New Bilinear Formulation to Semi-Supervised Classification Based on Kernel Spectral Clustering

Vilen Jumutc
KU Leuven, ESAT-STADIUS
Kasteelpark Arenberg 10
B-3001 Leuven, Belgium
Email: Vilen.Jumutc@esat.kuleuven.be

Johan A.K. Suykens
KU Leuven, ESAT-STADIUS
Kasteelpark Arenberg 10
B-3001 Leuven, Belgium
Email: Johan.Suykens@esat.kuleuven.be

***Abstract*—In this paper we present a novel semi-supervised classification approach which combines bilinear formulation for non-parallel binary classifiers based upon Kernel Spectral Clustering. The cornerstone of our approach is a bilinear term introduced into the primal formulation of semi-supervised classification problem. In addition we perform separate manifold regularization for each individual classifier. The latter relates to the Kernel Spectral Clustering unsupervised counterpart which helps to obtain more precise and generalizable classification boundaries. We derive the dual problem which can be effectively translated into a linear system of equations and then solved without introducing extra costs. In our experiments we show the usefulness and report considerable improvements in performance with respect to other semi-supervised approaches, like Laplacian SVMs and other KSC-based models.**

I. Introduction

For many decades classification and Semi-Supervised Learning (SSL) were among the most popular machine learning topics [1], [2]. The substantial interest is driven by the simple and very important observation: the amount of unlabeled data is instantly growing while resources for labeling and preprocessing are limited and not always easily accessible. Many existing semi-supervised approaches [3], [4] aim at utilizing either labeled or unlabeled data. While both strategies have numerous drawbacks it was recently stressed [5] that preserving them jointly might be important for obtaining good generalization in the presence of major unlabeled counterpart.

Laplacian Support Vector Machine (LapSVM) [5], [6] is one of the state-of-the-art techniques which addresses this problem and provides a natural out-of-sample extension for unseen unlabeled data. It is based on implicit manifold regularization introduced by the normalized Laplacian matrix [7]. Another possible approach is related to the Spectral Clustering techniques [8]–[10] which make use of the eigenspectrum of the Laplacian matrix to divide a dataset into the clusters such that points within the same cluster are similar and points from other clusters are dissimilar to each other. However though these techniques are of great interest for machine learning practitioners they often lack proper model selection and verification procedures in semi-supervised setting and cannot be applied blindly. Recently a kernel-based extension for semi-supervised learning was proposed known as Semi-Supervised Kernel Spectral Clustering [11], [12]. The main idea of the latter approach is to formulate the semi-supervised learning problem as a combination of a weighted kernelized Principal Component Analysis (PCA) and minimization of an empirical error associated with the labeled counterpart.

For the roots of the bilinear term minimization which we utilize in our approach we refer the interested reader to the recently proposed Supervised Novelty Detection (SND) approach [13], [14]. SND can be effectively used for multi-class classification and it supplements the class of SVM-based algorithms intended for the detection of outliers in the presence of several classes (distributions). We should mention that the SND approach is related to other non-parallel SVM classifiers, like Twin SVM (TWSVM), GEPSVM [15]–[17] or Non-Parallel Semi-Supervised KSC [18] etc., but it has one crucial difference in its primal formulation, i.e. a bilinear term which couples two non-parallel classifiers. In this paper we extend ideas given by [14] to the semi-supervised setting and show that it can be quite beneficial to couple non-parallel classifiers in the presence of unlabeled information. We compare our results with the aforementioned approaches and present some convincing and challenging toy-problems where we are violating a commonly referenced low-density assumption (between clusters) and performing better than other approaches.

The remainder of this paper is structured as follows. Section II outlines some existing semi-supervised learning approaches. Section III gives an overview of our method and the resulting optimization problem. Section IV provides the experimental setup and results, while Section V discusses some important issues and further directions of our research. Section VI concludes the paper.

II. Semi-supervised learning

In this section we present a brief outlook on the existing state-of-the-art methods in semi-supervised learning. But first we present some preliminary notations which will be used across this paper.

A. Preliminaries

We first introduce terminology and some essential conventions. We consider training data with the corresponding partial

labeling given as a set of pairs

$$(x_1, y_1), \ldots, (x_l, y_l), x_i \in \mathcal{X}, y_i \in \{-1, 1\},$$

where l is the number of corresponding labeled observations in the set $\mathcal{X}$ of size m. The unlabeled part is given by the set $\{x_{l+1}, \ldots, x_m\} = \mathcal{X}_u \in \mathcal{X}$. For simplicity we think of all our data as a compact subset of $\mathbb{R}^d$. Then define φ to be a feature map $\varphi : \mathbb{R}^d \to \mathbb{R}^h$ which is a mapping to a high-dimensional feature space of dimension h with the connection to a positive definite kernel [1]

$$k(x, y) = \langle \varphi(x), \varphi(y) \rangle, \tag{1}$$

where by using a kernel trick [1], [2] one can define a kernel matrix Ω as $\Omega_{ij} = k(x_i, x_j)$.

Further in the paper index i spans the range $\overline{1, m}$ if it is not declared explicitly. Greek letters α, λ without indices denote m-dimensional vectors.

B. Laplacian SVM [5]

The solution of the LapSVM problem proposed by Belkin *et al.* [5] is based on manifold learning. This problem explicitly incorporates the kernel matrix Ω and the Laplacian matrix L as follows:

$$\min_{\alpha \in \mathbb{R}^m; \xi \in \mathbb{R}^l} \gamma_A \alpha^T \Omega \alpha + \gamma_I \alpha^T \Omega L \Omega \alpha + \sum_{i=1}^{l} \xi_l \tag{2}$$

$$\begin{aligned} \text{s.t.} \quad & y_i(\textstyle\sum_{j=1}^{m} \alpha_j k(x_j, x_i) + b) \geq 1 - \xi_i, \quad i = 1, \ldots, l \\ & \xi_i \geq 0, \quad i = 1, \ldots, l \end{aligned} \tag{3}$$

This formulation directly implies regularization over two different terms and minimization of the empirical error ξ_i. The first term is establishing a regularization framework for function learning, given a kernel function $k(\cdot, \cdot)$, its associated Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}_k$ of functions $X \to \mathbb{R}$ with the corresponding norm $\| \cdot \|_A$ and a trade-off hyperparameter γ_A. The second term stands for the implicit manifold regularization with the corresponding norm $\|f\|_I = f^T L f$ defined as the Laplace-Beltrami operator [19] applied to the function f.

C. Semi-supervised KSC [11]

The following formulation to the semi-supervised learning problem is very close to the presented one in the next section. We provide it for the interested reader as a reference point in understanding our own extension to it. The proposed primal problem of Semi-Supervised Kernel Spectral Clustering (SemiKSC) in [11] is given by:

$$\begin{aligned} \min_{w \in \mathbb{R}^d; e \in \mathbb{R}^m} \quad & \tfrac{1}{2}\|w\|^2 + \tfrac{\rho}{2} \textstyle\sum_{i=1}^{l} (e_i - y_i)^2 - \tfrac{\gamma}{2} e^T V e \\ \text{s.t.} \quad & \Phi w + b 1_M = e, \end{aligned} \tag{4}$$

where $\Phi = [\varphi(x_1)^T; \ldots; \varphi(x_m)^T] \in \mathbb{R}^{m \times h}$ is a stacked feature map for all samples in $\mathcal{X}$, $e \in \mathbb{R}^m$ is the projection, b is the bias, V is the weighting diagonal matrix while γ and ρ are trade-off hyperparameters.

After eliminating primal variables and converting the above problem into a linear system of equations we obtain the following problem in terms of the dual variables $\alpha \in \mathbb{R}^m$:

$$(I_M - (\gamma D^{-1} - \rho A) M_S \Omega) \alpha = \rho M_S^T y, \tag{5}$$

where I_M is $m \times m$ identity matrix and $M_S \in \mathbb{R}^{m \times m}$ is a centering matrix defined as:

$$M_S = I_M - \frac{1}{c} 1_M 1_M^T (\gamma D^{-1} - \rho A)$$

with

$$A = \left[\begin{array}{c|c} 0_{p \times p} & 0_{p \times l} \\ \hline 0_{l \times p} & I_{l \times l} \end{array} \right] \in \mathbb{R}^{m \times m}$$

and $p = m - l$, $c = 1_M^T (\gamma D^{-1} - \rho A) 1_M$. If we omit bias term b in the latent clustering model for training points, we can now write it in terms of the dual variables α:

$$e = M_S \Omega \alpha - \frac{\rho}{c} 1_M 1_M^T y.$$

This model defines the binary cluster membership by $\text{sign}(e)$.

III. Bilinear Non-Parallel Kernel Spectral Semi-Supervised Learning

In this section we present an optimization problem in the primal form. Further we derive a dual formulation with respect to our constraints. Finally using Karush-Kuhn-Tucker (KKT) optimality conditions we boil-down the problem into a linear system of equations which can be efficiently solved without introducing extra costs.

A. Optimization problem

We start with a primal formulation for a simple binary semi-supervised classification problem where we introduce an additional bilinear term $\langle w_1, w_2 \rangle$ between classifiers:

$$\begin{aligned} \min_{w_1, w_2 \in \mathbb{R}^d; e_1, e_2 \in \mathbb{R}^m} \quad & \tfrac{\gamma_1}{2}(\|w_1\|^2 + \|w_2\|^2) + \langle w_1, w_2 \rangle \\ & + \tfrac{\gamma_2}{2} \textstyle\sum_{i=1}^{l} [(e_{1i} - y_i)^2 + (e_{2i} + y_i)^2] \\ & - \tfrac{\gamma_3}{2} e_1^T V e_1 - \tfrac{\gamma_4}{2} e_2^T V e_2 \end{aligned} \tag{6}$$

$$\begin{aligned} \text{s.t.} \quad & \Phi w_1 = e_1 \\ & \Phi w_2 = e_2 \end{aligned} \tag{7}$$

where $w_1, w_2 \in \mathbb{R}^h$ are the hyperplanes associated with each non-parallel classifier, $\Phi = [\varphi(x_1)^T; \ldots; \varphi(x_m)^T] \in \mathbb{R}^{m \times h}$ is a stacked feature map for all samples in $\mathcal{X}$, $e_1, e_2 \in \mathbb{R}^m$ are the projections, V is the weighting diagonal matrix and γ_1 through γ_4 denote trade-off hyperparameters. For the simplicity and because of the flexibility of RBF-kernel which is often used in the KSC-based models we omit in our formulation a bias term b.

In the primal optimization objective we have different terms which are related to either supervised or unsupervised counterpart. Terms $(e_{1i} - y_i)^2$ and $(e_{2i} + y_i)^2$ are trying to minimize empirical error *w.r.t.* each non-parallel classifier and provided set of labels y. Terms $e_1^T V e_1$ and $e_2^T V e_2$ are related

to the KSC model [10] and are trying to explain (maximize) variance *w.r.t.* involved classes.

Before proceeding to the formulation of the dual problem we explain the importance of the weighting matrix V. It can be shown that if we take V as an inverse of a degree matrix[1]

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

where $d_i = \sum_j k(x_i, x_j)$, the overall problem can be related to the random walk algorithms in spectral clustering [8], [9].

The final decision function is

$$c(x) = \begin{cases} \text{argmax}_i \ \langle w_i, \varphi(x) \rangle, & \text{if} \ \ \max_i \langle w_i, \varphi(x) \rangle > 0, \\ c_{out}, & \text{otherwise}, \end{cases} \tag{8}$$

where $i \in \{1, 2\}$ and c_{out} is representing the outliers' class. This decision rule can be advantageous when one tries to estimate a high-dimensional support of the underlying distributions (per class). This specific setting is discussed in detail in [14].

B. Dual formulation

Using α and λ vectors as Lagrange multipliers we introduce the following dual optimization problem

$$\begin{aligned} \mathcal{L}(w_1, w_2, e_1, e_2, \alpha, \lambda) &= \tfrac{\gamma}{2}(\|w_1\|^2 + \|w_2\|^2) + \langle w_1, w_2 \rangle \\ &+ \tfrac{\gamma_2}{2} \textstyle\sum_{i=1}^{l} [(e_{1i} - y_i)^2 + (e_{2i} + y_i)^2] \\ &- \tfrac{\gamma_3}{2} e_1^T V e_1 - \tfrac{\gamma_4}{2} e_2^T V e_2 \\ &+ \alpha^T (e_1 - \Phi w_1) + \lambda^T (e_2 - \Phi w_2). \end{aligned} \tag{9}$$

By setting the derivatives of the dual problem with respect to primal and dual variables to zero we obtain the following Karush-Kuhn-Tucker (KKT) optimality conditions:

$$\begin{cases} \frac{\partial \mathcal{L}}{\partial \alpha} = 0 \rightarrow e_1 - \Phi w_1 = 0, \\ \frac{\partial \mathcal{L}}{\partial \lambda} = 0 \rightarrow e_2 - \Phi w_2 = 0, \\ \frac{\partial \mathcal{L}}{\partial w_1} = 0 \rightarrow w_1 = \frac{1}{1-\gamma_1^2} \Phi^T (\lambda - \gamma_1 \alpha), \\ \frac{\partial \mathcal{L}}{\partial w_2} = 0 \rightarrow w_2 = \frac{1}{1-\gamma_1^2} \Phi^T (\alpha - \gamma_1 \lambda), \\ \frac{\partial \mathcal{L}}{\partial e_1} = 0 \rightarrow \alpha - \gamma_3 V e_1 + \gamma_2 G (e_1 - y^*) = 0, \\ \frac{\partial \mathcal{L}}{\partial e_2} = 0 \rightarrow \lambda - \gamma_4 V e_2 + \gamma_2 G (e_2 + y^*) = 0, \end{cases} \tag{10}$$

where defining $p = m - l$ as an auxiliary indexing upper bound G is given by

$$G = \left[\begin{array}{c|c} I_{l \times l} & 0_{l \times p} \\ \hline 0_{p \times l} & 0_{p \times p} \end{array} \right] \in \mathbb{R}^{m \times m}$$

and $y^* = [y_1, \ldots, y_l, 0, \ldots, 0]^T \in \mathbb{R}^m$, while $I_{l \times l}$ is an identity matrix.

[1] we assume degree of i-th data point in a neighborhood graph

C. Linear system

One can observe that our primal problem is non-convex but after eliminating the primal variables w_1, w_2 and e_1, e_2 we can directly work with the dual objective which is always concave in terms of the dual variables. Despite this fact it is more convenient to work with a linear system which we can obtain by plugging substituted primal variables back into the last two optimality conditions of Eq.(10) and deriving the following linear system of equations

$$\left[\begin{array}{c|c} A_d & A_{\text{off}} \\ \hline L_{\text{off}} & L_d \end{array} \right] \left[\frac{\alpha}{\lambda} \right] = \gamma_2 \left[\frac{y^*}{-y^*} \right], \tag{11}$$

where $A_d = \gamma_3 \gamma_1 \gamma_1' \Omega D^{-1} - \gamma_2 \gamma_1 \gamma_1' G + I$, $A_{\text{off}} = \gamma_2 \gamma_1' G - \gamma_3 \gamma_1' \Omega D^{-1}$, $L_d = \gamma_4 \gamma_1 \gamma_1' \Omega D^{-1} - \gamma_2 \gamma_1 \gamma_1' G + I$, $L_{\text{off}} = \gamma_2 \gamma_1' \Omega - \gamma_4 \gamma_1' G D^{-1}$, $\gamma_1' = 1/(1 - \gamma_1^2)$, Ω is a full-rank kernel matrix and G, D^{-1} were given in the previous subsection, while I is an identity matrix of size $m \times m$.

By solving this linear system of equations we obtain key components of our final decision function defined in Eq.(8). If we take a look again at KKT conditions in Eq.(10) we can notice that we have a closed form solution for w_1 and w_2 in terms of our dual variables α and λ. By plugging these variables and taking into account definition of $\langle \varphi(x), \varphi(y) \rangle$ in Eq.(1) we obtain a kernel expansion for the decision function as

$$c(x) = \begin{cases} \text{argmax}_i \ f_i(x), & \text{if} \ \ \max_i f_i(x) > 0 \\ c_{out}, & \text{otherwise}, \end{cases} \tag{12}$$

where $f_i(x)$ functions are defined as follows

$$\begin{aligned} f_1(x) &= \tfrac{1}{1-\gamma_1^2} \textstyle\sum_{i=1}^{m} k(x, x_i)(\lambda_i - \gamma_1 \alpha_i), \\ f_2(x) &= \tfrac{1}{1-\gamma_1^2} \textstyle\sum_{i=1}^{m} k(x, x_i)(\alpha_i - \gamma_1 \lambda_i). \end{aligned} \tag{13}$$

IV. Experiments

A. Experimental setup

In all our experiments we tested all semi-supervised kernel-based models using a 2-step procedure for tuning the hyper-parameters. This procedure consists of the Coupled Simulated Annealing [20] initialized with 5 and up to 30 random sets of parameters for the first step and the simplex method [21] for the second step. After CSA converges to some local minima we select the tuple of parameters that attains the lowest error and start the simplex procedure to refine our selection. On every iteration step for CSA and simplex method we proceed with a 5-fold cross-validation. Cross-validation is performed only with respect to the labeled information and classification accuracy. We do not impose any additional model selection criteria at this step, because most of them, like Silhouette index, Fisher index or Davies-Bouldin index [22], according to our empirical observations are resulting in worse generalization capabilities and might not reflect properly the non-linearity of clustering in some higher dimensional spaces

and manifolds. In our experiments we use RBF-kernel related to Eq.(1) and given by

$$k(x,y) = \langle \varphi(x), \varphi(y) \rangle = e^{-\frac{\|x-y\|^2}{2\sigma^2}}.$$

We use the aforementioned cross-validation tuning procedure to estimate kernel bandwidth σ together with other trade-off hyperparameters.

In our toy examples we experiment with two challenging problems when data are drawn from the mixture of overlapping distributions. In this case we do violate the low-density assumption between clusters and semi-supervised algorithms based upon spectral clustering techniques and normalized cuts [7] might fail in this case. We artificially created two datasets, namely "spirals" and "half-moons". In the first one we have two spirals which are separable in the beginning and are merging in the end. For the second dataset we have two banana-shaped distributions which are overlapping at their tails. Additionally to the overlapping scenarios we present the results for the noiseless case where classes are clearly separable. In every dataset only 20% of points were labeled randomly.

For the numerical benchmark tests we took some recognized datasets in semi-supervised learning mentioned or created for [23]. Each of these datasets[2] is provided in 12 splits where each split is represented by the labeled points and the remaining unlabeled counterpart. We perform cross-validation using entire data (only labeled information is considered) for the split and then report the average misclassification rate for all datapoints as we are provided with the ground truth for them. We average the results across all 12 splits for each dataset. This experimental setup is used in the original book by Chapelle *et al.* [23] and is considered in other research papers as well [11], [18]. For our experiments we use the splits with 100 labeled datapoints only. We perform additional experiments on public UCI datasets [24] where we randomly select 100 points to remain labeled. We run all algorithms for UCI datasets 50 times and report average misclassification rate evaluating the methods on the entire dataset, where we know the ground truth for the unlabeled points. All datasets were normalized by mapping each dimension to zero mean and one standard deviation as $(\mu_i, \sigma_i) = (0,1), i \in \overline{1,d}$. For the properties of UCI and SSL benchmark datasets used in this paper one can refer to the Table I.

We implemented Eq.(11) in MATLAB using the backslash operator. Because we are not using a bias term b in our method we decided to omit this term in the implementation of SemiKSC [11] approach as well, which would give us a fair comparison with the former. All experiments were run on Core i7 CPU with 8GB of available RAM.

B. Toy problems

In this subsection we present the results obtained on two artificial toy-problems. In Figure 1 we can see the classification boundaries of three methods applied to the "half-moons" problem. Figure 2 corresponds to the "spirals" dataset. In the noiseless case[3] all algorithms are performing reasonably good but SemiKSC approach is slightly inferior *w.r.t.* the other methods. Comparing Figures 1 - 2 (a-c) for the overlapping case we can clearly observe that Bilinear Semi-Supervised KSC (Bi-SemiKSC) is capturing the underlying manifolds better resulting in more precise and smoother classification boundaries. Although only 20% of datapoints are labeled in these examples Bi-SemiKSC method was able to converge to an acceptable solution while LapSVM is failing to generalize on Figures 1 - 2 (a-c). While for the "half-moons" problem original Semi-Supervised KSC (SemiKSC) is giving an acceptable result, for "spirals" problem it is obviously overfitting.

TABLE I
DATASETS

Dataset	# of attributes	# of classes	# of data points
Ionosphere	34	2	351
Parkinsons	23	2	197
Sonar	60	2	208
Ecoli	8	2	336
Arrhythmia	279	2	452
BCI	117	2	400
Text	11960	2	1500
g241c	241	2	1500
g241d	241	2	1500

C. Numerical experiments

In this subsection we are analyzing the performance of four algorithms on the benchmark UCI and SSL datasets. We are comparing our Bilinear Semi-Supervised KSC (Bi-SemiKSC) approach with LapSVM in primal (LapSVMp) [6], LapSVM in dual and the original Semi-Supervised KSC (SemiKSC) [11], [12]. As we can see from Tables II and III our approach is performing much better and the classification accuracy evaluated by the ground truth on the entire data suggests that Bi-SemiKSC with the help of an additional coupling term is able to find an appropriate embedding of data where classification is the most accurate.

TABLE II
AVERAGED MISCLASSIFICATION RATE FOR THE BENCHMARK SSL DATASETS

Dataset	Bi-SemiKSC	LapSVM [5]	LapSVMp [6]	SemiKSC [11]
g241c	**0.148**±0.044	0.232±0.052	0.248±0.063	0.194±0.102
g241d	**0.158**±0.034	0.278±0.062	0.276±0.053	0.191±0.069
BCI	**0.244**±0.056	0.376±0.049	0.324±0.057	0.315±0.078
Text	**0.382**±0.082	0.458±0.026	0.455±0.025	0.477±0.062

Additionally to the misclassification rate for UCI data we report in Table IV p-values of a pairwise t-test on the same rate. We do not report these values for SSL data because we averaged the results only across 12 runs/splits which is quite a small number for a reliable t-test. By evaluating

[2]i.e. BCI, Text, g241c and g241d

[3]Figures 1 - 2 (d-f)

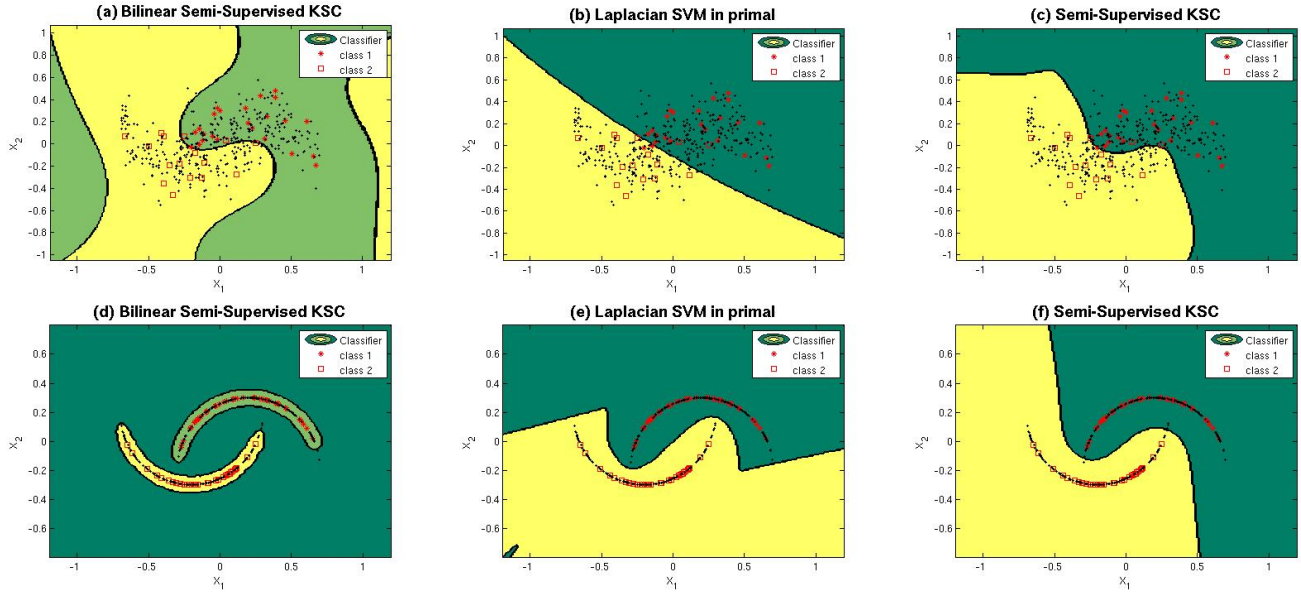


Fig. 1. Different approaches applied to the "half-moons" problem. From left to right: (a) Bilinear Semi-Supervised KSC, (b) Laplacian SVM in primal, (c) Semi-Supervised KSC in an overlapping scenario. Bottom from left to right: (d) Bilinear Semi-Supervised KSC, (e) Laplacian SVM in primal, (f) Semi-Supervised KSC in a noiseless scenario. With small black dots we denote unlabeled datapoints. Bigger red stars and squares represent labeled samples from two classes.

Table IV we can conclude that attained p-values are very small and the means of distributions for misclassification rates differ statistically significantly between our approach and the corresponding competing approaches.

TABLE III
AVERAGED MISCLASSIFICATION RATE FOR UCI DATASETS

Dataset	Bi-SemiKSC	LapSVM	LapSVMp	SemiKSC
Parkinsons	**0.039**±0.015	0.117±0.109	0.074±0.059	0.067±0.025
Sonar	**0.088**±0.018	0.131±0.085	0.122±0.086	0.102±0.053
Ionosphere	**0.099**±0.021	0.201±0.093	0.138±0.078	0.119±0.065
Ecoli	**0.036**±0.009	0.059±0.029	0.041±0.016	0.039±0.012
Arrhythmia	**0.257**±0.042	0.349±0.061	0.291±0.029	0.317±0.062

TABLE IV
P-VALUES OF A PAIRWISE T-TEST ON MISCLASSIFICATION RATE BETWEEN BILINEAR SEMIKSC AND OTHER METHODS

Dataset	to LapSVM	to LapSVMp	to SemiKSC
Parkinsons	2.4488E-06	0.00013319	1.1295E-09
Sonar	0.00064336	0.0077843	0.077444
Ionosphere	2.7248E-11	0.001315	0.044832
Ecoli	1.2486E-06	0.079189	0.11091
Arrhythmia	6.0693E-14	1.2068E-05	1.7722E-07

Additionally to the presented results we would like to compare some of the obtained results for SSL datasets with the reported values in the book of Chapelle *et al.*, Chapter 21.3 [23] and recently published paper on non-parallel KSC-based semi-supervised learning [18]. The experimental setup everywhere is the same but implementation details of cross-validation and hyperparameter tuning differ. Comparing our results with the book [23] we can clearly see that the best reported value for g241c dataset is 0.1741 which is by 2% worse than our result and it is related to Low-Density Separation technique (LDS) [25]. For BCI dataset our approach again is the best one and attained rates are much lower than those reported in the book. If we consider another relevant and non-parallel approach discussed in [18] we can see that authors report better result of 0.26 ± 0.01 for BCI dataset but it is still around 2% worse than we do.

V. DISCUSSION

A. Differences with other approaches

In this subsection we will briefly discuss some of the important differences between our approach and other non-parallel semi-supervised classifiers. As we aforementioned in Section I there exist several kernel-based non-parallel classifiers [15], [17] and only some of them explicitly utilize unlabeled information [18]. While being together with [11] and [18] based fundamentally on the same principles of KSC-implied modelling our method can be distinguished in a way we are coupling non-parallel classifiers. The reasoning and motivation behind this approach one can find in [14]. In brief it relates to the optimal separation between classes when *i.i.d.* assumption doesn't hold. From the modelling perspective in semi-supervised learning it might help to model each class and the underlying manifold better while preserving necessary discrimination between classes.

B. Future work

For the future research direction we are considering to explore different possible formulations of the Bi-SemiKSC model and apply them to large-scale datasets. While working

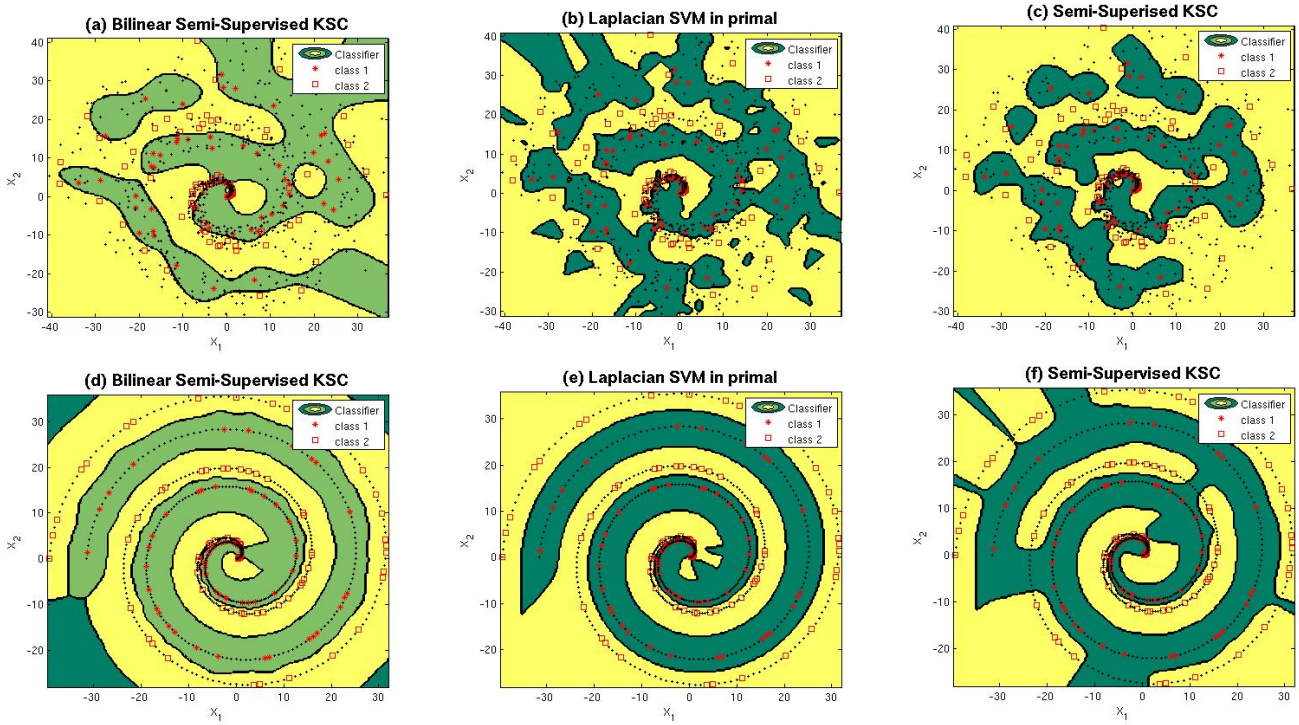


Fig. 2. Different approaches applied to the "spirals" problem. Top from left to right: (a) Bilinear Semi-Supervised KSC, (b) Laplacian SVM in primal, (c) Semi-Supervised KSC in an overlapping scenario. Bottom from left to right: (d) Bilinear Semi-Supervised KSC, (e) Laplacian SVM in primal, (f) Semi-Supervised KSC in a noiseless scenario. With small black dots we denote unlabeled datapoints. Bigger red stars and squares represent labeled samples from two classes.

on the millions of datapoints is not being directly feasibly in the dual formulation one can use Nyström approximation and Fixed-Size techniques [26]–[28] to go back for the primal formulation. This approach helps to devise a high-dimensional feature map which approximates well non-linearity while keeping first-order gradient-descent methods [29]–[31] applicable and reasonably effective for optimizing the objective function in the primal.

VI. Conclusion

In this paper we proposed a novel and promising way of handling Semi-Supervised Learning via incorporating a bilinear coupling term between non-parallel classifiers. This term enters a primal problem formulation and eventually boils-down to the coupled Lagrange multipliers comprising the solution of the dual problem. The fundamental part of our approach is a KSC-based model which helps to learn from the unlabeled datapoints and possesses a natural out-of-sample extension for the unseen ones. Our experimental validation for the artificial and real-life datasets confirms the usefulness and generalization capabilities of the proposed approach. We showed that even when a low-density assumption is broken we can learn from the unsupervised data and obtain acceptable results. For the future we consider extending our method to cope with large-scale data sources where proliferation of unlabeled information hampers the successful adoption of the supervised modelling techniques.

Acknowledgments

• EU: The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013) / ERC AdG A-DATADRIVE-B (290923). This paper reflects only the authors' views, the Union is not liable for any use that may be made of the contained information. • Research Council KUL: GOA/10/09 MaNet, CoE PFV/10/002 (OPTEC), BIL12/11T; PhD/Postdoc grants • Flemish Government: • FWO: projects: G.0377.12 (Structured systems), G.088114N (Tensor based data similarity); PhD/Postdoc grants • IWT: projects: SBO POM (100031); PhD/Postdoc grants • iMinds Medical Information Technologies SBO 2014 • Belgian Federal Science Policy Office: IUAP P7/19 (DYSCO, Dynamical systems, control and optimization, 2012-2017)

References

[1] B. Schölkopf, C. J. C. Burges, and A. J. Smola, Eds., *Advances in kernel methods: support vector learning*. Cambridge, MA, USA: MIT Press, 1999.

[2] J. A. K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, and J. Vandewalle, *Least Squares Support Vector Machines*. Singapore: World Scientific, 2002.

[3] M. Seeger, "Learning with labeled and unlabeled data," Tech. Rep., 2001.

[4] X. Zhu, "Semi-supervised learning literature survey," 2006.

[5] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples." *Journal of Machine Learning Research*, vol. 7, pp. 2399–2434, 2006.

[6] S. Melacci and M. Belkin, "Laplacian support vector machines trained in the primal," *Journal of Machine Learning Research*, vol. 12, pp. 1149–1184, 2011.

[7] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 888–905, 2000.

[8] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," *Advances in Neural Information Processing Systems*, vol. 14, pp. 849–856, 2001.

[9] F. R. K. Chung, *Spectral Graph Theory*. American Mathematical Society, 1997.

[10] 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, 2010.

[11] C. Alzate and J. A. K. Suykens, "A semi-supervised formulation to binary kernel spectral clustering," in *IJCNN*, 2012, pp. 1–8.

[12] S. Mehrkanoon, C. Alzate, R. Mall, R. Langone, and J. A. K. Suykens, "Multi-class semi-supervised learning based upon kernel spectral clustering," *IEEE Transactions on Neural Networks and Learning Systems, in press*, 2014.

[13] V. Jumutc and J. A. K. Suykens, "Supervised novelty detection," in *IEEE Symposium on Computational Intelligence and Data Mining*, 2013, pp. 143–149.

[14] V. Jumutc and J. A. K. Suykens, "Multi-class supervised novelty detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2014. [Online]. Available: http://dx.doi.org/10.1109/TPAMI.2014.2327984

[15] Jayadeva, R. Khemchandani, and S. Chandra, "Twin support vector machines for pattern classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 5, pp. 905–910, May 2007.

[16] Y.-H. Shao, C.-H. Zhang, X.-B. Wang, and N.-Y. Deng, "Improvements on twin support vector machines," *IEEE Transactions on Neural Networks*, vol. 22, no. 6, pp. 962–968, 2011.

[17] M. A. Kumar and M. Gopal, "Least squares twin support vector machines for pattern classification." *Expert Syst. Appl.*, vol. 36, no. 4, pp. 7535–7543, 2009.

[18] S. Mehrkanoon and J. A. K. Suykens, "Non-parallel semi-supervised classification based on kernel spectral clustering." in *IJCNN*. IEEE, 2013, pp. 1–8.

[19] M. Belkin and P. Niyogi, "Towards a theoretical foundation for laplacian-based manifold methods," *J. Comput. Syst. Sci.*, vol. 74, no. 8, pp. 1289–1308, 2008. [Online]. Available: http://dx.doi.org/10.1016/j.jcss.2007.08.006

[20] S. Xavier-De-Souza, J. A. K. Suykens, J. Vandewalle, and D. Bollé, "Coupled simulated annealing," *IEEE Trans. Sys. Man Cyber. Part B*, vol. 40, no. 2, pp. 320–335, Apr. 2010.

[21] J. A. Nelder and R. Mead, "A simplex method for function minimization," *Computer Journal*, vol. 7, pp. 308–313, 1965.

[22] J. C. Bezdek and N. R. Pal, "Some new indexes of cluster validity." *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, vol. 28, no. 3, pp. 301–315, 1998.

[23] O. Chapelle, B. Schölkopf, and A. Zien, *Semi-Supervised Learning*. MIT Press, 2006.

[24] A. Frank and A. Asuncion, "UCI machine learning repository," 2010. [Online]. Available: http://archive.ics.uci.edu/ml

[25] O. Chapelle and A. Zien, "Semi-supervised classification by low density separation," in *Proceedings of the International Workshop on Artificial Intelligence and Statistics*, 2005.

[26] A. Vedaldi and A. Zisserman, "Efficient additive kernels via explicit feature maps," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 3, pp. 480–492, 2012.

[27] K. De Brabanter, J. De Brabanter, J. A. K. Suykens, and B. De Moor, "Optimized fixed-size kernel models for large data sets," *Comput. Stat. and Data Anal.*, vol. 54, no. 6, pp. 1484–1504, Jun. 2010.

[28] S. Mehrkanoon and J. A. K. Suykens, "Large scale semi-supervised learning using KSC based model," in *International Joint Conference on Neural Networks (IJCNN 2014), accepted*, 2014.

[29] V. Jumutc, X. Huang, and J. A. K. Suykens, "Fixed-size Pegasos for hinge and pinball loss SVM," in *IJCNN*, 2013, pp. 1122–1128.

[30] S. Shalev-Shwartz, Y. Singer, and N. Srebro, "Pegasos: Primal Estimated sub-GrAdient SOlver for SVM," in *Proceedings of the 24th international conference on Machine learning*, ser. ICML '07, New York, NY, USA, 2007, pp. 807–814.

[31] T. Joachims, "Training linear SVMs in linear time," in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, ser. KDD '06. New York, NY, USA: ACM, 2006, pp. 217–226.