Mach Learn
DOI 10.1007/s10994-017-5626-8

Projected estimators for robust semi-supervised classification

Jesse H. Krijthe[1,2] · **Marco Loog**[1,3]

Received: 11 November 2013 / Accepted: 6 January 2017

Abstract For semi-supervised techniques to be applied safely in practice we at least want methods to outperform their supervised counterparts. We study this question for classification using the well-known quadratic surrogate loss function. Unlike other approaches to semi-supervised learning, the procedure proposed in this work does not rely on assumptions that are not intrinsic to the classifier at hand. Using a projection of the supervised estimate onto a set of constraints imposed by the unlabeled data, we find we can safely improve over the supervised solution in terms of this quadratic loss. More specifically, we prove that, measured on the labeled and unlabeled training data, this semi-supervised procedure never gives a lower quadratic loss than the supervised alternative. To our knowledge this is the first approach that offers such strong, albeit conservative, guarantees for improvement over the supervised solution. The characteristics of our approach are explicated using benchmark datasets to further understand the similarities and differences between the quadratic loss criterion used in the theoretical results and the classification accuracy typically considered in practice.

Keywords Semi-supervised learning · Least squares classification · Projection

1 Introduction

We consider the problem of semi-supervised classification using the quadratic loss function, which is also known as least squares classification or Fisher's linear discriminant classification (Hastie et al. 2009; Poggio and Smale 2003). Suppose we are given an $N_l \times d$ matrix with

Editor: Xiaojin Zhu.

✉ Jesse H. Krijthe
jkrijthe@gmail.com

1 Pattern Recognition and Bioinformatics, Delft University of Technology, Delft, The Netherlands

2 Department of Molecular Epidemiology, Leiden University Medical Center, Leiden, The Netherlands

3 The Image Group, University of Copenhagen, Copenhagen, Denmark

Published online: 03 April 2017

feature vectors $\mathbf{X}$, labels $\mathbf{y} \in \{0, 1\}^{N_l}$ and an $N_u \times d$ matrix with unlabeled objects $\mathbf{X}_u$ from the same distribution as the labeled objects. The goal of semi-supervised learning is to improve the classification decision function $f : \mathbb{R}^d \to \mathbb{R}$ using the unlabeled information in $\mathbf{X}_u$ as compared to the case where we do not have these unlabeled objects. In this work, we focus on linear classifiers where $f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}$.

Much work has been done on semi-supervised classification, in particular on what additional assumptions about the unlabeled data may help improve classification performance. These additional assumptions, while successful in some settings, are less successful in others where they do not hold. In effect they can greatly deteriorate performance when compared to a supervised alternative (Cozman and Cohen 2006). Since, in semi-supervised applications, the number of labeled objects may be small, the effect of these assumptions is often untestable. In this work, we introduce a conservative approach to training a semi-supervised version of the least squares classifier that is guaranteed to improve over the supervised least squares classifier, in terms of the quadratic loss on the labeled and unlabeled examples. It is the first procedure for which it is possible to give strong guarantees of non-degradation of this type (Theorem 1).

To guarantee these improvements, we avoid additional assumptions altogether. We introduce a constraint set of parameter vectors induced by the unlabeled data, which does not rely on additional assumptions about the data. Using a projection of the supervised solution vector onto this constraint set, we derive a method that can be proven to never degrade the surrogate loss evaluated on the labeled and unlabeled training data when compared to the supervised solution. Experimental results indicate that it not only never degrades, but often improves performance. Our experiments also indicate the results hold when performance is evaluated on objects in a test set that were not used as unlabeled objects during training.

The main contribution of this work is to prove that a semi-supervised learner that is guaranteed to outperform its supervised counterpart exists for some classifier. We do this by constructing one in the least squares classifier. This non-degradation property is important in practical applications, since one would like to be sure that the effort of the collection of, and computation with unlabeled data does not have an adverse effect. Our work is a conceptual step towards such methods. The goal of this work is to prove and illustrate this property.

Others have attempted to mitigate the problem of reduction in performance in semi-supervised learning by introducing safe versions of semi-supervised learners (Li and Zhou 2011; Loog 2010, 2014). These procedures do not offer any guarantees or only do so once particular assumptions about the data hold. Moreover, unlike some previous approaches, the proposed method can be formulated as a convex quadratic programming problem which can be solved using a simple gradient descent procedure.

The rest of this work is organized as follows. The next section discusses related work. Section 3 introduces our projection approach to semi-supervised learning. Section 4 discusses the theoretical performance guarantee and its implications. Section 5 provides some alternative interpretations of the method and relations to other approaches. In Sect. 6 empirical illustrations on benchmark datasets are presented to understand how the theoretical results in terms of quadratic loss in Sect. 4 relate to classification error on an unseen test set. We end with a discussion of the results and conclude.

2 Prior work and assumptions

Early work on semi-supervised learning dealt with the missing labels through the use of Expectation Maximization in generative models or closely related self-learning (McLachlan

1975). Self-learning is a simple wrapper method around any supervised procedure. Starting with a supervised learner trained only on the labeled objects, we predict labels for the unlabeled objects. Using the known labels and the predicted labels for the unlabeled objects, or potentially the predicted labels with highest confidence, we retrain the supervised learner. This process is iterated until the predicted labels converge. Although simple, this procedure has seen some practical success (Nigam et al. 2000).

Singh et al. (2008), among others, have argued that unlabeled data can *only* help if $P(\mathbf{x})$ and $P(y|\mathbf{x})$ are somehow linked. They show that when a specific cluster assumption holds, semi-supervised learning can be expected to outperform a supervised learner. The goal of our work is to show that in some cases (i.e. the least squares classifier) we do not need explicit assumptions about those links for semi-supervised learning to be possible. Instead, we leverage implicit assumptions, including possible model misspecification, that are already present in the supervised classifier. Similar to Singh et al. (2008), we also study the finite sample case.

Most recent work on semi-supervised methods considers what assumptions about this link between $P(\mathbf{x})$ and $P(y|\mathbf{x})$ allows for the effective use of unlabeled data. A lot of work involves either the assumption that the decision boundary is in a low-density region of the feature space, or that the data is concentrated on a low-dimensional manifold. A well-known procedure using the first assumption is the Transductive SVM (Joachims 1999). It can be interpreted as minimizing the following objective:

$$\min_{\mathbf{w}\in\mathbb{R}^d, \mathbf{y}_\mathrm{u}\in\{-1,+1\}^{N_u}} \sum_{i=1}^{N_l} \max(1 - y_i \mathbf{w}^\top \mathbf{x}_i, 0) + \lambda ||\mathbf{w}||^2 + \lambda_u \sum_{j=1}^{N_u} \max(1 - y_\mathrm{u}^{(j)} \mathbf{w}^\top \mathbf{x}_j, 0), \quad (1)$$

where class labels are encoded using $+1$ and -1. This leads to a hard to optimize, non-convex, problem, due to the dependence on the labels of the unlabeled objects $\mathbf{y}_\mathrm{u}$. Others, such as Sindhwani and Keerthi (2006), have proposed procedures to efficiently find a good local minimum of a related objective function. Similar low-density ideas have been proposed for other classifiers, such as entropy regularization for logistic regression (Grandvalet and Bengio 2005) and a method for Gaussian processes (Lawrence and Jordan 2004). One challenge with these procedures is setting the additional parameter λ_u that is introduced to control the effect of the unlabeled objects. This is both a computational problem, since minimizing (1) is already hard for a single choice of λ_u, as well as a estimation problem. If the parameter is incorrectly set using, for example, cross-validation on a limited set of labeled examples, the procedure may actually reduce performance as compared to a supervised SVM which disregards the unlabeled data. It is this behaviour that the procedure proposed in this work avoids. While it may be outperformed by the TSVM if the low-density assumption holds, robustness against deterioration would still constitute an important property in the cases when we are not sure whether it does hold.

Another oft-used assumption is that data is located on a lower dimensional manifold than the original dimensionality of the dataset. By estimating this manifold using unlabeled data we can improve the estimate of the classification boundary (Zhu et al. 2003). Theoretical results have shown that particular classes of problems can be constructed, where manifold regularization can solve classification problems (Niyogi 2013) that cannot be efficiently learned without knowing the manifold. For these classes of problem the objects actually do reside on a lower dimensional manifold and the distance between objects on this manifold is

essential for their classification. When a problem does not belong to such a class, Lafferty and Wasserman (2007) show that manifold regularization does not improve over supervised learning. In these cases, manifold regularization may actually lead to worse performance than the supervised alternative. In general, these methods require some domain knowledge by having to define a similarity matrix between objects. Again, if the manifold assumption does not hold, or the domain knowledge is not correctly specified, the semi-supervised classifier may be outperformed by the supervised classifier.

An attempt at safety in semi-supervised learning was introduced in Li and Zhou (2011), who propose a safe variant for semi-supervised support vector machines. By constructing a set of possible decision boundaries using the unlabeled and labeled data, the decision boundary is chosen that is least likely to degrade performance. While the goal of this work is similar, we do not rely on the existence of a low-density separator and obtain a much simpler optimization problem.

Another attempt at safety was proposed by Loog (2010, 2014), who introduce a semi-supervised version of linear discriminant analysis, which is closely related to the least squares classifier considered here. There, explicit constraints are proposed that take into account the unlabeled data. In our work, these constraints need not be explicitly derived, but follow directly from the choice of loss function and the data. While the impetus for these works is similar to ours, they provide no theory to guarantee no degradation in performance will occur similar to our results in Sect. 4.

3 Projection method

The proposed projection method works by forming a constraint set of parameter vectors Θ, informed by the labeled *and unlabeled* objects, that is guaranteed to include $\mathbf{w}_{\text{oracle}}$, the solution we would obtain if we had labels for all the training data. We will then find the closest projection of the supervised solution $\mathbf{w}_{\text{sup}}$ onto this set, using a chosen distance measure. This new estimate, $\mathbf{w}_{\text{semi}}$, will then be guaranteed to be closer to the oracle solution than the supervised solution $\mathbf{w}_{\text{sup}}$ in terms of this distance measure. For a particular choice of measure, it follows (Sect. 4) that $\mathbf{w}_{\text{semi}}$ will always have lower quadratic loss when measured on the labeled and unlabeled training data, as compared to $\mathbf{w}_{\text{sup}}$.

Before we move to the details of our particular contribution, we first introduce briefly the standard supervised least squares classifier.

3.1 Supervised solution

We consider classification using a quadratic surrogate loss (Hastie et al. 2009). In the supervised setting, the following objective is minimized for $\mathbf{w}$:

$$L(\mathbf{w}, \mathbf{X}, \mathbf{y}) = \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2. \tag{2}$$

The supervised solution $\mathbf{w}_{\text{sup}}$ is given by the minimization of (2) for $\mathbf{w}$. The well-known closed form solution to this problem is given by

$$\mathbf{w}_{\text{sup}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}. \tag{3}$$

If the true labels corresponding to the unlabeled objects, $\mathbf{y}_{\text{u}}^*$, would be given, we could incorporate these by extending the vector of labels $\mathbf{y}_{\text{e}}^{*\top} = \left[\mathbf{y}^\top \mathbf{y}_{\text{u}}^{*\top}\right]$ as well as the design

matrix $\mathbf{X}_e^\top = \left[\mathbf{X}^\top \mathbf{X}_u^\top\right]$ and minimize $L(\mathbf{w}, \mathbf{X}_e, \mathbf{y}_e^*)$ over the labeled as well as the unlabeled objects. We will refer to this oracle solution as $\mathbf{w}_{\text{oracle}}$.

3.2 Constraint set

Our proposed semi-supervised approach is to project the supervised solution $\mathbf{w}_{\text{sup}}$ onto the set of all possible classifiers we would be able to get from some labeling of the unlabeled data. To form this constraint set, consider all possible labels for the unlabeled objects $\mathbf{y}_u \in [0, 1]^{N_u}$. This includes fractional labelings, where an object is partly assigned to class 0 and partly to class 1. For instance, 0.5 indicates the object is assigned equally to both classes. For a particular labeling $\mathbf{y}_e^\top = \left[\mathbf{y}^\top \mathbf{y}_u^\top\right]$, we can find the corresponding parameter vector by minimizing $L(\mathbf{w}, \mathbf{X}_e, \mathbf{y}_e)$ for $\mathbf{w}$. This objective remains the same as (2) except that fractional labels are now also allowed. Minimizing the objective for all possible labelings generates the following set of solutions:

$$\Theta = \left\{ \left(\mathbf{X}_e^\top \mathbf{X}_e\right)^{-1} \mathbf{X}_e^\top \begin{bmatrix} \mathbf{y} \\ \mathbf{y}_u \end{bmatrix} \mid \mathbf{y}_u \in [0, 1]^{N_u} \right\}. \tag{4}$$

Note that this set, by construction, will also contain the solution $\mathbf{w}_{\text{oracle}}$, corresponding to the true but unknown labeling $\mathbf{y}_e^*$. Typically, $\mathbf{w}_{\text{oracle}}$ is a better solution than $\mathbf{w}_{\text{sup}}$ and so we would like to find a solution more similar to $\mathbf{w}_{\text{oracle}}$. This can be accomplished by projecting $\mathbf{w}_{\text{sup}}$ onto Θ.

3.3 Choice of metric

It remains to determine how to calculate the distance between $\mathbf{w}_{\text{sup}}$ and any other $\mathbf{w}$ in the space. We will consider the following metric:

$$\mathrm{d}(\mathbf{w}, \mathbf{w}') = \sqrt{(\mathbf{w} - \mathbf{w}')^\top \mathbf{X}_\circ^\top \mathbf{X}_\circ (\mathbf{w} - \mathbf{w}')}, \tag{5}$$

where we assume $\mathbf{X}_\circ^\top \mathbf{X}_\circ$ is a positive definite matrix. The projected estimator can now be found by minimizing this distance between the supervised solution and solutions in the constraint set:

$$\mathbf{w}_{\text{semi}} = \min_{\mathbf{w} \in \Theta} \mathrm{d}(\mathbf{w}, \mathbf{w}_{\text{sup}}). \tag{6}$$

Setting $\mathbf{X}_\circ = \mathbf{X}_e$ measures the distances using both the labeled and unlabeled data. This choice has the desirable theoretical properties leading us to the sought-after improvement guarantees as we will demonstrate in Sect. 4.

3.4 Optimization

By plugging into (6) the closed form solution of $\mathbf{w}_{\text{sup}}$ and $\mathbf{w}$ for a given $\mathbf{y}_u$, this problem can be written as a convex minimization problem in terms of $\mathbf{y}_u$, the unknown, fractional labels of the unlabeled data. This results in a quadratic programming problem, which can be solved using a simple gradient descent procedure that takes into account the constraint that the labels are within [0, 1]. The solution of this quadratic programming problem $\hat{\mathbf{y}}_u$ can then be used to find $\mathbf{w}_{\text{semi}}$ by treating these imputed labels as the true labels of the unlabeled objects and combining them with the labeled examples in Eq. (3).

4 Theoretical analysis

We start by stating and proving our main result which is a guarantee of non-degradation in performance of the proposed method compared to the supervised classifier. We then discuss extensions of this result to other settings and give an indication of when improvement over the supervised solution can be expected.

4.1 Robustness guarantee

Theorem 1 *Given* $\mathbf{X}$, $\mathbf{X}_\mathrm{u}$ *and* $\mathbf{y}$, $\mathbf{X}_\mathrm{e}^\top \mathbf{X}_\mathrm{e}$ *positive definite and* $\mathbf{w}_\mathrm{sup}$ *given by* (3). *For the projected estimator* $\mathbf{w}_\mathrm{semi}$ *proposed in* (6), *the following result holds:*

$$L(\mathbf{w}_\mathrm{semi}, \mathbf{X}_\mathrm{e}, \mathbf{y}_\mathrm{e}^*) \leq L(\mathbf{w}_\mathrm{sup}, \mathbf{X}_\mathrm{e}, \mathbf{y}_\mathrm{e}^*)$$

In other words: $\mathbf{w}_\mathrm{semi}$ will *always* be at least as good or better than $\mathbf{w}_\mathrm{sup}$, in terms of the quadratic surrogate loss on all, labeled and unlabeled, training data. While this claim does not prove, in general, that the semi-supervised solution improves in terms of the loss evaluated on the true distribution, or an unseen test set, we will consider why this is still a desirable property after the proof.

Proof The proof of this result follows from a geometric interpretation of our procedure. Consider the following inner product that induces the distance metric in Eq. (5):

$$\left\langle \mathbf{w}, \mathbf{w}' \right\rangle = \mathbf{w}^\top \mathbf{X}_\mathrm{e}^\top \mathbf{X}_\mathrm{e} \mathbf{w}' .$$

Let $\mathcal{H}_{\mathbf{X}_\mathrm{e}} = (\mathbb{R}^d, \langle ., . \rangle)$ be the inner product space corresponding with this inner product. As long as $\mathbf{X}_\mathrm{e}^\top \mathbf{X}_\mathrm{e}$ is positive definite, this is a Hilbert space. Next, note that the constraint space Θ is convex. More precisely, because, for any $k \in [0, 1]$ and $\mathbf{w}_1, \mathbf{w}_2 \in \Theta$ we have that

$$\begin{aligned}
(1-k)\mathbf{w}_1 + k\mathbf{w}_2 =& (1-k)\left(\mathbf{X}_\mathrm{e}^\top \mathbf{X}_\mathrm{e}\right)^{-1} \mathbf{X}_\mathrm{e}^\top \left[\mathbf{y}^\top \mathbf{y}_1^\top\right]^\top \\
&+ k\left(\mathbf{X}_\mathrm{e}^\top \mathbf{X}_\mathrm{e}\right)^{-1} \mathbf{X}_\mathrm{e}^\top \left[\mathbf{y}^\top \mathbf{y}_2^\top\right]^\top \\
=& \left(\mathbf{X}_\mathrm{e}^\top \mathbf{X}_\mathrm{e}\right)^{-1} \mathbf{X}_\mathrm{e}^\top \left[\mathbf{y}^\top \;\; k\mathbf{y}_1^\top + (1-k)\mathbf{y}_2^\top\right]^\top \\
\in& \Theta
\end{aligned}$$

where the last statement holds because $k\mathbf{y}_1 + (1-k)\mathbf{y}_2 \in [0, 1]^{N_u}$.

By construction $\mathbf{w}_\mathrm{semi}$ is the closest projection of $\mathbf{w}_\mathrm{sup}$ onto this convex constraint set Θ in $\mathcal{H}_{\mathbf{X}_\mathrm{e}}$. One of the properties for projections onto a convex subspace in a Hilbert space is (Aubin 2000, Proposition 1.4.1.) that

$$\mathrm{d}(\mathbf{w}_\mathrm{semi}, \mathbf{w}) \leq \mathrm{d}(\mathbf{w}_\mathrm{sup}, \mathbf{w}) \tag{7}$$

for any $\mathbf{w} \in \Theta$. In particular consider $\mathbf{w} = \mathbf{w}_\mathrm{oracle}$, which by construction is within Θ. That is, all possible labelings correspond to an element in Θ, so this also holds for the true labeling $\mathbf{y}_\mathrm{u}^*$. Plugging in the closed form solution of $\mathbf{w}_\mathrm{oracle}$ into (7) and squaring the distance we find:

$$\begin{aligned}
\mathrm{d}(\mathbf{w}_\mathrm{semi}, \mathbf{w}_\mathrm{oracle})^2 =& \mathbf{w}_\mathrm{semi}^\top \mathbf{X}_\mathrm{e}^\top \mathbf{X}_\mathrm{e} \mathbf{w}_\mathrm{semi} \\
&- 2\mathbf{w}_\mathrm{semi}^\top \mathbf{X}_\mathrm{e}^\top \mathbf{y}_\mathrm{e}^* + \mathbf{y}_\mathrm{e}^{*\top} \mathbf{y}_\mathrm{e}^* \\
&+ C \\
=& L(\mathbf{w}_\mathrm{semi}, \mathbf{X}_\mathrm{e}, \mathbf{y}_\mathrm{e}^*) + C
\end{aligned}$$

and

$$\begin{aligned} \mathrm{d}(\mathbf{w}_{\text{sup}}, \mathbf{w}_{\text{oracle}})^2 =& \mathbf{w}_{\text{sup}}^\top \mathbf{X}_{\text{e}}^\top \mathbf{X}_{\text{e}} \mathbf{w}_{\text{sup}} \\ & - 2\mathbf{w}_{\text{sup}}^\top \mathbf{X}_{\text{e}}^\top \mathbf{y}_{\text{e}}^* + \mathbf{y}_{\text{e}}^{*\top} \mathbf{y}_{\text{e}}^* \\ & + C \\ =& L(\mathbf{w}_{\text{sup}}, \mathbf{X}_{\text{e}}, \mathbf{y}_{\text{e}}^*) + C \end{aligned}$$

where C is the same constant in both cases. From this the result in Theorem 1 follows directly. □

4.2 Generalization performance

So far, we have considered the performance of the procedure evaluated on the labeled and unlabeled objects, instead of the out of sample performance on unseen test data. A different quantity of interest is the expected loss, which is based on the true underlying data distribution and is also referred to as the risk:

$$\sum_{y \in \{0,1\}} \int (y - \mathbf{x}^\top \mathbf{w})^2 p(\mathbf{x}, y) d\mathbf{x} \,.$$

The result does not prove that $\mathbf{w}_{\text{semi}}$ is, in general, better than $\mathbf{w}_{\text{sup}}$ in terms of this risk. In case $N_u \to \infty$, however, and $p(\mathbf{x})$ basically becomes known, $\mathbf{w}_{\text{semi}}$ is in fact guaranteed to be better in terms of the risk, since the risk becomes equal to the loss on the labeled and unlabeled data in this case.

When we have a finite number of unlabeled samples, the result presented in the theorem is still relevant because it proves that we at least get a better solution in terms of the empirical risk on the full data, i.e., the risk that is typically minimized if all labels are actually available.

Apart from this, one may be specifically interested in the performance on a given set of objects, the transductive learning setting, which we will address now.

4.3 Transduction and regularization

It is possible to derive a similar result for performance improvement on the unlabeled data alone by using $\mathbf{X}_\circ = \mathbf{X}_{\text{u}}$ in the distance measure and changing the constrained hypothesis space to:

$$\Theta_{\text{u}} = \left\{ (\mathbf{X}_{\text{u}}^\top \mathbf{X}_{\text{u}})^{-1} \mathbf{X}_{\text{u}}^\top \mathbf{y}_{\text{u}} \mid \mathbf{y}_{\text{u}} \in [0, 1]^{N_u} \right\}.$$

This would lead to a guarantee of the form:

$$L(\mathbf{w}_{\text{semi}}, \mathbf{X}_{\text{u}}, \mathbf{y}_{\text{u}}^*) \leq L(\mathbf{w}_{\text{sup}}, \mathbf{X}_{\text{u}}, \mathbf{y}_{\text{u}}^*) \,.$$

However, since we would not just like to perform well on the given unlabeled data, but on unseen data from the same distribution as well, we include the labeled data in the construction of the constrained hypothesis space.

The result in Theorem 1 also holds if we include regularization in the supervised classifier. Using L_2 regularization, the supervised solution becomes:

$$\mathbf{w}_{\text{sup}} = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\top \mathbf{y},$$

where λ is a regularization parameter and $\mathbf{I}$ a $d \times d$ identity matrix, potentially containing a 0 for the diagonal entry corresponding to the constant feature that encodes the bias. Theorem 1 also holds for this regularized supervised estimator.

4.4 Improved performance

Since the inequality in Theorem 1 is not necessarily a strict inequality, it is important to get an idea when we can expect improvement of the semi-supervised learner, rather than just equality of the losses. Consider a single unlabeled object. Improvement happens whenever $\mathbf{w}_{\text{sup}} \neq \mathbf{w}_{\text{semi}}$, which occurs if $\mathbf{w}_{\text{sup}} \notin \Theta$. For this to occur it needs to be impossible to assign a label y_u such that we can retrieve the $\mathbf{w}_{\text{sup}}$ by minimizing $L(\mathbf{w}, \mathbf{X}_e, \mathbf{y}_e)$. This in turn occurs when there is no $y_u \in [0, 1]$ for which the gradient

$$\nabla \|\mathbf{X}_e \mathbf{w} - \mathbf{y}_e\|^2 \Big|_{\mathbf{w}=\mathbf{w}_{\text{sup}}} = \mathbf{0}\,.$$

This happens only if $\mathbf{x}_u^\top \mathbf{w}_{\text{sup}} > 1$ or $\mathbf{x}_u^\top \mathbf{w}_{\text{sup}} < 0$. In other words, if observations $\mathbf{x}_u$ are possible with values that have a sufficiently large absolute value and $\mathbf{w}_{\text{sup}}$ is not small enough to mitigate this, an update will occur. This is especially likely to occur if the supervised solution is not sufficiently regularized, $\mathbf{x}_u^\top \mathbf{w}_{\text{sup}}$ can then easily be larger than 1 or smaller than 0. For more than a single unlabeled object, the conditions for a change are more complex, since the introduction of a non-zero gradient by one object can be compensated by other objects. The experiments in Sect. 6 confirm, however, that generally improvements can be expected by means of the proposed semi-supervised learning strategy.

5 Relation to other methods

The projection method in Eq. (6), using $\mathbf{X}_\circ = \mathbf{X}_e$ in the distance measure, can be rewritten in a different form:

$$\arg\min_{\mathbf{w}_{\text{semi}}} \max_{\mathbf{y}_u \in [0,1]^{N_u}} L(\mathbf{w}_{\text{semi}}, \mathbf{X}_e, \mathbf{y}_e) - L(\mathbf{w}_{\text{sup}}, \mathbf{X}_e, \mathbf{y}_e).$$

In other words, the procedure can be interpreted as a minimization of the difference in loss on the labeled and unlabeled data between the new solution and the supervised solution, over all possible labelings of the unlabeled data. From this perspective the projected estimator is similar to Maximum Contrastive Pessimistic Likelihood Estimation proposed by Loog (2016) who consider using log likelihood as the loss function. In this formulation it is apparent that the projected estimator is very conservative, since it has to have low loss for all possible labelings, even very unlikely ones.

In a similar way an alternative choice of distance function, $\mathbf{X}_\circ = \mathbf{X}$, has a different interpretation. It is the minimizer of the supervised loss function under the constraint that its solution has to be a minimizer for some labeling of the unlabeled data:

$$\arg\min_{\mathbf{w} \in \Theta} L(\mathbf{w}, \mathbf{X}, \mathbf{y}),$$

with Θ defined as in Eq. (4). This formulation corresponds to the Implicitly Constrained Least Squares Classifier (Krijthe and Loog 2015) and seems less conservative since the solution does not need to have a low loss for all possible labelings, it merely has to work well on the labeled examples. For this distance measure, the proof in Sect. 4 no longer holds, but empirical results indicate it may have better performance in practice, while it still protects against deterioration in performance by minimizing the loss over only the labeled objects.

Another interpretation of the projection procedure is that it minimizes the squared difference between the predictions of the supervised solution and a new semi-supervised solution

on the set of objects in $\mathbf{X}_\circ$, while ensuring the semi-supervised solution corresponds to a possible labeling of the unlabeled objects:

$$\min_{\mathbf{w}\in\Theta} \|\mathbf{X}_\circ \mathbf{w} - \mathbf{X}_\circ \mathbf{w}_{\text{sup}}\|^2 .$$

Since this comparison requires only the features in $\mathbf{X}_\circ$ and not the corresponding labels, this can be done either on the labeled data, when we choose $\mathbf{X}_\circ = \mathbf{X}$, but also on the labeled and unlabeled data combined when $\mathbf{X}_\circ = \mathbf{X}_\text{e}$. This interpretation is similar to the work of Schuurmans and Southey (2002), where the unlabeled objects are also used to measure the difference in predictions of two hypotheses.

6 Experimental analysis

For our experiments, we consider 16 classification datasets. six of these are the semi-supervised learning benchmark datasets proposed by Chapelle et al. (2006), while the other ten were retrieved from the UCI Machine Learning repository (Lichman 2013). All of the datasets are binary classification problems, or were turned into two-class problems by merging several similar classes. As a preprocessing step, missing input values were imputed using medians and modes for the Mammography and Diabetes datasets. The code to reproduce the results presented here is available from the first author's website.

The number of labeled examples is chosen such that $N_l > d$. This is necessarily to have a high probability that the matrix $\mathbf{X}_\text{e}^\top \mathbf{X}_\text{e}$ is positive definite, which was a requirement of Theorem 1. More importantly, this avoids peaking behaviour (Raudys and Duin 1998; Opper and Kinzel 1996), were the unregularized supervised least squares classifier has low performance when the matrix $\mathbf{X}^\top \mathbf{X}$ is not full-rank. For the SVM and TSVM implementations we made use of the SVMlin software (Sindhwani and Keerthi 2006). For these we used parameter settings $\lambda = 0.01$ and $\lambda_u = 1$.

6.1 Robustness

To illustrate Theorem 1 experimentally, as well as study the performance of the proposed procedure on a test set, we set up the following experiment. For each of the 16 datasets, we randomly select $2d$ labeled objects. We then randomly sample, with replacement, 1000 objects as the unlabeled objects from the dataset. In addition, a test set of 1000 objects is also sampled with replacement. This procedure is repeated 100 times and the ratio between the average quadratic losses for the supervised and the semi-supervised procedure $\frac{L(\mathbf{w}_\text{semi}, \mathbf{X}_\text{e}, \mathbf{y}_\text{e}^*)}{L(\mathbf{w}_\text{sup}, \mathbf{X}_\text{e}, \mathbf{y}_\text{e}^*)}$ is calculated. As stated by Theorem 1, this quantity should be smaller than 1 for the Projection procedure. We do the same for self-learning applied to the least squares classifier and to an L_2-Transductive SVM, which we compare to the supervised L_2-SVM. The results are shown in Figure 1.

On the labeled and unlabeled data the loss of the projection method is lower than that of the supervised classifier in all of the resamplings taken from the original dataset. Compare this to the behaviour of the self-learner. While on average, the performance is quite similar on these datasets, on a particular sample from a dataset, self-learning may lead to a higher quadratic loss than the supervised solution. It is favourable to have no deterioration in every resampling because in practice one does not deal with resamplings from an empirical distribution, but rather with a single dataset. A semi-supervised procedure should ideally work on this particular dataset, rather than in expectation over all datasets that one might have observed.

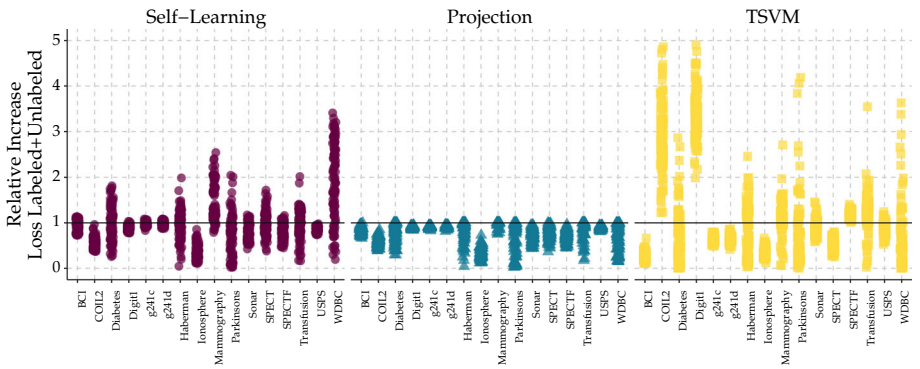

Fig. 1 Ratio of the loss in terms of surrogate loss of semi-supervised and supervised solutions measured on the labeled and unlabeled instances. Values smaller than 1 indicate that the semi-supervised method gives a lower average surrogate loss than its supervised counterpart. For both the projected estimator and self-learning this supervised counterpart is the supervised least squares classifier and loss is in terms of quadratic loss. For the L_2-Transductive SVM, quadratic hinge loss is used and compared to the quadratic hinge loss of a supervised L_2-SVM. Unlike the other semi-supervised procedures, the projection method, evaluated on labeled and unlabeled data, never has higher loss than the supervised procedure, as was proven in Theorem 1

We see similar behaviour as self-learning for the difference in squared hinge loss between the L_2-SVM and the L_2-TSVM. While better parameter choices may improve the number of resamplings with improvements, this experiment illustrates that while semi-supervised methods may improve performance on average, for a particular sample from a dataset there is no guarantee like Theorem 1 for the projected estimator. When looking at the difference in loss on an unseen test set, we find a similar results (not shown).

6.2 Learning curves

To illustrate the behaviour of the procedure with increasing amounts of unlabeled data and to explore the relationship between the quadratic surrogate loss and classification accuracy on an unseen test set we generate learning curves in the following manner. For each of three illustrative datasets (Ionosphere, SPECT and USPS), we randomly sample $2d$ objects as labeled objects. The remaining objects are used as a test set. For increasing subsets of the unlabeled data $(2, 4, 8, \ldots, 512)$, randomly sampled without replacement, we train the supervised and semi-supervised learners and evaluate their performance on the test objects, in terms of classification accuracy as well as in terms of quadratic loss. We consider both the projection procedure where the distance measure is based on the labeled and the unlabeled data (denoted as Projection) as well as the projected estimator that only uses the labeled data in the distance measure (denoted as ICLS). The resampling is repeated 1000 times and averages and standard errors are reported in Figure 2.

The first dataset (Ionosphere) in Figure 2 is an example where the error of the self-learning procedure starts to increase once we add larger amounts of unlabeled data. In terms of the loss, however, the performance continues to increase. This illustrates that a decrease in the surrogate loss does not necessarily translate into a lower classification error. The projected estimators do not suffer from decreases in performance for larger numbers of unlabeled data in this example. In terms of the loss, however, there seems to be little difference between the three methods.

The second dataset (SPECT) is an example where both the self-learning procedure and the conservative projected estimator are not able to get any improvement out of the data, while the

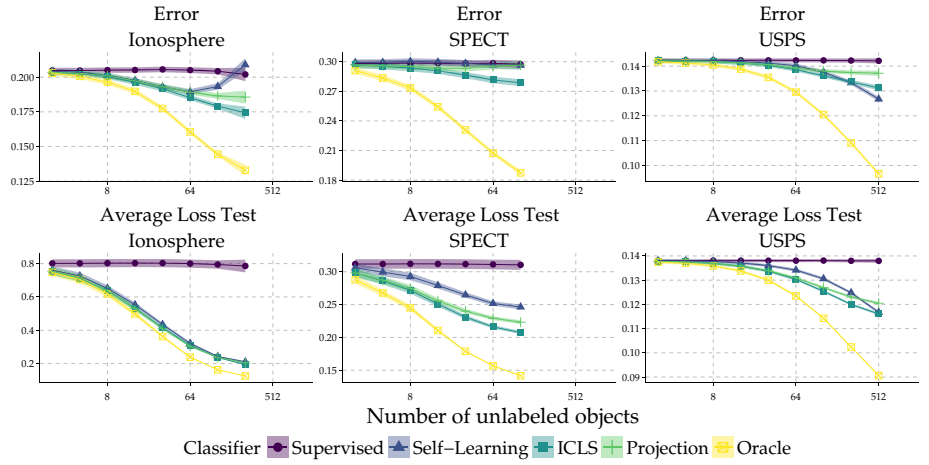

Fig. 2 Learning curves in terms of classification errors (*top*) and quadratic loss (*bottom*) on the test set for increasing numbers of unlabeled data on three illustrative datasets. The *lines* indicate average errors respectively losses on the test set, averaged over 1000 repeats. The *shaded bars* indicate ± 2 standard errors around the mean

less conservative projection (ICLS) does show some improvement in terms of classification error.

On the USPS dataset the self-learning assumptions do seem to hold and it is able to attain a larger performance improvement as the amount of unlabeled data grows. Both in terms of the error and in terms of the loss, the projected estimators show smaller, but significant improvements.

6.3 Cross-validation

In a third experiment, we apply a cross-validation procedure to compare the performance increase in terms of the classification error of semi-supervised classifiers when compared to their supervised counterpart. The cross-validation experiments were set up as follows. For each dataset, the objects were split into 10-folds. Subsequently leaving out each fold, we combine the other 9 folds and randomly select $d + 5$ labeled objects while the rest is used as unlabeled objects. We end up with a single prediction for each object, for which we evaluate the misclassification error. This procedure is repeated 20 times and the averages are reported in Table 1.

The results indicate that in terms of classification errors, the projection procedure never significantly reduces performance over the supervised solution. This is in contrast to the self-learner, which does significantly increase classification error on 2 of the datasets. The price the projected estimator pays for this robustness, is smaller improvements over the supervised classifier than the less conservative self-learner. The Transductive SVM shows similar behaviour as the self-learner: it shows large improvements over the supervised alternative, but is also prone to degradation in performance on other datasets. The ICLS procedure is, as expected, less conservative than the projection method based on the labeled and unlabeled observations, which leads to larger improvements on all of the datasets.

Table 1 20 repeat 10-fold cross-validation results for 16 datasets for the supervised least squares classifier, the projected least squares classifier (Projected), the projection based on only the labeled data (ICLS), the self-learned least squares classifier, the supervised L_2-SVM and the L_2-TSVM

Dataset	Supervised	Self-Learning	ICLS	Projection	SVM	TSVM
BCI	0.40 ± 0.03	$\mathbf{0.35 \pm 0.02}$	$\mathbf{0.28 \pm 0.02}$	$\mathbf{0.36 \pm 0.03}$	0.30 ± 0.02	0.31 ± 0.02
COIL2	0.39 ± 0.01	$\mathbf{0.26 \pm 0.01}$	$\mathbf{0.19 \pm 0.01}$	$\mathbf{0.34 \pm 0.01}$	0.14 ± 0.01	0.15 ± 0.01
Diabetes	0.31 ± 0.02	0.34 ± 0.01	$\mathbf{0.30 \pm 0.02}$	$\mathbf{0.31 \pm 0.02}$	0.36 ± 0.02	0.38 ± 0.02
Digit1	0.42 ± 0.02	$\mathbf{0.35 \pm 0.02}$	$\mathbf{0.20 \pm 0.01}$	$\mathbf{0.38 \pm 0.01}$	0.06 ± 0.00	0.06 ± 0.01
g241c	0.46 ± 0.01	$\mathbf{0.39 \pm 0.01}$	$\mathbf{0.28 \pm 0.01}$	$\mathbf{0.42 \pm 0.02}$	0.22 ± 0.01	$\mathbf{0.21 \pm 0.01}$
g241d	0.44 ± 0.02	$\mathbf{0.38 \pm 0.01}$	$\mathbf{0.29 \pm 0.01}$	$\mathbf{0.41 \pm 0.02}$	0.23 ± 0.01	$\mathbf{0.22 \pm 0.01}$
Haberman	0.29 ± 0.02	0.28 ± 0.02	0.29 ± 0.02	0.29 ± 0.02	0.29 ± 0.02	0.31 ± 0.03
Ionosphere	0.28 ± 0.03	$\mathbf{0.24 \pm 0.01}$	$\mathbf{0.19 \pm 0.02}$	$\mathbf{0.22 \pm 0.03}$	0.20 ± 0.02	0.19 ± 0.02
Mammography	0.30 ± 0.03	0.30 ± 0.02	$\mathbf{0.29 \pm 0.03}$	0.30 ± 0.03	0.30 ± 0.03	0.28 ± 0.02
Parkinsons	0.25 ± 0.02	0.23 ± 0.03	0.24 ± 0.03	0.25 ± 0.03	0.22 ± 0.02	0.23 ± 0.02
Sonar	0.44 ± 0.04	$\mathbf{0.38 \pm 0.04}$	$\mathbf{0.33 \pm 0.02}$	$\mathbf{0.39 \pm 0.02}$	0.26 ± 0.02	0.33 ± 0.03
SPECT	0.39 ± 0.04	0.38 ± 0.02	$\mathbf{0.33 \pm 0.03}$	0.39 ± 0.03	0.25 ± 0.03	$\mathbf{0.20 \pm 0.02}$
SPECTF	0.44 ± 0.03	$\mathbf{0.40 \pm 0.04}$	$\mathbf{0.36 \pm 0.03}$	$\mathbf{0.42 \pm 0.03}$	0.25 ± 0.02	$\mathbf{0.21 \pm 0.01}$
Transfusion	0.26 ± 0.02	0.28 ± 0.03	0.26 ± 0.02	0.26 ± 0.02	0.27 ± 0.01	0.28 ± 0.02
USPS	0.42 ± 0.02	$\mathbf{0.34 \pm 0.02}$	$\mathbf{0.20 \pm 0.01}$	$\mathbf{0.38 \pm 0.02}$	0.11 ± 0.01	$\mathbf{0.10 \pm 0.00}$
WDBC	0.09 ± 0.01	0.13 ± 0.03	$\mathbf{0.08 \pm 0.01}$	0.09 ± 0.01	0.10 ± 0.01	0.11 ± 0.02
Total		9 / 5 / 2	13 / 3 / 0	10 / 6 / 0		5 / 8 / 3

Bold respectively underlined values indicate whether the performance of a semi-supervised solution is significantly better or worse than the supervised alternative as evaluated by a one-sided Wilcoxon signed rank test with family wise error rate of 0.05. The Win/Draw/Loss indicates on how many datasets a semi-supervised learner performs significantly better, equal or worse than the supervised alternative

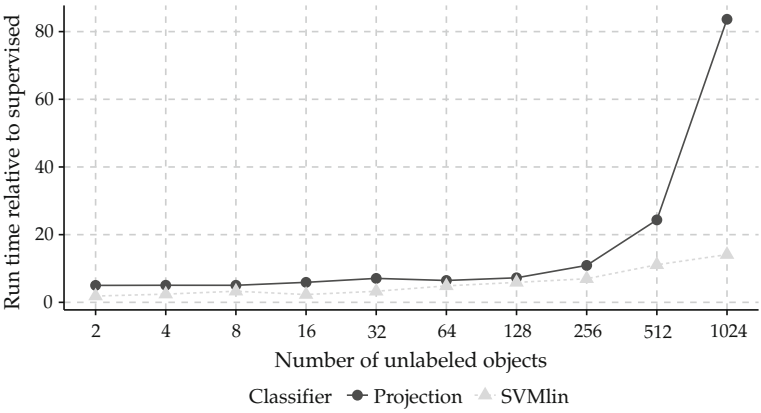

Fig. 3 Training time of semi-supervised methods relative to the training time of the supervised classifier for increasing amounts of unlabeled data on a simulated dataset with 2 Gaussian classes with 100 features and $N_l = 200$

6.4 Computational considerations

Since the projection proposed in Eq. (6) can be formulated as a quadratic programming problem with a positive definite matrix, the worst case complexity is $O(N_u^3)$. Comparing this to Transductive SVM solvers, for instance, the CCCP procedure for Transductive SVMs by Collobert et al. (2006) has a worst case complexity of $O((N_l + 2N_u)^3)$. The SVMlin implementation of Sindhwani and Keerthi (2006) we compare to here makes few claims about its theoretical complexity. As Collobert et al. (2006) and Sindhwani and Keerthi (2006) note, however, the practical complexity is often much lower than the worst case complexity and should be evaluated empirically. Figure 3 shows the computational time relative to the supervised classifier as we increase the number of unlabeled samples for our implementation of the Projection estimator and for the L_2-TSVM implementation of Sindhwani and Keerthi (2006). The figure shows that the SVMlin implementation scales much better as the number of unlabeled examples is increased. SVMlin's solution does not, however, guarantee that the solution is a global optimum, as the projection approach does, or guarantee any safe improvements over supervised learning. Whereas the explicit goal of SVMlin is to scale TSVM to larger datasets, we have not attempted to more efficiently solve the quadratic programming problem posed by our approach and leave this as an open problem.

7 Discussion

The main result of this work is summarized in Theorem 1 and illustrated in Figure 1: the proposed semi-supervised classifier is guaranteed to improve over the supervised classifier in terms of the quadratic loss on all training data, labeled and unlabeled. The results from the experiments indicate that on average, both the projected estimator and other semi-supervised approaches often show improved performance, while on individual samples from the datasets, the projected estimator never reduces performance in terms of the surrogate loss. This is an important property since, in practical settings, one only has a single sample (i.e. dataset) from

a classification problem, and it is important to know that performance will not be degraded when applying a semi-supervised version of a supervised procedure on that particular dataset. Even if we do not have enough labeled objects to accurately estimate this performance, Theorem 1 guarantees we will not perform worse than the supervised alternative on the labeled and unlabeled data in terms of the surrogate loss.

7.1 Surrogate loss

Theorem 1 is limited to showing improvement in terms of quadratic loss. As the experiments also indicate, good properties in terms of this loss do not necessarily translate into good properties in terms of the error rate. In the empirical risk minimization framework, however, classifiers are constructed by minimizing surrogate losses. This particular semi-supervised learner is effective in terms of this objective. In this sense, it can be considered a proper semi-supervised version of the supervised quadratic loss minimizer.

One could question whether the quadratic loss is a good choice as surrogate loss (Ben-David et al. 2012). In practice, however, it can perform very well and is often on par and sometimes better than, for instance, an SVM employing hinge loss (Rasmussen and Williams 2005; Hastie et al. 2009; Poggio and Smale 2003). Moreover, the main result in this work basically demonstrates that strong improvement guarantees are at all possible for *some* surrogate loss function. Whether and when an increase in performance in terms of this surrogate loss translates into improved classification accuracy is, like in the supervised setting, unclear. Much work is currently being done to understand the relationship between surrogate losses and $0-1$ loss (Bartlett et al. 2006; Ben-David et al. 2012).

7.2 Conservatism

Arguably, a robust semi-supervised learning procedure could also be arrived at by very conservatively setting the parameters controlling the influence of unlabeled data in semi-supervised learner procedures such as the TSVM. There are two reasons why this is difficult to achieve in practice. The first reason is a computational one. Most semi-supervised procedures are computationally intensive. Doing a grid search over both a regularization parameter as well as the parameter controlling the influence of the unlabeled objects using cross-validation is time-consuming. Secondly, and perhaps more importantly, it may be very difficult to choose a good parameter using limited labeled data. Goldberg and Zhu (2009) study this problem in more detail. While their conclusion suggests otherwise, their results indicate that performance degradation occurs on a significant number of datasets.

The projected estimator presented here tries to alleviate these problems in two ways. Firstly, unlike many semi-supervised procedures, it can be formulated as a quadratic programming problem in terms of the unknown labels which has a global optimum (which is unique in terms of $\mathbf{w}$) and there are no hyper-parameters involved. Secondly, at least in terms of its surrogate loss, there is a guarantee performance will not be worse than the alternative of discarding the unlabeled data.

As our results indicate, however, the proposed procedure is very conservative. The projection with $\mathbf{X}_\circ = \mathbf{X}$ (ICLS) is a classifier which is less conservative than the projection based on all data, and offers larger improvement in the experiments while still being robust to degradation of performance. For this procedure Theorem 1 does not hold. Better understanding in what way we can still prove other robustness properties for this classifier is an open issue.

An alternative way to derive less conservative approaches could be by changing the constraint set Θ. The purpose of this work has been to show that if we choose Θ conservatively, such that we can guarantee it contains the oracle solution $\mathbf{w}_{\text{oracle}}$, we can guarantee non-degradation, while still allowing for improved performance over the supervised solution in many cases. To construct a method with wider applicability, an interesting question is how to restrict Θ based on additional assumptions, while ensuring that $\mathbf{w}_{\text{oracle}} \in \Theta$ with high probability.

7.3 Other losses

Another open question is for what other losses we can apply the projection procedure presented here. Apart from the issue of defining the metric in these cases, for some other loss functions the current definition of the constraint set might not constrain the parameter space at all. In the case of hinge loss or logistic loss, empirical results seem to indicate that the constraint set Θ always includes $\mathbf{w}_{\text{sup}}$. The lack of a closed-form solution, however, hampers a detailed theoretical analysis of these settings. Therefore, an exact characterization of the kinds of losses for which the procedure is amenable has to be left as an open problem.

8 Conclusion

We introduced and analyzed an approach to semi-supervised learning with quadratic surrogate loss that has the interesting theoretical property of never decreasing performance when measured on the full, labeled and unlabeled, training set in terms of this surrogate loss when compared to the supervised classifier. This is achieved by projecting the solution vector of the supervised least squares classifier onto a constraint set of solutions defined by the unlabeled data. As we have illustrated through simulation experiments, the safe improvements in terms of the surrogate loss on the labeled and unlabeled data also partially translate into safe improvements in terms of the classification errors on an unseen test set. Moreover, the procedure can be formulated as a standard quadratic programming problem, leading to a simple optimization procedure. An open problem is how to apply this procedure or a procedure with similar theoretical performance guarantees, to other loss functions.

References

Aubin, J. P. (2000). *Applied functional analysis* (2nd ed.). Hoboken: Wiley. ISBN 9780471179764.

Bartlett, Peter L., Jordan, Michael I., & McAuliffe, Jon D. (2006). Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, *101*(473), 138–156. ISSN 0162-1459.

Ben-David, S., Loker, D., Srebro, N., & Sridharan, K. (2012). Minimizing the misclassification error rate using a surrogate convex loss. In *Proceedings of the 29th international conference on machine learning*, (pp. 1863–1870).

Chapelle, Olivier, Schölkopf, Bernhard, & Zien, Alexander. (2006). *Semi-supervised learning*. Cambridge: MIT press. ISBN 9780262033589.

Collobert, Ronan, Sinz, Fabian, Weston, Jason, & Bottou, Leon. (2006). Large scale transductive SVMs. *Journal of Machine Learning Research*, *7*, 1687–1712.

Cozman, F., & Cohen, I. (2006). Risks of semi-supervised learning. In O. Chapelle, B. Schölkopf, & A. Zien (Eds.), *Semi-supervised learning, chapter 4* (pp. 56–72). Cambridge: MIT press.

Goldberg, A. B., & Zhu, X. (2009). Keepin'it real: semi-supervised learning with realistic tuning. *NAACL HLT 2009 Workshop on semi-supervised learning for natural language processing*.

Grandvalet, Y., & Bengio, Y. (2005). Semi-supervised learning by entropy minimization. In L. K. Saul, Y. Weiss, & L. Bottou (Eds.), *Advances in neural information processing systems 17* (pp. 529–536). Cambridge: MIT Press.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning* (2nd ed.). New York: Spinger. ISBN 0387848576.

Joachims, T. (1999). Transductive inference for text classification using support vector machines. In *Proceedings of the 16th international conference on machine learning*, (pp. 200–209). Burlington: Morgan Kaufmann Publishers.

Krijthe, J. H. & Loog, M. (2015). Implicitly constrained semi-supervised least squares classification. In Fromont, E., De Bie, T., & van Leeuwen, M. (Eds.), *14th international symposium on advances in intelligent data analysis XIV (Lecture Notes in Computer Science Volume 9385)*, (pp. 158–169), Saint Étienne. France.

Lafferty, John D., & Wasserman, Larry. (2007). Statistical analysis of semi-supervised regression. *In Advances in Neural Information Processing Systems*, *20*, 801–808.

Lawrence, N. D., & Jordan, M. I. (2004). Semi-supervised learning via Gaussian processes. In *Advances in Neural Information Processing Systems*, (pp. 753–760).

Li, YF., & Zhou, ZH. (2011). Towards making unlabeled data never hurt. In *Proceedings of the 28th international conference on machine learning*, (pp. 1081–1088).

Lichman, M. (2013). UCI Machine Learning Repository. URL http://archive.ics.uci.edu/ml.

Loog, M. (2010). Constrained parameter estimation for semi-supervised learning: the case of the nearest mean classifier. In *Proceedings of the 2010 European conference on machine learning and knowledge discovery in databases*, (pp. 291–304).

Loog, Marco. (2014). Semi-supervised linear discriminant analysis through moment-constraint parameter estimation. *Pattern Recognition Letters*, *37*, 24–31.

Loog, Marco. (2016). Contrastive pessimistic likelihood estimation for semi-supervised classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *38*(3), 462–475.

McLachlan, G. J. (1975). Iterative reclassification procedure for constructing an asymptotically optimal rule of allocation in discriminant analysis. *Journal of the American Statistical Association*, *70*(350), 365–369.

Nigam, Kamal, McCallum, Andrew, K., Thrun, Sebastian, & Mitchell, Tom. (2000). Text classification from labeled and unlabeled documents using EM. *Machine Learning*, *34*, 1–34.

Niyogi, Partha. (2013). Manifold regularization and semi-supervised Learning : Some theoretical analyses. *Journal of Machine Learning Research*, *14*, 1229–1250.

Opper, M., & Kinzel, W. (1996). Statistical mechanics of generalization. In E. Domany, J. L. Hemmen, & K. Schulten (Eds.), *Models of neural networks III* (pp. 151–209). New York: Springer.

Poggio, T., & Smale, S. (2003). The mathematics of learning: Dealing with data. *Notices of the AMS*, *50*(5), 537–544.

Rasmussen, C. E., & Williams, C. K. I. (2005). *Gaussian processes for machine learning*. The MIT Press.

Raudys, Sarunas, & Duin, Robert P. W. (1998). Expected classification error of the Fisher linear classifier with pseudo-inverse covariance matrix. *Pattern Recognition Letters*, *19*(5–6), 385–392.

Schuurmans, Dale, & Southey, Finnegan. (2002). Metric-based methods for adaptive model selection and regularization. *Machine Learning*, *48*, 51–84.

Sindhwani, V., & Keerthi, S. S. (2006). Large scale semi-supervised linear SVMs. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, (p. 477), New York, ACM Press. ISBN 1595933697.

Singh, A., Nowak, R., & Zhu, X. (2008). Unlabeled data: Now it helps, now it doesn't. In *Advances in Neural Information Processing Systems*, (pp. 1513–1520).

Zhu, X., Ghahramani, Z., & Lafferty, J. (2003). Semi-supervised learning using gaussian fields and harmonic functions. In *Proceedings of the 20th international conference on machine learning*, (pp. 912–919).