SPARSE QUASI-NEWTON OPTIMIZATION FOR SEMI-SUPERVISED SUPPORT VECTOR MACHINES

Fabian Gieseke\textsuperscript{1}, Antti Airola\textsuperscript{2}, Tapio Pahikkala\textsuperscript{2}, and Oliver Kramer\textsuperscript{1}

\textsuperscript{1}Computer Science Department, Carl von Ossietzky Universität Oldenburg, 26111 Oldenburg, Germany
\textsuperscript{2}Turku Centre for Computer Science, Department of Information Technology, University of Turku, 20520 Turku, Finland
\{f.gieseke,oliver.kramer\}@uni-oldenburg.de, \{antti.airola,tapio.pahikkala\}@utu.fi

Keywords: Semi-Supervised Support Vector Machines, Non-Convex Optimization, Quasi-Newton Methods, Sparse Data, Nyström Approximation.

Abstract: In real-world scenarios, labeled data is often rare while unlabeled data can be obtained in huge quantities. A current research direction in machine learning is the concept of semi-supervised support vector machines. This type of binary classification approach aims at taking the additional information provided by unlabeled patterns into account to reveal more information about the structure of the data and, hence, to yield models with a better classification performance. However, generating these semi-supervised models requires solving difficult optimization tasks. In this work, we present a simple but effective approach to address the induced optimization task, which is based on a special instance of the quasi-Newton family of optimization schemes. The resulting framework can be implemented easily using black box optimization engines and yields excellent classification and runtime results on both artificial and real-world data sets that are superior (or at least competitive) to the ones obtained by competing state-of-the-art methods.

1 INTRODUCTION

One of the most important machine learning tasks is classification. If sufficient labeled training data is given, there exists a variety of techniques like the $k$-nearest neighbor-classifier or support vector machines (SVMs) (Hastie et al., 2009; Steinwart and Christmann, 2008) to address such a task. However, labeled data is often rare in real-world applications. One active research field in machine learning is semi-supervised learning (Chapelle et al., 2006b; Zhu and Goldberg, 2009). In contrast to supervised methods, the latter class of techniques takes both labeled and unlabeled data into account to construct appropriate models. A well-known concept in this field are semi-supervised support vector machines ($S^3$VMs) (Bennett and Demiriz, 1999; Joachims, 1999; Vapnik and Sterin, 1977) which depict the direct extension of support vector machines to semi-supervised learning scenarios. The key idea is depicted in Figure 1: The aim of a standard support vector machine consists in finding a hyperplane which separates both classes well such that the margin is maximized. It is obvious that, in case of lack of labeled data, suboptimal models might be obtained. Its semi-supervised variant aims at taking the unlabeled patterns into account by searching for a partition (of these patterns into two classes) such that a subsequent application of a (modified) support vector machine leads to the best result.

1.1 Related Work

The original problem formulation of semi-supervised support vector machines was given by Vapnik and Sterin (Vapnik and Sterin, 1977) under the name of transductive support vector machines. From an optimization point of view, the first approaches have been proposed in the late nineties by Joachims (Joachims, 1999) and Bennett and Demiriz (Bennett and Demiriz, 1999). In general, there exist two lines of research, namely (a) combinatorial and (b) continuous optimization schemes. The naive brute-force approach (which tests every possible partition), for instance, is among the combinatorial schemes since it aims at directly finding a good assignment for the unknown labels. The continuous optimization perspective (see below) leads to a real-valued but non-convex task. For both research directions, a variety of different techniques has been proposed in recent years that are based on semi-definite programming (Bie and Cristianini, 2004; Xu and Schuurmans, 2005), the continuation method (Chapelle et al., 2006a), deterministic
annealing (Sindhwani et al., 2006), the (constrained) concave-convex procedure (Collobert et al., 2006; Fung and Mangasarian, 2001; Zhao et al., 2008), and other strategies (Adankon et al., 2009; Chapelle and Zien, 2005; Mierswa, 2009; Sindhwani and Keerthi, 2006; Zhang et al., 2009). A related approach, also based on a quasi-Newton framework, is proposed by Reddy et al. (Reddy et al., 2010); however, they do not consider differentiable surrogate functions and therefore apply more complicated subgradient methods. Many other approaches exist, and we refer to Chapelle et al. (Chapelle et al., 2006b; Chapelle et al., 2008) and Zhu et al. (Zhu and Goldberg, 2009) for comprehensive surveys.

1.2 Notations

We use \([m]\) to denote the set \(\{1, \ldots, m\}\). Given a vector \(y \in \mathbb{R}^n\), we use \(y_i\) to denote its \(i\)-th coordinate. Further, the set of all \(m \times n\) matrices with real coefficients is denoted by \(\mathbb{R}^{m \times n}\). Given a matrix \(M \in \mathbb{R}^{m \times n}\), we denote the element in the \(i\)-th row and \(j\)-th column by \([M]_{i,j}\). For two sets \(R = \{i_1, \ldots, i_r\} \subseteq [m]\) and \(S = \{k_1, \ldots, k_s\} \subseteq [n]\) of indices, we use \(M_{RS}\) to denote the matrix that contains only the rows and columns of \(M\) that are indexed by \(R\) and \(S\), respectively. Moreover, we set \(M_{R[n]} = M_R\).

2 CLASSIFICATION TASK

In supervised scenarios, we are given a training set \(T_l = \{(x_i, y_i^l) \mid x_i \in \mathcal{X}\}\) of labeled patterns \(x_i\) belonging to a set \(\mathcal{X}\). The general goal of classification approaches consists in building good models which can predict valuable labels for unseen patterns (Hastie et al., 2009). In semi-supervised learning frameworks, we are additionally given a set \(T_u = \{x_i \mid x_i \in \mathcal{X}\}\) of unlabeled training patterns. Here, the goal is to improve the quality of the models by taking both the labeled and the unlabeled part of the data into account.

2.1 Support Vector Machines

The concept of support vector machines can be seen as instance of regularization problems of the form

\[
\min_{f \in \mathcal{H}} \frac{1}{L} \sum_{i=1}^{L} L(y_i, f(x_i)) + \lambda ||f||_2^2, \tag{1}
\]

where \(\lambda > 0\) is a fixed real number, \(L : Y \times \mathbb{R} \to [0, \infty)\) is a loss function and \(||f||_2^2\) is the squared norm in a so-called reproducing kernel Hilbert space \(\mathcal{H} \subseteq \mathbb{R}^k = \{f : X \to \mathbb{R}\}\) induced by a kernel function \(k : X \times X \to \mathbb{R}\) (Steinwart and Christmann, 2008). Here, the first term measures the loss caused by the prediction function on the labeled training set and the second one penalizes complex functions. Plugging in different loss functions leads to various models; one of the most popular choices is the hinge loss \(L(y, f) = \max(0, 1 - yf)\) which leads to the original definition of support vector machines (Schölkopf et al., 2001; Steinwart and Christmann, 2008), see Figure 2 (a).

2.2 Semi-Supervised SVMs

Given the additional set \(T_u = \{x_{i+1}, \ldots, x_{i+u}\} \subseteq \mathcal{X}\) of unlabeled training patterns, semi-supervised support vector machines (Bennett and Demiriz, 1999; Joachims, 1999; Vapnik and Sterin, 1977) aim at finding an optimal prediction function for unseen data based on both the labeled and the unlabeled part of the data. More precisely, we search for a function \(f^* \in \mathcal{H}\) and a labeling vector \(y^* = (y_1^*, \ldots, y_u^*)^T \in \{-1, +1\}^u\) that are optimal with respect to

\[
\begin{align*}
\min_{f \in \mathcal{H}, y \in \{-1, +1\}^u} & \frac{1}{L} \sum_{i=1}^{L} L(y_i, f(x_i)) + \lambda ||f||_2^2, \\
& + \lambda' \frac{1}{u} \sum_{i=1}^{u} L(y_i, f(x_{i+1})) + \lambda ||f||_2^2
\end{align*} \tag{2}
\]

where \(\lambda', \lambda > 0\) are user-defined parameters. Thus, the main task consists in finding the optimal assignment vector \(y^*\) for the unlabeled part; the combinatorial nature of this task renders the optimization problem difficult to solve.

2.3 Continuous Optimization

As mentioned above, one can derive an equivalent continuous optimization task: Using the hinge loss,
the optimal assignments for the vector \( y \) for a fixed \( f \in \mathcal{H} \) are given by \( y_i = \text{sgn}(f(x_i)) \) (Chapelle and Zien, 2005).\(^2\) This yields

\[
\text{minimize}_{f \in \mathcal{H}} \frac{1}{l} \sum_{i=1}^{l} \max \left( 0, 1 - y_i f(x_i) \right) + \frac{\lambda'}{u} \sum_{i=1}^{u} \max \left( 0, 1 - |f(x_{l+i})| \right) + \lambda |f|_{2\frac{1}{2}}.
\]

Note that the effective loss on the unlabeled patterns penalizes predictions around the origin; thus, the overall loss increases if the decision function \( f \) passes through these patterns, see Figure 2 (b). By applying the representer theorem (Schölkopf et al., 2001) for latter task, it follows that an optimal solution \( f \in \mathcal{H} \) is of the form

\[
f(\cdot) = \sum_{i=1}^{l} c_i k(x_{l+i}, \cdot) + \sum_{j=1}^{u} c_j k(x_{l+u+j}, \cdot)
\]

with coefficients \( c = (c_1', \ldots, c_l', c_1, \ldots, c_u)^T \in \mathbb{R}^{l+u} \). Thus, one gets a continuous optimization task which consists in finding the optimal coefficient vector \( c \in \mathbb{R}^n \) with \( n = l + u \). Note that the effective loss renders the task non-convex and non-differentiable (since the partial functions are non-differentiable).

### 3 QUASI-NEWTON SCHEME

We will consider a special instance of the quasi-Newton optimization framework (Nocedal and Wright, 2000). Besides the objective function itself, methods belonging to this class of schemes only require the gradient to be supplied.

\(^2\)Note that the latter observation does only hold without a balance constraint of the form

\[
\frac{1}{n} \sum_{i=1}^{n} \max(0, y_i) - b_c < \epsilon
\]

for small \( \epsilon > 0 \) and \( b_c \in (0, 1) \).

#### 3.1 Differentiable Surrogates

Aiming at the application of the latter type of schemes, we introduce the following differentiable surrogate loss functions depicted in Figure 2. Here, the differentiable replacement for the hinge loss is the modified logistic loss (Zhang and Oles, 2001); the replacement for the effective loss for the unlabeled part is a well-known candidate in this field (Chapelle and Zien, 2005). Thus, the new overall surrogate objective is given by

\[
F_{l,t}(c) = \frac{1}{l} \sum_{i=1}^{l} \log \left( 1 + \exp \left( \gamma \left( 1 - y_i f(x_i) \right) \right) \right) + \frac{\lambda'}{u} \sum_{i=1}^{u} \exp \left( -3(f(x_{l+i}))^2 \right) + \lambda \sum_{i=1}^{n} \sum_{j=1}^{n} c_j c_k k(x_i, x_j)
\]

with \( f(\cdot) = \sum_{p=1}^{n} c_p k(x_p, \cdot) \) and using \( ||f||_{\mathcal{H}}^2 = \sum_{p=1}^{n} \sum_{q=1}^{n} c_p c_q k(x_p, x_q) \) (Schölkopf et al., 2001). The next lemma shows that both a function and a gradient call can be performed efficiently.

**Lemma 1.** For a given \( c \in \mathbb{R}^n \), one can compute both the objective \( F_{l,t}(c) \) and the gradient \( \nabla F_{l,t}(c) \) in \( O(n^2) \) time. The overall space consumption is \( O(n^2) \).

**Proof.** The gradient is given by

\[
\nabla F_{l,t}(c) = Ka + 2\lambda Kc
\]

with \( a \in \mathbb{R}^n \) and

\[
a_i = \begin{cases} 
- \frac{1}{l} \frac{\exp(\gamma(1-f(x_i)y_i))}{1+\exp(\gamma(1-f(x_i)y_i))} \cdot y_i & \text{for } i \leq l \\
- 6\lambda' \frac{\exp(-3(f(x_i))^2)}{u} \cdot f(x_i) & \text{for } i > l
\end{cases}
\]

Since all predictions \( f(x_1), \ldots, f(x_n) \) can be computed in \( O(n^2) \) total time, one can compute the vector \( a \in \mathbb{R}^n \) and therefore the objective and the gradient in \( O(n^2) \). The space requirements are dominated by the kernel matrix \( K \in \mathbb{R}^{n \times n} \).

Note that numerical instabilities might occur when evaluating \( \exp(\gamma(1-f(x_i)y_i)) \) for a function or a gradient call. However, one can deal with these degeneracies in a safe way since \( \log(1 + \exp(t)) - t \to 0 \) and \( \frac{\exp(t)}{1+\exp(t)} - 1 \to 0 \) converge rapidly for \( t \to \infty \). Thus, each function and gradient evaluation can be performed spending \( O(n^2) \) time in a numerically stable manner.
Algorithm 1 QN-S³VM

Require: A labeled training set $T_l = \{(x_1, y_l^1), \ldots, (x_t, y_l^t)\}$, an unlabeled training set $T_u = \{x_{t+1}, \ldots, x_u\}$, model parameters $\lambda', \lambda$, an initial (positive definite) inverse Hessian approximation $H_0$, and a sequence $0 < \alpha_1 < \ldots < \alpha_t$.

1: Initialize $c_0$ via supervised model.
2: for $i = 1$ to $\tau$ do
3: $k = 0$
4: while termination criteria not fulfilled do
5: Compute search direction $p_k$ via (7)
6: Update $c_{k+1} = c_k + \beta_k p_k$
7: Update $H_{k+1}$ via (8)
8: $k = k + 1$
9: end while
10: $c_0 = c_k$
11: end for

3.2 Quasi-Newton Framework

One of the most popular quasi-Newton schemes is the Broyden-Fletcher-Goldfarb-Shanno (BFGS) (Nocedal and Wright, 2000) method, which we will now sketch in the context of the given task. The overall algorithmic framework is given in Algorithm 1: The initial candidate solution is obtained via Equation (5) while ignoring the (non-convex) unlabeled part (i.e., $\lambda' = 0$). The influence of the unlabeled part is then increased gradually via the sequence $\alpha_1, \ldots, \alpha_t$.

For each parameter $\alpha_i$, a standard BFGS optimization phase is performed, i.e., a sequence $c_{k+1} = c_k + \beta_k p_k$ of candidate solutions is generated, where $p_k$ is computed via

$$p_k = -H_k \nabla F_{\alpha_i, \lambda'}(c_k)$$

and where the step length $\beta_k$ is computed via line search. The approximation $H_k$ of the inverse Hessian is then updated via

$$H_{k+1} = (I - \rho_k s_k z_k^T) H_k (I - \rho_k z_k s_k^T) + \rho_k s_k s_k^T$$

with $z_k = \nabla F_{\alpha_i, \lambda'}(c_{k+1}) - \nabla F_{\alpha_i, \lambda'}(c_k)$, $s_k = c_{k+1} - c_k$, and $\rho_k = (z_k^T z_k)^{-1}$. New candidate solutions are generated as long as a convergence criterion is fulfilled (e.g., as long as $\|\nabla F_{\alpha_i, \lambda'}(c_k)\| > \epsilon$ is fulfilled for a small $\epsilon > 0$ or as long as the number of iterations is smaller than a used-defined number). As initial approximation, one usually resorts to $H_0 = \gamma I$ for $\gamma > 0$; an important property of the update scheme is that it preserves the positive definiteness of the inverse Hessian approximations (Nocedal and Wright, 2000).

3.3 Computational Speed-Ups

Two main computational bottlenecks arise for large-scale settings: Firstly, the recurrent computation of the objective and gradient needed by the quasi-Newton framework is cumbersome. Secondly, the approximation of the Hessian’s inverse is, in general, not sparse, which leads to quadratic-time operations for the quasi-Newton framework itself (Nocedal and Wright, 2000). In the following, we will show how to alleviate these two problems.

3.3.1 Limited Memory Quasi-Newton

The non-sparse approximation of the Hessian’s inverse leads to a $O(n^3)$ time and to a $O(n^2)$ space consumption. To reduce these computational costs, we consider the L-BFGS methods (Nocedal and Wright, 2000), which depicts a memory and time saving variant of the original BFGS scheme. In a nutshell, the idea consists in generating the approximations $H_0, H_1, \ldots$ only based on the last $m \ll n$ iterations and to perform low-rank updates on the fly without storing the involved matrices explicitly. This leads to an update time of $O(mn)$ for all operations related to the intermediate optimization phases (not counting the time for function and gradient calls). As pointed out by Nocedal and Wright, small values for $m$ are usually sufficient in practice (ranging from, e.g., $m = 3$ to $m = 50$). Thus, assuming $m$ to be a relatively small constant, the operations needed by the optimization engine essentially scale linearly with the number $n$ of optimization variables.

3.3.2 Low-Dimensional Search Space

It remains to show how to reduce the second bottleneck, i.e., the recurrent computation of both the objective and the gradient. For this sake, one can resort to the subset of regressors method (Rifkin, 2002) to reduce these computational costs, i.e., one approximates the original hypothesis $f(\cdot)$ via

$$\hat{f}(\cdot) = \sum_{k=1}^{r} \tilde{c}_{jk} k(x_k, \cdot),$$

where $R = \{j_1, \ldots, j_r\} \subset \{1, \ldots, n\}$ is a subset of indices. Using this approximation scheme leads to a slightly modified objective $\hat{F}(\tilde{c})$ for $\tilde{c} \in \mathbb{R}^r$, where
the predictions \( f(x_1), \ldots, f(x_n) \) are replaced by their corresponding approximations \( \hat{f}(x_1), \ldots, \hat{f}(x_n) \) in the objective (5). Similar derivations as for the non- approximation case show that the gradient \( \nabla F_{\hat{c}}(\hat{c}) \) is then given as

\[
\nabla F_{\hat{c}}(\hat{c}) = K_R \alpha + 2\lambda K_{RR} \hat{c},
\]

(10)

where \( f \) has to be replaced by \( \hat{f} \) in the former definition of the vector \( \alpha \in \mathbb{R}^n \). It is easy to see that one can compute both the new objective as well as its gradient in an efficient kind of way:

**Lemma 2.** For \( \hat{c} \in \mathbb{R}^p \), the approximated objective \( \hat{F}_A(\hat{c}) \) and the gradient \( \nabla \hat{F}_A(\hat{c}) \) can be computed in \( O(nr) \) time spending \( O(nr) \) space.

**Proof.** All predictions \( \hat{f}(x_1), \ldots, \hat{f}(x_n) \) can be predicted in \( O(nr) \) time for \( \alpha \in \mathbb{R}^p \). Given these predictions, one can compute the modified vector \( \alpha \in \mathbb{R}^n \) in \( O(n) \) time. The remaining operations for obtaining the new objective \( \hat{F}_A(\hat{c}) \) and its gradient \( \nabla \hat{F}_A(\hat{c}) \) can be performed in \( O(nr + r^2) = O(nr) \) time. The space consumption, dominated by \( K_R \), is \( O(nr) \). \( \square \)

Thus, in combination with the L-BFGS scheme depicted above, both the runtime as well as the space consumption are reduced significantly. Here, the parameter \( r \in \{1, \ldots, n\} \) determines a trade-off between the achieved speed-up and the accuracy of the approximation. Another way to obtain considerable speed-ups can be achieved for the special case of a linear kernel, which we will describe next.

### 3.3.3 Linear Kernel and Sparse Data

Assume that we are given patterns in \( X = \mathbb{R}^d \) and let \( X \in \mathbb{R}^{n \times d} \) denote the data matrix containing the training patterns as rows. In case of the linear kernel, one can write the kernel matrix as \( K = XX^T \in \mathbb{R}^{n \times n} \) and can achieve substantial computational savings by avoiding its explicit construction. This is the case, for instance, if the data resides in a low-dimensional feature space (i.e., \( d \ll n \)) or due to the data matrix being sparse, meaning that it contains only few nonzero entries.

**Lemma 3.** For a linear kernel with patterns in \( X = \mathbb{R}^d \), one can compute the objective \( F_A(c) \) and the gradient \( \nabla F_A(c) \) in \( O(nd) \) time using \( O(s) \) space for a given candidate solution \( c \in \mathbb{R}^n \).

**Proof.** Due to the linear kernel, one can compute

\[
Kc = X(X^Tc)
\]

(11)

and thus all predictions \( f(x_1), \ldots, f(x_n) \) in \( O(nd) \) time. In the same manner, one can obtain \( c^TKc \) and \( Ka \) in \( O(nd) \) time (where the vector \( \alpha \in \mathbb{R}^n \) can be computed in \( O(n) \) time given the predictions). Thus, both the objective \( F_A(c) \) and the gradient \( \nabla F_A(c) \) can be obtained in \( O(nd) \) time. The space requirements are bounded by the space needed to store the data matrix \( X \in \mathbb{R}^{n \times d} \), which is \( O(nd) \).

For high-dimensional but sparse data (i.e., if the matrix \( X \in \mathbb{R}^{n \times d} \) contains only \( s \ll nd \) nonzero entries), one can further reduce the computational cost in the following kind of way:

**Lemma 4.** For a linear kernel with patterns in \( X = \mathbb{R}^d \) and data matrix \( X \in \mathbb{R}^{n \times d} \) with \( s \ll nd \) nonzero entries, one can compute the objective \( F_A(c) \) and the gradient \( \nabla F_A(c) \) in \( O(s) \) time using \( O(s) \) space for a given candidate solution \( c \in \mathbb{R}^n \).

**Proof.** Without loss of generality, we assume that \( s \geq n - 1 \) holds. Similar to the derivations above, one can compute \( Kc = X(X^Tc) \) and therefore the predictions \( f(x_1), \ldots, f(x_n) \) as well as \( \alpha \in \mathbb{R}^n \) in \( O(s) \) time using standard sparse matrix multiplication techniques. In the same way, one can compute \( \alpha^TKc \) and \( Ka \) in \( O(s) \) time. Hence, both the objective \( F_A(c) \) and the gradient \( \nabla F_A(c) \) can be obtained in \( O(s) \) time spending \( O(s) \) space.

\[
\square
\]

### 4 EXPERIMENTS

We will now describe the experimental setup and the outcome of our experimental evaluation.

#### 4.1 Experimental Setup

The runtime analysis is performed on a 3 GHZ Intel CoreTM Duo PC running Ubuntu 10.04. We start by providing details related to the experimental setup.

##### 4.1.1 Implementation Details

Our implementation is based on Python, the Scipy package (using the L-BFGS implementation optimize.fmin_l_bfgs_b (Byrd et al., 1995) with \( m = 50 \)), and the Numpy package. The function and gradient evaluations are based on efficient matrix operations provided by the Numpy package. As pointed out above, a direct implementation of the latter ones might suffer from numerical instabilities. To avoid these instabilities, we make use of \( \log(1 + \exp(t)) \approx t \) and \( \frac{\exp(t)}{\exp(t)} \approx 1 \) for \( t \geq 500 \). We denote the resulting implementation by \( QN-S^3\text{VM} \).

\( ^4 \text{Note that the term } s \text{ is sometimes used to denote the average number of nonzero entries per pattern } x \in X \text{ in the training set.} \)
4.1.2 Data Sets

We consider several artificial and real-world data sets, where the first half of each data set is used as training and the second half as test set. To induce semi-supervised scenarios, we split each training set into a labeled and an unlabeled part and use different ratios for the particular setting (where \(l, u, t\) denotes the number of labeled, unlabeled, and test patterns, respectively).

**Artificial Data Sets:** The first artificial data set is composed of two Gaussian clusters; to generate it, we draw \(n/2\) points from each of two multivariate Gaussian distributions \(X_i \sim \mathcal{N}(\mathbf{m}_i, \mathbf{I})\), where \(\mathbf{m}_1 = (-2.5, 0.0, \ldots, 0.0) \in \mathbb{R}^d\) and \(\mathbf{m}_2 = (+2.5, 0.0, \ldots, 0.0) \in \mathbb{R}^d\). The class label of a point corresponds to the distribution it was drawn from, see Figure 3 (a). If not noted otherwise, we use \(n = 500\) and \(d = 500\) and denote the induced data set by \(\text{Gaussian2C}\). The second artificial data set aims at generating a (possibly) misleading structure: Here, we draw \(n/4\) points from each of four multivariate Gaussian distributions \(X_i \sim \mathcal{N}(\mathbf{m}_i, \mathbf{I})\), where

\[
\begin{align*}
\mathbf{m}_1 &= (-2.5, -5.0, 0.0, \ldots, 0.0) \in \mathbb{R}^d, \\
\mathbf{m}_2 &= (-2.5, +5.0, 0.0, \ldots, 0.0) \in \mathbb{R}^d, \\
\mathbf{m}_3 &= (+2.5, -5.0, 0.0, \ldots, 0.0) \in \mathbb{R}^d, \\
\mathbf{m}_4 &= (+2.5, +5.0, 0.0, \ldots, 0.0) \in \mathbb{R}^d,
\end{align*}
\]

see Figure 3 (b). The points drawn from the first two distributions belong to the first class and the remaining one to the second class. Again, we fix \(n = 500\) and \(d = 500\) and denote the corresponding data set by \(\text{Gaussian4C}\). Finally, we consider the well-known two-dimensional \(\text{Moons}\) data set with \(n = 500\) points, see Figure 3 (c).

**Real-World Data Sets:** In addition to these artificial data sets, we consider several real-world data sets including the \(\text{COIL}\) (Nene et al., 1996) and the \(\text{USPS}\) (Hastie et al., 2009) data sets (consisting of both the training and test set of the original data set). For the \(\text{COIL}\) data set, we reduce the input dimensions of each image from \(128 \times 128\) to \(20 \times 20\) and use \(\text{COIL}(i,j)\) to denote the binary classification task induced by the objects \(i\) and \(j\) out of the available 20 objects (using the ordering given in the data set). A similar notation is used for the binary classification tasks induced by the 10 classes present in the \(\text{USPS}\) data set. For both the \(\text{COIL}\) and the \(\text{USPS}\) data set, we rescaled all pixels such that the resulting values lie between 0.0 and 1.0. Further, we focus on those pairs of objects/digits which are difficult to separate. Finally, we consider the \(\text{Newsgroup20}\) and the \(\text{TEXT}\) data set. Let latter one is composed of the \text{mac} and \text{mwindows} classes of the \(\text{Newsgroup20}\) data set (Chapelle and Zien, 2005).

4.1.3 Model Selection

In semi-supervised settings, model selection can be unreliable due to the lack of labeled data and is widely considered to be an open issue (Chapelle et al., 2006b). Due to this model selection problem, we consider two scenarios to select (non-fixed) parameters. The first one is a non-realistic scenario where we make use of the test set to evaluate the model performance.\(^5\) The second one is a realistic scenario where only the labels of the labeled part of the training set are used for model evaluation (via 5-fold cross-validation). The reason for the non-realistic scenario is the following: By making use of the test set (with a large amount of labels), we can first evaluate the flexibility of the model, i.e., we can first investigate if the model is in principle capable of adapting to the inherent structure of the data while ignoring the (possible) problems caused by small validation sets.

In both scenarios, we first tune the non-fixed parameters via grid search and subsequently retrain the final model on the training set with the best parameters.

\(^5\)This setup is often considered in related evaluations (Chapelle et al., 2006b).
performing set of parameters. As similarity measures we consider a linear \( k(x_i, x_j) = \langle x_i, x_j \rangle \) and a radial basis function (RBF) kernel \( k(x_i, x_j) = \exp(-\frac{(2\sigma^2)^{-1}}{2}||x_i - x_j||^2) \) with kernel width \( \sigma \). To select the kernel width \( \sigma \) for the RBF kernel, we consider the set \{0.01s, 0.1s, 1s, 10s, 100s\} of possible assignments, where the value \( s \) is a rough estimate of the maximum distance between any pair of samples. \(^6\) The cost parameters \( \lambda \) and \( \lambda' \) are tuned on a small grid \( \{2^{-10}, \ldots, 2^{10}\} \times \{0.01, 1, 100\} \) of possible parameters. Further, a short sequence of annealing steps is used (\( \alpha_1 = 0.01, \alpha_2 = 0.1, \alpha_3 = 1.0 \)).

### 4.1.4 Competing Approaches

We use the LIBSVM (Chang and Lin, 2001) as supervised competitor with \( C \in \{2^{-10}, \ldots, 2^{10}\} \). As semi-supervised competitor, we consider the UniverSVM approach (Collobert et al., 2006). Again, we perform grid search for tuning the involved parameters \( (C, C') \in \{2^{-10}, \ldots, 2^{10}\} \times \{0.01, 1, 10, 100\} \). The ratio between the two classes is provided to the algorithm via the \(-w\) option. Except for the option

\[ s = \sqrt{\sum_{k=1}^{d} (\max([x_{ik}], \ldots, [x_{nk}]) - \min([x_{ik}], \ldots, [x_{nk}]))^2} \]

\(^6\)A detailed comparison of the UniverSVM approach with other semi-supervised optimization schemes (like the TSVM approach (Joachims, 1999)) has been conducted by Collobert et al. (Collobert et al., 2006). Their results indicate the superior performance of UniverSVM compared to related methods. For the sake of exposition, we will therefore focus on a comparison of our approach with UniverSVM.

### 4.2 Experimental Results

We will now depict the outcome of several experiments demonstrating the potential of our approach.
4.2.1 Model Flexibility

The well-known Moons data set is said to be a difficult training instance for semi-supervised support vector machines due to its non-linear structure. In Figure 4, the results for the LIBSVM (top row), the UniverSVM (middle row), and the QN-3VM implementation (bottom row) are shown given slightly varying distributions (using the RBF kernel). To select the model parameters, we make use of the test set (non-realistic scenario). For all figures, the average test error (with one standard deviation) over 10 random partitions into labeled, unlabeled, and test patterns, is given. It can be clearly seen that the supervised approach is not able to generate reasonable models. Further, the semi-supervised approaches can successfully incorporate the additional information provided by the unlabeled data, whereas the QN-3VM scheme seems to perform better on this particular data set instances.

4.2.2 Amount of Data

As motivated above, sufficient labeled data is essential for supervised learning approaches to yield reasonable models. For semi-supervised approaches, the amount of unlabeled data used for training is an important issue as well. To analyze how much labeled and unlabeled data is needed for our approach, we consider the Gaussian2C data set and vary the amount of labeled and unlabeled data. For this experiment, we make use of the non-realistic scenario and resort to the LIBSVM implementation as baseline. First, we vary the amount of labeled data from 5% to 80% with respect to (the size of) the training set; the remaining part the training set is used as unlabeled data. In Figure 5 (a), the result of this experiment is shown: Given more than 20% labeled data, the semi-supervised approach performs clearly better. Now, we fix the amount of labeled data to 20% and vary the amount of unlabeled data from 5% to 80% with respect to (the size of) the training set, see Figure 5 (b). Clearly, the semi-supervised approach needs sufficient unlabeled data to yield appropriate models in a reliable manner.

4.2.3 Classification Performance

We consider both the realistic and the non-realistic scenario to evaluate the classification performance of our approach. For each data set, we analyze the behavior of all competing approach given up to three amounts of labeled, unlabeled, and test patterns. For all data sets and for all competing approaches, a linear kernel is used. In Table 1, the test errors (and one standard deviations) averaged over 10 random partitions are given for both scenarios. It can be clearly seen that, for the non-realistic scenario, the semi-supervised approaches mostly yield better results compared to LIBSVM, even if only few labeled patterns are given. Thus, they can successfully incorporate the additional information provided by the unlabeled data. Due to lack of labeled data for model selection, the results for the realistic scenario are worse compared to the non-realistic one. Still, except for the COIL data sets, the results for QN-3VM are at least as good as the ones of LIBSVM.

4.2.4 Computational Considerations

We will finally focus on the runtime behavior. To simplify the setup, we fix the model parameters (i.e., $\lambda = 1$ and $\lambda' = 1$) and use a linear kernel.

Medium-Scale Scenarios: To give an idea of the runtime needed to obtain the results provided in Table 1, we consider the Gaussian2C, the USPS(8,0) and the sparse TEXT data set (with $l = 25$, 22, and 48 labeled patterns, respectively). Further, we vary the amount of unlabeled patterns as shown in Figure 6; the average runtime over 10 runs is provided. The plots indicate a comparable runtime behavior on the non-sparse data sets. On the sparse text data set, however, the QN-3VM is considerably faster; note that even with 1,000 unlabeled examples, the practical runtime is less than 0.1 seconds per single execution.
5 CONCLUSIONS

We proposed a quasi-Newton optimization framework for the non-convex task induced by semi-supervised support vector machines. It seems that this type of optimization schemes is well suited for the task at hand since it (a) can be implemented easily due to its conceptual simplicity and (b) admits direct accelerations for sparse and non-sparse data. The experiments indicate that the resulting approach can successfully incorporate unlabeled data, even in realistic scenarios where the lack of labeled data complicates the model selection phase.

ACKNOWLEDGEMENTS

This work has been supported in part by funds of the Deutsche Forschungsgemeinschaft (Fabian Gieseke, grant KR 3695) and by the Academy of Finland (Tapio Pahikkala, grant 134020).
REFERENCES


