Multiclass Semisupervised Learning Based Upon Kernel Spectral Clustering

Siamak Mehrkanoon, Carlos Alzate, Raghvendra Mall, Rocco Langone, and Johan A. K. Suykens, *Senior Member, IEEE*

***Abstract*— This paper proposes a multiclass semisupervised learning algorithm by using kernel spectral clustering (KSC) as a core model. A regularized KSC is formulated to estimate the class memberships of data points in a semisupervised setting using the one-versus-all strategy while both labeled and unlabeled data points are present in the learning process. The propagation of the labels to a large amount of unlabeled data points is achieved by adding the regularization terms to the cost function of the KSC formulation. In other words, imposing the regularization term enforces certain desired memberships. The model is then obtained by solving a linear system in the dual. Furthermore, the optimal embedding dimension is designed for semisupervised clustering. This plays a key role when one deals with a large number of clusters.**

***Index Terms*— Kernel spectral clustering (KSC), low embedding dimension for clustering, multiclass problem, semisupervised learning.**

I. Introduction

THE incorporation of some form of prior knowledge of the problem at hand into the learning process is a key element that allows an increase of performance in many applications.

In many contexts, ranging from data mining to machine perception, obtaining the labels of input data is often difficult and expensive. Therefore in many cases one deals with a huge amount of unlabeled data, while the fraction of labeled data points will typically be small.

Manuscript received August 20, 2013; revised April 13, 2014; accepted April 30, 2014. Date of publication May 29, 2014; date of current version March 16, 2015. This work was supported in part by the Research Council KUL through MaNet under Grant GOA/10/09, OPTEC under Grant PFV/10/002, and the several Ph.D./Post-Doctoral and Fellow grants, in part by the Flemish Government under Grant IOF/KP/SCORES4CHEM, in part by FWO through the Ph.D./Post-Doctoral Grant under Project G.0320.08 through the convex MPC, Robust MHE under Grant G.0558.08, Glycemia2 under Grant G.0557.08, Brain-machine Grant G.0588.09, Mechatronics MPC under G.0377.09, Structured Systems Research Community under Grant G.0377.12, in part by the IWT through the Ph.D. Grants through the Projects such as Eureka-Flite+, SBO LeCoPro, SBO Climaqs, SBO POM, O&O-Dsquare, in part by the Belgian Federal Science Policy Office through the Dynamical Systems, Control and Optimization under Grant IUAP P7, in part by IBBT, in part by EU through the ERNSI, FP7-EMBOCON under Grant ICT-248940, FP7-SADCO under Grant MC ITN-264735, ERC ST HIGHWIND under Grant 259 166, ERC AdG A-DATADRIVE-B, in part by COST Action under Grant ICO806 through the IntelliCIS, and in part by the Contract Research through AMINAL. The work of J. A. K. Suykens was supported by KU Leuven, Belgium.

S. Mehrkanoon, R. Mall, R. Langone, and J. A. K. Suykens are with the Department of Electrical Engineering, ESAT-STADIUS, Katholieke Universiteit Leuven, Leuven B-3001, Belgium (e-mail: siamak.mehrkanoon@esat.kuleuven.be; rocco.langone@esat.kuleuven.be; raghvendra.mall@esat.kuleuven.be; johan.suykens@esat.kuleuven.be).

C. Alzate is with the Smarter Cities Technology Center, IBM Research, Dublin, Ireland (e-mail: carlos.alzate@ie.ibm.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNNLS.2014.2322377

Semisupervised algorithms aim at learning from both labeled and unlabeled data points. In fact in semisupervised learning one tries to incorporate the labels (prior knowledge) in the learning process to enhance the clustering/classification performance. The semisupervised learning can be classified into two categories: transductive and inductive learning. Transductive learning aims at predicting the labels for a specified set of test data by considering both labeled and unlabeled data together in the learning process. In contrast, in inductive learning the goal is to learn a decision function from a training set consisting of labeled and unlabeled data for future unseen test data points. Throughout this paper we refer to semisupervised inductive learning as semisupervised learning.

The semisupervised inductive learning itself can be categorized into semisupervised clustering and classification. The former addresses the problem of exploiting additional labeled data to adjust the cluster memberships of the unlabeled data. The latter aims at using both unlabeled and labeled data to obtain a better classification model, and higher quality predictions on unseen test data points.

In some classical semisupervised techniques, a classifier is first trained using the available labeled data points and then the labels for the unlabeled data points are predicted using out-of-extension. In the second step, unlabeled data that are classified with the highest confidence score are added incrementally to the training set, and the process is repeated until the convergence is satisfactory [1]–[3]. Several semisupervised algorithms have been proposed in the literature, see [4]–[10]. For instance, the Laplacian support vector machine (LapSVM) [6], is one of the graph-based methods with a data-dependent geometric regularization which provides a natural out-of-sample extension. Xiang *et al*. [7] used local spline regression for semisupervised classification by introducing splines developed in Sobolev space to map the data points to class labels. A transductive semisupervised algorithm called ranking with local regression and global alignment (LRGA) to learn a robust Laplacian matrix for data ranking is proposed in [8]. In this paper, for each data point, the ranking scores of neighboring points are estimated using a local linear regression model. A label propagation approach in graph-based semisupervised learning has been introduced in [9]. Wang *et al*. [10] developed a semisupervised classification method with class membership, motivated by the fact that similar instances should share similar label memberships.

Spectral clustering methods belong to a family of unsupervised learning algorithms that make use of the eigenspectrum

2162-237X © 2014 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

of the Laplacian matrix of the data to divide a data set into natural groups such that points within the same group are similar and points in different groups are dissimilar to each other [11]–[13].

Kernel spectral clustering (KSC) is an unsupervised algorithm that represents a spectral clustering formulation as a weighted kernel PCA problem, cast in the LS-SVM framework [14]. In contrast to classical spectral clustering, there is a systematic model selection scheme for tuning the parameters and also the extension of the clustering model to out-of-sample points is possible.

In [15], for dimensionality reduction, kernel maps with a reference point are generated from a least squares support vector machine core model via an additional regularization term to preserve local mutual distances together with reference point constraints. In contrast with the class of kernel eigenmap methods, the solution (coordinates in the low dimensional space) is characterized using a linear system instead of an eigenvalue problem.

Recently Alzate and Suykens [16] have extended the kernel spectral clustering to binary semisupervised learning (semi-KSC) by incorporating the information of labeled data points in the learning process. Therefore, the problem formulation is a combination of unsupervised and binary classification approaches. Contrary to the approach described in [16], a nonparallel semisupervised classification (NP-Semi-KSC) is introduced in [17]. It generates two nonparallel hyperplanes which are then used for out-of-sample extension.

This paper develops a new multiclass semisupervised KSC-based algorithm (MSS-KSC) using a one-versus-all strategy. In contrast to the methods described in [1]–[3] and [6]–[9], in the proposed approach we start with a purely unsupervised algorithm as a core model and the available side information is incorporated via a regularization term. Given Q labels, the approach is not restricted to find just Q classes (semisupervised classification) and instead it is able to uncover up to 2^Q hidden clusters (semisupervised clustering). In addition, it uses low embedding dimension to reveal the existing number of clusters which is important when one deals with large number of clusters. There is a systematic model selection scheme for tuning the parameters and it is provided with the out-of-sample extension property. Furthermore, the formulation is constructed for the multiclass semisupervised classification and clustering.

Here KSC [14] is used as the core model. In this case because of the discriminative property of KSC one can benefit from unlabeled data points. Unlike the KSC approach that projects the data to a $k-1$ dimensional space for being able to group the data into k clusters, in this paper the embedding dimension is rather equal to the number of available class-labels in the semisupervised learning framework. Therefore, the highlights of this manuscript can be summarized as follows.

1) Using an unsupervised model as the core model and incorporating the available side-information (labels) through a regularization term.
2) Addressing both multiclass semisupervised classification and semisupervised clustering.
3) Extension of the binary case to multiclass case and addressing the encoding schemes.
4) Realizing low embedding dimension to reveal the existing number of clusters.

This paper is organized as follows. In Section II, a brief review about kernel spectral clustering is given. In Section III, we formulate our multiclass semisupervised classification algorithm using a one-versus-all strategy. In Section IV, the semisupervised clustering algorithm is discussed. The model selection of the proposed method is discussed in Section V. In Section VI, numerical experiments are carried out to demonstrate the viability of the proposed method. Both synthetic and real-life data sets in different application domains such as in image segmentations and community detection in networks are considered.

II. Brief Overview of KSC

The KSC method corresponds to a weighted kernel PCA formulation providing a natural extension to out-of-sample data, that is, the possibility to apply the trained clustering model to out-of-sample points. Given training data $\mathcal{D} = \{x_i\}_{i=1}^M$, $x_i \in \mathbb{R}^d$, the primal problem of kernel spectral clustering is formulated as follows [14]:

$$\begin{aligned} \min_{w_\ell, b_\ell, e_\ell} \quad & \frac{1}{2}\sum_{\ell=1}^{k-1} w^{(\ell)^T} w^{(\ell)} - \frac{1}{2M}\sum_{\ell=1}^{k-1} \gamma_\ell e^{(\ell)^T} V e^{(\ell)} \\ \text{s.t.} \quad & e^{(\ell)} = \Phi w^{(\ell)} + b^{(\ell)} 1_M, \ \ell = 1, \ldots, k-1 \end{aligned} \tag{1}$$

where $e^{(\ell)} = [e_1^\ell, \ldots, e_M^\ell]^T$ are the projected variables, $\ell = 1, \ldots, k-1$ is the number of score variables required to encode the k clusters, $\gamma_\ell \in \mathbb{R}^+$ are the regularization constants, and $b^{(\ell)}$ is the bias term which is a scalar.

Here $\Phi = \left[\varphi(x_1), \ldots, \varphi(x_M)\right]^T$ and a vector of all ones with size M is denoted by 1_M. $\varphi(\cdot) : \mathbb{R}^d \to \mathbb{R}^h$ is the feature map and h is the dimension of the feature space which can be infinite dimensional. $w^{(\ell)}$ is the model parameters vector in the primal. $V = \text{diag}(v_1, \ldots, v_M)$ with $v_i \in \mathbb{R}^+$ is a user defined weighting matrix.

Applying the Karush–Kuhn–Tucker (KKT) optimality conditions one can show that the solution in the dual can be obtained by solving an eigenvalue problem of the following form:

$$V P_v \Omega \alpha^{(\ell)} = \lambda \alpha^{(\ell)} \tag{2}$$

where $\lambda = M/\gamma_\ell$, $\alpha^{(\ell)}$ are the Lagrange multipliers and P_v is the weighted centering matrix

$$P_v = I_M - \frac{1}{1_M^T V 1_M} 1_M 1_M^T V$$

where I_M is the $M \times M$ identity matrix and Ω is the kernel matrix with ijth entry $\Omega_{ij} = K(x_i, x_j) = \varphi(x_i)^T \varphi(x_j)$. In the ideal case of k well-separated clusters, for a properly chosen kernel parameter, the matrix $V P_v \Omega$ has $k-1$ piecewise constant eigenvectors with eigenvalue 1.

It should be noted that no assumption about the data is made upon applying the KSC algorithm. Because of the bias term b^ℓ, as follows from one of the KKT conditions associated with the

primal optimization (1), the kernel matrix gets automatically centered and it will be premultiplied by the centering matrix P_v in the dual (for more details see [14]).

The eigenvalue problem (2) is related to spectral clustering with random walk Laplacian. In this case, the clustering problem can be interpreted as finding a partition of the graph in such a way that the random walker remains most of the time in the same cluster with few jumps to other clusters, minimizing the probability of transitions between clusters. It is shown that if

$$V = D^{-1} = \text{diag}\left(\frac{1}{d_1}, \ldots, \frac{1}{d_M}\right)$$

where $d_i = \sum_{j=1}^{M} K(x_i, x_j)$ is the degree of the ith data point, the dual problem is related to the random walk algorithm for spectral clustering.

With the KKT optimality conditions, one can show that the score variables can be written as follows:

$$\begin{aligned} e^{(\ell)} &= \Phi w^{(\ell)} + b^{(\ell)} \\ &= \Phi\Phi^T \alpha^{(\ell)} + b^{(\ell)} = \Omega\alpha^{(\ell)} + b^{(\ell)}, \quad \ell = 1, \ldots, k-1. \end{aligned}$$

For the model selection, that is, selection of the number of clusters k and kernel parameter, several criteria have been proposed in literature including a Fisher [18], balanced line fit (BLF) [14], or Silhouette criterion [19]. These criteria use the special structure of the projected out-of-sample points to estimate the out-of-sample eigenvectors for selecting the model parameters. When the clusters are well separated the out-of-sample eigenvectors show a localized structure in the eigenspace.

The out-of-sample extensions to test points $\{x_i\}_{i=1}^{N\text{test}}$ is done by an error-correcting output coding (ECOC) decoding scheme. First the cluster indicators are obtained by binarizing the score variables for test data points as follows:

$$\begin{aligned} q_{\text{test}}^{\ell} &= \text{sign}\left(e_{\text{test}}^{\ell}\right) = \text{sign}\left(\Phi_{\text{test}} w^{(\ell)} + b^{(\ell)}\right) \\ &= \text{sign}(\Omega_{\text{test}}\alpha^{(\ell)} + b^{(\ell)} 1_{Ntest}) \end{aligned}$$

where $\Phi_{\text{test}} = [\varphi(x_1), \ldots, \varphi(x_{N\text{test}})]^T$ and $\Omega_{\text{test}} = \Phi_{\text{test}}\Phi^T$. The decoding scheme consists of comparing the cluster indicators obtained in the test stage with the codebook (which is obtained in the training stage) and selecting the nearest codeword in terms of Hamming distance.

In what follows we study two scenarios.

1) In the first case the number of available class labels is equal to the actual number of existing classes.
2) The second case corresponds to the situation where the number of available class labels is less than both the number of existing classes and the number of existing clusters.

In this paper, the terminology semisupervised classification is used for referring to the first case and the problem of the second case is referred to as semisupervised clustering.

III. Semisupervised Classification

In this section, we assume that there is a total number of Q classes ($\mathcal{C}_j$, $j = 1, \ldots, Q$). The corresponding number of available class labels is also equal to Q. Suppose the training data set $\mathcal{D}$ consists of M data points and is defined as follows:

$$\mathcal{D} = \{\underbrace{x_1, \ldots, x_N}_{\substack{\text{Unlabeled data} \\ (\mathcal{D}_U)}}, \quad \underbrace{x_{N+1}, \ldots, x_M}_{\substack{\text{Labeled data} \\ (\mathcal{D}_L)}}\}$$

where $\{x_i\}_{i=1}^{M} \in \mathbb{R}^d$. The labels are available for the last $N_L = M - N$ data points in $\mathcal{D}_L$ and are denoted by

$$Z = \left[z_{N+1}^T, \ldots, z_M^T\right]^T \in \mathbb{R}^{(M-N)\times Q}$$

with $z_i \in \{+1, -1\}^Q$ is the encoding vector for the training point x_i.

In the proposed method, we start with an unsupervised algorithm as a core model. Then by introducing a regularization term, we incorporate the available side information, which in this case are the labels, to the core model. Here the kernel spectral clustering is used as the core model. As it has been shown in [14] in contrast to classical spectral clustering, KSC has a systematic model selection scheme for tuning the parameters and it is provided with the out-of-sample extension property.

The one-versus-all strategy is utilized to build the codebook, that is, the training points belonging to the ith class are labeled by $+1$ and all the remaining data from the rest of the classes are considered to have negative labels. Both the labeled and unlabeled data points are arranged such that the top N data points are the unlabeled ones and the rest, that is, N_L, are the labeled data points. We consider the labels of the unlabeled data points to be zero as in [16]. In our formulation, unlabeled data points are only regularized using the KSC core model.

A. Primal-Dual Formulation of the Method

We formulate the multiclass semisupervised learning in the primal as the following optimization problem:

$$\begin{aligned} \min_{w^{(\ell)},\, b^{(\ell)},\, e^{(\ell)}} \quad & \frac{1}{2}\sum_{\ell=1}^{Q} w^{(\ell)T} w^{(\ell)} - \frac{\gamma_1}{2}\sum_{\ell=1}^{Q} e^{(\ell)T} V e^{(\ell)} \\ & + \frac{\gamma_2}{2}\sum_{\ell=1}^{Q} (e^{(\ell)} - c^{(\ell)})^T A (e^{(\ell)} - c^{(\ell)}) \\ \text{s.t.} \quad & e^{(\ell)} = \Phi w^{(\ell)} + b^{(\ell)} 1_M, \ \ell = 1, \ldots, Q \end{aligned} \tag{3}$$

where c^{ℓ} is the ℓth column of the matrix C defined as

$$C = [c^{(1)}, \ldots, c^{(Q)}]_{M\times Q} = \left[\frac{0_{N\times Q}}{Z}\right]_{M\times Q} \tag{4}$$

where $0_{N\times Q}$ is a zero matrix of size $N \times Q$ and Z is defined as previously. $b^{(\ell)}$ is a bias term which is a scalar. The matrix A is defined as follows:

$$A = \left[\begin{array}{c|c} 0_{N\times N} & 0_{N\times N_L} \\ \hline 0_{N_L\times N} & I_{N_L\times N_L} \end{array}\right]$$

where $I_{N_L\times N_L}$ is the identity matrix of size $N_L \times N_L$.

The available prior knowledge, that is, the labels, is added to the KSC model through the third term in the objective function of (3). This term aims at minimizing the difference between the score variables of the labeled data points, that is, e_i for $i \in \mathcal{D}_L$, and the actual labels provided by the user.

Therefore, it enforces the e_i values for the labeled data points to be close enough to the actual labels in the projection space. Furthermore, since we do not intend to prejudge about the memberships of the unlabeled data points, the matrix A is taking place in the third term in the objective function.

Lemma 3.1: Given a positive definite kernel function $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ with $K(x,z) = \varphi(x)^T\varphi(z)$ and regularization constants $\gamma_1, \gamma_2 \in \mathbb{R}^+$, the solution to (3) is obtained by solving the following dual problem:

$$(I_M - RS\Omega)\alpha^{(\ell)} = \gamma_2 S^T c^{(\ell)}, \quad \ell = 1, \ldots, Q \tag{5}$$

where $R = \gamma_1 V - \gamma_2 A$, $\alpha^{(\ell)} = [\alpha_1^{(\ell)}, \ldots, \alpha_M^{(\ell)}]^T$ are the Lagrange multipliers and $S = I_M - (1/1_M^T R 1_M) 1_M 1_M^T R$. Ω and I_M are defined as previously.

Proof: The Lagrangian of the constrained optimization problem (3) becomes

$$\begin{aligned}\mathcal{L}(w^{(\ell)}, b^{(\ell)}, e^{(\ell)}, \alpha^{(\ell)}) &= \frac{1}{2}\sum_{\ell=1}^{Q} w^{(\ell)^T} w^{(\ell)} - \frac{\gamma_1}{2}\sum_{\ell=1}^{Q} e^{(\ell)^T} V e^{(\ell)} \\ &+ \frac{\gamma_2}{2}\sum_{\ell=1}^{Q} (e^{(\ell)} - c^{(\ell)})^T A (e^{(\ell)} - c^{(\ell)}) \\ &+ \sum_{\ell=1}^{Q} \alpha^{(\ell)^T} (e^{(\ell)} - \Phi w^{(\ell)} - b^{(\ell)} 1_M)\end{aligned}$$

where $\alpha^{(\ell)}$ is the vector of Lagrange multipliers. Then the KKT optimality conditions are as follows:

$$\begin{cases} \dfrac{\partial \mathcal{L}}{\partial w^{(\ell)}} = 0 \to w^{(\ell)} = \Phi^T \alpha^{(\ell)}, \quad \ell = 1, \ldots, Q, \\ \dfrac{\partial \mathcal{L}}{\partial b^{(\ell)}} = 0 \to 1_M^T \alpha^{(\ell)} = 0, \quad \ell = 1, \ldots, Q, \\ \dfrac{\partial \mathcal{L}}{\partial e^{(\ell)}} = 0 \to \alpha^{(\ell)} = (\gamma_1 V - \gamma_2 A) e^{(\ell)} + \gamma_2 c^{(\ell)}, \ell = 1, \ldots, Q, \\ \dfrac{\partial \mathcal{L}}{\partial \alpha^{(\ell)}} = 0 \to e^{(\ell)} = \Phi w^{(\ell)} + b^{(\ell)} 1_M, \ell = 1, \ldots, Q. \end{cases} \tag{6}$$

Elimination of the primal variables $w^{(\ell)}, e^{(\ell)}$ and making use of Mercer's theorem [20], results in the following equation:

$$R\Omega\alpha^{(\ell)} + b^{(\ell)} R 1_M = \alpha^{(\ell)} - \gamma_2 c^{(\ell)}, \quad \ell = 1, \ldots, Q \tag{7}$$

where $R = \gamma_1 V - \gamma_2 A$. From the second KKT optimality condition and (7), the bias term becomes

$$b^{(\ell)} = (1/1_M^T R 1_M)(-1_M^T \gamma_2 c^{(\ell)} - 1_M^T R \Omega \alpha^{(\ell)}), \quad \ell = 1, \ldots, Q. \tag{8}$$

Substituting the obtained expression for the bias term $b^{(\ell)}$ into (7) along with some algebraic manipulation one can obtain the solution in dual as the following linear system:

$$\gamma_2 \left(I_M - \frac{R 1_M 1_M^T}{1_M^T R 1_M} \right) c^{(\ell)} = \alpha^{(\ell)} - R \left(I_M - \frac{1_M 1_M^T R}{1_M^T R 1_M} \right) \Omega \alpha^{(\ell)}.$$

■

Remark 3.1: It should be noted that since the optimization problem (3) does have equality constraints therefore the KKT conditions include the primal equality constraints and the gradient of the Lagrangian with respect to the primal variables (see [21, Ch. 5]). In (6), the first three equations correspond to the derivative of the Lagrangian with respect to primal variables and the primal equality constraints are equivalently obtained by taking the derivative of the Lagrangian with respect to dual variables.

It should be noticed that one can also obtain the following linear system when the primal variables $w^{(\ell)}, e^{(\ell)}$ are eliminated from (KKT) optimality conditions in (6):

$$\left[\begin{array}{c|c} \Omega - R^{-1} & 1_M \\ \hline 1_M^T & 0 \end{array}\right] \left[\begin{array}{c} \alpha^{(\ell)} \\ \hline b^{(\ell)} \end{array}\right] = \left[\begin{array}{c} -R^{-1}\gamma_2 c^{(\ell)} \\ \hline 0 \end{array}\right], \ \ell = 1, \ldots, Q \tag{9}$$

where $\alpha^{(\ell)} = [\alpha_1^{(\ell)}, \ldots, \alpha_M^{(\ell)}]^T$ and $\Omega = \Phi\Phi^T$ is the kernel matrix. Matrix R is a diagonal matrix and it is invertible if and only if $\gamma_1 v_i \neq \gamma_2$ for $i = 1, \ldots, M$.

The linear systems (5) and (9) have a unique solution when the associated coefficient matrix is full-rank which depends on the regularization parameters.

B. Encoding/Decoding Scheme

In semisupervised classification, the encoding scheme is chosen in advance since the number of existing classes is known beforehand. The codebook $\mathcal{CB}$ used for out-of-sample extension is defined using the encoding vectors for the training points. If $Z = [z_{N+1}^T, \ldots, z_M^T]^T$ is the encoding matrix for the training points, the $\mathcal{CB} = \{c_q\}_{q=1}^{Q}$, where $c_q \in \{-1, 1\}^Q$, is defined by the unique rows of Z (i.e., from identical rows of Z one selects one row). Considering the test set $\mathcal{D}^{\text{test}} = \{x_i^{\text{test}}\}_{i=1}^{N_{\text{test}}}$ the score variables evaluated at the test points become

$$\begin{aligned} e_{\text{test}}^{(\ell)} &= \Phi_{\text{test}} w^{\ell} + b^{(\ell)} 1_{N_{\text{test}}} \\ &= \Omega_{\text{test}} \alpha^{(\ell)} + b^{(\ell)} 1_{N_{\text{test}}}, \quad \ell = 1, \ldots, Q \end{aligned} \tag{10}$$

where $\Omega_{\text{test}} = \Phi_{\text{test}} \Phi^T$. The procedure for the multiclass semisupervised classification is summarized in Algorithm 1.

IV. Semisupervised Clustering

In what follows we assume that there is a total number of T clusters and a few labels from Q of the clusters are available ($Q \leq T$). Therefore, we are dealing with the case that some of the clusters are partially labeled. The aim is to incorporate these labels in the learning process to guide the clustering algorithm to adjust the membership of the unlabeled data. Next we will show how one can use the approach described in Section III in this setting.

A. From Solution of Linear Systems to Clusters: Encoding

Because the number of existing clusters is not known *a priori*, we cannot use the predefined codebook as in semisupervised classification. Therefore, a new scheme is developed for generating a codebook to be used in the learning process.

Algorithm 1: Multiclass Semisupervised Classification

Input: Training data set $\mathcal{D}$, labels Z, tuning parameters $\{\gamma_i\}_{i=1}^2$, kernel parameter (if any), test set $\mathcal{D}^{\text{test}} = \{x_i^{\text{test}}\}_{i=1}^{N_{\text{test}}}$, and codebook $\mathcal{CB} = \{c_q\}_{q=1}^{Q}$

Output: Class membership of test data points $\mathcal{D}^{\text{test}}$

1 Solve the dual linear system (5) to obtain $\{\alpha^\ell\}_{\ell=1}^{Q}$ and compute the bias term $\{b^\ell\}_{\ell=1}^{Q}$ using (8).
2 Estimate the test data projections $\{e_{\text{test}}^{(\ell)}\}_{\ell=1}^{Q}$ using (10).
3 Binarize the test projections and form the encoding matrix $[\text{sign}(e_{\text{test}}^{(1)}), \ldots, \text{sign}(e_{\text{test}}^{(Q)})]_{N_{\text{test}} \times Q}$ for the test points. (Here $e_{\text{test}}^{(\ell)} = [e_{\text{test},1}^{(\ell)}, \ldots, e_{\text{test},N_{\text{test}}}^{(\ell)}]^T$).
4 $\forall i$, assign x_i^{test} to class q^*, where $q^* = \underset{q}{\text{argmin}}\, d_H(e_{\text{test},i}^{\ell}, c_q)$ and $d_H(\cdot,\cdot)$ is the Hamming distance.

Algorithm 2: Semisupervised Clustering

Input: Training data set $\mathcal{D}$, labels Z, the tuning parameters $\{\gamma_i\}_{i=1}^2$, the kernel parameter (if any), number of clusters k, the test set $\mathcal{D}^{\text{test}} = \{x_i^{\text{test}}\}_{i=1}^{N_{\text{test}}}$, and number of available class labels, that is, Q

Output: Cluster membership of test data points $\mathcal{D}^{\text{test}}$

1 Solve the dual linear system (5) to obtain $\{\alpha^\ell\}_{\ell=1}^{Q}$ and compute the bias term $\{b^\ell\}_{\ell=1}^{Q}$ using (8).
2 Binarize the solution matrix $S_\alpha = [\text{sign}(\alpha^{(1)}), \ldots, \text{sign}(\alpha^{(Q)})]_{M \times Q}$, where $\alpha^\ell = [\alpha_1^\ell, \ldots, \alpha_M^\ell]^T$.
3 Form the codebook $\mathcal{CB} = \{c_q\}_{q=1}^{p}$, where $c_q \in \{-1, 1\}^Q$, using the k most frequently occurring encodings from unique rows of solution matrix S_α.
4 Estimate the test data projections $\{e_{\text{test}}^{(\ell)}\}_{\ell=1}^{Q}$ using (10).
5 Binarize the test projections and form the encoding matrix $[\text{sign}(e_{\text{test}}^{(1)}), \ldots, \text{sign}(e_{\text{test}}^{(Q)})]_{N_{\text{test}} \times Q}$ for the test points (Here $e_{\text{test}}^{\ell} = [e_{\text{test},1}^{\ell}, \ldots, e_{\text{test},N_{\text{test}}}^{\ell}]^T$).
6 $\forall i$, assign x_i^{test} to class/cluster q^*, where $q^* = \underset{q}{\text{argmin}}\, d_H(e_{\text{test},i}^{\ell}, c_q)$ and $d_H(\cdot,\cdot)$ is the Hamming distance.

It has been observed that the solution vector $\alpha^{(\ell)}$, $\ell = 1, \ldots, Q$ of the dual linear system (5) has a piecewise constant property when there is an underlying cluster structure in the data [see Fig. 2(d)]. Once the solution to (5) is found, the codebook $\mathcal{CB} \in \{-1, 1\}^{p \times Q}$ is formed by the unique rows of the binarized solution matrix (i.e., $[\text{sign}(\alpha^{(1)}), \ldots, \text{sign}(\alpha^{(Q)})]$). The maximum number of clusters that can be decoded is 2^Q since the maximum value that p can take is 2^Q. In our approach the number of encodings, that is, p, is tuned along with the model selection procedure. Therefore, a grid search on the interval $[Q, 2^Q]$ is conducted to determine the number of clusters.

It should be noted that in Algorithm 1, the static codebook $\mathcal{CB}$ (static in a sense that the number of codewords is fixed and only depends on Q) is known beforehand and is of size $Q \times Q$. On the other hand in Algorithm 2, the codebook $\mathcal{CB}$ is no longer a static codebook and is of size $p \times Q$, where p can be maximally 2^Q. Furthermore, it is obtained based on the solution matrix S_α (see steps 2 and 3 in Algorithm 2).

B. Low-Dimensional Spectral Embedding

One may notice that as opposed to kernel spectral clustering [14] where the score variables lie in a $T-1$ (where T is the actual number of clusters) dimensional space, in our formulation the embedding dimension is Q which can be smaller than T. This can also be seen as the optimized embedding dimension for clustering which plays an important role when the number of existing clusters is large. In fact one only requires $Q = \log T$ solution vectors to uncover T clusters. Therefore, one is able to deal with a larger number of clusters in a more compact way. In contrast with the KSC approach where one needs to solve an eigenvalue problem, in our formulation we solve a linear system. It should be noted that although the two approaches share almost the same computational complexity, the quality of the solution vector obtained by the proposed algorithm is higher that that of KSC as shown in Figs. 5 and 6. This demonstrates the advantage of prior knowledge incorporation. The proposed semisupervised clustering is summarized in Algorithm 2.

V. Model Selection

The performance of the multiclass semisupervised model depends on the choice of the tuning parameters. In the case of RBF kernel, the optimal values of γ_1, γ_2 and the kernel parameter σ can be obtained by evaluating the performance of the model (classification accuracy) on the validation set using a grid search over the parameters. One may also consider to use coupled simulated annealing (CSA) to minimize the misclassification error in the cross-validation process. CSA leads to an improved optimization efficiency as it reduces the sensitivity of the algorithm with respect to the initialization of the parameters while guiding the optimization process to quasi-optimal runs [22].

In our experiments, with the analysis given in [16, Sec. III.C] we set $\gamma_1 = 1$. Then we tune γ_2 and σ through a grid search. The range in which the search is made is discussed for each of the experiments in Section VI. In general in our experiments we observed that a good value for γ_2, most of the times, is selected from the range $[0, 1]$.

Since labeled and unlabeled data points are involved in the learning process, it is natural to have a model selection criterion that makes use of both. Therefore, for semisupervised classification, one may combine two criteria where one of them evaluates the performance of the model on the unlabeled data points (evaluation of clustering results) and the other one maximizes the classification accuracy [16], [17].

A common approach for evaluating the quality of the clustering results consists of using internal cluster validity indices [23] such as Silhouette, Fisher, and Davies-Bouldin index (DB) criteria. In this paper, the Silhouette index is used to assess the clustering results. The Silhouette technique

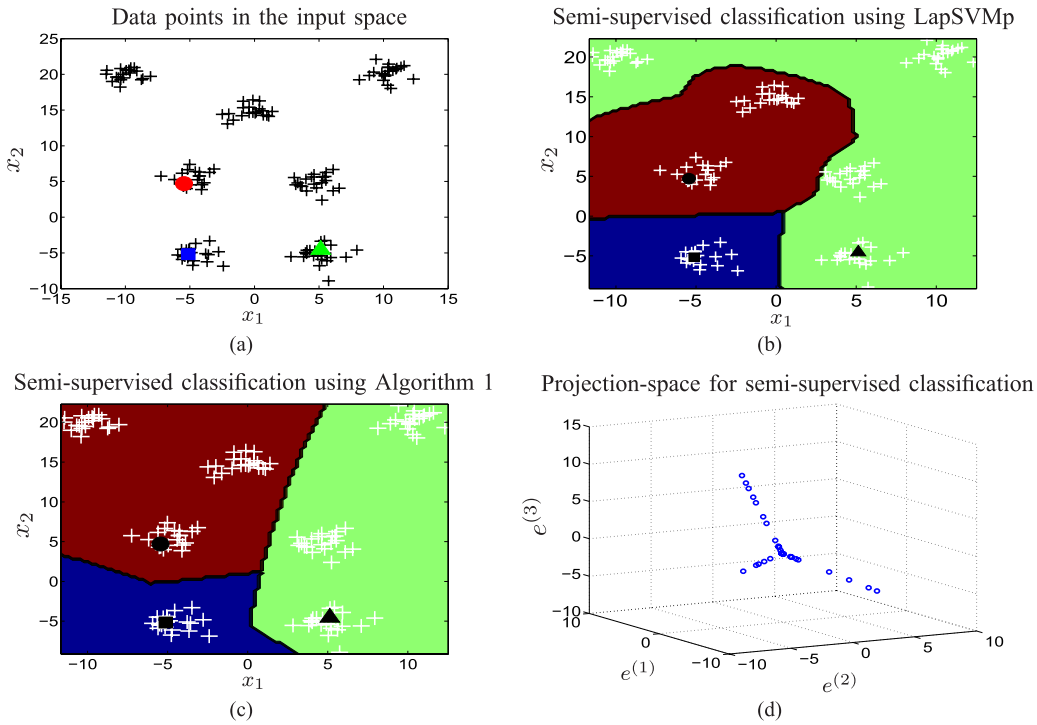


Fig. 1. Toy problem: seven well-separated Gaussians. The labeled data points of only three classes are available and are depicted by the blue squares (■), green triangles (▲), and red circles (●). (a) Data points in the original space. (b) Result of multiclass semisupervised classification using LapSVMp with RBF kernel. (c) Result of the proposed multiclass semisupervised classification with RBF kernel (note that the algorithm detected three classes. The first class consists of one cluster whereas the second and third classes consist of three clusters each). (d) Projections of the validation data points when the proposed semisupervised classification algorithm is used (indicating the line structure in the projection-space).

assigns to the ith sample of jth class, $\mathcal{C}_j$, a quality measure $s(i)$ which is defined as

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

where $a(i)$ is the average distance between the ith sample and all of the samples included in $\mathcal{C}_j$. b_i is the minimum average distance from the ith sample to points in different clusters. The silhouette value for each sample is a measure of how similar that sample is to samples in its own cluster versus samples in other clusters, and is in the range of $[-1, 1]$.

The proposed model selection criterion for semisupervised learning, with kernel parameter σ, can be expressed as follows:

$$\max_{\gamma_1,\gamma_2,\sigma,k} \eta\, \mathrm{Sil}(\gamma_1, \gamma_2, \sigma, k) + (1 - \eta)\mathrm{Acc}(\gamma_1, \gamma_2, \sigma, k). \quad (11)$$

It is a convex combination of Silhouette (Sil) and classification accuracy (Acc). $\eta \in [0, 1]$ is a user-defined parameter that controls the trade off between the importance given to unlabeled and labeled instances. If few labeled data points are available, one may give more weight to Silhouette criterion and vice versa.

The silhouette criterion is evaluated on the unlabeled data points in the validation set. One can also consider to evaluate it on the out-of-sample solution vectors.

In (11), k denotes the number of clusters that is unknown beforehand. In the case of semisupervised classification where the number of classes is known *a priori*, one does not need to tune k and thus it can be removed from the list of decision variables of the aforementioned model selection criterion.

In any unsupervised learning algorithm, one has to find the right number of existing clusters over the specified range which is provided by the user. When there is a form of prior knowledge about the data under study, the search space is reduced. In our semisupervised clustering, the lower bound of the range in which the number of clusters are sought is Q (assuming that Q cluster labels are available). Therefore, applying the proposed MSS-KSC algorithm will make it easier to reveal the lower level of the cluster hierarchy. In the proposed MSS-KSC approach, one requires to solve a linear system. Therefore, the complexity of the proposed MSS-KSC algorithm in the worst case scenario is $\mathcal{O}(M^3)$ where M is the number of training data points.

VI. Experimental Results

In this section, some experimental results are presented to illustrate the applicability of the proposed semisupervised classification and clustering approaches. We start with a toy problem and show the differences between the obtained results when semisupervised classification and semisupervised clustering are applied on the same data (see Figs. 1 and 2).

The performance of the proposed algorithms is also tested on two moons and two spirals data sets which are standard benchmark for semisupervised learning algorithms used in the literature [24].

Next we apply the proposed semisupervised classification on some benchmark data sets taken from the UCI machine learning repository and the performance is compared with Laplacian SVM [6] and MeanS3VM [25]. Then, we have tested the performance of the semisupervised clustering on

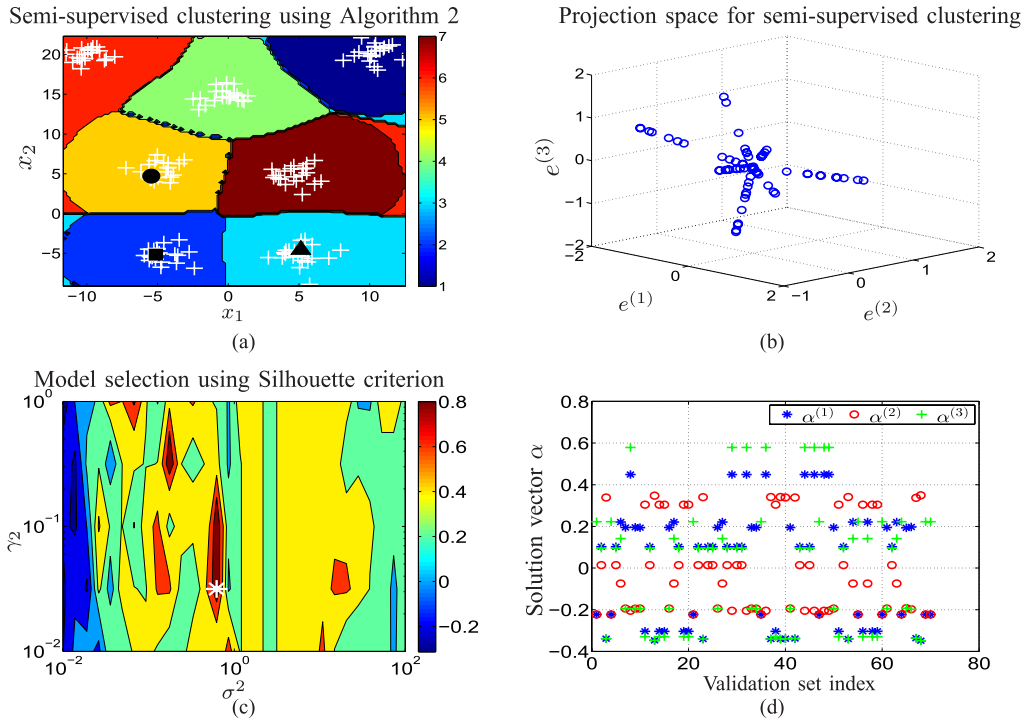


Fig. 2. Toy problem: seven well-separated Gaussians. The labeled data points of only three classes are available and are depicted by the blue squares (■), green triangles (▲), and red circles (●). (a) Result of the proposed multiclass semisupervised clustering with RBF kernel. (b) Projections of the validation data points when semisupervised clustering algorithm is used (indicating the line structure in the projection-space). (c) Model selection for semisupervised clustering using Silhouette validity index corresponding to the best case $T = 7$. The asterisk (*) marks the optimal model. (d) Piecewise constant property of the solution vector.

image segmentation tasks, and the obtained results are compared with the kernel spectral clustering algorithm [14]. Finally the application of the semisupervised classification is also shown in community detection of real-world networks.

A. Toy Problems

The performance of the proposed semisupervised classification and clustering algorithms are shown on a synthetic data set consisting of seven well-separated Gaussians. Some labeled data points from three of them are available [see Fig. 1(a)]. When the semisupervised classification algorithm is used the data are grouped into three classes as the codebook used in semisupervised classification is a static codebook and it consists of three codewords. On the other hand, in semisupervised clustering algorithm the codebook is designed using the solution vector of the associated linear system and is not static, that is, the number of codewords is not fixed and is tuned. Therefore, by applying the semisupervised clustering one is able to partition the data into seven clusters. As it can be seen from Figs. 1(d) and 2(b), the projected data points are embedded in 3-D space and yet we are able to cluster them in contrast with kernel spectral clustering algorithm [14] which requires an embedding space with dimension 6 to be able to group the given data sets into seven clusters.

We also conducted experiments on nonlinear toy problems such as two moons and two spirals and the obtained results are shown in Fig. 3. For two spirals data set, two scenarios are tested corresponding to different positions of the labeled data point. A comparison is made with LRGA algorithm[1] proposed in [8]. The LRGA algorithm has two parameters k and λ. In these experiments, the parameter k (size of the neighborhood) is set to 10 and λ is searched within $[1, 10^{16}]$ using a logarithmic scale. As Fig. 3 shows, for the two moons data set the results of both method are comparable. However, the results of the two spirals data set indicate that our proposed algorithm is less sensitive to the position of labeled data points[2] compared with LRGA algorithm.

In these experiments, γ_2 and σ are tuned through a grid search. The range in which the search (using a logarithmic scale) is made for γ_2 and σ is shown in Figs. 2(c) and 3(g)–(i). From these figures, it is apparent that there exist a range of γ_2 and σ for which the value of the utilized model selection criterion is quite high on the validation set.

B. Real-Life Benchmark Data Sets

Four benchmark data sets used in the following experiments are chosen from the UCI machine learning repository [26]. The benchmark consists of Wine, Iris, Zoo and Seeds data sets. In all cases, the data points are divided using proportion 80% and 20% into training and test counterparts, respectively. Then one fourth of randomly selected data points in the training set are considered to be labeled and the remaining three fourths are unlabeled. The performance of the proposed semisupervised classification approach (MSS-KSC), is compared with Laplacian SVM (LapSVMp)[3] [6] and MeanS3VM [25] using the one-versus-all strategy.

[1] Available at: http://www.cs.cmu.edu/∼yiyang/LRGA_ranking.m

[2] The equivalent of the query provided by the user.

[3] Available at: http://www.dii.unisi.it/∼melacci/lapsvmp/

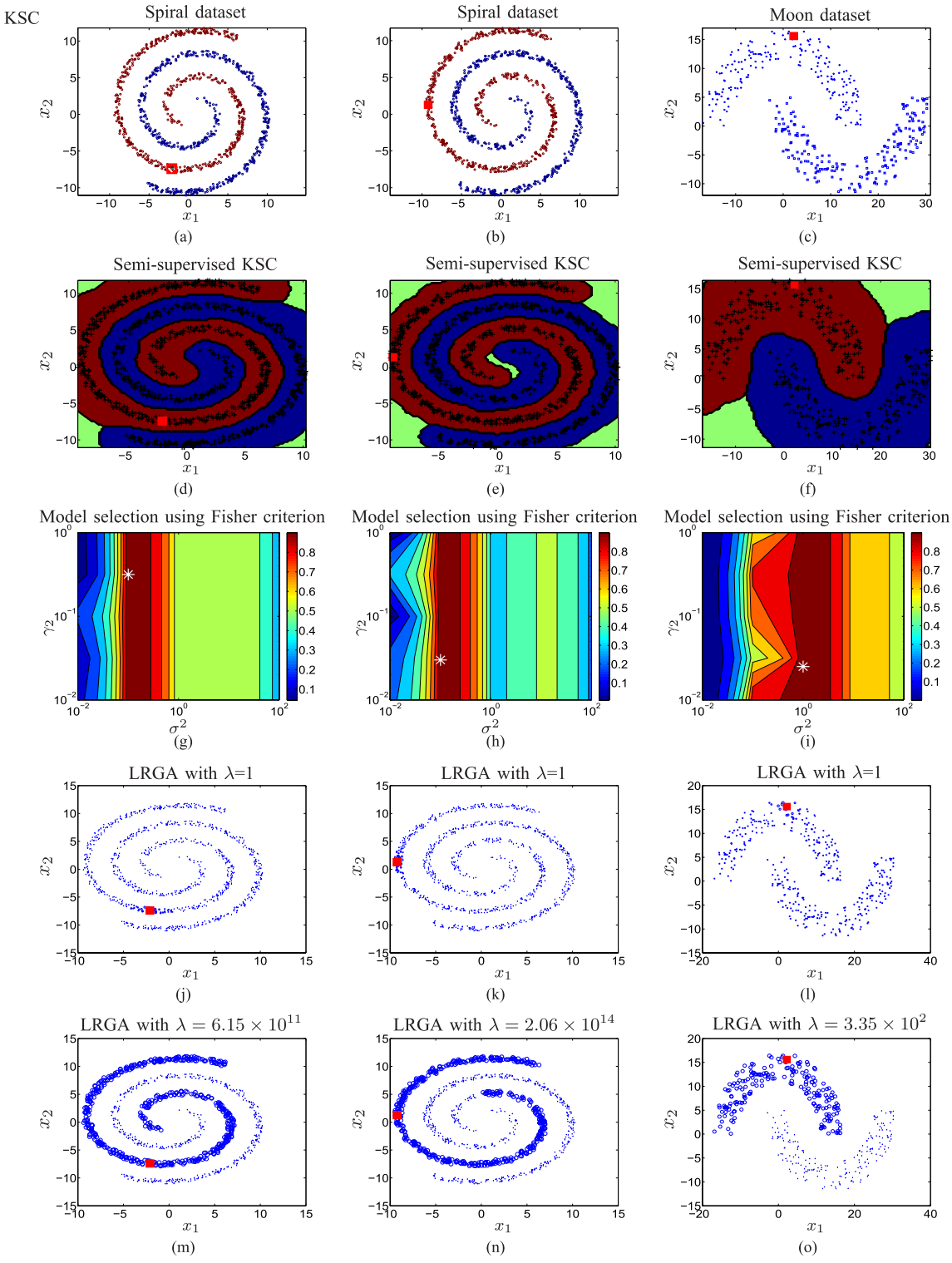


Fig. 3. Toy problems: two spiral and two moon data sets. The labeled data point is depicted by the red squares (■). (a)–(c) Data points in the original space. (d)–(f) Result of the proposed semisupervised algorithm with RBF kernel. (g)–(i) Model selection of the proposed algorithm. [The asterisk (*) marks the optimal model for these examples.] (j)–(l) Result of the LRGA algorithm corresponding to the worst case when the parameter k (size of the neighborhood) is set to 10 and λ is searched within $[1, 10^{16}]$ using a logarithmic scale. (m)–(o) Result of the LRGA algorithm corresponding to the best case when the parameter k (size of the neighborhood) is set to 10 and λ is searched within $[1, 10^{16}]$ using a logarithmic scale.

In this paper, the procedure used for model selection is a two-step procedure which consists of CSA [22] initialized with random sets of parameters for the first step and the simplex method [27] for the second step. After CSA converges to some local minima, the parameters that obtained the lowest misclassification error are used for initialization of the simplex procedure to refine our selection. At every iteration for CSA method a tenfold cross-validation

TABLE I

AVERAGE ACCURACY AND THE STANDARD DEVIATION OF THE LAPSVMP [6], MEANS3VM-ITER [25], MEANS3VM-MKL [25], AND THE PROPOSED MSS-KSC APPROACH ON FOUR REAL DATA SETS FROM UCI REPOSITORY [26]

Dataset	# of attributes	# of classes	# of data points	$\mathcal{D}^{\text{train}}_{\text{Labeled}}/\mathcal{D}^{\text{train}}_{\text{Unlabeled}}/\mathcal{D}^{\text{test}}$	Method			
					MSS-KSC	**LapSVMp**	**means3vm-iter**	**means3vm-mkl**
Wine	13	3	178	36/107/35	$\mathbf{0.96 \pm 0.02}$	0.94 ± 0.03	0.95 ± 0.02	0.94 ± 0.07
Iris	4	3	150	30/90/30	0.89 ± 0.08	0.88 ± 0.05	$\mathbf{0.90 \pm 0.03}$	0.89 ± 0.01
Zoo	16	7	101	21/60/20	$\mathbf{0.93 \pm 0.05}$	0.90 ± 0.06	0.88 ± 0.02	0.89 ± 0.07
Seeds	7	3	210	42/126/42	$\mathbf{0.90 \pm 0.04}$	0.89 ± 0.03	0.88 ± 0.07	0.89 ± 0.02

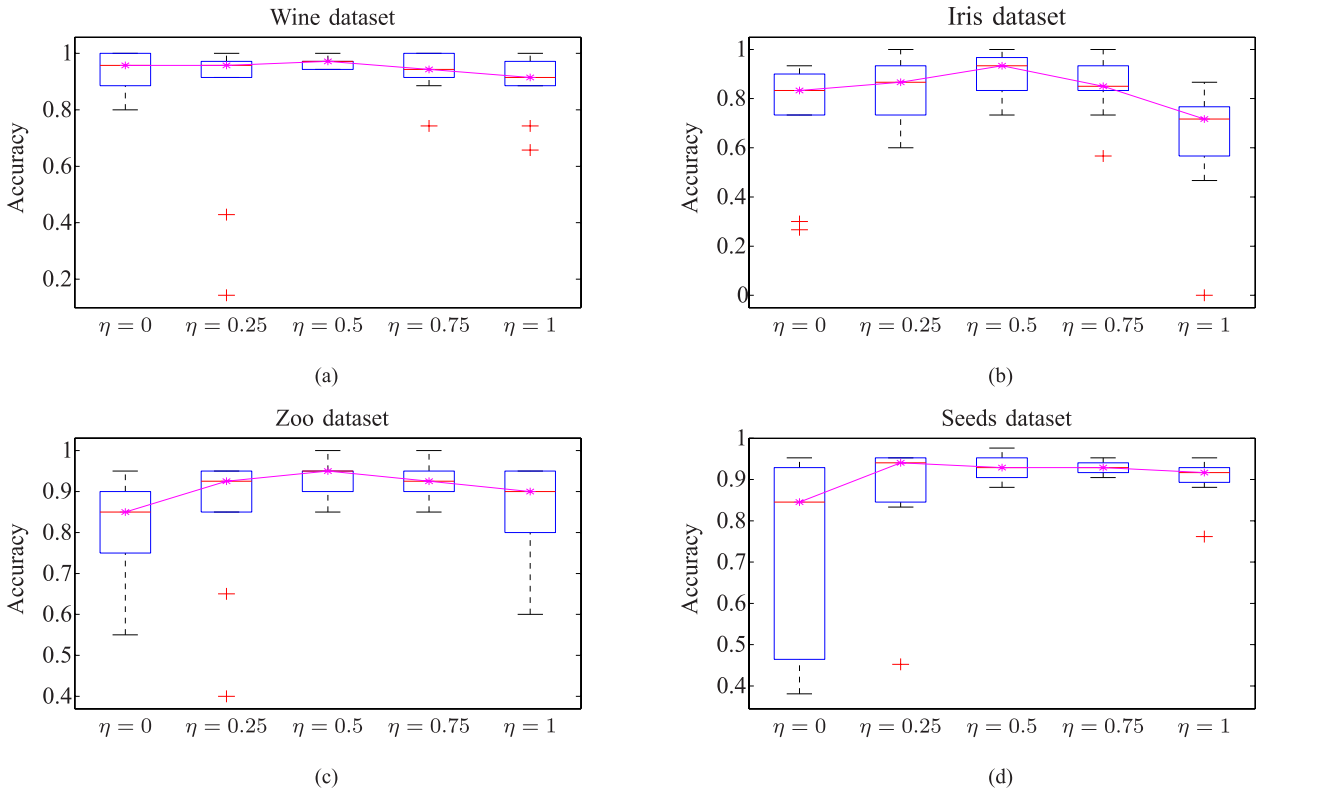


Fig. 4. Obtained accuracy of the proposed MSS-KSC approach, with respect to different η value, over ten simulation runs. The outliers are denoted by red $+$. (a) Wine dataset. (b) Iris dataset. (c) Zoo dataset. (d) Seeds dataset.

is used. In all the experiments, the RBF kernel is used.

For MeanS3VM method, the regularization parameters $C1$ and $C2$ are fixed to 1 and 0.1, respectively, and the width parameter in RBF kernel is tuned with respect to the accuracy on the validation set. For the Laplacian SVMs, we tuned the kernel parameter and γ_A with respect to the accuracy on the validation set. The remaining parameters, that is, γ_I and NN, are set to their default values ($\gamma_I = 1$ and $NN = 6$).

The mean and standard deviation of the accuracy rates on test data points with respect to ten random splits are reported in Table I. Table I shows that the proposed MSS-KSC approach outperforms in most cases the other approaches on these tested problems. The effect of changing the value of the user defined parameter η, used for model selection, on the performance of the proposed algorithm with respect to ten random splits can be seen in Fig. 4.

C. Image Segmentation

In this section, the task is to segment the given image using the proposed semisupervised clustering. Here the aim is to show that by incorporating the side-information (labels in this case) to the unsupervised model, it is possible to improve the result of the unsupervised algorithm.

Experimental results on two synthetic images and some color images from the Berkeley image data set [28] are shown in Figs. 5 and 6. For each image, a local color histogram with a 5×5 local window around each pixel is computed using minimum variance color quantization of eight levels. A subset of 500 unlabeled pixels together with some labeled pixels (see Table II) are used for training and the whole image for testing.

For the synthetic images, we provided a qualitative evaluation of both approaches, since the ground truth of these images were not available. For the Berkeley images data set for which the ground truth segmentations are known, the segmentations obtained by MSS-KSC and KSC are compared with the ground truth in Table II. Two evaluation criteria are used.

1) F-measure, that is, $(2 \times \text{Precision} \times \text{Recall})/\text{Precision}+\text{Recall}$ with respect to human ground-truth boundaries.
2) Variation of information (VI): it measures the distance between two segmentations in terms of their average

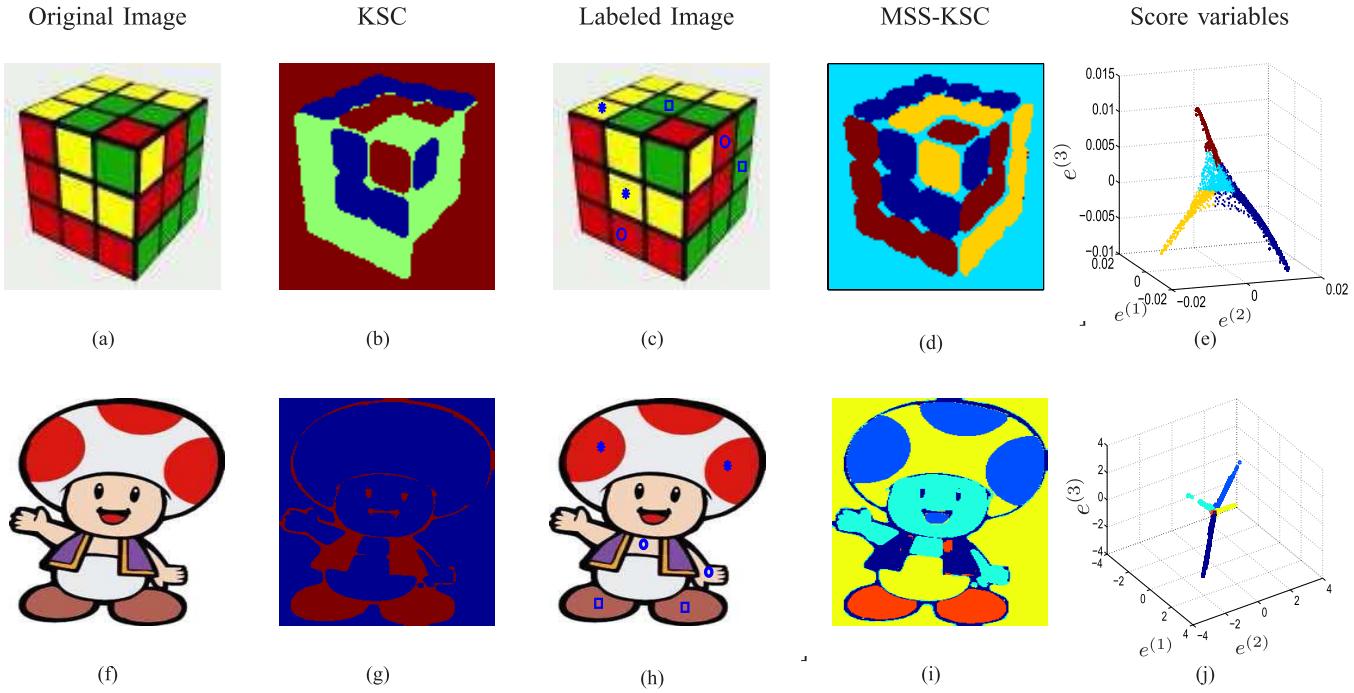


Fig. 5. (a) and (f) Original image used for the KSC algorithm. (b) and (g) Segmented image using the KSC approach. (c) and (h) Labeled image used for the proposed semisupervised clustering algorithm. (d) and (i) Segmented image using the MSS-KSC approach. (e) and (j) Score variables in the projection space.

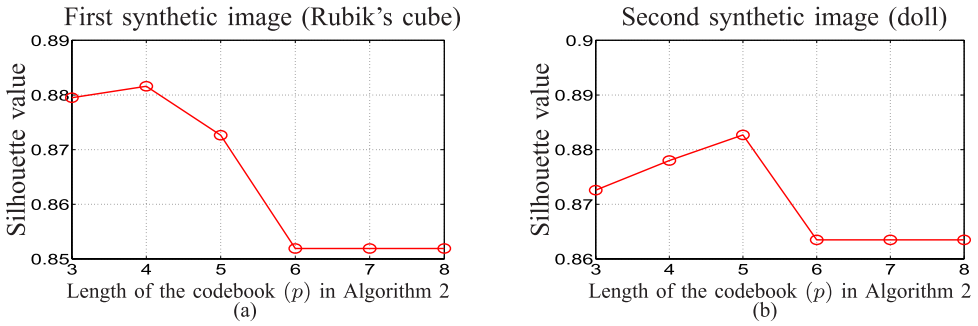


Fig. 6. Model selection curves corresponding to the obtained Silhouette value for different p value. (a) Three labels were provided ($Q = 3$). The maximum Silhouette value for the first syntactic image, while σ and γ_2 are tuned, over a range of $p \in \{Q, \ldots, 2^Q\}$. (b) Three labels were provided ($Q = 3$). The maximum Silhouette value for the second syntactic image, while σ and γ_2 are tuned, over a range of $p \in \{Q, \ldots, 2^Q\}$.

TABLE II
COMPARISON OF KSC AND MSS-KSC FOR IMAGE SEGMENTATIONS IN TERMS OF F-MEASURE AND VARIATION OF INFORMATION INDICES

Image ID	Q	$\mathcal{D}$		$\mathcal{D}^{val}$		F-measure		Variation of information	
		$\mathcal{D}_u$	$\mathcal{D}_L$	$\mathcal{D}_u$	$\mathcal{D}_L$	KSC	MSS-KSC	KSC	MSS-KSC
100007	4	500	8	3000	8	0.57	**0.62**	**1.64**	1.95
295087	4	500	8	3000	5	0.59	**0.62**	**2.54**	2.88
372019	3	500	6	3000	6	0.40	**0.44**	2.83	**2.44**
385039	5	500	14	3000	12	0.48	**0.48**	3.20	**3.18**
388067	3	500	6	3000	6	0.60	**0.74**	4.61	**4.50**
8049	3	500	6	3000	7	0.70	**0.75**	2.22	**2.07**

Note: For variation of information the lower the value the better, whereas for F-measure the higher value the better the segmentation is.

conditional entropy. Low values indicate good match between the segmentations [29].

In these experiments, the range in which the search (using a logarithmic scale) is made for tuning the parameters γ_2 and σ are $[0, 1]$ and $[10^{-3}, 10^1]$, respectively. The length of the codebook p is also tuned on the interval $[Q, 2^Q]$. The score variables obtained by the proposed MSS-KSC algorithm for two images are shown in Fig. 5 when Silhouette criterion is used. As it can be seen, the embedding dimension (spectral embedding) is three and yet we can detect more than four clusters from the given image. Unlike the toy example 1 for these images, as clusters are not well separated, the line

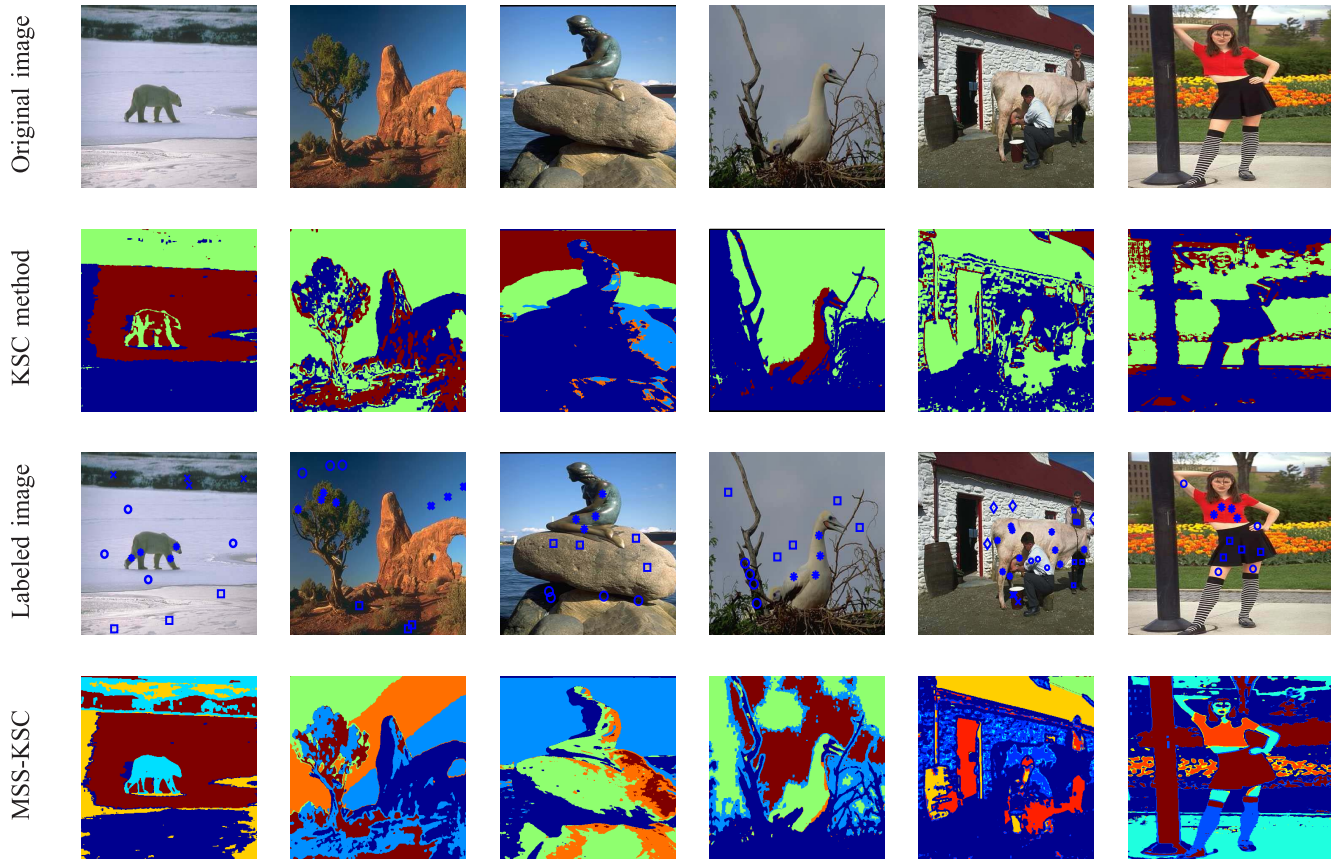


Fig. 7. Image segmentation results using the proposed method and the KSC [14]. A subset of 500 and 3000 randomly chosen pixel histograms (unlabeled data points) together with some labeled data points are used for training and validation, respectively. The whole image is used for testing. The original image is shown in the first row. The segmentation results obtained by KSC using the original images are shown in the second row. The third row shows the images labeled by human. The results of the proposed semisupervised clustering algorithm applied on the labeled images are shown in the fourth row.

TABLE III
KARATE NETWORK. COMPARISON OF MSS-KSC AND METHODS DESCRIBED IN [36] IN TERMS OF AVERAGED NMI

Methods in [36]						**The proposed method**	
pairs constraints %	# **pairs** [r,t]	**NMF-LSE**	**NMF-KL**	**SNMF**	**SP**	# **nodes**	**MSS-KSC**
2%	4 [4,8]	0.98	0.73	0.51	0.90	4	0.91
4%	6 [6,12]	0.99	0.85	0.60	0.96	6	0.95
5%	8 [8,16]	0.99	0.89	0.53	0.95	8	0.98
10%	16 [16,32]	1.00	0.89	0.57	1.00	10	1.00
20%	31 [31,34]	1.00	0.98	0.56	1.00	12	1.00

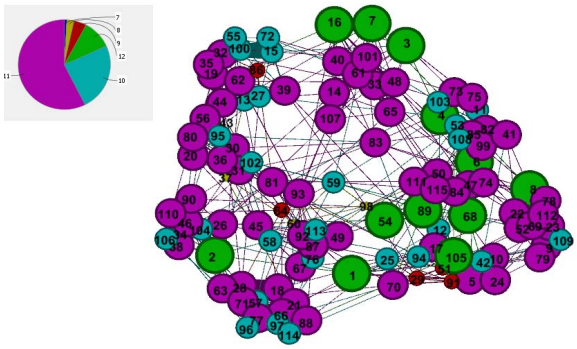
Note: The minimum and maximum number of nodes that could results in the given number of pairs are denoted by r and t.

structure of the score variables is less clear. In Fig. 6, we plotted the maximum value of the Silhouette criterion for each p (length of the codebook) while tuning γ and σ. Therefore, the predicted number of clusters is equal to p for which the Silhouette value is maximum. The obtained results are shown in Figs. 5 and 7 which reveal that incorporating the prior knowledge (labels provided by human), can potentially increase the performance in the segmentation task with respect to a genuinely unsupervised approach.

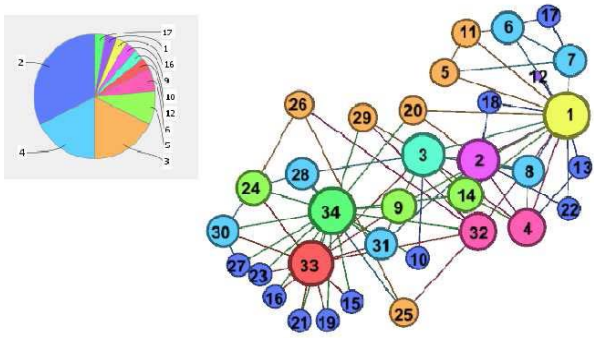
D. Community Detection

Community detection is an important topic related to complex networks [30]. It consist of finding clusters of strongly connected nodes such that nodes in the same community share more connections than nodes in different communities. Once properly identified, the community structure can help to shed light on the whole functioning of the network. Community detection is an unsupervised technique. However, if some form of prior knowledge of the community structure is present, semisupervised techniques could in principle be used to improve the results [31], [32].

In this section, the performance of the proposed method is analyzed in the community detection problems when there exist some form of prior knowledge about the existing communities. We conduct the experiments on two well-known real-world networks: Karate and Football data



(a)



(b)

Fig. 8. Visualization of the networks when nodes are colored according to their degree value. (a) American college football undirected graph. (b) Zachary's karate club undirected graph.

sets shown in Fig. 7, which are described briefly as follows.

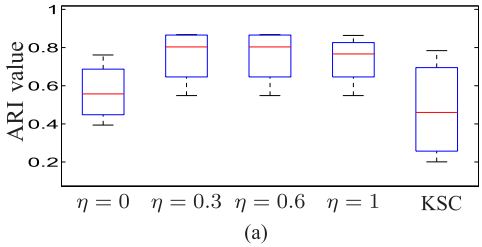
1) Karate: The Zachary's karate club network [33] consists of 34 member nodes, and splits into two smaller clubs after a dispute emerged during the course of Zachary's study between the administrator and the instructor.

2) Football: This network [34] describes American college football games and is formed by 115 nodes (the teams) and 616 edges (the games). It can be divided into 12 communities according to athletic conferences.

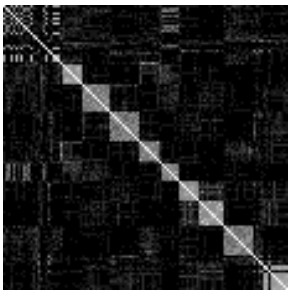
Concerning the Karate network, a comparison with the methods described in [35] is performed. In [35], a percentage of node pairs, which then are determined whether they belong to must-link or cannot-link groups, is used in the learning process. The reported results in [36, Table I] for different percentages of node pairs are tabulated in Table III.

Since in the proposed approach we work with the labeled nodes, not pairs, we randomly select some nodes and labeled them according to the true community to which they belong. The averaged normalized mutual information (NMI) over ten simulation runs for Karate network is reported in Table III. One can observe that the proposed method is able to achieve the maximum performance using less labeled nodes than the other algorithms. In particular with 10 labeled nodes the maximum value of NMI is achieved.

Concerning the Football network, we conducted the semi-supervised classification task. The training set consists of both



(a)



(b)

Fig. 9. American college football network. (a) Obtained ARI value when KSC and MSS-KSC algorithm are used. (b) Kernel matrix showing the partitioning related to $\eta = 0.6$. A clear block structure revealing the presence of the 12 communities can be noticed.

labeled and unlabeled nodes. 40% of each class (community) is randomly selected and form the labeled training nodes and another 40% randomly selected nodes form the unlabeled nodes. The whole network is considered as the test set and the obtained result is compared with KSC approach. The partitions found by KSC and MSS-KSC are evaluated according to the adjusted Rand index (ARI) [35]. The obtained ARI values on the test set after ten runs are shown in Fig. 9. We can observe that, the prior knowledge incorporation helps to improve the performance with respect to KSC.

VII. Conclusion

In this paper, a multiclass semisupervised formulation based on kernel spectral clustering has been proposed. The method is able to handle both semisupervised classification and clustering. In the semisupervised clustering case, an optimal embedding dimension is designed and used. The validity and applicability of the proposed method are shown on synthetic examples as well as on real benchmark data sets in different areas including semisupervised classification, image segmentation, and community detection problems.

References

[1] O. Chapelle, B. Schölkopf, and A. Zien, *Semi-Supervised Learning*, vol. 2. Cambridge, MA, USA: MIT Press, 2006.

[2] X. Zhu, "Semi-supervised learning literature survey," Ph.D. dissertation, Dept. Comput. Sci., Univ. Wisconsin-Madison, Madison, WI, USA, 2006.

[3] M. M. Adankon, M. Cheriet, and A. Biem, "Semisupervised least squares support vector machine," *IEEE Trans. Neural Netw.*, vol. 20, no. 12, pp. 1858–1870, Dec. 2009.

[4] Y. Huang, D. Xu, and F. Nie, "Semi-supervised dimension reduction using trace ratio criterion," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 3, pp. 519–526, Mar. 2012.

[5] F. Nie, Z. Zeng, I. W. Tsang, D. Xu, and C. Zhang, "Spectral embedded clustering: A framework for in-sample and out-of-sample spectral clustering," *IEEE Trans. Neural Netw.*, vol. 22, no. 11, pp. 1796–1808, Nov. 2011.
[6] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples," *J. Mach. Learn. Res.*, vol. 7, pp. 2399–2434, Nov. 2006.
[7] S. Xiang, F. Nie, and C. Zhang, "Semi-supervised classification via local spline regression," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 11, pp. 2039–2053, Nov. 2010.
[8] Y. Yang, F. Nie, D. Xu, J. Luo, Y. Zhuang, and Y. Pan, "A multimedia retrieval framework based on semi-supervised ranking and relevance feedback," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 4, pp. 723–742, Apr. 2012.
[9] M. Karasuyama and H. Mamitsuka, "Multiple graph label propagation by sparse integration," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 12, pp. 1999–2012, Dec. 2013.
[10] Y. Wang, S. Chen, and Z.-H. Zhou, "New semi-supervised classification method based on modified cluster assumption," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 5, pp. 689–702, May 2012.
[11] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Proc. Adv. NIPS*, vol. 2. 2002, pp. 849–856.
[12] U. von Luxburg, "A tutorial on spectral clustering," *Statist. Comput.*, vol. 17, no. 4, pp. 395–416, 2007.
[13] F. R. Chung, *Spectral Graph Theory*, vol. 92. Providence, RI, USA: AMS, 1997.
[14] C. Alzate and J. A. K. Suykens, "Multiway spectral clustering with out-of-sample extensions through weighted kernel PCA," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 2, pp. 335–347, Feb. 2010.
[15] J. A. K. Suykens, "Data visualization and dimensionality reduction using kernel maps with a reference point," *IEEE Trans. Neural Netw.*, vol. 19, no. 9, pp. 1501–1517, Sep. 2008.
[16] C. Alzate and J. A. K. Suykens, "A semi-supervised formulation to binary kernel spectral clustering," in *Proc. IEEE IJCNN*, Jun. 2012, pp. 1992–1999.
[17] S. Mehrkanoon and J. A. K. Suykens, "Non-parallel semi-supervised classification based on kernel spectral clustering," in *Proc. IEEE IJCNN*, 2012, pp. 2311–2318.
[18] C. M. Bishop and N. M. Nasrabadi, *Pattern Recognition and Machine Learning*, vol. 1. New York, NY, USA: Springer-Verlag, 2006.
[19] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, no. 1, pp. 53–65, 1987.
[20] V. Vapnik, *Statistical Learning Theory*. New York, NY, USA: Wiley, 1998.
[21] S. Boyd and L. Vandenberghe, *Convex Optimization*. New York, NY, USA: Cambridge Univ. Press, 2004.
[22] S. Xavier-De-Souza, J. A. K. Suykens, J. Vandewalle, and D. Bollé, "Coupled simulated annealing," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 40, no. 2, pp. 320–335, Apr. 2010.
[23] J. C. Bezdek and N. R. Pal, "Some new indexes of cluster validity," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 28, no. 3, pp. 301–315, Jun. 1998.
[24] O. Chapelle, V. Sindhwani, and S. S. Keerthi, "Branch and bound for semi-supervised support vector machines," in *Proc. NIPS*, 2006, pp. 217–224.
[25] Y.-F. Li, J. T. Kwok, and Z.-H. Zhou, "Semi-supervised learning using label mean," in *Proc. 26th Annu. Int. Conf. Mach. Learn.*, 2009, pp. 633–640.
[26] A. Asuncion and D. J. Newman, "UCI machine learning repository," Dept. Inf. Comput. Sci., Univ. California, Irvine, CA, USA, 2007.
[27] J. A. Nelder and R. Mead, "A simplex method for function minimization," *Comput. J.*, vol. 7, no. 4, pp. 308–313, 1965.
[28] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *Proc. 8th Int. Conf. Comput. Vis.*, vol. 2. 2001, pp. 416–423.
[29] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik, "Contour detection and hierarchical image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 5, pp. 898–916, May 2011.
[30] S. Fortunato, "Community detection in graphs," *Phys. Rep.*, vol. 486, no. 3, pp. 75–174, 2010.
[31] X. Ma, L. Gao, X. Yong, and L. Fu, "Semi-supervised clustering algorithm for community structure detection in complex networks," *Phys. A, Statist. Mech. Appl.*, vol. 389, no. 1, pp. 187–197, 2010.
[32] S. A. Macskassy and F. Provost, "Classification in networked data: A toolkit and a univariate case study," *J. Mach. Learn. Res.*, vol. 8, pp. 935–983, May 2007.
[33] W. W. Zachary, "An information flow modelfor conflict and fission in small groups," *J. Anthropological Res.*, vol. 33, no. 4, pp. 452–473, 1977.
[34] M. Girvan and M. E. Newman, "Community structure in social and biological networks," *Proc. Nat. Acad. Sci.*, vol. 99, no. 12, pp. 7821–7826, 2002.
[35] N. X. Vinh, J. Epps, and J. Bailey, "Information theoretic measures for clusterings comparison: Is a correction for chance necessary?" in *Proc. 26th Annu. Int. Conf. Mach. Learn.*, 2009, pp. 1073–1080.
[36] Z.-Y. Zhang, "Community structure detection in complex networks with partial background information," *Europhys. Lett.*, vol. 101, no. 4, p. 48005, 2013.



Siamak Mehrkanoon received the B.Sc. degree in pure mathematics in 2005 and the M.Sc. degree in applied mathematics from the Iran University of Science and Technology, Tehran, Iran, in 2007. He is currently pursuing the Ph.D. degree with the STADIUS Research Division, Department of Electrical Engineering, Katholieke Universiteit Leuven, Leuven, Belgium.

He was developing numerical methods for the simulation of large-scale dynamical systems using message passing interface from 2007 to 2010. He is involved in the development of kernel-based approaches for semisupervised clustering, simulation of dynamical systems, and parameter estimation. His current research interests include machine learning, unsupervised and semi-supervised classification/clustering, pattern recognition, system identification, numerical algorithms, and optimization.



Carlos Alzate received the B.Sc degree in electronic engineering from Universidad Nacional de Colombia, Bogotá, Colombia, in 2002, and the master's degree in artificial intelligence and the Ph.D. degree in engineering from Katholieke Universiteit Leuven (KU Leuven), Leuven, Belgium, in 2004 and 2009, respectively.

He was a Research Foundation Flanders Post-Doctoral Fellow of the Department of Electrical Engineering at KU Leuven from 2010 to 2012. He is currently a Research Scientist with the Smarter Cities Technology Centre, IBM Research–Ireland, Dublin, Ireland. His current research interests include machine learning, kernel methods, unsupervised and semisupervised learning, optimization, and natural language processing.



Raghvendra Mall was born in Kolkata, India, in 1988. He received the B.Tech. degree and the M.S. degree in computer science from IIIT Hyderabad, Hyderabad, India, in 2010 and 2011, respectively. He is currently pursuing the Ph.D. degree with the STADIUS Research Division, Department of Electrical Engineering, Katholieke Universiteit Leuven (KU Leuven), Leuven, Belgium.

He is currently developing techniques to explore sparsity in large-scale machine learning problems. His current research interests include sampling and community detection in complex networks, exploring the role of sparsity in classification, clustering, and regression.



Rocco Langone was born in Potenza, Italy, in 1983. He received the bachelor's degree in physics and information technology, the master's degree in physics with the thesis titled "A Neural Network Model for Studying the Attribution of Global Circulation Atmospheric Patterns on the Climate at a Local Scale," and the second master's degree in scientific computing with the thesis titled "Stochastic Volatility Models for European Calls Option Pricing" from the Sapienza University of Rome, Rome, Italy, in 2002, 2008, and 2010, respectively. He is currently pursuing the Ph.D. degree in machine learning with the STADIUS Research Division, Department of Electrical Engineering, KU Leuven.

He was a Researcher with the National Research Council, Rome, until 2008. His current research interests include kernel methods, neural networks, optimization, unsupervised learning (clustering and community detection), big data, climate, finance, and fault detection.



Johan A. K. Suykens (SM'05) was born in Willebroek, Belgium, in 1966. He received the M.S. degree in electromechanical engineering and the Ph.D. degree in applied sciences from Katholieke Universiteit Leuven (KU Leuven), Leuven, Belgium, in 1989 and 1995, respectively.

He was a Visiting Post-Doctoral Researcher with the University of California at Berkeley, Berkeley, CA, USA, in 1996. He has been a Post-Doctoral Researcher with the Fund for Scientific Research FWO Flanders, and is currently a Professor (Hoogleraar) at KU Leuven. He has authored the books *Artificial Neural Networks for Modelling and Control of Non-linear Systems* (Kluwer Academic Publishers) and *Least Squares Support Vector Machines* (World Scientific), co-authored the book *Cellular Neural Networks, MultiScroll Chaos and Synchronization* (World Scientific), and is an Editor of the books *Nonlinear Modeling: Advanced Black-Box Techniques* (Kluwer Academic Publishers) and *Advances in Learning Theory: Methods, Models and Applications* (IOS Press).

Prof. Suykens organized an International Workshop on Nonlinear Modelling with Time-Series Prediction Competition in 1998. He has served as an Associate Editor of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS from 1997 to 1999 and from 2004 to 2007, and the IEEE TRANSACTIONS ON NEURAL NETWORKS from 1998 to 2009. He was a recipient of the IEEE Signal Processing Society Best Paper (Senior) Award in 1999, several Best Paper Awards at international conferences, and the International Neural Networks Society INNS Young Investigator Award for significant contributions in the field of neural networks in 2000. He served as the Director and an Organizer of the NATO Advanced Study Institute on Learning Theory and Practice (Leuven 2002), a Program Co-Chair of the International Joint Conference on Neural Networks in 2004 and the International Symposium on Nonlinear Theory and its Applications in 2005, an Organizer of the International Symposium on Synchronization in Complex Networks in 2007, and a Co-Organizer of the NIPS Workshop on Tensors, Kernels and Machine Learning in 2010. He was awarded an ERC Advanced Grant in 2011.