Semi-supervised learning: avoiding zero label assumptions in kernel based classifiers

Jan Luts [a,*], Johan A.K. Suykens [a], Sabine Van Huffel [a]

[a] *Katholieke Universiteit Leuven, Department of Electrical Engineering, ESAT-SCD (SISTA), Kasteelpark Arenberg 10, B-3001 Leuven-Heverlee, Belgium*

Abstract

In some recent papers in the area of semi-supervised and transductive learning an assumption is made that the true class labels for unlabeled data equal zero. In this paper we discuss an approach to overcome this assumption by adding extra variables and extra constraints in the problem formulation. We apply it to least squares support vector machine classifiers. The method is illustrated on both artificial and real-life data sets. The results are comparable with the other studies that assume zero values for true class labels. However, when only a limited amount of labeled data is available, an increased performance is observed for the proposed method on some data sets.

Key words: Classification, semi-supervised learning, support vector machines, kernel methods, least squares

* Corresponding author. Tel.: +32 16 321065; fax: +32 16 321970
Email addresses: jan.luts@esat.kuleuven.be (Jan Luts), johan.suykens@esat.kuleuven.be (Johan A.K. Suykens), sabine.vanhuffel@esat.kuleuven.be (Sabine Van Huffel).

19 July 2007

1 Introduction

This communication concerns the subject of semi-supervised learning. Within the area of semi-supervised learning, or the related topic of transductive learning, the goal is to take unlabeled data points into account when making predictions and estimating a classifier. For this purpose, certain assumptions have to hold (Zhou, Bousquet, Lal, Weston, & Schölkopf, 2004; Chapelle, Schölkopf, & Zien, 2006). First there is the smoothness assumption; cases that are close in the input space, often tend to have similar class labels. The cluster assumption means that points from the same cluster are likely to belong to the same class. Low density separation supposes that the decision boundary should be in a low density region. Finally, the manifold assumption indicates that data lie on a low-dimensional manifold. Nowadays, semi-supervised learning is an extensively studied topic. This can be explained by the fact that unlabeled data are available in larger amounts than fully labeled data as they are in general easier to collect. An overview of the field can be found in (Zhu, 2005). Methods include expectation maximization algorithms (Nigam, McCallum, Thrun, & Mitchell, 2000), transductive support vector machines (SVMs) (Vapnik, V. N., 1998), self-training (Yarowsky, 1995), co-training (Blum & Mitchell, 1998) and graph based approaches (Blum & Chawla, 2001; Chapelle, Weston, & Schölkopf, 2003).

In this paper an assumption is addressed that is made in a number of studies, solving the semi-supervised learning problem (Tsuda, Shin, & Schölkopf, 2005; Belkin, Niyogi, & Sindhwani, 2006). These studies assume that unlabeled data have a true class label to be equal to zero which is not further motivated. We provide a solution to overcome this assumption. It is formulated here as an ex-

tension of a least squares support vector machine (LS-SVM) classifier (Suykens & Vandewalle, 1999) by additional constraints. LS-SVM methodology, as discussed in (Suykens, Van Gestel, De Brabanter, De Moor, & Vandewalle, 2002), enables to extend standard SVMs to a wider range of problems with primal-dual formulations for kernel methods in supervised and unsupervised learning and beyond.

This paper is organized as follows. First in Section 2 the approaches in (Tsuda et al., 2005) and (Belkin et al., 2006) are reviewed and the problem is stated. In Section 3, it is illustrated how to overcome the unrealistic assumption by taking additional constraints in LS-SVM classifiers. In Section 4 the method is tested on a number of artificial and real-life data sets.

2 Semi-supervised learning methods

In this Section two existing approaches to handle the semi-supervised learning problem are reviewed. First the method by (Tsuda et al., 2005) is explained, next, the approach by (Belkin et al., 2006) is briefly described. In order to be consistent with the sequel, the notations from the approach by (Tsuda et al., 2005) are slightly modified.

2.1 Tsuda et al.

Consider a data set $\{x_i, y_i\}_{i=1}^N$, with $x_i \in \mathbb{R}^d$ being the input vectors and $y_i \in \{-1, +1\}$ the class labels. In a semi-supervised setting part of the y_i values are unknown, resulting into a number of unlabeled data points x_i. In (Tsuda et al., 2005) this problem is tackled by a graph-based method,

inspired by (Zhou et al., 2004). A weighted graph is assumed and the strength of linkage is represented by an adjacency matrix V with elements v_{ij}. The elements v_{ij} are assumed to be nonnegative and equal to zero in case there is no edge between node i and node j. The first p data points are supposed to be labeled, the last q vectors remain unlabeled such that $N = p + q$. The goal of the approach in (Tsuda et al., 2005) is to make predictions $\hat{y}_{p+1}, ..., \hat{y}_N$ by exploiting the structure of the graph. As such, nodes with a strong linkage (i.e. a high v_{ij} value) tend to originate from the same class. The following criterion is proposed in (Tsuda et al., 2005)

$$\min_{\hat{y}} \mathcal{J}_1(\hat{y}) = \sum_{i=1}^{p} (\hat{y}_i - y_i)^2 + \sum_{i=p+1}^{N} \hat{y}_i^2 + \eta \sum_{i,j=1}^{N} v_{ij} (\hat{y}_i - \hat{y}_j)^2, \tag{1}$$

where $\hat{y} = [\hat{y}_1 \ ... \ \hat{y}_N]^T$ is the vector with class labels. Final class predictions are obtained by thresholding the $\hat{y}_i$ values. The first term in this minimization problem is the standard error on the training data. The second term imposes a reasonable range for the predictions of unlabeled data and the last term regularizes local smoothness where η denotes a positive regularization constant. Next, the problem is reformulated as

$$\min_{\hat{y}} \mathcal{J}_1(\hat{y}) = (\hat{y} - y)^T (\hat{y} - y) + \eta \ \hat{y}^T L \hat{y}, \tag{2}$$

obtaining the solution

$$\hat{y} = (I + \eta L)^{-1} y, \tag{3}$$

with L the graph Laplacian matrix (Chung, F. R. K., 1997) defined as $L =$

$D - V$ where $D = \text{diag}(d_1, \ ... \ , d_N)$, $d_i = \sum_{j=1}^{N} v_{ij}$ and $y = [y_1 \ ... \ y_p \ 0 \ ... \ 0]^T$. As such, in this approach it is explicitly assumed that the true labels for unlabeled data are equal to 0 while they are in fact completely unknown.

2.2 Belkin et al.

The study in (Belkin et al., 2006) is closely related to (Tsuda et al., 2005) and establishes a general framework for semi-supervised learning incorporating labeled and unlabeled data. Specifically, an extension to regularized least squares is proposed. The Laplacian regularized least squares (RLS) method minimizes

$$\min_{f \in \mathcal{H}_K} \mathcal{J}_2(f) = \frac{1}{2}\|f\|_K^2 + \gamma \frac{1}{2} \sum_{i=1}^{p} (y_i - f(x_i))^2 + \eta \mathbf{f}^T L \mathbf{f}, \tag{4}$$

where f belongs to a reproducing kernel Hilbert space $\mathcal{H}_K$. The first term is a regularization term to impose smoothness on the estimated function, K is the kernel function and $\mathbf{f} = [f(x_1) \ ... \ f(x_N)]^T$. The representer theorem is used to obtain the solution with an expansion of kernel functions with coefficients α_i such that

$$f(x) = \sum_{i=1}^{N} \alpha_i K(x, x_i). \tag{5}$$

Next, the following convex quadratic objective function is minimized

$$\min_{\alpha} \mathcal{J}_2(\alpha) = \gamma \frac{1}{2}(y - JK\alpha)^T (y - JK\alpha) + \frac{1}{2}\alpha^T K \alpha + \eta \alpha^T K L K \alpha, \tag{6}$$

resulting in the following solution (Belkin et al., 2006)

$$\alpha = (\gamma^{-1} I + JK + 2\eta\gamma^{-1} LK)^{-1} y, \tag{7}$$

where $J = \text{diag}(1, \ ... \ , 1, 0, \ ... \ , 0)$ with the first p elements equal to 1 and the last q equal to 0, $y = [y_1 \ ... \ y_p \ 0 \ ... \ 0]^T$ and the coefficients $\alpha = [\alpha_1 \ ... \ \alpha_N]^T$.

Hence, in this methodology it is also assumed that the targets y_i for unlabeled data equal 0. In the following Section, it is shown how to overcome this assumption and this is illustrated by incorporating it within an LS-SVM formulation.

3 A semi-supervised LS-SVM formulation

In (Suykens & Vandewalle, 1999) least squares support vector machines classifiers were proposed as

$$\min_{w,e,b} \mathcal{J}_3(w, e) = \frac{1}{2} w^T w + \gamma \frac{1}{2} \sum_{i=1}^{N} e_i^2, \tag{8}$$

subject to

$$y_i(w^T \varphi(x_i) + b) = 1 - e_i, \quad i = 1, ..., N, \tag{9}$$

with $e = [e_1 \ ... \ e_N]^T$, $\varphi : \mathbb{R}^d \to \mathbb{R}^{d_h}$ a mapping from the input space into a high-dimensional (potentially infinite dimensional) feature space of dimension d_h, w a vector of the same dimension as φ, γ a positive regularization constant

and b a bias term. In this way, the standard SVM formulation is changed by taking a least squares loss function with error variables e_i and modifying the inequality constraints into equality constraints. The value 1 in the equality constraints is a target value instead of a threshold value. Therefore, the method is related to kernel Fisher discriminant analysis (Suykens et al., 2002). This formulation results in solving a set of linear equations instead of solving a quadratic programming problem. The primal problem is expressed in terms of the feature map, the dual problem in terms of the kernel function.

We now extend this method towards semi-supervised problems by the following expression

$$\min_{w,e,b,\hat{y}} \mathcal{J}_4(w,e,\hat{y}) = \frac{1}{2}w^T w + \gamma \frac{1}{2}\sum_{i=1}^{N} e_i^2 + \eta \frac{1}{2}\sum_{i,j=1}^{N} v_{ij}(\hat{y}_i - \hat{y}_j)^2, \qquad (10)$$

such that

$$\begin{cases} \hat{y}_i = w^T\varphi(x_i) + b, \quad i = 1,...,N, \\ \\ \hat{y}_i = \nu_i y_i - e_i, \quad \nu_i \in \{0,1\}, \quad i = 1,...,N. \end{cases} \qquad (11)$$

In contrast with the approaches in (Tsuda et al., 2005; Belkin et al., 2006), this method does not assume a target value of 0 for unlabeled cases. In fact, no assumptions are made about the y_i values of the unlabeled data vectors. In our approach the encoding of labels is fully handled by the use of the ν_i variables. In case x_i is a labeled data point, the ν_i value is set to 1, otherwise 0 is assigned. This results in the standard squared loss function if a case is labeled. When dealing with unlabeled data an extra regularization term applies as in the

graph-based learning method presented in (Tsuda et al., 2005). The $\hat{y}_i$ values are minimized, which is less restrictive than assuming zero values for the true labels of the unlabeled data points.

Lemma 3.1 *The dual solution to the semi-supervised learning problem (10)-(11) is given by*

$$\begin{bmatrix} \gamma^{-1}I + K + 2\eta\gamma^{-1}LK & 1_N \\ 1_N^T & 0 \end{bmatrix} \begin{bmatrix} \alpha \\ b \end{bmatrix} = \begin{bmatrix} z \\ 0 \end{bmatrix}, \tag{12}$$

with $z = [\nu_1 y_1 \ ... \ \nu_N y_N]^T$, $1_N = [1 \ ... \ 1]^T$, K *the kernel matrix with* $K_{ij} = \varphi(x_i)^T\varphi(x_j) = K(x_i, x_j)$, L *the graph Laplacian matrix and*

$$(LK)_{ij} = \sum_{l=1}^{N} v_{il}(K_{ji} - K_{jl}), \ i, j = 1, ...N. \tag{13}$$

The resulting classifier is given by

$$\hat{y}(x) = \text{sign}\left(\sum_{i=1}^{N} \alpha_i K(x_i, x) + b\right). \tag{14}$$

Proof The Lagrangian for the problem (10) is

$$\mathcal{L}(w, b, e, \hat{y}; \alpha, \beta) = \mathcal{J}_4(w, e, \hat{y}) + \sum_{i=1}^{N} \alpha_i(\hat{y}_i - w^T\varphi(x_i) - b) + \sum_{i=1}^{N} \beta_i(\hat{y}_i - \nu_i y_i + e_i), \tag{15}$$

with α_i and β_i the Lagrange multipliers, $\alpha = [\alpha_1 \ ... \ \alpha_N]^T$ and $\beta = [\beta_1 \ ... \ \beta_N]^T$. The conditions for optimality yield

$$
\begin{cases}
\frac{\partial \mathcal{L}}{\partial w} = 0 \rightarrow w = \sum_{i=1}^{N} \alpha_i \varphi(x_i) \\
\frac{\partial \mathcal{L}}{\partial b} = 0 \rightarrow \sum_{i=1}^{N} \alpha_i = 0 \\
\frac{\partial \mathcal{L}}{\partial e_i} = 0 \rightarrow e_i = -\gamma^{-1} \beta_i, \ i = 1, ..., N \\
\frac{\partial \mathcal{L}}{\partial \alpha_i} = 0 \rightarrow \hat{y}_i = w^T \varphi(x_i) + b, \ i = 1, ..., N \\
\frac{\partial \mathcal{L}}{\partial \beta_i} = 0 \rightarrow \hat{y}_i = \nu_i y_i - e_i, \ i = 1, ..., N \\
\frac{\partial \mathcal{L}}{\partial \hat{y}_i} = 0 \rightarrow \alpha_i + \beta_i + 2\eta \sum_{j=1}^{N} v_{ij} (\hat{y}_i - \hat{y}_j), \ i = 1, ..., N.
\end{cases}
\tag{16}
$$

Elimination of w and e gives

$$
\begin{cases}
\sum_{i=1}^{N} \alpha_i = 0 \\
\sum_{j=1}^{N} \alpha_j \varphi(x_j)^T \varphi(x_i) + b = \nu_i y_i + \gamma^{-1} \beta_i, \ i = 1, ..., N \\
\beta_i = -\left(\alpha_i + 2\eta \sum_{j=1}^{N} v_{ij} (\hat{y}_i - \hat{y}_j) \right), \ i = 1, ..., N
\end{cases}
\tag{17}
$$

which gives the result (12).

□

The linear system in (12) is closely related to (3) and (7). Compared to the other formulations, the assumption that unlabeled cases x_i have corresponding y_i equal to 0 is omitted. Furthermore, the difference with (3) is that a kernel based approach applies in (12). Other than in (7), an extra regularization term keeping predicted $\hat{y}_i$ values small is added in (12). As a result, the matrix J is excluded from the equation in (12). This observation also suggests that when only a limited number of labeled data is available, the proposed

method might achieve an increased performance compared to Laplacian RLS since more information is used. This is further empirically studied in the next Section. Finally, an extra bias term b and an entry in the linear system that forces the Lagrange multipliers α_i to sum to 1 has been included.

4 Examples

In this Section both artificial and real-life data sets are used to illustrate the proposed model. The method is compared with standard LS-SVM classifiers. Also, the performance of semi-supervised LS-SVMs is compared with the one of Laplacian RLS on various problems.

4.1 Artificial examples

In the first example the proposed method is illustrated on the two moons benchmark classification problem, which is also considered in (Zhou et al., 2004; Belkin et al., 2006). The data set is made available by the study in (Belkin et al., 2006) and comprises 200 data points from which one case is labeled per class (i.e. moon). In Figure 1 the data are plotted in the upper left panel. Based on the labeled data point in each moon, the goal is to predict the class value for the unlabeled points in the grid. For this first toy problem optimal parameters are used, though these observations hold for a wider range of parameter values. The standard LS-SVM classifier with radial basis function kernel (RBF), depicted in the upper right panel of Figure 1, constructs a linear classifier which does not recognize the two moons. Next, all the unlabeled data points are included in the analysis by solving the linear system proposed

in (12) using an RBF kernel ($\gamma = 0.0001$, $\eta = 1$, $\sigma^2 = 0.2$), producing the lower left panel. Further, in the lower right panel, the same classifier is used to predict the label of 200 new data points. It can be seen that the inclusion of the term regularizing local smoothness improves classifier performance.

To verify the effect of the number of labeled data for standard LS-SVM classifiers, semi-supervised LS-SVM classifiers and Laplacian RLS, two overlapping Gaussians are used in the example, depicted in Figure 2. Each of the Gaussians includes 100 data points. Various amounts of labeled data are chosen to check the error for the different methods. Since the model selection procedure is rather extensive, the experiments are performed in the transductive setting. This does not require an extra subset to be held out from training. Transductive classifiers also have the potential to include more unlabeled data compared to inductive classifiers. The reported error rates on the unlabeled cases are averaged over 1000 runs, whereas in each run a fixed number of random data points from this data set are labeled in order to reduce any coincidence.

Model selection for the different classifiers is done on a separate data set as follows:

- For all methods an RBF kernel is used. The same bandwidth is used for all three methods. This bandwidth is tuned by cross-validation on the separate data set of 200 data points using the standard LS-SVM classifier.
- Model selection for the standard LS-SVM classifier ($\gamma = 14.5875$, $\sigma^2 = 2.28296$) is done only once by cross-validation on the separate set (i.e. 200 cases). For each of the different amounts of labeled data, these tuning parameters are used for the standard LS-SVM classifier. This is because model

selection is difficult for standard LS-SVMs when only a few cases are labeled.

- On the contrary, for Laplacian RLS and semi-supervised LS-SVMs the parameters (i.e. $\gamma \in \{10^{-8}, 10^{-7}, 10^{-6}, 10^{-5}, ..., 1, 2, ..., 30\}$, $\eta \in \{0.01, 0.2, 0.4, 0.6, 1, 2, ..., 30\}$) are tuned on the separate data set, for every different number of labeled data. Also, in this transductive setting 1000 random selections of labeled data are used. Each time an amount of points are randomly chosen, they are assumed to be labeled. All other points are assumed to be unlabeled. Finally, the performance is evaluated on the unlabeled points.
- The graph Laplacian matrices for the semi-supervised LS-SVM classifiers and Laplacian RLS are equal. The adjacency matrices contain binary weights, calculated based on the nearest neighbours. For each case the six nearest neighbours are determined. The v_{ij} values for these neighbours are set to 1, all others equal 0. Next, the adjacency matrices are symmetrized.

These models are now used on the independent data set and produce the results, depicted in the lower panel of Figure 2. One can observe that the averaged error rates of semi-supervised LS-SVMs and Laplacian RLS are smaller compared to the ones of standard LS-SVMs. When only a small amount of data is labeled this difference is larger. Increasing the number of labeled data points, makes the error rates converge. Another observation is that for smaller amounts of labeled data the performance of the semi-supervised LS-SVM classifier is slightly better than the one of Laplacian RLS.

In the example in Figure 3 the dimension of the input space is increased to 100 and more overlap is introduced. The experimental setup is identical to the previous one. In Figure 3 one can observe that the difference between Laplacian RLS and semi-supervised LS-SVMs is larger. Semi-supervised LS-SVMs show an increased performance compared to standard LS-SVMs and Laplacian RLS.

It is observed that standard LS-SVMs sometimes produce smaller error rates than Laplacian RLS. However, one has to keep in mind that model selection for standard LS-SVMs was done using all data. An explanation for the difference in performance between Laplacian RLS and semi-supervised LS-SVMs is the introduction of more overlap. Since there is more overlap between the Gaussians, more v_{ij} values indicate false relations between cases. This suggests that local smoothness is less decisive. As such, the semi-supervised LS-SVMs might achieve an increased performance when a smaller amount of data is labeled.

4.2 Real-life data sets

Here we further study the effect of the number of labeled data on two real-life data sets. As in (Belkin et al., 2006), the different methods are compared based on the USPS handwritten digits data set. Because of the extensive model selection process the analysis is restricted to a single binary problem, separating digit *6* from digit *8*. As in (Belkin et al., 2006), the first 400 images of each of the two digits from the USPS data set, are preprocessed by PCA allowing 100 dimensions. The first 200 cases for each of the two classes are used for model selection, the remaining are left for evaluation. Model selection is done in a similar way as before. An RBF kernel is used for standard LS-SVMs, semi-supervised LS-SVMs and Laplacian RLS. Model selection for the standard LS-SVM is done once by cross-validation ($\gamma = 193.8082$, $\sigma^2 = 144.5378$) on the first 400 images. Again, these parameters are used for all different amounts of labeled data. For Laplacian RLS and semi-supervised LS-SVMs the $\gamma \in \{10^{-8}, 10^{-7}, 10^{-6}, 10^{-5}, ..., 1, 2, ..., 30\}$ and $\eta \in \{0.01, 0.2, 0.4, 0.6, 1, 2, ..., 30\}$ are tuned on the first 400 data points

for each amount of labeled data. Also, in each run, 1000 random selections of labeled data are performed. Again, the Laplacian matrices for semi-supervised LS-SVMs and Laplacian RLS are equal. The adjacency matrices contain binary weights, selected based on the six nearest neighbours. For each case the v_{ij} values for the six nearest neighbours are set to 1. The performance on the independent data set, the last 400 images, is summarized in Figure 4. In the upper panel the performance of the standard LS-SVM classifier is decreased compared to the methods that take into account unlabeled data. When more data become labeled, their performances converge. The difference in performance between semi-supervised LS-SVMs and Laplacian RLS is larger for small amounts of labeled data. Also, boxplots are provided for the smallest amounts of labeled data in the lower panel. In this example, the spread for semi-supervised LS-SVMs seems to be a bit smaller.

Next, the analysis is repeated on the Isolet database, containing the letters of the English alphabet spoken in isolation. This data set contains expressions of 150 subjects, who spoke all letters of the alphabet twice. Like in (Belkin et al., 2006), 5 sets of 30 subjects each are constructed. In this comparison the focus is on distinguishing between the letters a and b. For model selection the first 300 cases, coming from the first three sets, are used while for evaluation purposes the last 300 cases, selected from the last three sets, are used. An RBF kernel is used for all methods. The bandwidth of this kernel is tuned by cross-validation on the first 300 cases using standard LS-SVMs. Model selection for the standard LS-SVM classifier is also done once by cross-validation ($\gamma = 4.0381$, $\sigma^2 = 607.9657$) on the first 300 data points. This standard LS-SVM model is used for all amounts of labeled data. Model selection for Laplacian RLS and semi-supervised LS-SVMs is performed for each

number of labeled data (i.e. $\gamma \in \{10^{-8}, 10^{-7}, 10^{-6}, 10^{-5}, ..., 1, 2, ..., 30\}$ and $\eta \in \{0.01, 0.2, 0.4, 0.6, 1, 2, ..., 30\}$), including 1000 random selections of labeled data as before. Laplacian matrices for semi-supervised LS-SVMs and Laplacian RLS are constructed based on the six nearest neighbours. Only binary weights are used in the adjacency matrices. The results of the selected models on the independent test set (i.e. 300 cases) are depicted in Figure 5. For small numbers of labeled data, the performance of Laplacian RLS and semi-supervised LS-SVMs is increased compared to standard LS-SVMs. These performances converge when more labeled data become involved. Also, for small numbers of labeled data semi-supervised LS-SVMs seem to result into less errors with respect to Laplacian RLS.

5 Conclusion

A semi-supervised version of the LS-SVM classifier is discussed that does not depend on the assumption that true class labels for unlabeled data equal zero. The problem is solved by the introduction of a set of variables and additional equality constraints into the classifier formulation. The solution is given by a linear system that is a modification to existing semi-supervised methods. It is illustrated that the inclusion of unlabeled information into the LS-SVM classifier increases the performance. The proposed method achieves a good performance on both artificial data sets and real-life problems that is comparable with the one of Laplacian RLS. For some data sets an increase in performance, compared to Laplacian RLS, is observed when only limited numbers of labeled data are available.

Acknowledgements

This research is funded by a PhD grant of the Institute for the Promotion of Innovation through Science and Technology in Flanders (IWT-Vlaanderen); Research supported by Research Council KUL: GOA-AMBioRICS, several PhD/postdoc & fellow grants, Centers-of-excellence optimisation, IDO 05/010 EEG-fMRI; Flemish Government: FWO: PhD/postdoc grants, projects, G.0360.05 (Advanced EEG analysis techniques for epilepsy monitoring), G.0519.06 (Noninvasive brain oxygenation), G.0341.07 (Data fusion), G.0321.06 (Numerical tensor techniques for spectral analysis), G.0302.07 (Support vector machines and kernel methods), research communities (ICCoS, ANMMM); IWT: PhD Grants; Belgian Federal Government: DWTC (IUAP IV-02 (1996-2001), IUAP V-22 (2002-2006): Dynamical Systems and Control: Computation, Identification & Modelling) and Belgian Federal Science Policy Office IUAP P6/04 (Dynamical systems, control and optimization, 2007-2011); EU: BIOPATTERN (contract no. FP6-2002-IST 508803), eTUMOUR (contract no. FP6-2002-LIFESCIHEALTH 503094), HealthAgents (contract no. FP6-2005-IST 027213), FAST (contract no. FP6-019279-2), ESA: Cardiovascular Control (Prodex-8 C90242).

References

Belkin, M., Niyogi, P., & Sindhwani, V. (2006). Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research*, *7*, 2399-2434.

Blum, A., & Chawla, S. (2001). Learning from labeled and unlabeled data using graph mincuts. In *Icml '01: Proceedings of the eighteenth international conference on machine learning* (pp. 19–26). San Francisco: Morgan Kaufmann Publishers Inc.

Blum, A., & Mitchell, T. (1998). Combining labeled and unlabeled data with co-training. In *COLT '98: Proceedings of the eleventh annual conference on computational learning theory* (pp. 92–100). New York: ACM Press.

Chapelle, O., Schölkopf, B., & Zien, A. (2006). *Semi-supervised learning*. Cambridge: MIT Press.

Chapelle, O., Weston, J., & Schölkopf, B. (2003). Cluster kernels for semi-supervised learning. *Neural Information Processing Systems*, *15*, 585-592.

Chung, F. R. K. (1997). *Spectral graph theory*. Providence, RI: American Mathematical Society.

Nigam, K., McCallum, A. K., Thrun, S., & Mitchell, T. M. (2000). Text classification from labeled and unlabeled documents using EM. *Machine Learning*, *39*, 103-134.

Suykens, J. A. K., Van Gestel, T., De Brabanter, J., De Moor, B., & Vandewalle, J. (2002). *Least squares support vector machines*. Singapore: World Scientific.

Suykens, J. A. K., & Vandewalle, J. (1999). Least squares support vector machine classifiers. *Neural Processing Letters*, *9*(3), 293-300.

Tsuda, K., Shin, H., & Schölkopf, B. (2005). Fast protein classification with multiple networks. *Bioinformatics*, *21*, 59-65.

Vapnik, V. N. (1998). *Statistical learning theory.* New York: Wiley-Interscience.

Yarowsky, D. (1995). Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd annual meeting on association for computational linguistics* (pp. 189–196). Morristown, NJ: Association for Computational Linguistics.

Zhou, D., Bousquet, O., Lal, T. N., Weston, J., & Schölkopf, B. (2004). Learning with local and global consistency. *Advances in Neural Information Processing Systems*, *16*, 321-328.

Zhu, X. (2005). *Semi-supervised learning literature survey* (Tech. Rep. No. 1530). Computer Sciences, University of Wisconsin-Madison.

Captions of figures

Figure 1: The two moons benchmark classification problem. The upper left panel contains the training data. Only one example of each class is labeled. The use of standard LS-SVMs, upper right panel, results in a linear decision boundary, not recognizing the structure of the two moons. By incorporating unlabeled data, based on the semi-supervised LS-SVM model, the two moons are clearly recognized as is depicted in the lower left panel. The performance of the same semi-supervised LS-SVM classifier on 200 new data points is provided in the lower right panel. These results are similar to the ones from earlier studies (Zhou et al., 2004; Belkin et al., 2006).

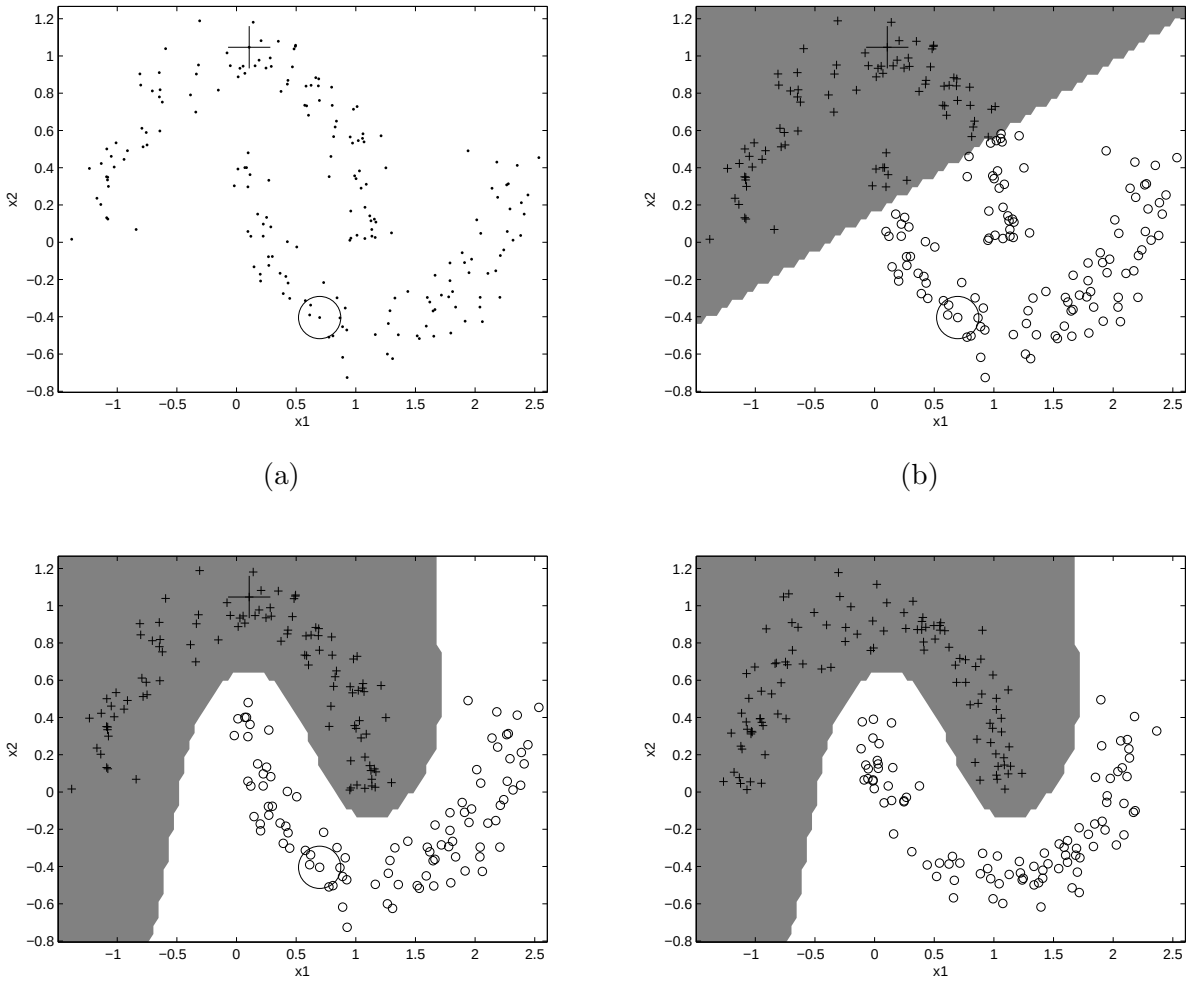
Figure 2: The effect of increasing the number of labeled data for standard LS-SVMs, semi-supervised LS-SVMs and Laplacian RLS is shown on an artificial data set. In the upper panel two overlapping Gaussians are depicted. For various amounts of labeled data the test error rates, averaged over 1000 runs of randomly selecting labeled points, is plotted. Semi-supervised LS-SVMs and Laplacian RLS produce less errors than standard LS-SVMs. When including more labeled data, the performances of all methods converge. For a smaller amount of labeled data, the semi-supervised LS-SVMs show an increased performance compared to Laplacian RLS.

Figure 3: The effect of increasing the number of labeled data for standard LS-SVMs, semi-supervised LS-SVMs and Laplacian RLS is shown for two overlapping Gaussians. Compared to the previous example, the input dimension is increased to 100 and there is more overlap between the Gaussians. The test error rates, averaged over 1000 runs of randomly selecting labeled points, is plotted for various amounts of labeled data. The differences between

semi-supervised LS-SVMs and Laplacian RLS are larger.

Figure 4: The effect of increasing the number of labeled data for standard LS-SVMs, semi-supervised LS-SVMs and Laplacian RLS when classifying between digit *6* and *8*, selected from the USPS data set, is shown. The test set error rate is plotted for different amounts of labeled data. In the upper panel the results are averaged over 1000 runs of randomly selecting labeled data. For small numbers of labeled data semi-supervised LS-SVMs and Laplacian RLS increase the performance with respect to standard LS-SVMs. When more labeled data are included their performances converge. The difference between semi-supervised LS-SVMs and Laplacian RLS is larger for a small amount of labeled data. In the lower panel boxplots are provided for the smallest amounts of labeled data for Laplacian RLS and semi-supervised LS-SVMs.

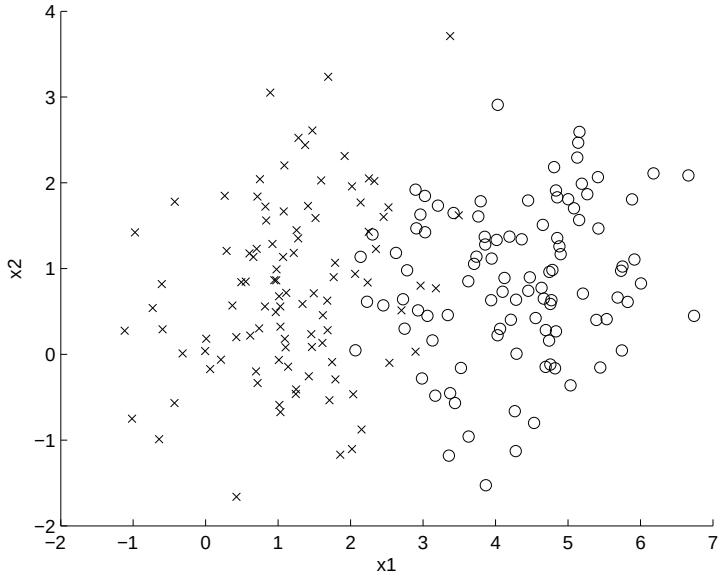
Figure 5: The effect of increasing the number of labeled data for standard LS-SVMs, semi-supervised LS-SVMs and Laplacian RLS when classifying between the letters a and b, chosen from the Isolet database, is shown. For little amounts of labeled data the average test error rates, over 1000 runs of randomly selecting labeled data, of semi-supervised LS-SVMs is smaller compared to the one of Laplacian RLS. Including more labeled data makes the error rates of standard LS-SVMs, semi-supervised LS-SVMs and Laplacian RLS converge.



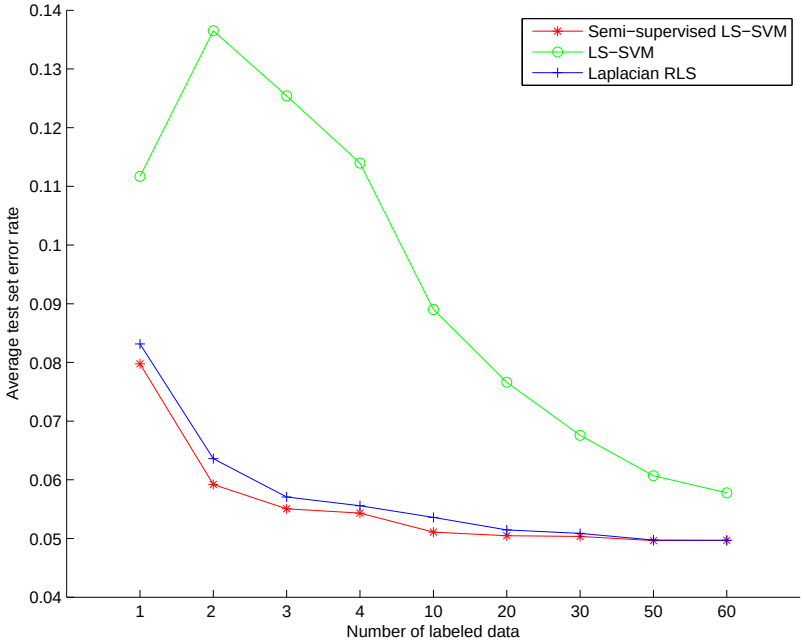
(a) (b)

(c) (d)

Fig. 1. Luts *et al.*



(a)



(b)

Fig. 2. Luts *et al.*

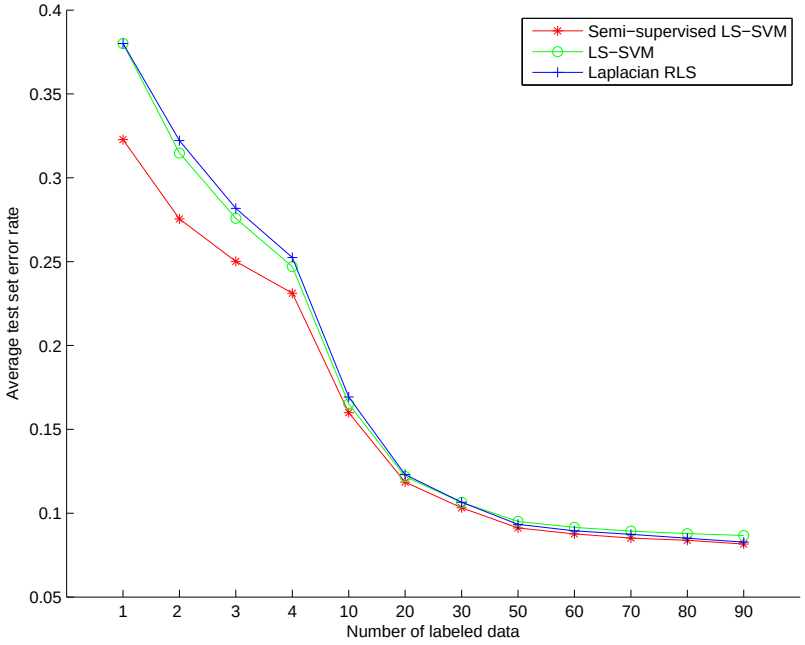
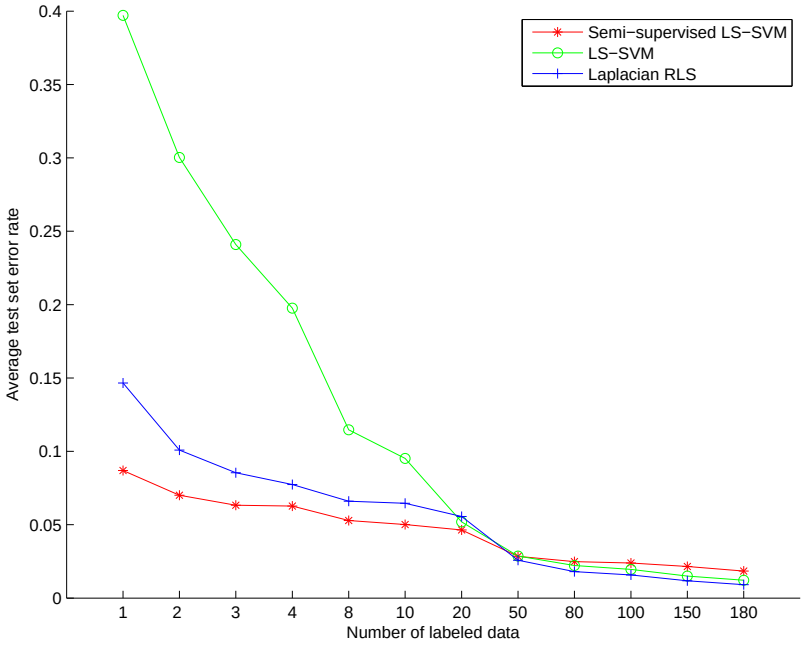
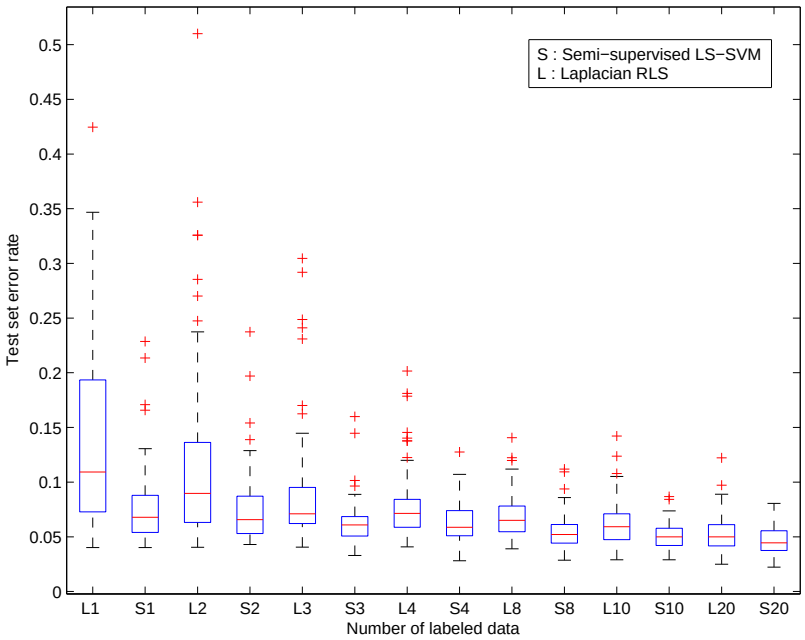


Fig. 3. Luts *et al.*



(a)



(b)

Fig. 4. Luts *et al.*

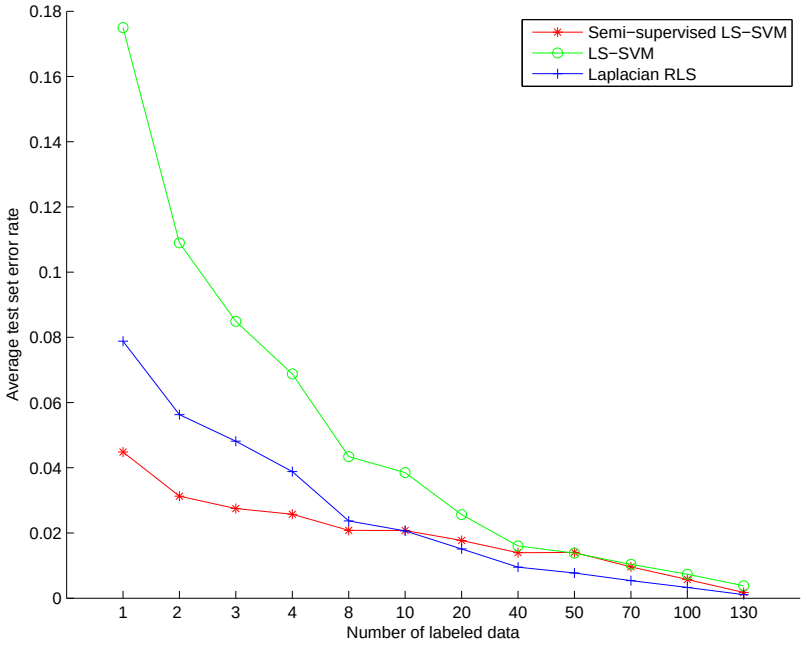


Fig. 5. Luts *et al.*