
Convex Transduction with the Normalized Cut

Tijl De Bie

ESAT - SCD, K.U.Leuven
Kasteelpark Arenberg 10
3001 Leuven, Belgium

tijl.debie@esat.kuleuven.ac.be

Nello Cristianini

Dept. of Statistics, U.C.Davis
360 Kerr Hall, One Shields Ave.
Davis, CA 95616, USA

nello@support-vector.net

Abstract

We discuss approaches to transduction based on graph cut cost functions. More specifically, we focus on the normalized cut, which is the cost function of choice in many clustering applications, notably in image segmentation. Since optimizing the normalized cut cost is an NP-complete problem, much of the research attention so far has gone to relaxing the problem of normalized cut clustering to tractable problems, producing so far a spectral relaxation and a more recently a tighter but computationally much tougher semi-definite programming (SDP) relaxation. In this paper we deliver two main contributions: first, we show how an alternative SDP relaxation yields a much more tractable optimization problem, and we show how scalability and speed can further be increased by making a principled approximation. Second, we show how it is possible to efficiently optimize the normalized cut cost in a transduction setting using our newly proposed approaches. Positive empirical results are reported.

1 Introduction

The machine learning paradigm to divide learning algorithms into supervised and unsupervised types of methods is arguably artificial and too limitative. In many practical cases, label information is expensive and scarce. On the other hand while unlabeled data may be available or easier to obtain, the same data can often be clustered in different plausible ways, each of which is probably more appropriate for a different task. For this reason, it is meaningless to seek a general purpose similarity measure or clustering algorithm in the completely unsupervised case.

Therefore, much recent research has focused on the very practical scenario of partially labeled data sets, in a setting sometimes called transduction, interpolating between supervised and unsupervised learning.

One class of approaches makes use of traditional clustering or classification methods, but uses the labels to learn a metric in which the data clusters better ([1, 2, 3, 4] and more).

Another class of approaches doesn't change the metric, but optimizes some classical clustering cost function over all possible labelings of the data compatible with the available training labels ([5, 6, 7] and more). While conceptually this approach seems more natural, the exponential number of possible label assignments for the unlabeled part of the data set, makes good heuristics or approximation techniques indispensable.

In this paper, we will pursue this second line of research. As first investigated in [7] in the context of SVM-transduction, SDP promises to be an important tool to relax the combinatorial problem associated with transduction settings. Here we will zoom in on a different cost function, namely the normalized cut cost associated with a similarity matrix of the samples, and study how a new tight SDP relaxation of the associated optimization problem can be tractably solved in the transduction setting.

1.1 Cut, average cut and normalized cut cost functions

The problem setting we consider is the following. Given is a sample \mathcal{S} of size n , consisting of a (labeled) *training set* \mathcal{S}_t and an (unlabeled) *working set* \mathcal{S}_w of size n_t and n_w respectively. Between every pair of samples $(\mathbf{x}_i, \mathbf{x}_j)$, an affinity measure $a_{ij} = a(\mathbf{x}_i, \mathbf{x}_j)$ is given, such that we are able to make an affinity matrix \mathbf{A} containing a_{ij} as its i th row and j th column. We assume the function a is symmetric and positive, however, no positive definiteness of \mathbf{A} will be necessary, making the application domain larger than that of kernel based methods as discussed in e.g. [2, 7].

In this setting, the problem of transduction can be approached as a constrained graph cut problem on a fully connected graph, where the nodes in the graph represent the samples, and the edges between them are assigned weights equal to the affinities, and the constraints specify that training samples with the same label can not be separated by the cut. Several graph cut cost functions have been proposed in literature in the context of clustering, among which the *cut cost*, the *average cut cost (ACut)* and the *normalized cut cost (NCut)* [8].

The Cut cost is computationally the easiest to handle in a transduction setting [9], however as clearly motivated in [5], it often leads to degenerate results. This problem could largely be solved by using the ACut or NCut cost functions, of which the ACut cost seems to be more vulnerable to outliers (distant samples, meaning that they have low affinity to the rest of the sample). However, both optimizing the ACut and NCut costs are NP-complete problems. To get around this, *spectral* relaxations of the ACut and NCut optimization problems have been proposed in a clustering [8, 10, 11] and more recently also in a transduction setting [12, 5, 13]. A recent paper [14] also proposes an interesting SDP relaxation for the NCut in a multiclass *clustering* setting, however, the computational cost to solve this relaxation turns out to be too high to cluster data sets of more than about 150 samples, and hence impractical in real situations.

1.2 Outline of the paper

To put the paper in the right context, in section 3 we provide a short derivation of the well-known *spectral* relaxation of the NCut optimization problem, as well as of an (impractical) *SDP* relaxation that is similar to but different from the one obtained in [14] for *clustering*.

In section 4, we develop a practically viable SDP-based transduction algorithm that scales up to real size problems, the main goal in this paper. To achieve this, we make three new contributions. First, we propose a novel SDP relaxation of the NCut optimization problem that is computationally much more tractable than the one proposed in [14]. Second, we show how this relaxation can efficiently deal with label information to operate in a *transduction* setting. Third, we show how this *SDP* relaxation can be approximated using the (looser) *spectral* relaxation, to scale up even further. With the resulting algorithm, one can scale up to thousands of samples. We conclude the paper with empirical results in section 5, clearly showing the scalability and accuracy of the method.

Notation. Scalars are standard lower case; vectors bold face lower case; matrices bold face upper case; sets calligraphic upper case. The identity matrix is represented by \mathbf{I} ; $\mathbf{0}$ is a matrix or vector containing all zeros; the vector containing all ones is \mathbf{e} . A transpose is $'$.

2 NCut transduction

The NCut cost function for a partitioning of the sample \mathcal{S} into a positive \mathcal{P} and a negative \mathcal{N} set is given by (as originally denoted in [8]):

$$\frac{\text{cut}(\mathcal{P}, \mathcal{N})}{\text{assoc}(\mathcal{P}, \mathcal{S})} + \frac{\text{cut}(\mathcal{N}, \mathcal{P})}{\text{assoc}(\mathcal{N}, \mathcal{S})} = \left(\frac{1}{\text{assoc}(\mathcal{P}, \mathcal{S})} + \frac{1}{\text{assoc}(\mathcal{N}, \mathcal{S})} \right) \cdot \text{cut}(\mathcal{P}, \mathcal{N}), \quad (1)$$

where $\text{cut}(\mathcal{P}, \mathcal{N}) = \text{cut}(\mathcal{N}, \mathcal{P}) = \sum_{i:\mathbf{x}_i \in \mathcal{P}, j:\mathbf{x}_j \in \mathcal{N}} a_{ij}$ is the cut between sets \mathcal{P} and \mathcal{N} , and $\text{assoc}(\mathcal{P}, \mathcal{S}) = \sum_{i:\mathbf{x}_i \in \mathcal{P}, j:\mathbf{x}_j \in \mathcal{S}} a_{ij}$ the association between sets \mathcal{P} and the full sample \mathcal{S} . (Note that in fact $\text{cut}(\mathcal{P}, \mathcal{N}) = \text{assoc}(\mathcal{P}, \mathcal{N})$.) We do not discuss here its statistical properties, which would be interesting to investigate in a separate paper. Intuitively however, it is clear that the second factor $\text{cut}(\mathcal{P}, \mathcal{N})$ defines how well the two clusters separate.

The first factor $\left(\frac{1}{\text{assoc}(\mathcal{P}, \mathcal{S})} + \frac{1}{\text{assoc}(\mathcal{N}, \mathcal{S})} \right)$ measures how well the clusters are balanced. This specific measure of balancedness can be seen to be relatively insensitive to distant samples:¹ such outliers have a small cut cost with the other samples, making it beneficial to separate them out into a cluster of their own, which would lead to a useless result in our 2-class setting. However, they also have a small association with the rest of the sample \mathcal{S} , which on the other hand increases the cost function. In other words, the NCut cost function promotes partitions that are balanced in the sense that both clusters are roughly equally ‘coherent’ It is this feature that makes it preferable over the ACut cost function.

To optimize this cost function, we reformulate it into algebraic terms using the unknown label vector $\mathbf{y} \in \{-1, 1\}^n$, the affinity matrix \mathbf{A} , the degree vector $\mathbf{d} = \mathbf{A}\mathbf{e}$ and associated matrix $\mathbf{D} = \text{diag}(\mathbf{d})$, and shorthand notations $s_+ = \text{assoc}(\mathcal{P}, \mathcal{S})$ and $s_- = \text{assoc}(\mathcal{N}, \mathcal{S})$.

Observe that $\text{cut}(\mathcal{P}, \mathcal{N}) = \frac{(\mathbf{e}+\mathbf{y})' \mathbf{A} (\mathbf{e}-\mathbf{y})'}{2} = \frac{1}{4} (-\mathbf{y}' \mathbf{A} \mathbf{y} + \mathbf{e}' \mathbf{A} \mathbf{e}) = \frac{1}{4} \mathbf{y}' (\mathbf{D} - \mathbf{A}) \mathbf{y}$. Furthermore, $s_+ = \text{assoc}(\mathcal{P}, \mathcal{S}) = \frac{1}{2} \mathbf{e}' \mathbf{A} (\mathbf{e} + \mathbf{y}) = \frac{1}{2} \mathbf{d}' (\mathbf{e} + \mathbf{y})$ and $s_- = \frac{1}{2} \mathbf{d}' (\mathbf{e} - \mathbf{y})$. Then we can write the combinatorial optimization problem as

$$\begin{aligned} \min_{\mathbf{y}, s_+, s_-} \quad & \frac{1}{4} \left(\frac{1}{s_+} + \frac{1}{s_-} \right) \cdot \mathbf{y}' (\mathbf{D} - \mathbf{A}) \mathbf{y} \\ \text{s.t.} \quad & \mathbf{y} \in \{-1, 1\}^n, \\ & \begin{cases} s_+ = \frac{1}{2} \mathbf{d}' (\mathbf{e} + \mathbf{y}) \\ s_- = \frac{1}{2} \mathbf{d}' (\mathbf{e} - \mathbf{y}) \end{cases} \Leftrightarrow \begin{cases} \mathbf{d}' \mathbf{y} = s_+ - s_- \\ \mathbf{d}' \mathbf{e} = s_+ + s_- \end{cases} \end{aligned} \quad (2)$$

Minimizing this cost function with respect to additional constraints on the labels \mathbf{y} as specified by the training labels is equivalent to performing transduction with this cost function.

3 A spectral and a first SDP relaxation of NCut clustering

A spectral relaxation. We now provide a short derivation of the NCut *spectral* relaxation as first given in [8]. We introduce the variable $\tilde{\mathbf{y}} = \frac{1}{2} \left(\mathbf{y} - \mathbf{e} \frac{s_+ - s_-}{s_+ + s_-} \right) \cdot \sqrt{\frac{1}{s_+} + \frac{1}{s_-}}$ and rewrite the optimization problem in terms of this variable, s_+ and s_- :

$$\begin{aligned} \min_{\tilde{\mathbf{y}}, s_+, s_-} \quad & \tilde{\mathbf{y}}' (\mathbf{D} - \mathbf{A}) \tilde{\mathbf{y}} \\ \text{s.t.} \quad & \tilde{\mathbf{y}} \in \left\{ -\sqrt{\frac{s_+}{s_-}} \sqrt{\frac{1}{s_+ + s_-}}, \sqrt{\frac{s_-}{s_+}} \sqrt{\frac{1}{s_+ + s_-}} \right\}^n, \\ & \mathbf{d}' \tilde{\mathbf{y}} = 0 \quad \text{and} \quad \mathbf{d}' \mathbf{e} = s_+ + s_-. \end{aligned} \quad (3)$$

¹This property seems even more important in the relaxations of NCut based methods: the variables then have even more freedom, often making the methods more vulnerable to outliers.

Now, observe that the \mathbf{D} weighted 2-norm of $\tilde{\mathbf{y}}$ is constant here, and equal to $\tilde{\mathbf{y}}' \mathbf{D} \tilde{\mathbf{y}} = 1$. The spectral relaxation is obtained by relaxing the combinatorial constraint on $\tilde{\mathbf{y}}$ to this norm constraint that is implied by it. The result is

$$\text{Spectral} : \begin{cases} \min_{\tilde{\mathbf{y}}} & \tilde{\mathbf{y}}' (\mathbf{D} - \mathbf{A}) \tilde{\mathbf{y}} \\ \text{s.t.} & \tilde{\mathbf{y}}' \mathbf{D} \tilde{\mathbf{y}} = 1 \text{ and } \mathbf{d}' \tilde{\mathbf{y}} = 0, \end{cases} \quad (4)$$

which can be solved by taking the eigenvector corresponding to the second smallest generalized eigenvalue of $(\mathbf{D} - \mathbf{A}) \tilde{\mathbf{y}} = \lambda \mathbf{D} \tilde{\mathbf{y}}$.

Recently, several methods have been proposed to use spectral clustering relaxations of graph cut problems in a transduction setting [12, 5, 13]. In this paper we will later on make use of method proposed in [13] by one of us.

An SDP relaxation. Starting from (3), and using a matrix $\tilde{\mathbf{\Gamma}} = \tilde{\mathbf{y}} \tilde{\mathbf{y}}'$, we will now show how a (tighter but more expensive) SDP relaxation can be obtained. First rewrite (3) as:

$$\begin{aligned} \min_{\tilde{\mathbf{\Gamma}}, s_+, s_-} & \langle \tilde{\mathbf{\Gamma}}, \mathbf{D} - \mathbf{A} \rangle \\ \text{s.t.} & \tilde{\mathbf{\Gamma}} = \tilde{\mathbf{y}} \tilde{\mathbf{y}}' \text{ with } \tilde{\mathbf{y}} \in \left\{ -\sqrt{\frac{s_+}{s_-}} \sqrt{\frac{1}{s_+ + s_-}}, \sqrt{\frac{s_-}{s_+}} \sqrt{\frac{1}{s_+ + s_-}} \right\}^n, \\ & \tilde{\mathbf{\Gamma}} \mathbf{d} = \mathbf{0} \text{ and } \mathbf{d}' \mathbf{e} = s_+ + s_-. \end{aligned} \quad (5)$$

The pair of constraints (5) on $\tilde{\mathbf{y}}$ are the hard ones. So we will relax it, by finding a constraint set that is convex while as tight as possible. By inspection, we see that for $\tilde{\mathbf{y}}$ satisfying (5): $\tilde{\mathbf{\Gamma}} \succeq \mathbf{0}$ and $\tilde{\mathbf{\Gamma}} \geq -\frac{1}{s_+ + s_-}$. Furthermore, the \mathbf{D} -norm constraint we used in the spectral clustering relaxation translates here into $\langle \tilde{\mathbf{\Gamma}}, \mathbf{D} \rangle = 1$. A last constraint that is slightly more difficult to observe is that $\mathbf{I} \succeq \mathbf{D}^{1/2} \left(\tilde{\mathbf{\Gamma}} + \frac{\mathbf{e} \mathbf{e}'}{s_+ + s_-} \right) \mathbf{D}^{1/2}$ (to see this, note that given (5), the matrix on the right hand side is of rank 2 with two eigenvalues equal to 1). As a result we get the relaxed problem (\mathbf{P} stands for primal, the dual \mathbf{D} is stated without derivation):

$$\mathbf{P}_{\text{SDP1}}^{\text{clust}} : \begin{cases} \min_{\tilde{\mathbf{\Gamma}}} & \langle \tilde{\mathbf{\Gamma}}, \mathbf{D} - \mathbf{A} \rangle \\ \text{s.t.} & \mathbf{D} - \frac{\mathbf{e} \mathbf{e}'}{s_+ + s_-} \\ & \succeq \tilde{\mathbf{\Gamma}}, \\ & \tilde{\mathbf{\Gamma}} \succeq \mathbf{0}, \\ & \tilde{\mathbf{\Gamma}} \geq -\frac{\mathbf{e} \mathbf{e}'}{s_+ + s_-} \\ & \tilde{\mathbf{\Gamma}} \mathbf{e} = \mathbf{0} \\ & \langle \tilde{\mathbf{\Gamma}}, \mathbf{D} \rangle = 1 \end{cases} \quad \mathbf{D}_{\text{SDP1}}^{\text{clust}} : \begin{cases} \min_{\Lambda_1, \Lambda_2, \lambda, \mu} & \langle \Lambda_1, \frac{\mathbf{e} \mathbf{e}'}{s_+ + s_-} \rangle - \langle \Lambda_2, \mathbf{D} \rangle \\ & -\frac{n}{s_+ + s_-} \mathbf{e}' \lambda \\ & + \mu \left(\langle \frac{\mathbf{e} \mathbf{e}'}{s_+ + s_-}, \mathbf{D} \rangle + 1 \right) \\ \text{s.t.} & \Lambda_1 \succeq \mathbf{0} \\ & \Lambda_2 \succeq \mathbf{0} \\ & (\mathbf{D} - \mathbf{A}) - \Lambda_1 + \Lambda_2 \\ & -\mu \mathbf{D} - \mathbf{e}' \lambda \geq \mathbf{0}. \end{cases}$$

(We want to point out that a very similar result was obtained in [14].) As an SDP problem, it can be solved in polynomial time. However, the time and space complexity are still prohibitively large, making these results impractical. This is due to the two large SDP constraints and the n^2 inequality constraints on the elements of $\tilde{\mathbf{\Gamma}}$.

The label constraints —for this method to work in a transduction setting— could be imposed on $\tilde{\mathbf{\Gamma}}$ directly: its entries $\tilde{\mathbf{\Gamma}}_{i,j}$ with $\mathbf{x}_i, \mathbf{x}_j \in \mathcal{P}_t$ can be constrained to be equal to each other, and similarly for the entries $\tilde{\mathbf{\Gamma}}_{i,j}$ with $\mathbf{x}_i, \mathbf{x}_j \in \mathcal{N}_t$. However it is not clear for this relaxation how to impose constraints reflecting the fact that two samples are from the opposite class. We will not investigate this further here however, because the basic clustering problem is already intractable in most practical cases.

4 Two tractable SDP relaxations for NCut transduction

We have derived the well-known spectral relaxation as well as a first SDP relaxation of NCut clustering. We will now present an alternative way to relax the NCut optimization

problem and show how it can be effectively used in the transduction setting. While this approach will lead to a computationally much more tractable optimization problem already, we will also show how the method can be sped up even further by performing a principled approximation, ultimately leading to a method that easily handles thousands of samples, which is the goal of this paper.

4.1 A first practically useful SDP relaxation

The approach taken here starts from formulation (2). We introduce the notation $\mathbf{\Gamma} = \mathbf{y}\mathbf{y}'$. Then, we can write the equivalent optimization problem:

$$\begin{aligned} \min_{\mathbf{\Gamma}, s_+, s_-} \quad & \frac{1}{4} \left(\frac{1}{s_+} + \frac{1}{s_-} \right) \langle \mathbf{\Gamma}, \mathbf{D} - \mathbf{A} \rangle = \frac{s_+ + s_-}{4s_+s_-} \langle \mathbf{\Gamma}, \mathbf{D} - \mathbf{A} \rangle \\ \text{s.t.} \quad & \mathbf{\Gamma} = \mathbf{y}\mathbf{y}' \text{ and } \mathbf{y} \in \{-1, 1\}^n, \\ & \langle \mathbf{\Gamma}, \mathbf{d}\mathbf{d}' \rangle = (s_+ - s_-)^2 = (s_+ + s_-)^2 - 4s_+s_- \text{ and } \mathbf{d}'\mathbf{e} = s_+ + s_-. \end{aligned} \quad (6)$$

Now we can relax the combinatorial constraint by replacing it with $\mathbf{\Gamma} \succeq \mathbf{0}$ and $\text{diag}(\mathbf{\Gamma}) = \mathbf{e}$ (while this is a tight relaxation, tighter relaxations are possible at higher computational cost, see [15]). If we further use the notation $p = 4s_+s_-$, and the shorthand notation $s = s_+ + s_- = \mathbf{d}'\mathbf{e}$, we get:

$$\begin{aligned} \min_{\mathbf{\Gamma}, p} \quad & \frac{s}{p} \langle \mathbf{\Gamma}, \mathbf{D} - \mathbf{A} \rangle \\ \text{s.t.} \quad & \mathbf{\Gamma} \succeq \mathbf{0} \text{ and } \text{diag}(\mathbf{\Gamma}) = \mathbf{e}, \\ & \langle \mathbf{\Gamma}, \mathbf{d}\mathbf{d}' \rangle = s^2 - p \text{ and } 0 < p \leq s^2. \end{aligned}$$

By once again reparameterizing with $\hat{\mathbf{\Gamma}} = \frac{\mathbf{\Gamma}}{p}$ and $q = 1/p$, we obtain an SDP problem as shown below (P) along with its dual (D) (without derivation due to space constraints):

$$\mathbf{P}_{\text{SDP2}}^{\text{clust}} : \begin{cases} \min_{\hat{\mathbf{\Gamma}}, q} & s \langle \hat{\mathbf{\Gamma}}, \mathbf{D} - \mathbf{A} \rangle \\ \text{s.t.} & \hat{\mathbf{\Gamma}} \succeq \mathbf{0}, \\ & \text{diag}(\hat{\mathbf{\Gamma}}) = q\mathbf{e}, \\ & \langle \hat{\mathbf{\Gamma}}, \mathbf{d}\mathbf{d}' \rangle = qs^2 - 1, \\ & q \geq \frac{1}{s^2}. \end{cases} \quad \mathbf{D}_{\text{SDP2}}^{\text{clust}} : \begin{cases} \max_{\boldsymbol{\lambda}, \mu} & \frac{1}{s^2} \mathbf{e}'\boldsymbol{\lambda}, \\ \text{s.t.} & s(\mathbf{D} - \mathbf{A}) - \text{diag}(\boldsymbol{\lambda}), \\ & -\mu\mathbf{d}\mathbf{d}' \succeq \mathbf{0}, \\ & \mu s^2 + \mathbf{e}'\boldsymbol{\lambda} \geq 0. \end{cases}$$

Importantly, this relaxation contains much less constraints than \mathbf{P}_{SDP1} . Furthermore, the dual contains only $n + 1$ variables. It is this difference that makes this relaxation much more efficiently solvable, for example by using self-dual SDP solvers like SeDuMi [16].

To impose label constraints for the transductive version, we define the label constraint matrix $\mathbf{L} = \begin{pmatrix} \mathbf{y}_t & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{pmatrix}$, where we assume without loss of generality that the samples are ordered such that the training samples precede the unlabeled samples. \mathbf{y}_t is a column vector containing the training labels. Then the label constraints can be imposed by observing that any valid $\hat{\mathbf{\Gamma}}$ must satisfy $\hat{\mathbf{\Gamma}} = \mathbf{L}\hat{\mathbf{\Gamma}}_c\mathbf{L}'$. Thus the transductive NCut relaxation becomes:

$$\mathbf{P}_{\text{SDP2}}^{\text{trans}} : \begin{cases} \min_{\hat{\mathbf{\Gamma}}_c, q} & s \langle \hat{\mathbf{\Gamma}}_c, \mathbf{L}'(\mathbf{D} - \mathbf{A})\mathbf{L} \rangle \\ \text{s.t.} & \hat{\mathbf{\Gamma}}_c \succeq \mathbf{0}, \\ & \text{diag}(\hat{\mathbf{\Gamma}}_c) = q\mathbf{e}, \\ & \langle \hat{\mathbf{\Gamma}}_c, \mathbf{L}'\mathbf{d}\mathbf{d}'\mathbf{L} \rangle \\ & \quad = qs^2 - 1, \\ & q \geq \frac{1}{s^2}. \end{cases} \quad \mathbf{D}_{\text{SDP2}}^{\text{trans}} : \begin{cases} \max_{\boldsymbol{\lambda}, \mu} & \frac{1}{s^2} \mathbf{e}'\boldsymbol{\lambda}, \\ \text{s.t.} & s\mathbf{L}'(\mathbf{D} - \mathbf{A})\mathbf{L} \\ & \quad - \text{diag}(\boldsymbol{\lambda}) \\ & \quad - \mu\mathbf{L}'\mathbf{d}\mathbf{d}'\mathbf{L} \succeq \mathbf{0}, \\ & \mu s^2 + \mathbf{e}'\boldsymbol{\lambda} \geq 0, \end{cases}$$

which is computationally even easier to solve. (Note that we can handle more general label constraints as in [13], by using the appropriate matrix \mathbf{L} .) This is the first main result of this paper.

It turns out for small n_t/n —and in particular for the unsupervised case—that the problem is often badly conditioned. To solve this, the original objective (2) can be slightly altered by replacing the factor $\left(\frac{1}{s_+} + \frac{1}{s_-}\right)$ with $\left(\frac{1}{s_+ - \epsilon s} + \frac{1}{s_- - \epsilon s}\right)$. This will give rise to solutions that are slightly more biased towards balanced partitionings. The primal and dual formulations above can be changed according to this modification by simply multiplying s by $1 - 2\epsilon$. In practice, $\epsilon = 0.1$ seems to suffice, corresponding to disallowing partitions with one of s_+ and s_- smaller than $0.1s$, thus having a minor effect on the quality of the result.

Note that the methods thus obtained are fully automatic and require no parameter-tuning whatsoever. In conclusion, this approach is very suitable for transduction, and easily allows to handle data sets containing more than 500 unlabeled samples and a much more labeled samples on a 2GHz pentium computer with 0.5Gb RAM. For larger data sets further approximations will have to be made. This is the subject of the following subsection.

4.2 A fast approximation

In practice, if we assume the sample is drawn randomly from the population, we can estimate the imbalance parameter q from the training data if n_t is large enough. Then, we can fix q or equivalently $s_+ - s_-$ in (3) to its estimate. (Of course, one could also try several values for the imbalance, thus essentially performing a line search for this one parameter.)

If we do this, we can make the following approximation. Assuming that the spectral transduction method performs well, we know that the label vector will be close to the space spanned by the eigenvectors corresponding to the d smallest eigenvalues, stored in the columns of the matrix $\mathbf{V} \in \mathbb{R}^{n \times d}$. Then, we can approximate $\mathbf{\Gamma}$ in (6) by $\mathbf{\Gamma} \approx \mathbf{V}\mathbf{M}\mathbf{V}'$. Without derivational details, we state the thus obtained approximated SDP relaxation:

$$\mathbf{P}_{\text{SDP3}} : \begin{cases} \max_{\mathbf{M}} & \langle \mathbf{M}, \mathbf{V}'\mathbf{A}\mathbf{V} \rangle \\ \text{s.t.} & \mathbf{M} \succeq \mathbf{0}, \\ & \text{diag}(\mathbf{V}\mathbf{M}\mathbf{V}') \leq \mathbf{e}, \\ & \langle \mathbf{M}, \mathbf{V}'\mathbf{d}\mathbf{d}'\mathbf{V} \rangle \\ & \leq (s_+ - s_-)^2. \end{cases} \quad \mathbf{D}_{\text{SDP3}} : \begin{cases} \min_{\boldsymbol{\lambda}, \mu} & \mathbf{e}'\boldsymbol{\lambda} + (s_+ - s_-)^2\mu \\ \text{s.t.} & \mathbf{V}'\mathbf{A}\mathbf{V} - \mathbf{V}'\text{diag}(\boldsymbol{\lambda})\mathbf{V} \\ & - \mu\mathbf{V}'\mathbf{d}\mathbf{d}'\mathbf{V} \preceq \mathbf{0}, \\ & \boldsymbol{\lambda} \geq \mathbf{0}, \\ & \mu \geq 0. \end{cases}$$

Depending on whether the matrix \mathbf{V} is obtained using a standard or using a transductive spectral relaxation of NCut as in [13], this allows to do approximate NCut clustering or transduction respectively. Note that the number of primal variables as well as the size of the primal and dual constraints can be drastically reduced by taking d small enough. We will see that this is often possible in practice. Using this formulation, data sets of several thousands of samples can be dealt with. This is the second main result of the paper.

5 Empirical results

All SDP optimization problems are implemented using SeDuMi [16]. The algorithms described here compute a good label matrix $\mathbf{\Gamma}$. To find an approximate label vector \mathbf{y} from $\mathbf{\Gamma}$ one can use several techniques, including thresholding its dominant eigenvector, or simply picking a column corresponding to one of the training samples (which is what we will do in this paper). Other techniques are described in literature [15].

First experiment. We show empirical experiments on a data set extracted from the Swiss constitution, available in English, French, German and Italian. The data set contains 195 articles per language (so $n = 780$), which are organized in so-called Titles. The affinity matrix used here is the bag of words kernel after stop word removal and stemming. More

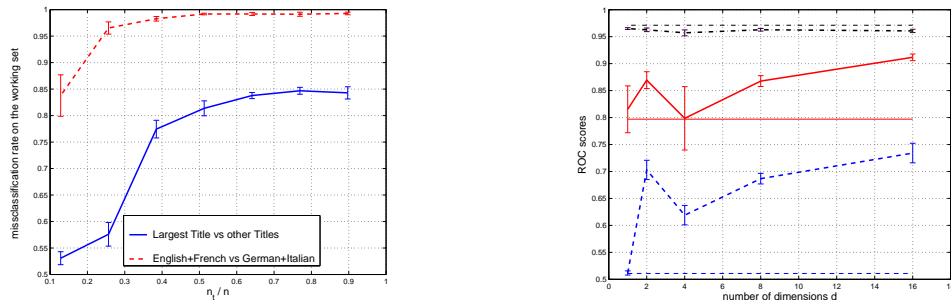


Figure 1: The left picture (a) shows the average working set error over 10 randomizations for the two Swiss constitution experiments. Method $\mathbf{P}/\mathbf{D}_{\text{SDP}2}$ was used. Clearly, the larger the training set size n_t , the better the performance. In the right picture (b), method $\mathbf{P}/\mathbf{D}_{\text{SDP}3}$ was used on the USPS test set, classifying the digits up to 4 in one class, the other 5 digits in the other class. This transduction task was carried out for 3 different amounts of labeled samples n_t , respectively 1% (dashed), 5% (full) and 20% (dash-dotted) of the total sample size n . The picture shows average ROC-scores on the working set, as a function of the dimensionality d used in the approximation. For comparison, the fainter straight horizontal lines give the average scores obtained by the method described in [12].

info can be found in [17]. To demonstrate the ability of the system to exploit partial label information, we try two different partitions: first, all English plus French articles are classified versus all German plus Italian articles; second, all articles in the largest of the 7 Titles, versus all other articles, no matter what language (note: the size of the Title is not directly reflected in the articles themselves). Figure (1a) shows the error rates for both problems for increasing n_t with fixed n , showing that label information is effectively exploited to find the right bipartitioning of the data. Obviously, more natural splits should require less labels. Indeed, as one could expect, a small n_t already gives a good performance for the language partition, while the partitioning by Title needs much more label information.

Second experiment. Here we use the test set of the USPS data set: the positive class contains all digits from 0 through 4, the negative class contains the other digits. Since the number of samples in this data set ($n = 2007$) is too large for the unapproximated relaxation to be practically solvable, we resort to the approximated method $\mathbf{P}/\mathbf{D}_{\text{SDP}3}$. The ROC-score as a function of the dimensionality d is shown in figure (1b), and this for three different sizes n_t of the training set. A 20-nearest neighbor affinity matrix is used (with nearest in the Euclidian sense). For comparison, the performance of the method described in [12], shown by the authors to operate well on nearest neighbor affinity matrices, is shown as well. From the figure, we can conclude that generally a better performance is achieved for larger n_t and d . Furthermore, for small n_t the method performs significantly better than the approach proposed in [12]. For larger n_t , the performance seems to be slightly worse.

6 Conclusions and further work

We addressed the use of the NCut cost function in the transduction framework. Whereas this cost function empirically leads to good results in clustering applications, unfortunately optimizing it is NP-complete. Therefore, the two main results in this paper try to solve this computational problem: an efficiently solvable SDP relaxation, and an approximation to speed it up even further. Both approaches are shown to deal with label information in a natural way. Positive empirical results show the validity of the approach.

Further work includes further experimentation to empirically validate the method. The statistical study of the NCut cost function remains an interesting open research topic. Another interesting question is if there are other good (or better) ways to efficiently find a good V .

Acknowledgments

TDB is a Research Assistant with the Fund for Scientific Research – Flanders (F.W.O.–Vlaanderen).

References

- [1] N. Shental, T. Hertz, D. Weinshall, and M. Pavel. Adjustment learning and relevant component analysis. In *Proceedings of the 7th European Conference of Computer Vision*, May, 2002.
- [2] O. Chapelle, J. Weston, and B. Schölkopf. Cluster kernels for semi-supervised learning. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15*. MIT Press, Cambridge, MA, 2003.
- [3] E. P. Xing, A. Y. Ng, M. I. Jordan, and S. Russell. Distance metric learning, with application to clustering with side-information. In *Advances in Neural Information Processing Systems 15*, Cambridge, MA, 2003. MIT Press.
- [4] T. De Bie, M. Momma, and N. Cristianini. Efficiently learning the metric with side-information. In *Proceedings of the International Conference on Algorithmic Learning Theory (ALT)*, Nara, Japan, April, 2003.
- [5] T. Joachims. Transductive learning via spectral graph partitioning. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2003.
- [6] Noam Shental, Aharon Bar-Hillel, Tomer Hertz, and Daphna Weinshall. Computing gaussian mixture models with em using side-information. In *Proc. of workshop The Continuum from labeled to unlabeled data in machine learning and data mining (ICML)*, 2003.
- [7] T. De Bie and N. Cristianini. Convex methods for transduction. In *Advances in Neural Information Processing Systems 16*. MIT Press, Cambridge, MA, 2004.
- [8] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.
- [9] A. Blum and S. Chawla. Learning from labeled and unlabeled data using graph mincuts. In *Proc. of the 18th International Conf. on Machine Learning (ICML)*, 2001.
- [10] A. Ng, M. I. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, Cambridge, MA, 2002. MIT Press.
- [11] Nello Cristianini, John Shawe-Taylor, and Jaz Kandola. Spectral kernel methods for clustering. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, Cambridge, MA, 2002. MIT Press.
- [12] S. D. Kamvar, D. Klein, and C. D. Manning. Spectral learning. In *IJCAI*, 2003.
- [13] T. De Bie, J. Suykens, and B. De Moor. Learning from general label constraints. In *Proceedings of IAPR International Workshop on Statistical Pattern Recognition (SPR)*. 2004.
- [14] E.P. Xing and M.I. Jordan. On semidefinite relaxation for normalized k-cut and connections to spectral clustering. Technical Report CSD-03-1265, Division of Computer Science, University of California, Berkeley, 2003.
- [15] C. Helmberg. Semidefinite programming for combinatorial optimization. Habilitationsschrift ZIB-Report ZR-00-34, TU Berlin, Konrad-Zuse-Zentrum Berlin, 2000.
- [16] J.F. Sturm. Using sedumi 1.02, a matlab toolbox for optimization over symmetric cones. *Optimization Methods and Software, Special issue on Interior Point Methods (CD supplement with software)*, 11-12:625–653, 1999.
- [17] T. De Bie and N. Cristianini. Kernel methods for exploratory data analysis: a demonstration on text data. In *Proceedings of the International Workshop on Statistical Pattern Recognition (SPR2004)*. Lisbon, Portugal, August 2004.