Learning from general label constraints

Tijl De Bie, Johan Suykens, and Bart De Moor

Katholieke Universiteit Leuven, ESAT-SCD
Kasteelpark Arenberg 10, 3001 Leuven, Belgium
{tijl.debie, johan.suykens, bart.demoor}@esat.kuleuven.ac.be,
WWW home page: www.esat.kuleuven.ac.be/sista-cosic-docarch

Abstract. Most machine learning algorithms are designed either for supervised or for unsupervised learning, notably classification and clustering. Practical problems in bioinformatics and in vision however show that this setting often is an oversimplification of reality. While label information is of course invaluable in most cases, it would be a huge waste to ignore the information on the cluster structure that is present in an (often much larger) unlabeled sample set. Several recent contributions deal with this topic: given partially labeled data, exploit all information available. In this paper, we present an elegant and efficient algorithm that allows to deal with very general types of label constraints in class learning problems. The approach is based on spectral clustering, and leads to an efficient algorithm based on the simple eigenvalue problem.

1 Introduction

We address the clustering problem where general information on the class labels y_i of some of the samples $\mathbf{x}_i$ $(i = 1, \ldots n)$ is given. This problem of taking general label information into account has received increasing attention in recent literature, and is know under different names as side information learning, semi-supervised learning, transductive learning (in a more restrictive setting), learning from (in)equivalence constraints, and more.

Roughly two ways of addressing the problem can be distinguished. Some of the methods try to learn a metric that is in accordance with the side information given, after which a standard clustering method can be applied, such as in [7, 11, 2]. Other methods propose actual adaptations of algorithms that were originally designed for clustering or for classification, such as in [9, 12, 3, 4, 1, 6, 10, 13]. These adaptations make it possible to take general types of label information into account. In this paper, we describe a novel fast, principled and highly general method that belongs to the second category of algorithms.

The label information to be dealt with can be of two general forms: in the first setting subsets of samples are given for which is specified that they belong to the same class; in the second setting, similarly subsets of samples with the same label are given, but now additionally, for some pairs of such subsets, it is given that they contain samples that do not belong to the same class. Note that the

03-106

De Bie T., Suykens J.A.K., De Moor B., ``Learning from general label constraints", in *Proc. of the joint IAPR international workshops on Syntactical and Structural Pattern Recognition and Statistical Pattern Recognition (SSSPR 2004)*, Lisbon, Portugal,, vol. 3138, Aug. 2004., Lirias number: 60124.

standard transduction setting, where part of the samples is labeled, is in fact a special case of this type of label information.

In this paper, we present an elegant way to handle the first type of label information in both the two class and the multi class learning settings. Furthermore, we show how the second type of label information can be dealt with in full generality for the two class case.

In a first section we will review spectral clustering as a relaxation of a combinatorial problem. In the second section, we will show how to enforce the constraints to the spectral clustering method, first for the two class case, and subsequently for the multi class case. Then, without going into detail, we will point out how the constraints can be imposed in a soft way as well. Finally, empirical results are reported and compared to a recently proposed approach [4] that is able to deal with similar settings and has comparable computational cost.

General notation: $\mathbf{1}$ is a column vector containing all ones, sometimes its size n is indicated as a subscript: $\mathbf{1}_n$. The identity matrix is denoted by $\mathbf{I}$. A transpose will be denoted by a prime $'$. The matrix containing all zeros is denoted by $\mathbf{0}$. Matrices are denoted by upper case bold face, vectors by lower case bold face, and scalars by standard lower case symbols.

2 Spectral clustering

Spectral clustering methods can best be seen as relaxations of graph cut problems, as clearly presented in [8]. Below a detailed derivation will be discussed only for the two class setting.

2.1 Two class clustering

Consider a weighted graph over the nodes each representing a sample $\mathbf{x}_i$. The edge weights correspond to some similarity measure to be defined in an appropriate way. These similarities can be arranged in a symmetric *affinity* matrix $\mathbf{K}$: its entry at row i and column j, denoted by K_{ij}, represents the similarity between sample $\mathbf{x}_i$ and $\mathbf{x}_j$.[1]

A graph cut algorithm searches for a partition of the nodes in two sets (corresponding clusters of the samples $\mathbf{x}_i$) such that a certain cost function is minimized. Several cost functions are proposed in literature, among which the *average cut cost* and the *normalized cut cost* are best known and most widely used.

For the normalized cut cost, the discrete optimization problem can be written in the form (see e.g. [8]):

$$\min_{\mathbf{y}} \frac{\mathbf{y}'(\mathbf{D}-\mathbf{K})\mathbf{y}}{\mathbf{y}'\mathbf{D}\mathbf{y}} \tag{1}$$

$$\text{s.t.} \quad \mathbf{1}'\mathbf{D}\mathbf{y} = 0 \tag{2}$$

$$y_i \in \{y_+, y_-\} \tag{3}$$

[1] In many practical cases this similarity will be given by a kernel function $k(\mathbf{x}_i, \mathbf{x}_j) = K_{ij}$ in which case $\mathbf{K}$ is semi positive definite, often named the kernel matrix.

where y_+ and y_- are the two possible values the y_i take depending on the class $\mathbf{x}_i$ is assigned to, and $\mathbf{D} = \text{diag}(\mathbf{K1})$ is a diagonal matrix containing all row sums d_i of $\mathbf{K}$ as its diagonal entries. The matrix $\mathbf{D} - \mathbf{K}$ is generally known as the Laplacian of the graph associated with $\mathbf{K}$. Note that the Laplacian is always semi positive definite.

It is constraint (3) that causes this problem to be combinatorial. However, the relaxed problem obtained by dropping this constraint can be solved very easily as we will show now. Furthermore, using the resulting vector $\mathbf{y}$ as an approximation has been observed to be very effective in practical problems. Note that since the scale of $\mathbf{y}$ in fact does not matter, we can as well solve

$$\begin{aligned} &\min_{\mathbf{y}} \ \mathbf{y}'(\mathbf{D}-\mathbf{K})\mathbf{y} \\ &\text{s.t.} \quad \mathbf{y}'\mathbf{D}\mathbf{y} = 1 \\ &\qquad \mathbf{1}'\mathbf{D}\mathbf{y} = 0 \end{aligned} \tag{4}$$

If we would drop constraint (4), the minimization becomes equivalent to solving for the minimal eigenvalue of

$$(\mathbf{D}-\mathbf{K})\mathbf{y} = \lambda \mathbf{D}\mathbf{y} \tag{5}$$

or, after left multiplication with $\mathbf{D}^{-1/2}$

$$\begin{aligned} &\mathbf{D}^{-1/2}(\mathbf{D}-\mathbf{K})\mathbf{D}^{-1/2}\mathbf{v} = \lambda \mathbf{v} \\ &\text{with} \qquad \mathbf{v} = \mathbf{D}^{1/2}\mathbf{y}. \end{aligned} \tag{6}$$

Now note that this is an ordinary symmetric eigenvalue problem, of which the eigenvectors are orthogonal. Since the Laplacian and thus also $\mathbf{D}^{-1/2}(\mathbf{D}-\mathbf{K})\mathbf{D}^{-1/2}$ is always semi positive definite, none of its eigenvalues can be smaller than 0. We can see immediately that a 0 eigenvalue is achieved by the eigenvector $\mathbf{v}_0 = \mathbf{D}^{1/2}\mathbf{1}$. This means that all other eigenvectors $\mathbf{v}_i$ of $\mathbf{D}^{-1/2}(\mathbf{D}-\mathbf{K})\mathbf{D}^{-1/2}$ are orthogonal to $\mathbf{v}_0$, such that for all other eigenvectors $\mathbf{v}_i$ and thus for $\mathbf{y}_i = \mathbf{D}^{-1/2}\mathbf{v}_i$, we have that $\mathbf{v}_0'\mathbf{v}_i = \mathbf{1}'\mathbf{D}\mathbf{y}_i = 0$. It thus follows that constraint (4) is automatically taken into account by simply solving for the *second* smallest eigenvalue of (6) or (5). This is the final version of the spectral clustering method as a relaxation of the normalized cut problem.[2]

2.2 Multi class clustering

In the k-class case, one usually extracts the eigenvectors $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_{k-1}$ corresponding to the smallest $k-1$ eigenvalues (excluding the 0 eigenvalue). Then, these vectors are put next to each other in a matrix $\mathbf{Y} = \begin{pmatrix} \mathbf{y}_1 & \mathbf{y}_2 & \cdots & \mathbf{y}_{k-1} \end{pmatrix}$, and subsequently any clustering algorithm can be applied to the rows of this matrix.[3] Every sample $\mathbf{x}_i$ is then assigned a label corresponding to which cluster row i of $\mathbf{Y}$ is assigned to.

[2] We can follow a similar derivation for the average cut cost function, ultimately leading to solving for the second smallest eigenvalue of $(\mathbf{D}-\mathbf{K})\mathbf{y} = \lambda \mathbf{y}$. All results presented in this paper can immediately be transferred to the average cut variant of spectral clustering.

[3] In [5] it is suggested to first normalize the rows of $\mathbf{Y}$ before performing the clustering.

3 Constrained spectral clustering

The results in this paper derive from the observations that

- constraining the labels according to the information as specified in the introduction can be seen as constraining the label vector $\mathbf{y}$ to some subspace;
- it is easy, in principle and computationally, to constrain the vector $\mathbf{y}$ to this subspace, while optimizing the Rayleigh quotient (1) subject to (2).

We will first tackle the two class learning problem subject to general label equality and inequality constraints. Afterwards, we show how equality constraints can be handled in the multi class setting.

3.1 Two class learning

Consider again the unrelaxed graph cut problem (1),(2),(3). We would now like to solve it with respect to the label information as additional constraints. For this we introduce the label constraint matrix $\mathbf{L} \in \{0,1\}^{n \times m}$ (with $n \geq m$) associated with the label equality and inequality constraints:

$$\mathbf{L} = \begin{pmatrix} 1_{s_1} & 1_{s_1} & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ 1_{s_2} & -1_{s_2} & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ 1_{s_3} & 0 & 1_{s_3} & \cdots & 0 & 0 & 0 & \cdots & 0 \\ 1_{s_4} & 0 & -1_{s_4} & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots & \vdots & \cdots & \vdots \\ 1_{s_{2p-1}} & 0 & 0 & \cdots & 1_{s_{2p-1}} & 0 & 0 & \cdots & 0 \\ 1_{s_{2p}} & 0 & 0 & \cdots & -1_{s_{2p}} & 0 & 0 & \cdots & 0 \\ 1_{s_{2p+1}} & 0 & 0 & \cdots & 0 & 1_{s_{2p+1}} & 0 & \cdots & 0 \\ 1_{s_{2p+2}} & 0 & 0 & \cdots & 0 & 0 & 1_{s_{2p+2}} & \cdots & 0 \\ \cdots & \vdots & \vdots & \cdots & \vdots & \vdots & \vdots & \cdots & \vdots \\ 1_{s_c} & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 1_{s_c} \end{pmatrix}.$$

Hereby, every row i of $\mathbf{L}$ corresponds to sample $\mathbf{x}_i$, in such a way that samples corresponding to one block row of size s_k are given to belong to the same class (i.e., for ease of notation the samples are sorted accordingly; of course also the rows of $\mathbf{K}$ need to be sorted in the same way). On the other hand, inequality constraints are encoded by the first $2p$ block rows: for all $k \leq p$, samples from block row k are given to belong to a different class as samples from block row $k+1$. For the last $c-2p$ blocks no inequality constraints are given. Note that in most practical cases, many block row heights s_k will be equal to 1, indicating that no constraint for the corresponding sample is given.

Using the label constraint matrix $\mathbf{L}$, it is possible to impose the label constraints explicitly, by introducing an auxiliary vector $\mathbf{z}$ and equating

$$\mathbf{y} = \mathbf{L}\mathbf{z}.$$

Then again constraint (3) is dropped, leading to

$$\min_{\mathbf{z}} \frac{\mathbf{z}'\mathbf{L}'(\mathbf{D}-\mathbf{K})\mathbf{L}\mathbf{z}}{\mathbf{z}'\mathbf{L}'\mathbf{D}\mathbf{L}\mathbf{z}} \quad \text{s.t.} \quad \mathbf{1}'\mathbf{D}\mathbf{L}\mathbf{z}=0$$

or equivalently

$$\begin{aligned} &\min_{\mathbf{z}} \mathbf{z}'\mathbf{L}'(\mathbf{D}-\mathbf{K})\mathbf{L}\mathbf{z} \\ &\text{s.t.} \quad \mathbf{z}'\mathbf{L}'\mathbf{D}\mathbf{L}\mathbf{z} \\ &\qquad \mathbf{1}'\mathbf{D}\mathbf{L}\mathbf{z}=0 \end{aligned} \tag{7}$$

Note that (similarly as in the derivation on standard spectral clustering above) after dropping the constraint (7), we would only have to solve the following eigenvalue problem

$$\mathbf{L}'(\mathbf{D}-\mathbf{K})\mathbf{L}\mathbf{z}=\lambda\mathbf{L}'\mathbf{D}\mathbf{L}\mathbf{z} \tag{8}$$

or, by left multiplication with $(\mathbf{L}'\mathbf{D}\mathbf{L})^{-1/2}$ and an appropriate substitution:

$$\begin{aligned} (\mathbf{L}'\mathbf{D}\mathbf{L})^{-1/2}[\mathbf{L}'(\mathbf{D}-\mathbf{K})\mathbf{L}](\mathbf{L}'\mathbf{D}\mathbf{L})^{-1/2}\mathbf{v} &= \lambda\mathbf{v} \\ \text{with} \qquad \mathbf{v} &= (\mathbf{L}'\mathbf{D}\mathbf{L})^{1/2}\mathbf{z} \end{aligned} \tag{9}$$

Again, one can see that the extra constraint is taken into account automatically by picking the *second* smallest eigenvalue and associated eigenvector of this eigenvalue problem. To this end, note that $(\mathbf{L}'\mathbf{D}\mathbf{L})^{-1/2}[\mathbf{L}'(\mathbf{D}-\mathbf{K})\mathbf{L}](\mathbf{L}'\mathbf{D}\mathbf{L})^{-1/2}$ is semi positive definite, such that its smallest eigenvalue is larger than or equal to 0. Now, note that the 0 eigenvalue is actually achieved for[4]

$$\mathbf{v}_0 = (\mathbf{L}'\mathbf{D}\mathbf{L})^{1/2} \cdot \begin{pmatrix} 1 & 0 & \cdots & 0 \end{pmatrix}'.$$

Thus, since (9) is an ordinary symmetric eigenvalue problem, all other eigenvectors $\mathbf{v}_i$ have to be orthogonal: $\mathbf{v}_0'\mathbf{v}_i = 0$. This means that

$$\begin{pmatrix} 1 & 0 & \cdots & 0 \end{pmatrix} (\mathbf{L}'\mathbf{D}\mathbf{L})^{1/2} \cdot (\mathbf{L}'\mathbf{D}\mathbf{L})^{1/2}\mathbf{z} = 0$$

and thus $\mathbf{1}'\mathbf{D}\mathbf{L}\mathbf{z}=0$.

As a result, it suffices to solve (9) or equivalently (8) for its second smallest eigenvalue, and constraint (7) is taken into account automatically.

In summary, the procedure is as follows

- Compute the affinity matrix $\mathbf{K}$ and the matrix $\mathbf{D}=\mathbf{K}\mathbf{1}$.
- Compute the label constraint matrix $\mathbf{L}$.
- Compute the eigenvector $\mathbf{z}$ corresponding to the second smallest eigenvalue of the eigenvalue problem $\mathbf{L}'(\mathbf{D}-\mathbf{K})\mathbf{L}\mathbf{z}=\lambda\mathbf{L}'\mathbf{D}\mathbf{L}\mathbf{z}$.
- Compute the resulting relaxed label vector as $\mathbf{y}=\mathbf{L}\mathbf{z}$.

[4] To see this note that $\mathbf{L} \cdot \begin{pmatrix} 1 & 0 & \cdots & 0 \end{pmatrix}' = \mathbf{1}$.

All operations can be carried out very efficiently, thanks to the sparsity of $\mathbf{L}$. The most expensive step is the eigenvalue problem, which is even smaller than in the unconstrained spectral clustering algorithm: the size of the matrices is only $m \times m$ instead of $n \times n$ (where m is the number of columns of $\mathbf{L}$).

As a last remark in this section, note that we are actually not interested in the component of $\mathbf{y}$ along $\mathbf{1}$. Thus, we could choose to take $\widetilde{\mathbf{y}} = \mathbf{L} \cdot \begin{pmatrix} 0 & z_2 & \cdots & z_m \end{pmatrix}'$ as an estimate for the labels, instead of $\mathbf{y} = \mathbf{L}\mathbf{z}$. This results in the fact that estimates for labels $\widetilde{y}_i$ that were specified to be different are actually *opposite in sign*. Therefore thresholding the vector $\widetilde{\mathbf{y}}$ around 0 in fact makes more sense than thresholding $\mathbf{y}$ around 0.

3.2 Multi class learning

In the multi class setting, it is not possible anymore to include label inequality constraints in the same straightforward elegant way. The reason is that the true values of the labels can not be made equal to 1 and -1 anymore.

We can still take the equality constraints into account however. This means we would use a label constraint matrix of the form

$$\mathbf{L} = \begin{pmatrix} \mathbf{1}_{s_1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_{s_2} & \cdots & \mathbf{0} \\ \vdots & \vdots & \cdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{1}_{s_c} \end{pmatrix},$$

constructed in a similar way. Note that this time we don't need a column containing all ones, as such vector $\mathbf{1}$ is included in its column space already.

As in the unconstrained spectral clustering algorithm, often $k-1$ eigenvectors will be calculated when k clusters are expected, leading to $\mathbf{Y} = \begin{pmatrix} \mathbf{y}_1 & \mathbf{y}_2 & \cdots & \mathbf{y}_{k-1} \end{pmatrix}$. Finally, the clustering of $\mathbf{x}_i$ is obtained by clustering the rows of $\mathbf{Y}$.

4 Softly constrained spectral clustering

In both the two class case and the multi class case, the constraints could be imposed in a soft way as well. This can be done by adding a cost term to the cost function that penalizes the distance between the weight vector and the column space of $\mathbf{L}$, in the following way (we give it without derivation or empirical results due to space restrictions):

$$\begin{aligned} &\min_{\mathbf{y}} \gamma \mathbf{y}'(\mathbf{D} - \mathbf{K})\mathbf{y} + (1-\gamma)\mathbf{y}'(\mathbf{D} - \mathbf{D}\mathbf{L}(\mathbf{L}'\mathbf{D}\mathbf{L})^{-1}\mathbf{L}'\mathbf{D})\mathbf{y} \\ &\text{s.t.} \quad \mathbf{y}'\mathbf{D}\mathbf{y} = 1 \quad \text{and} \quad \mathbf{1}'\mathbf{D}\mathbf{y} = 0 \end{aligned}$$

where γ is called the regularization parameter. Again the same reasoning can be applied, leading to the conclusion that one needs to solve for the second smallest eigenvalue (i.e. the smallest eigenvalue different from 0) of the eigenvalue problem:

$$\left[\gamma(\mathbf{D} - \mathbf{K}) + (1-\gamma)(\mathbf{D} - \mathbf{D}\mathbf{L}(\mathbf{L}'\mathbf{D}\mathbf{L})^{-1}\mathbf{L}'\mathbf{D})\right] \mathbf{y} = \lambda \mathbf{D}\mathbf{y} \tag{10}$$

For γ close to 0, the side-information is enforced very strongly, and $\mathbf{y}$ will satisfy the constraints nearly exactly. In the limit for $\gamma \to 0$, the soft constraints become hard constraints.

5 Empirical results

We report experiments for the transduction setting, and this for the three newsgroup dataset (2995 samples) as also used in [4], and for the training subset of the USPS dataset (7291 samples). Comparisons are shown with the method proposed in [4]. This method works in similar settings and has a similar computational cost. We construct the affinity matrix in the same way as in that paper, namely by equating the entries $\mathbf{K}_{ij}$ equal to 1 if j is among the 20 samples lying closest (in euclidian distance) to i or vice versa.

Since spectral clustering methods provide eigenvectors on which subsequently a clustering of the rows has to be performed, and since we are only interested in evaluating the spectral clustering part, we used a cost function defined on the eigenvectors themselves (without doing the actual clustering step). Specifically, the within cluster variance divided by the total variance in the eigenvectors is used as a quality measure, attaining values in between 0 and 1. All experiments are averaged over 10 randomizations of the labeled part of the training set; each time the standard deviation on the estimated average is shown on the figures.

Figures 1 show that the performance of both methods effectively increases for increasing fractions of labeled samples on the three newsgroup dataset as well as on the USPS dataset. Moreover, for small fractions of labeled samples (which is when side information methods are most useful in practice), the newly proposed performs slightly but significantly better.

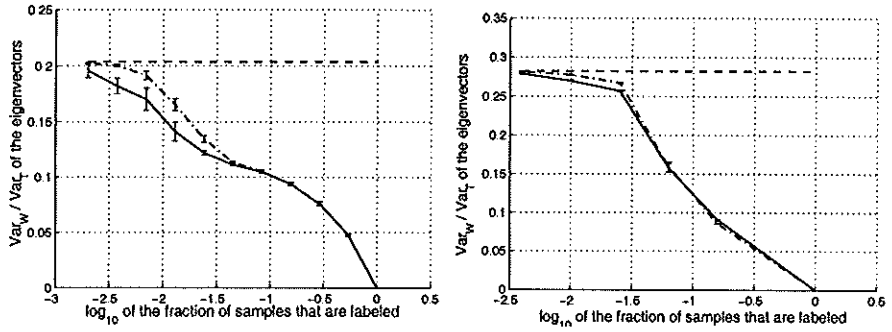

Fig. 1. The cost (within class variance divided by total variance of the eigenvectors) for spectral learning [4] in dash-dotted line, as compared with our method in full line as a function of the fraction of labeled samples (on a log scale), on the left for the three newsgroup dataset, on the right for the 10-class USPS dataset. For reference, the unconstrained cost is plotted as a horizontal dashed line.

Subsequently, we solve a binary classification problem derived from the USPS dataset, where one class contains the samples representing numbers from 0 up to 4, and the other class from 5 up to 9. In figure 2 we see that the performance of both methods is indistinguishable in this case. Note however that relatively few information is already sufficient to provide a significant improvement over the clustering with no label information at all.

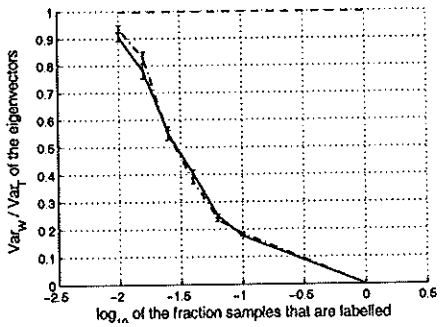

Fig. 2. Similar experiment as in figures 1, but now in the binary classification setting. One class contains all handwritten digits from 0 to 4, the other class contains the digits from 5 to 9. As can be expected the score for no labeled data at all is really bad.

6 Conclusions and further work

We have presented an efficient, performant and natural method to incorporate general constraints on the labels in class learning problems. The performance of the method compares well with a recently proposed approach that has a similar computational cost and that is designed to deal with a similar generality of learning settings. However further empirical investigation would be useful.

As compared to other related approaches in literature, the constrained spectral clustering method compares favorably in two respects. First, computationally the method is very attractive since basically it only requires the computation of a few dominant eigenvectors of a matrix with less than n rows and columns (n being the number of samples). Second, the method not only deals with the transductive learning setting, but addresses more general side information learning in the same natural way, both for two class and multi class problems.

Note that the softly constrained version can be seen as the application of the spectral clustering method to a sum of two affinity matrices, where one of both is derived from the label constraints. In principle, one may be able to construct a label affinity matrix for very general label information, also for the the multi class case. This will be subject of a later paper.

Acknowledgements This work was supported by grants and projects for the Research Council K.U.L (GOA-Mefisto 666, IDO, PhD/Postdocs & fellow grants), the Flemish Government (FWO: PhD/Postdocs grants, projects G.0240.99, G.0407.02, G.0197.02, G.0141.03, G.0491.03, G.0120.03, ICCoS, ANMMM; AWI;IWT:PhD grants, Soft4s), the Belgian Federal Government (DWTC: IUAP IV-02, IUAP V-22; PODO-II CP/40), the EU(CAGE, ERNSI, Eureka 2063-Impact;Eureka 2419-FLiTE) and Contracts Research/Agreements (Data4s, Electrabel, Elia, LMS, IPCOS, VIB). T. De Bie is a Research Assistant with the Fund for Scientific Research – Flanders (F.W.O–Vlaanderen). J. Suykens and B. De Moor are an assistant professor and a full professor at the K.U.Leuven, Belgium, respectively. The scientific responsibility is assumed by its authors.

References

1. T. De Bie and N. Cristianini. Convex methods for transduction. In *Advances in Neural Information Processing Systems 16*. MIT Press, Cambridge, MA, 2004.
2. 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.
3. T. Joachims. Transductive learning via spectral graph partitioning. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2003.
4. S. D. Kamvar, D. Klein, and C. D. Manning. Spectral learning. In *IJCAI*, 2003.
5. 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.
6. N. Shental, A. Bar-Hillel, T. Hertz, and D. Weinshall. Computing gaussian mixture models with em using equivalence constraints. In *Advances in Neural Information Processing Systems 16*. MIT Press, Cambridge, MA, 2004.
7. 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.
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. V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 2nd edition, 1999.
10. J. Weston, C. Leslie, D. Zhou, A. Elisseeff, and W. Noble. Semi-supervised protein classification using cluster kernels. In *Advances in Neural Information Processing Systems 16*. MIT Press, Cambridge, MA, 2004.
11. E.P. Xing, A.Y. Ng, M.I. Jordan, and S. Russell. Distance metric learning, with application to clustering with side-information. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15*, Cambridge, MA, 2003. MIT Press.
12. Stella X. Yu and Jianbo Shi. Grouping with bias. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, Cambridge, MA, 2002. MIT Press.
13. D. Zhou, O. Bousquet, T. Navin Lal, J. Weston, and B. Schölkopf. Learning with local and global consistency. In *Advances in Neural Information Processing Systems 16*. MIT Press, Cambridge, MA, 2004.