generalized by the *Multiple Normalized Cut* (MNCut):

$$MNCut(C) = \sum_{k=1}^{K} \frac{Cut(C_k, C \backslash C_k)}{Vol(C_k)} = \sum_{k=1}^{K}\left(1 - \frac{Cut(C_k, C_k)}{Vol(C_k)}\right) \tag{2}$$

Many authors of spectral clustering algorithms have shown that the minimization of MNCut criterion can be achieved by solving an eigenvalue system (or generalized eigenvalue system). Their optimal clustering processing can be resumed in three steps:

1. Computation and normalization of the similarity matrix S. The result is generally a normalized Laplacian matrix $\overline{L}$.
2. Spectral Mapping. Some K vector solutions of an eigenvalue system such as $\overline{L}z_k = \lambda_k z_k$ based on the matrix issued from Step 1, are computed to form the matrix $Z = [z_1, z_2, \ldots, z_K]$. If the eigenvalues are not distinct, the eigenvectors are chosen such that $z_i^T D z_j = 0 \;\; for\ i \neq j$. Z is then normalized into a matrix U, whose each i-th row is used to map object x_i.
3. Partioning. A grouping algorithm like K-means clusters the points in the spectral space, and assigns the obtained clusters to the corresponding objects.

Now, some usual spectral algorithms are described, in order to illustrate both paradoxal aspects: the quasi-equivalence of their solutions, and the difference between the formalisms they adopt.

$K = 2$. **Shi and Malik.** In their paper (Shi and Malik, 2000), the authors define the indicator vector of cluster C_1 as $u \in \{-1, 1\}^N$: $u_i = 1 \Leftrightarrow x_i \in C_1$. NCut criterion is then written as:

$$NCut(G, u) = \frac{\sum_{x_i > 0, x_j < 0} -u_i u_j S_{i,j}}{\sum_{x_i > 0} D_{i,i}} + \frac{\sum_{x_i < 0, x_j > 0} -u_i u_j S_{i,j}}{\sum_{x_i < 0} D_{i,i}}. \tag{3}$$

With variable change $v = (\mathbf{1} + u) - b(\mathbf{1} - u)$ with $b = \sum_{x_i > 0} D_{ii} / \sum_{x_i < 0} D_{ii}$, infering both conditions $v_i \in \{1, -b\}$ and $v^T D \mathbf{1} = 0$, the above equation becomes a Rayleigh quotient:

$$\min_v NCut(G, v) = \min_v \frac{v^T (D - S) v}{v^T D v}. \tag{4}$$

By relaxing the constraints on indicator vector u' to take on real values, the minimization is obtained by solving the generalized eigenvalue system: $(D - S)v = \lambda D v$ that satisfies the constraint $v^T D \mathbf{1} = 0$. By setting $z = D^{\frac{1}{2}} v$, a standard eigensystem, easier to solve, is derived: $D^{-\frac{1}{2}}(D - S)^{-\frac{1}{2}} z = \lambda z$.

So in Step 1, Shi and Malik compute the Laplacian matrix $L = D - S$ and its normalized variant $\overline{L} = I - D^{-\frac{1}{2}} S D^{-\frac{1}{2}}$.

In Step 2, they extract the second smallest eigenvector z of $\overline{L}$, which is then transformed to approximate the optimal vector indicator looked for: $v = D^{-\frac{1}{2}} z$. First eigenvector z_0, collinear to $D^{\frac{1}{2}}\mathbf{1}$, is left in order to satisfy condition $v^T D\mathbf{1} = 0$.

In Step 3, the objects are split into two clusters based on the values of v (the optimal NCut splitting value being looked for).

$K = 2$. **Von Luxburg.** In his tutorial (Luxburg, 2007), the author defines the indicator vector of cluster C_1 as $u \in \{a, -a^{-1}\}^N$: $u_i = a \Leftrightarrow x_i \in C_1$, with $a = \sqrt{\frac{Vol(C_2)}{Vol(C_1)}}$. NCut criterion is then written as a quadratic function of u:

$$\begin{aligned} NCut(G,u) =& \frac{1}{2} \sum_{i,j} (u_i - u_j)^2 S_{ij} \\ =& u^T (D - S) u = u^T L u. \end{aligned} \tag{5}$$

The problem solved is the same than Shi and Malik's one:

$$\min_z NCut(G,z) = \min_z z^T \overline{L} z, \ \textbf{s.t. } z^T z = 1, \tag{6}$$

with exatly the same formal condition $u^T D\mathbf{1} = 0$.

The same steps are then followed.

$K >= 2$. **Shi and Malik, Von Luxburg.** These authors(Shi and Malik, 2000; Luxburg, 2007) generalize the NCut criterion to the Multiple-NCut (MNCut) criterion, by proposing an average criterion:

$$MNCut(G,U) = \sum_{k=1}^{K} NCut(G,u_k),$$

whose K vectors u_k, denote indicator vectors partitioning $\mathcal{X}$ in K clusters.

Two authors((Meila and Shi, 2000; Luxburg, 2007)) propose to solve this problem, by considering: $u_k \in \{0, \frac{1}{\sqrt{Vol(C_k)}}\}^N$, and $u_{ik} = \frac{1}{\sqrt{Vol(C_k)}} \Leftrightarrow x_i \in C_k$. These indicator vectors are column-wise gathered in matrix U.

They finally express their problem, in a way similar to case $K = 2$:

$$\min_Z MNCut(G,Z) = \min_Z \sum_{k=1}^{K} {z_k}^T \overline{L} z_k, \ \textbf{s.t. } {z_k}^T z_k = 1, \tag{7}$$

with additional formal condition $U = D^{-\frac{1}{2}} Z$: $U^T D U = I$. Let's note that condition $u_k D\mathbf{1}$ will be verified, although it is no more justified.

Consequently, the first K eigenvectors of $\overline{L}$ (i.e. with the K smallest eigenvalues) minimize the criterion and allow to estimate the K cluster indicator vectors. In order to retrieve discrete cluster indicator values, the eigenvector extraction is followed by a K-means step on the row of $U = D^{-\frac{1}{2}} Z$.

Shi and Malik (Shi and Malik, 2000) describe the same solution, but from a direct generalization from case $K = 2$.

$K >= 2$. **Ng et al.'s.** The authors (Ng et al., 2002) proposed an other algorithm based on Weiss (Weiss, 1999) and Meila and Shi (Meila and Shi, 2000) that also solved the spectral problem (eq. 7), but without formulating any optimization problem in terms of indicator vectors.

They proposed to modify the initial similarity matrix: $S_{ii} = 0$, and to use the K highest eigenvectors z_k of $L_{Ng} = D^{-\frac{1}{2}} S D^{-\frac{1}{2}}$, orthogonal to each others, to map data. Let's remark that these eigenvectors are the K lowest eigenvectors of $I - L_{Ng} = \overline{L}$. Then, instead of computing a matrix $U = D^{-\frac{1}{2}} Z$ from matrix Z stacking the extracted eigenvectors, they rather project data points in the spectral space on the unit-sphere, by normalizing Z into U: $U_{ij} = Z_{ij} / \sqrt{\sum_j Z_{ij}^2}$.

Step 3 is K-means too, initialized by points at most orthogonal.

Despite the diversity of the formalisms used to define the indicator vectors, all these authors finally solve the same objective function (eq. 7), which involves the same normalized Laplacian matrix $\overline{L}$.

2.2 Spectral Clustering Methods using Pairwise Constraints

2.2.1 Pairwise Constraints Information

We now focus on additional knowledge, formalized as pairwise constraints. The set of objects $\mathcal{X}$ and its similarity matrix S is now completed with the following two sets of pairs of objects (Wagstaff and Cardie, 2000):

- pairs of points that must belong to different clusters: $\{x_i, x_j\} \in \mathcal{CL}$, with $\{x_i, x_j\} \subseteq \mathcal{X}$, the *Cannot-Link* set of pairs;
- pairs of points that must belong to the same cluster: $\{x_i, x_j\} \in \mathcal{ML}$, with $\{x_i, x_j\} \subseteq \mathcal{X}$, the *Must-Link* set of pairs.

Spectral clustering methods integrating this type of information has previously been proposed, first by Kamvar et al. (Kamvar et al., 2003), and more recently by Wang and Davidson (Wang and Davidson, 2010). Both methods are now presented, while hightlighting some of their weakness.

2.2.2 *Spectral Learning* Method

In (Kamvar et al., 2003), the constrained spectral clustering method described is built as a basic spectral clustering method, in which two steps are modified:

- the similarity matrix S, built by applying a gaussian kernel on a set of N points describing the objects in $\mathcal{X}$, is modified in the following way: for each pair $\{x_i, x_j\} \in \mathcal{ML}$, elements $S_{ij} = S_{ji}$ are set to 1; and for each pair $\{x_i, x_j\} \in \mathcal{CL}$, elements $S_{ij} = S_{ji}$ are set to 0;
- then, similarity matrix S is not normalized as in the *MNCut*-graph paradigm, but in a *normalized additive* way: $(S + d_{max}I - D)/d_{max}$, with d_{max} the maximal rowsum of S; the obtained matrix is a symmetric Markov transition probabilities matrix; the authors underline that must-linked pairs have a higher mutual transition value than other pairs; eigenvectors are then extracted from this normalized S, and their rows are unit-length normalized.

The main weakness of this variant is that must-linked (respectively, cannot-linked) similarities are arbitrarly set to their maximal (r., minimal) theoritical values: 1 and 0. About the maximal value, and even for the minimal value (although the paper is focused on Markov's probability matrix formalism), this choice may be discussed: greater or smaller values could have been prefered. With such *a priori* values, it's difficult to know if the constraint on pairs of points is excessive, weak, or well balanced.

2.2.3 *Flexible Constrained Spectral Clustering* Method

In their paper (Wang and Davidson, 2010), Wang and Davidson express their constrained spectral clustering problem, as a constrained optimization problem, which is solved by an eigenvector extraction. Their approach is consequently less empirical than the previous one, and it gives an answer to the problem of tuning the strength of the constraints.

The semi-supervised spectral clustering problem is detailed with $K = 2$. The indicator vector looked for is denoted $u \in \{-1, +1\}^N$, and the satisfaction of pairwise constraints is measured thanks to a matrix Q:

$$Q_{ij} = Q_{ji} = \begin{cases} -1 & \text{if } \{x_i, x_j\} \in \mathcal{CL}, \\ +1 & \text{if } \{x_i, x_j\} \in \mathcal{ML}, \\ 0 & \text{else.} \end{cases} \quad (8)$$

With such a Q matrix, the measure $u^T Q u$ increases with the number of satisfied constraints.

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

$$\begin{aligned} &\min_z z^T \overline{L} z, \\ &\textbf{s.t. } z^T \overline{Q} z \geq \alpha, z^T z = Vol(G), z \neq D^{\frac{1}{2}} \mathbf{1}. \end{aligned} \quad (9)$$

The first constraint lower-bounds the satisfaction of constraints, the second one normalizes the indicator vector, and the last one is intented to avoid the trivial solution of spectral clustering (i.e. the "constant" indicator vector).

The problem is finally solved using Lagrange multipliers, but the infinite set of solutions has to be reduced by constraining this multipliers.

A feasible set of eigenvectors z, is then set as the solutions of the following generalized eigenproblem whose eigenvalues λ are strictly positive (because of the constraints satisfaction):

$$\overline{L} z = \lambda (\overline{Q} - \frac{\theta}{Vol(G)} I) z. \quad (10)$$

And the optimal z is then selected as the one minimizing the MNCut measure $z^T \overline{L} z$, while differing from the trivial solution $D^{\frac{1}{2}} \mathbf{1}$. Final indicator vector solution u is then obtained from the usual: $u = D^{-\frac{1}{2}} z$.

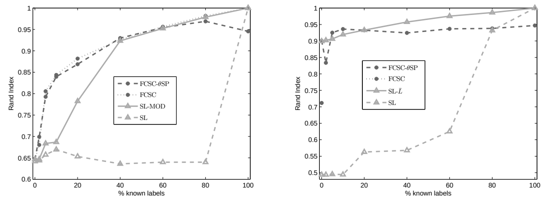
Parameter θ is used to weight the constraints impact: $\theta < \lambda_{\max} Vol(G)$, with $\lambda_{\max}$ the largest eigenvalue of $\overline{Q}$. The authors propose the following *a priori* value:

$$\theta = \lambda_{\max} \times Vol(G) \times \left(0.5 + 0.4 \times \frac{\text{\# Constraints}}{N^2} \right).$$

As shown in their paper (Wang and Davidson, 2010) in case $K = 2$, this algorithm outperforms Kamvar's method, which directly modifies the similarity matrix using 0 and 1 values.

In case $K > 2$, although the authors generalize the method by selecting not only the first, but the top-K generalized eigenvectors corresponding to the positive eigenvalues, we generally observe lower performances on UCI benchmarks, sometimes even lower than Kamvar's method ones.

As a possible explanation of these differences, we remark that the K-dimensional spectral subspace is not built as in the original spectral clustering method:



(a) Glass ($K = 2$). (b) Wine ($K = 3$).

Figure 1: Rand Index on two UCI datasets, functions of the percentage of known labels. (FCSC-θSP: modified version of FCSC, FCSC: original version of FCSC, SL-$\overline{L}$: modified version of SL, SL: original version of SL).

in particular, properties $u_k^T D\mathbf{1} = 0$ and $u_k^T D u_{k'} = 0$ are generally not satisfied. Although they are not always constrained in the original *MNCut* minimization problem (it depends on the formalism used), they could favour better clustering.

Let's finally remark that, on contrary to Von Luxburg's approach, the conditions verified by the eigenvectors are not justified by the formalism used ($u_k \in \{-1,1\}^N$): neither equations $z_k^T L z_{k'} = 0 \Leftrightarrow k \neq k'$ and $z_k^T(\overline{Q} - \frac{\theta}{Vol(G)})z_k \Leftrightarrow k \neq k'$, nor equation $u_k^T D\mathbf{1}$.

3 SEMI-SUPERVISED K-WAY SPECTRAL CLUSTERING ALGORITHM

Our problem formulation consists in a MNCut problem, where the objective function is modified, in such a way to penalize the non-respect of constraints. Unlike to FCSC method, the spectral subspace is obtained from a basic spectral clustering algorithm.

3.1 Penalty Cost

This penalty cost could be expressed on the indicator vector u_k. First, we should have to decide which binary domain of values $\{a,b\}$ to use, such that $u_k \in \{a,b\}^N$. But we prefer here to consider that this domain choice does not matter a lot: all the spectral clustering methods presented in Section 2 including Wang's one, whatever this domain is, finally define the spectral subspace from the top-K eigenvectors z_k from Laplacian matrix $\overline{L} = D^{-\frac{1}{2}} S D^{-\frac{1}{2}}$, i.e. the ones minimizing $\overline{L} = I - D^{-\frac{1}{2}} S D^{-\frac{1}{2}}$. The penalty cost will then depend on these eigenvectors z_k, stacked in matrix Z.

Then, previous methods post-transform these vectors, either by a $D^{-\frac{1}{2}}$ pre-multiplication, or by a projection on the unit-sphere. We consider here this last choice, as in different previously presented methods (Ng et al., 2002)(Kamvar et al., 2003).

Because of this final projection, we decide to make the penalty cost depend on the angles between spectral projections given by the K eigenvectors. Penalty term PC is defined by dot products between constrained points, considering that this measure suits well to the alteration of angles:

$$\begin{aligned} & PC = PC(\mathcal{CL}, \mathcal{ML}, \alpha, \beta, Z) \\ = \quad & \frac{-\alpha}{|\mathcal{CL}|} \sum_{\{x_i,x_j\}\in\mathcal{CL}} \sum_{k=1}^{K} z_{ik}.z_{jk} + \frac{\beta}{|\mathcal{ML}|} \sum_{\{x_i,x_j\}\in\mathcal{ML}} \sum_{k=1}^{K} z_{ik}.z_{jk} \\ = \quad & \sum_{k=1}^{K} \left[-\frac{\alpha}{|\mathcal{CL}|} \sum_{\{x_i,x_j\}\in\mathcal{CL}} z_{ik}.z_{jk} + \frac{\beta}{|\mathcal{ML}|} \sum_{\{x_i,x_j\}\in\mathcal{ML}} z_{ik}.z_{jk} \right]. \end{aligned}$$

Weights α and β are used to balance the contributions of the must-linked and cannot-linked constraints. Zhang et al. incorporate a quite similar Pairwise-Constraints penalty cost in a PCA method (Zhang et al., 2007), but with an Euclidean distance measure. As they do, we now express penalty cost PC as a matrix product, using a more general cost matrix Q than Wang's one:

$$Q_{ij} = Q_{ji} = \begin{cases} -\frac{\alpha}{|\mathcal{CL}|} & \text{if } \{x_i,x_j\} \in \mathcal{CL}, \\ +\frac{\beta}{|\mathcal{ML}|} & \text{if } \{x_i,x_j\} \in \mathcal{ML}, \\ 0 & \text{else.} \end{cases} \qquad (11)$$

PC term is then written in the following way:

$$PC = \frac{1}{2} \sum_{i,j} \sum_{k=1}^{K} z_{ik} z_{jk} Q_{ij} = \sum_{k=1}^{K} z_k^T Q z_k. \qquad (12)$$

3.2 Penalized MNCut Cost Function

This penalizing term is now combined to the MNCut criterion, so as to build a pairwise constrained spectral clustering optimization problem:

$$\begin{aligned} J &= J(G, \mathcal{CL}, \mathcal{ML}, Z) \\ &= MNCut(G,Z) + PC(\mathcal{CL}, \mathcal{ML}, \alpha, \beta, Z). \end{aligned} \qquad (13)$$

Minimizing this objective function allows to characterize a spectral projection reflecting both the original structure of data and the constraints proposed. We now want to reveal the criterion *PC* as a Rayleigh quotient, in order to set our problem as an eigenproblem.

MNCut and PC costs are now introduced in Equation 13:

$$J = \sum_{k=1}^{K} z_k^T \overline{L} z_k + \sum_{k=1}^{K} z_k^T Q z_k = \sum_{k=1}^{K} z_k^T (\overline{L} + Q) z_k. \qquad (14)$$

The penalized optimization problem can then set as:

$$\min_{Z} \sum_{k=1}^{K} z_k^T(\overline{L}+Q)z_k, \ \textbf{s.t. } {z_k}^T z_k = 1. \qquad (15)$$

This problem is clearly related to the basic spectral clustering's one Equation 7, except that the normalized Laplacian matrix $\overline{L}$ is penalized by matrix Q carrying the set of pairwise constraints.

3.3 Setting the Balance between the Two Parts of Criterion J

Considering that a *ML* information has the same importance as a *CL* information, and that the necessary strength to force them may be equal, we set $\alpha = \beta$; in the following part, these weights will be tuned by variable γ.

In addition, we propose a normalization making J easier to interpret. The MNCut expression $z_k^T \overline{L} z_k$ belonging to $[0,1]$ and the penalty one $z_k^T Q z_k$ belonging to $[\lambda_{Qmin}, \lambda_{Qmax}]$, we propose to normalize matrix Q using its minimal and maximal eigenvalues λ_{Qmin} and λ_{Qmax}: $\overline{Q} = \frac{Q-\lambda_{Qmin}}{\lambda_{Qmax}-\lambda_{Qmin}}$.

Thanks to balancing term γ, criterion J now belongs to $[0,1]$, and the final problem is set as:

$$\min_{Z} \sum_{k=1}^{K} ((1-\gamma).{z_k}^T \overline{L} z_k + \gamma.{z_k}^T \overline{Q} z_k), \qquad (16)$$
$$\textbf{s.t. } {z_k}^T z_k = 1.$$

3.4 "Mono-cluster" Solution $u_0 = D^{\frac{1}{2}}\mathbf{1}$

Because of the penalty term used, this vector is not solution of our optimization problem for most Q matrix, on contrary to basic spectral clustering's one or even to Wang's constrained spectral clustering problem. This can be seen as a weakness, because it's make mono-cluster vectors more difficult to recognize and to reject: in basic spectral clustering, all the eigenvectors orthogonal to z_0 (the smallest eigenvector of $\overline{L}$) are necessarily valid solutions.

To overcome this problem, a simple Euclidean distance can be used instead of the dot product penalty measure: matrix Q would then be modified by the substraction of a diagonal matrix R composed of its rowsums: $R_{ii} = \sum_j Q_{ij}$. With this penalty measure used on $U = D^{-\frac{1}{2}}Z$ rather than on Z, mono-cluster vector u_0 becomes a solution of the obtained eigensystem, quite similar to the one proposed; so it can be easily rejected. But in practice, the obtained results on all the benchmarks tested were less performant; that is why this solution was left.

In case $K = 2$, we then decide to reject the mono-cluster solution obtained from vectors u containing only positive (or negative) values.

In case $K > 2$, we maintain the usage of K eigenvectors, considering that this mono-cluster vector has high chance to take part in the subspace building. All the experiments made did not appear to be penalized by this point, as it will be shown in the next section.

The algorithm in its K-way variant is resumed below (cf. Algorithm 1).

Algorithm 1: Semi-Supervised K-way Spectral Clustering.

```
Spectral projection step
  1. For a given data matrix $\mathcal{X} \in \Re^{N\times P}$, with $N$ points described in a $P$-features space, compute a similarity matrix $S$ between these points ; for example: $S_{ij} = e^{-\frac{d^2(x_i,x_j)}{2\sigma^2}}$, with $\sigma$ a scale parameter, and $d$ a distance measure.
  2. Set $S_{ii} = 0$.
  3. Compute the constraints weighting matrix $Q$:
       $Q_{ij} = \begin{cases} -\frac{1}{|\mathcal{CL}|} & \text{if } \{x_i,x_j\} \in \mathcal{CL}, \\ +\frac{1}{|\mathcal{ML}|} & \text{if } \{x_i,x_j\} \in \mathcal{ML}, \\ 0 & \text{else.} \end{cases}$
  4. Compute the minimum and maximum eigenvalues (denoted $\lambda_{Qmin}$ and $\lambda_{Qmax}$) of $Q$.
  5. Compute the constraints weighting matrix $\overline{Q}$: $\overline{Q} = \frac{Q-\lambda_{Qmin}}{\lambda_{Qmax}-\lambda_{Qmin}}$
  6. Compute the degree diagonal matrix $D \in \Re^{N\times N}$: $D_{ii} = \sum_j S_{ij}$.
  7. Compute the normalized Laplacian matrix: $\overline{L} = I - D^{-\frac{1}{2}} S D^{-\frac{1}{2}}$.
  8. Find, the $K$ lowest eigenvectors $\{z_1,\ldots,z_K\}$ of matrix:
       $(1-\gamma)\overline{L} + \gamma\overline{Q}$,
     and form the matrix $Z = [z_1,\ldots,z_K] \in \Re^{N\times K}$.
  9. Normalize the rows of $Z$ to be unit-lengthed (projection on the unit-sphere).
Spectral clustering step
  1. Apply a $K$-means clustering on the data matrix $Z$.
  2. Cluster each point of $\mathcal{X}$ as its corresponding point in $Z$ was clustered.
```

4 EXPERIMENTAL RESULTS

In this section, our Semi-Supervised Spectral Clustering method (denoted SSSC) is applied first on some illustrative synthetic examples, then on public benchmarks belonging to the UCI repository. For each dataset, some pairwise constraints are generated from the known labels, and results are analyzed using objective evaluation measures like MNCut, satisfied constraints rates, or Rand Index. These results are then compared with outputs of a set of similar methods (like Kamvar's and Davidson's ones).

4.1 Algorithms for Comparison

For all experiments, the proposed algorithm is compared with the following seven clustering methods:

- SC: the basic *Spectral Clustering* Ng's algorithm (cf. 2.1.2), as a control reference unsupervised method, in order to assess the impact of the added pairwise constraints on the initial clustering;
- SL: the original semi-supervised *Spectral Learning* algorithm introduced in Section 2.2.2;
- SL-$\overline{L}$: a modified version of the SL algorithm, whose Laplacian matrix is replaced by the one used in our SSSC method (i.e. $\overline{L} = I - D^{-\frac{1}{2}}SD^{-\frac{1}{2}}$);
- FCSC: the original *Flexible Constrained Spectral Clustering* method introduced in Section 2.2.3, weighted by the value θ obtained from the rule given by the authors;
- FCSC-θ: a variant of FCSC, where the weight θ is *a posteriori* choosed in the range $(\lambda_{\min}Vol(G), \lambda_{\max}Vol(G))$ introduced by the authors, using an exhaustive search;
- FCSC-θSP: a variant of FCSC-θ, which consists in incorporating the projection on the unit-sphere step;
- FCSC-θ_2SP: a variant of FCSC-θSP, where parameter θ is looked inside a range larger than the one proposed by the authors.

In order to facilitate the comparison of the methods, without promoting our SSSC method, some homogenisations were done. Except for methods FCSC and FCSC-θ, the projection step on the unit-sphere is applied. The integration of this step in the algorithms facilitates the comparison and allows to not promote our SSSC method.

In all FCSC variants except the original one, the weighting matrix used for experiments is the one defined in Algorithm1. The weights of each kind of constraints are then similar and depend on the number of contraints defined.

For SSSC and FCSC variants (except the original), the weight of the penalty term θ or γ is *a posteriori* optimized, by discretizing their definition interval into 100 equidistant values, and choosing the one which maximizes the criterion:

$$E = (1 - MNCut) + \mathcal{ML}_{satisfied} + \mathcal{CL}_{satisfied}, \quad (17)$$

where $\mathcal{ML}_{satisfied}$ and $\mathcal{CL}_{satisfied}$ are the respective rates of satisfied $\mathcal{ML}$ and $\mathcal{CL}$ constraints.

For FCSC-θ and FCSC-θSP, the optimal θ in searched in the range $[\lambda_{\min}Vol(G), \lambda_{\max}Vol(G)]$. The authors show that this range is sufficient to assure the existence of K vectors satisfying the constraint of their optimization problem; moreover, it contains the values in which the constraints are at most satisfied (Wang and Davidson, 2010).

For FCSC-θ_2SP, we decide to enlarge the range used: $[-100 \times \max(|\lambda_{min}|, |\lambda_{max}|) \times vol(G), \lambda_{max}]$. The lower bound is an empirical value choosed in order to make their constraint problem converge to the unconstrained spectral clustering method, like in our method.

4.2 Illustrative Example

To study the effect of constraints in clustering, we propose to use pairwise constraints in a multiclass problem.

The dataset is composed of 400 data samples drawn from a mixture of five bivariate Gaussian distributions, as shown in Figure 2(a). The proportion of each Gaussian distribution is set to $\frac{1}{5}$. In this case, the desired number of clusters K is set to 4.

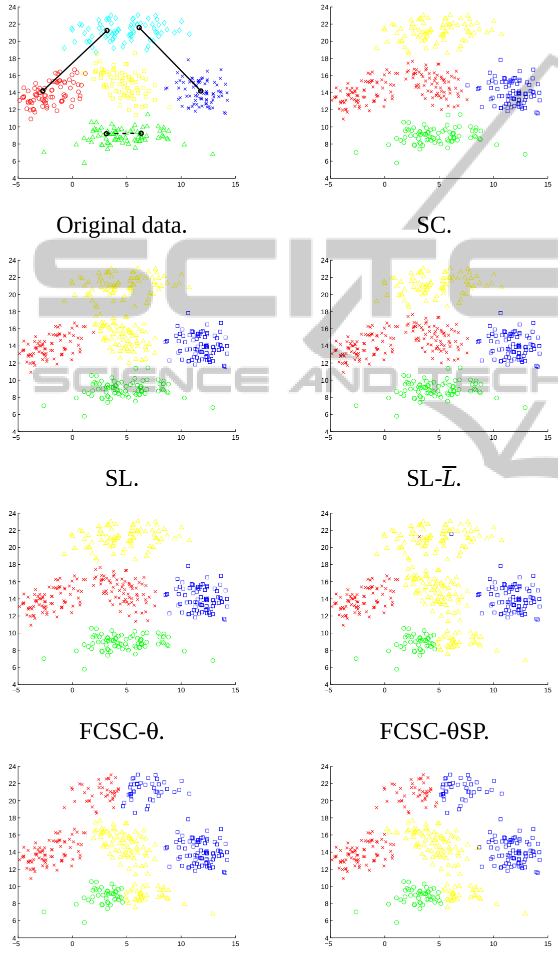
Three pairwise informations are considered: two *ML* constraints between data points from different clusters, and one *CL* constraint between two data points from the same Gaussian cluster (cf. Figure 2(a)). These pairwise constraints were deliberately chosen so as to make the expected clustering differ from the natural minimal cut obtained by the Spectral Clustering algorithm (SC) (i.e. we try to break the natural cut of the dataset).

For this example, the similarity matrix is built from a Gaussian kernel with a scale parameter σ set to 1, and with d set to the Euclidean distance.

Figure 2 shows the clusterings resulted for the eight methods tested. Here, FCSC clustering is not shown because its optimization problem can not be solved for the given value of θ; in fact, the proposed rule is clearly not suitable to case $K > 2$.

While all others methods fail to break the natural cut, the proposed SSSC, FCSC-θSP and FCSC-θ_2SP succeed in imposing the three constraints, as shown in Figure 2(f), (g) and (h). The combination of the three pairwise constraints succeeds in affecting the clustering, even with a “non-natural” *CL* constraint.

In order to complete the analyse of these clustering results, some performance indicators such as MNCut values and the total proportion of satisfied constraints (*ML* and *CL*) are shown in Table 1.



Original data. SC.

SL. SL-$\overline{L}$.

FCSC-θ. FCSC-θSP.

FCSC-θ_2SP. SSSC.

Figure 2: Clustering results on Bivariate Gaussian clusters with 2 *ML* and 1 *CL* constraints.

Worst MNCUT values are obtained from SSSC, FCSC-θSP and FCSC-θ_2SP methods, but they are the only ones which satisfy the pairwise constraints, necessarily at the expense of MNCut. The MNCut for SSSC is smaller than for FCSC-θSP and FCSC-θ_2SP, as shown in Table 1.

In this case, FCSC-θSP does not appear very performant: despite its high rate of satisfied constraints, it tends to isolate the data points linked by these pairwise constraints, in contrary of SSSC and FCSC-θ_2SP. Weights of proposed interval $[\lambda_{\min} Vol(G), \lambda_{\max} Vol(G)]$ appear too high in this case.

Table 1: *MNCut* values and percentage of satisfied constraints, for the different methods with 2 *ML* and 1 *CL* constraints.

Methods	*MNCut*	$\%(ML+CL)$
SC	**0.004**	0.0
SL	0.0151	0.0
SL-$\overline{L}$	0.013	0.0
FCSC	/	/
FCSC-θ	0.030	0.0
FCSC-θSP	0.048	**100.0**
FCSC-θ_2SP	0.042	**100.0**
SSSC	0.031	**100.0**

This experiment shows that the introduction of prior knowledge is well managed by SSSC method and FCSC modified method (variant FCSC-θ_2SP). The comparison with the basic *Spectral Clustering* method shows that supplying prior information, in the form of pairwise constraints, allows to improve the clustering accuracy.

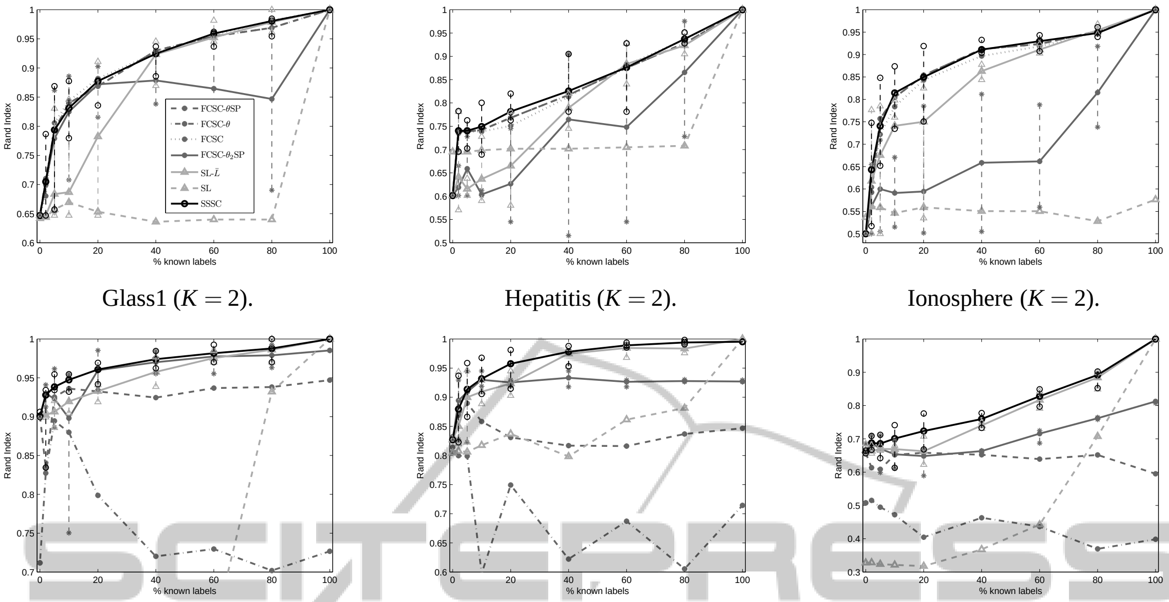
Moreover, in this example, the proposed SSSC method succeeds in conjointly satisfying both constraints and minimal MNCut score, in a more efficient way than all other algorithms.

4.3 Application to UCI Datasets

In this section, our Semi-Supervised K-way Spectral Clustering method is applied to some datasets well-known in the classification world (UCI datasets). For each example, some given proportions of objets are randomly selected, so as to build sets of labelled objects. Then, they are used to deduce both $\mathcal{CL}$ and $\mathcal{ML}$ constraints sets. For each percentage tested, we enlarge the previous sets of constraints with new informations. The quality of the clusterings obtained is measured by the Rand index, which reflects the similarity between the complete known partition (ground truth) and the one obtained, depending on the number of pairs of points classified similarly in the two partitions (Wagstaff and Cardie, 2000). The performance scores are averaged over 10 repetitions of the constraints generation process.

Table 2 shows the six datasets used. The data preprocessing is described in (Wang and Davidson, 2010). For each example, the similarity matrix is built using a Gaussian kernel: $S_{ij} = \exp(-\frac{||x_i - x_j||^2}{2\sigma^2})$ where σ is the scale parameter equal to the mean of the variances of features.

Figure 3 shows the performance measures of all the methods applied on these UCI datasets, in terms



Glass1 ($K = 2$).

Hepatitis ($K = 2$).

Ionosphere ($K = 2$).

Wine ($K = 3$).

Dermatology ($K = 6$).

Glass2 ($K = 6$).

Figure 3: Rand Index (mean, maximum and minimum), functions of the percentage of known labels, on UCI datasets.

Table 2: UCI datasets.

	Nb. Objects	Nb. Features	Nb. Classes
Glass1	214	9	2
Hepatitis	80	19	2
Ionosphere	351	34	2
Wine	178	13	3
Dermatology	366	34	6
Glass2	214	9	6

of Rand index, i.e. the rate of pairwise relations equal to the real ones. As it can be observed:

- Globally, methods like SSSC and some FCSC variants achieve to significantly improves the basic spectral clustering (corresponding to abscissa 0). Increasing the number of constraints globally improves the performances, and this increase is faster between abscissa 0% and 5%. This means that best methods are able to improve the clustering with small amongs of pairwise constraints.

- For $K = 2$, the best results are obtained from methods SSSC and all FCSC variants except FCSC-θ_2SP: their Rand indexes are the highest and the more stable: they do not decrease with the number of constraints added.

  SL-$\overline{L}$ and FCSC-θ_2SP show quite lower performances. The superiority of FCSC over FCSC-θ_2SP may be explained by the fact that FCSC-θ_2SP searches the optimal value of θ in a larger range than FCSC-θSP, but with the same discretization step (100 values): some interesting values may consequently be omitted. This tends to show that the choice of this parameter is not so obvious in FCSC method. SL-$\overline{L}$ becomes interesting, only with high numbers of constraints: weigths 0 and 1 seem too low (in absolute value) to impact the clustering.

  Then SL gives the lowest Rand indexes: the Laplacian used does not achieve to minimize the NCut measure.

- For $K > 2$, SSSC gets better performances than all others methods. FCSC-θ_2SP and SL-$\overline{L}$ give second best results. FCSC-θSP's ones are lower (the range of θ being too small). Then the methods FCSC-θ and SL give very low Rand indexes: both weigths and projection step are required to assure good performances. FCSC original method does not appear, because the constrained problem is not solved with the proposed θ value.

Table 3 shows some performance indicators of the different methods applied on a specific example, *Dermatology*, whose number of clusters K is set to 6. In each category, *percentage of known labels* by *performance indicator*, the best result is printed in bold type.

The proposed method thus appears to be very competitive versus the other methods tested. Indeed, for these datasets, SSSC method frequently reaches the highest rates of satisfied constraints (over 99% for each case), while keeping a satisfactory MNCut value for each percentage of known labels (almost always

Table 3: Evaluation measures on ”Dermatology” dataset ($K = 6$) with different numbers of constraints.

% known labels	Methods	% *ML*	% *CL*	% Total	*MNCut*	Rand Index
0	SL	/	/	/	0.245	0.805
	SL-$\overline{L}$	/	/	/	0.013	**0.827**
	FCSC	/	/	/	**0.011**	0.814
	FCSC-θ	/	/	/	**0.011**	0.814
	FCSC-θSP	/	/	/	0.013	**0.827**
	FCSC-θ_2SP	/	/	/	0.013	**0.827**
	SSSC	/	/	/	0.013	**0.827**
2	SL	**100.0**	87.1	93.5	0.251	0.808
	SL-$\overline{L}$	**100.0**	94.1	97.1	0.059	0.850
	FCSC	/	/	/	/	/
	FCSC-θ	48.8	70.7	59.7	0.085	0.800
	FCSC-θSP	37.4	80.7	59.1	0.109	0.869
	FCSC-θ_2SP	**100.0**	**100.0**	**100.0**	0.036	**0.894**
	SSSC	**100.0**	99.7	99.9	**0.013**	0.880
5	SL	**100.0**	84.1	92.1	0.273	0.806
	SL-$\overline{L}$	**100.0**	95.7	97.9	0.038	0.900
	FCSC	/	/	/	/	/
	FCSC-θ	65.0	77.6	71.3	0.102	0.799
	FCSC-θSP	62.8	92.3	77.6	0.139	0.890
	FCSC-θ_2SP	96.7	95.0	95.9	0.040	0.909
	SSSC	**100.0**	**98.4**	**99.2**	**0.018**	**0.914**
100	SL	**100.0**	**100.0**	**100.0**	0.063	**1.000**
	SL-$\overline{L}$	**100.0**	**100.0**	**100.0**	0.063	**1.000**
	FCSC	/	/	/	/	/
	FCSC-θ	75.8	70.3	73.0	0.334	0.714
	FCSC-θSP	72.5	87.7	80.1	0.095	0.847
	FCSC-θ_2SP	87.6	93.9	90.8	**0.037**	0.927
	SSSC	**100.0**	**100.0**	**100.0**	0.045	**1.000**

lower than other methods).

For example, for a small percentage of known labels (5%), the total proportion of satisfied constraints (*ML* and *CL*) for SSSC is better than for the others methods (99.2%) and the MNCut value is small (0.018). Moreover, this value is coherent with the one obtained for the basic spectral clustering (corresponding to 0% of known labels and equal to 0.013) and is smaller than for SL, SL-$\overline{L}$ and the four FCSC methods. Best Rand index is achieved too (0.914): final result for SSSC is then closer to the optimal clustering than other methods.

For a lower percentage (2%), SSSC method satisfies not exactly all constraints (99.9%), contrary to FCSC-θ_2SP. But its MNCut is the lowest (0.13 versus 0.36).

5 CONCLUSIONS

In this paper, we proposed a new efficient K-way spectral clustering algorithm, using *Cannot-Link* and *Must-Link* as semi-supervised information. Like in its unsupervised version, the clustering problem is set as an optimization problem, consisting in minimizing an objective function proportional to the *Multiple Normalized Cut* measure. This measure is here balanced by a weighted penalty term assessing the non-satisfaction of the given constraints.

Some comparisons with similar methods have been carried on synthetic samples and some UCI benchmarks. Different variants of the compared methods have been proposed, in order to make the methods more comparable, so as to get fair conclusions. In all cases, the results illustrated that the most performing methods, ours and the modified Wang's algorithms, are able to rapidly adjust the initial clustering to a more convenient one, satisfying the given constraints, even with quite low numbers of constraints. Our method seems to be part of this head group of methods, its clusterings often achieving the lowest MNCut values, and the highest satisfied constraints rates in the two-class and multi-class cases. These experiments highlighted the importance of two steps in this kind of semi-supervised spectral clustering methods: first, the usual projection step of basic spectral clustering appears as crucial; then, a lot of efforts have to be done to tune the constraints weight.

REFERENCES

Han, J. and Kamber, M. (2006). *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers.

Kamvar, S., Klein, D., and Manning, C. (2003). Spectral learning. In *IJCAI, International Joint Conference on Artificial Intelligence*, pages 561–566.

Luxburg, U. (2007). A tutorial on spectral clustering. In *Statistics and Computing*, pages 395–416.

Meila, M. and Shi, J. (2000). Learning segmentation by random walks. In *NIPS12, Neural Information Processing Systems*, pages 873–879.

Ng, A., Jordan, M., and Weiss, Y. (2002). On spectral clustering: Analysis and an algorithm. In *NIPS14, Neural Information Processing Systems*, pages 849–856.

Shi, J. and Malik, J. (2000). Normalized cuts and image segmentation. In *PAMI, Transactions on Pattern Analysis and Machine Intelligence*, pages 888–905.

Wagstaff, K. and Cardie, C. (2000). Clustering with instance-level constraints. In *ICML, International Conference on Machine Learning*, pages 1103–1110.

Wang, X. and Davidson, I. (2010). Flexible constrained spectral clustering. In *KDD, International Conference on Knowledge Discovery and Data Mining*, pages 563–572.

Weiss, Y. (1999). Segmentation using eigenvectors: an unifying view. In *IEEE, International Conference on Computer Vision*, pages 975–982.

Zhang, D., Zhou, Z., and Chen, S. (2007). Semi-supervised dimensionality reduction. In *SIAM, 7th International Conference on Data Mining*, pages 629–634.