Pattern Recognition Letters 37 (2014) 24–31

Contents lists available at SciVerse ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Semi-supervised linear discriminant analysis through moment-constraint parameter estimation

Marco Loog *

Pattern Recognition Laboratory, Delft University of Technology, Mekelweg 4, 2628 CD, Delft, The Netherlands
The Image Group, University of Copenhagen, Universitetsparken 5, DK-2100, Copenhagen Ø, Denmark

ARTICLE INFO

Article history:
Available online 18 March 2013
Communicated by G. Sanniti di Baja

Keywords:
Linear discriminant analysis
Semi-supervised learning
Moment constraints
Affine invariant
Classification

ABSTRACT

A semi-supervised version of classical linear discriminant analysis is presented. As opposed to most current approaches to semi-supervised learning, no additional extrinsic assumptions are made to tie information coming from labeled and unlabeled data together. Our approach exploits the fact that the parameters that are to be estimated fulfill particular relations, intrinsic to the classifier, that link label-dependent with label-independent quantities. In this way, the latter type of parameters, which can be estimated based on unlabeled data, impose constraints on the former and lead to a reduction in variability of the label dependent estimates. As a result, the performance of our semi-supervised linear discriminant is typically expected to improve over that of its regular supervised match. Possibly more important, our semi-supervised linear discriminant analysis does not show the severe deteriorations other approaches frequently display with increasing numbers of unlabeled data. This work recapitulates, corrects, extends, and revises our previous work that has been published as part of the First IAPR TC3 Workshop on Partially Supervised Learning. The main novelty it provides over our earlier work is an affine invariant approach to semi-supervised learning befitting linear discriminant analysis. Besides, more elaborate and convincing experimental evidence of the potential of our general approach is provided. We essentially believe that the general principle of intrinsic constraints is of interest as such and may inspire other novel semi-supervised methods.

1. Introduction

Supervised learning aims to learn from examples: given a limited number of instances of a particular input–output relation, its goal is to generalize this relationship to new and unseen data in order to enable the prediction of the associated output given new input. Specifically, supervised classification seeks to infer an unknown feature vector-class label relation from a finite number of feature vectors and their associated, desired class labels. Now, an elementary question is whether and, if so, how additional unlabeled data can significantly improve the training of such classifier. This is what constitutes the problem of semi-supervised learning (Chapelle et al., 2006; Zhu and Goldberg, 2009).

The expectation is that semi-supervised learning can indeed bring considerable improvements to many research and application areas in which classification problems play a key role by simply exploiting the often enormous amounts of unlabeled data available (think image analysis, computer vision, natural language processing, medical diagnostics, but also the social and environmental sciences and various metrics). The matter of the fact, however, is that current semi-supervised methods have not been widely accepted outside of the realms of computer science. Part of the reason for this may be that current methods offer no performance guarantees (Ben-David et al., 2008; Singh et al., 2009) and often deteriorate when confronted with large amounts of unlabeled samples (Cohen et al., 2004; Cozman and Cohen, 2006; Mann and McCallum, 2010; Nigam et al., 1998).

Earlier, we identified as main reason for the frequent failure of semi-supervised learning the fact that current semi-supervised approaches typically rely on assumptions extraneous to the classifier being considered (Loog, 2010, 2012). Indeed, the main current approaches to semi-supervised learning stress the need for extrinsic assumptions such as the cluster assumption: points from the same class cluster, the smoothness assumption: neighboring point have the same label, the assumption of low density separation: the decision boundary is located in low density areas, and the like (Chapelle et al., 2006; Zhu and Goldberg, 2009). Given a particular assumption holds, one is able to extract relevant information not only from the labeled, but especially from the unlabeled examples. While it is undeniably true that having more precise knowledge on the distribution of data could, or even should, help in training a better classifier, in many real-world settings it may be questionable if one can

* Address: Pattern Recognition Laboratory, Delft University of Technology, Mekelweg 4, 2628 CD, Delft, The Netherlands. Tel.: +31 15 27 89395.
E-mail address: m.loog@tudelft.nl
URL: http://prlab.tudelft.nl

0167-8655/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.patrec.2013.03.004

at all check if such conditions are indeed met. Moreover, as soon as these additional model assumptions do not fit the data, there obviously is the real risk that adding unlabeled data actually leads to a severe deterioration of classification performance (Cohen et al., 2004; Cozman and Cohen, 2006; Loog, 2010, 2012; Nigam et al., 1998). Note that this is in contrast with the supervised setting, where most classifiers, generative or not, are capable of handling mismatched data assumptions rather well, in the sense that adding more training data generally improves the performance of the classifier (but cf. Loog and Duin, 2012).

The current work devises a specific semi-supervised scheme tailored to classical linear discriminant analysis (LDA, Hastie et al., 2001; Ripley, 1996), which is sometimes referred to as a normal-based linear discriminant function (see McLachlan, 1992). Regarding LDA, we would like to stress that it still is a widely employed classifier and therefore, also from a practical perspective, the current investigation is of interest and could have consequences beyond the mere academical. Also, like any other classifier, LDA has its validity and cannot be put aside as being outdated or not state-of-the-art. In this respect, we would also like to refer the reader to insightful contributions such at the ones by Hand (2006) and Efron (2001) (see also Duin et al., 2010).

Now, the underlying, more general idea presented in this paper is that the class-specific parameters to be estimated in the learning phase are related to each other and, more importantly, to certain label-independent statistics. These relations can be seen as intrinsic—rather than additional, extrinsic—constraints between particular estimates coming from labeled data and those derived from unlabeled instances. Enforcing these constraints during semi-supervised learning yields label-dependent estimates that are in a sense closer to the true parameter values, which, in turn, often lead to reduced classification errors.

Though the focus in this work is specifically on LDA, we do believe that the general, underlying principle of searching for intrinsically motivated semi-supervised learning is of interest in its own respect and we hope that it this work will inspire further research in this direction.

On the conceptual side, this work continuous in the spirit of the earlier research reported on in Loog (2010, 2012). Methodologically, the paper presents a revised version of the latter contribution and extends and corrects parts of the method proposed there. In particular, it dwells on an important shortcoming of the technique from Loog (2012), which is the lack of invariance (or, if one prefers, covariance) of the parameters of semi-supervised LDA under nonsingular affine transformations of the feature space. Such affine transformations do not only comprise translation, rotation, reflection, and isotropic scaling but also anisotropic scaling, and shearing. Classification methods have this invariance property if, for any two nonsingular affine transformation A and B, the classifiers C_A and C_B trained on data transformed by A and B, respectively, deliver the same classification outcomes on a similarly transformed test object x, i.e., $C_A(A(x)) = C_B(B(x))$. Note that there are many classifiers for which this noticeable property does not hold, e.g. neither k-nearest neighbors, nor Parzen classifiers, nearest mean classifiers, or support vector machines are invariant to anisotropic scaling and shearing transforms. The regular counterpart to semi-supervised LDA, however, does enjoy this invariance property (Fukunaga, 1990; McLachlan, 1992). It therefore seems nothing more than reasonable to retain the same invariance for the semi-supervised case. This paper demonstrates how such elementary characteristic can also be enforced in semi-supervised learning.

1.1. Outline

Following the next section, which presents an overview of further related work, Section 3 briefly recapitulates some relevant details of the approach presented in Loog (2010) for semi-supervised nearest mean classification. The main focus in that section will, however, be on semi-supervised LDA as presented in Loog (2012) and its novel variation that satisfies the earlier sketched invariance property. Section 4 provides experimental results on various real-world and benchmark data sets in which our constrained approach is mainly compared to regular LDA and so-called self-learned LDA (the latter of which is briefly explained in the next section as well). Additional comparisons are made with logistic regression, nearest neighbor classification, an entropy regularization method, and transductive SVM. Subsequently, Section 5 completes the paper, providing a discussion and conclusions.

2. Additional related works

There are few works that focus on semi-supervised LDA (see Efron (2001)'s rule 1). Most relevant contributions come from statistics and have been published mainly by the end of the 1960s and halfway the 1970s. Hartley and Rao (1968) suggests to maximize the likelihood over all permutations of possible labelings of unlabeled objects. A computationally more feasible approach has been proposed by McLachlan (1975, 1977), which follow an iterative procedure. Firstly, the linear discriminant is trained on the labeled data only and used to label all unlabeled instances. Using the now-labeled data, the classifier is retrained and employed to relabel the initially unlabeled data. This process of relabeling originally unlabeled data is repeated until none of the samples changes label.

The above approach to semi-supervised learning is basically a form of so-called self-training or self-learning, which has been presented in different guises and at different levels of complication (see, for instance, Basu et al., 2002; McLachlan, 1975; McLachlan and Ganesalingam, 1982; Nigam et al., 1998; Titterington, 1976; Vittaut et al., 2002; Yarowsky, 1995; Zhou and Li, 2010). This iterative method also relates directly to the well-known approach to semi-supervised learning based on expectation maximization (see Nigam et al. (1998) and the discussion papers related to Dempster et al. (1977)). The similarity between self-learning and expectation maximization (in some cases equivalence even) has been noted in various papers, e.g. by Abney (2004), Basu et al. (2002), and it is to no surprise that such approaches suffer from the same drawback: as soon as the underlying model assumptions do not fit the data, there is the real risk that adding too much unlabeled data leads to a substantial decrease of classification performance (Cohen et al., 2004; Cozman and Cohen, 2006; Nigam et al., 1998).

An approach seemingly different from self-learning is, among others, known as label propagation. It relies on the smoothness assumption, assuming that data points close to each other tend to belong to the same class. Various versions of this idea have been studied, most of which are related to graph-based techniques, manifold learning, or spectral clustering methods (Bengio et al., 2006; Szummer and Jaakkola, 2002; Zhu and Ghahramani, 2002). The propagation of label information through such graph structure can also be thought of as a particular instantiation of the iterative expectation maximization or self-learning methods. A more explicit connection between self-learning and graph-based propagation methods can be found in Culp and Michailidis (2008).

We finally remark that there are also semi-supervised approaches to LDA as a dimensionality reduction technique. As we consider LDA as a classifier, we do not discuss these approach in any detail. As it comes in some sense close to our work, the single paper we do like to mention is by Fan et al. (2009). The work notes that the Fisher criterion, which typically employs the between-class and within-class covariance matrices, can also be expressed in such a way that the total covariance matrix replaces one of

the other two (cf. Fukunaga, 1990; Loog, 2007). Clearly, the total covariance can be better estimated using all data, both labeled and unlabeled, which in turn might result in better performance of the dimensionality reduction scheme. Our work, however, aims at LDA for classification in which the total covariance does not directly play a role and therefore we cannot resort to the simple and straightforward suggestion made in Fan et al. (2009).

3. Semi-supervised learning through moment constraints

Loog (2010) introduces a semi-supervised version of the plain nearest mean classifier (NMC, Duda and Hart, 1973; McLachlan, 1992). Though simple indeed, this classifier is still topical (see, for example Liu et al., 2009b,a; Mira et al., 2009; Roepman et al., 2009; Salazar et al., 2011). Similarly, semi-supervised NMC in some cases provides error rates that are competitive with state-of-the-art methods (compare Loog (2010) and Chapelle et al. (2006)).

To start with, note that when employing a regular supervised NMC, the K class means, m_i with $i \in \{1,\ldots,K\}$, and the overall mean of the data, m, satisfy the linear constraint (Fukunaga, 1990)

$$m = \sum_{i=1}^{K} p_i m_i, \tag{1}$$

where p_i is the prior of class i. Having additional unlabeled data, one can improve the estimate of m because it does not depend on any labels. Having a better estimate of m, however, the constraint in Eq. (1) will typically be violated. The core idea in Loog (2010) is that one can get improved estimates of the class means as well by adapting them such that the constraint is satisfied again. The basic solution chosen is to simply alter the K sample class means m_i by the same shift such that the new total sample mean

$$m' = \sum_{i=1}^{K} p_i m_i' \tag{2}$$

of the shifted class means m_i' coincides with the total sample mean μ. The total mean m' has been obtained using all data available. Ultimately, the following update of the class means is suggested:

$$m_i' = m_i - \sum_{i=1}^{K} p_i m_i + \mu. \tag{3}$$

3.1. Constrained linear discriminant analysis: the variant case

For LDA, next to Eq. (1), an additional known constraint equates the sum of the estimates of the between-class covariance matrix $\mathbf{B}$ and within-class covariance $\mathbf{W}$ to the total covariance over all data $\mathbf{T}$ (cf. Fukunaga, 1990). That is,

$$\mathbf{T} = \mathbf{W} + \mathbf{B}, \tag{4}$$

where

$$\mathbf{T} = \frac{1}{N}\sum_{i=1}^{K}\sum_{j=1}^{N_i}(x_{ij} - m)(x_{ij} - m)^{\mathrm{t}}, \tag{5}$$

in which x_{ij} is the jth feature vector from class i, m is the estimated overall mean, N_i is the number of samples from class i, and N is the total number of samples. The other variables in the equation have the following definitions:

$$\mathbf{W} = \sum_{i=1}^{K} p_i \mathbf{C}_i, \tag{6}$$

where $\mathbf{C}_i$ is the sample covariance matrix for class i, and

$$\mathbf{B} = \sum_{i=1}^{K} p_i (m_i - m)(m_i - m)^{\mathrm{t}}. \tag{7}$$

The parameters of interest are the class means m_i, the within-class covariance matrix $\mathbf{W}$, and the priors p_i. These parameters should be estimated from both labeled and unlabeled data under the constraints provided in Eqs. (1) and (4), in which the left hand sides are fixed and determined by all data available.

Now, denote the estimated total mean based on *all* the data by μ, as in the beginning of Section 3, and let the corresponding total covariance matrix be denoted by $\mathbf{\Theta}$. The corresponding mean m and covariance matrix $\mathbf{T}$ are based merely on the labeled data. Loog (2012) now suggests the following easy and effective solution in order to meet the constraints. To start with, transform every labeled datum x into a new feature vector x' as follows:

$$x' = \mathbf{\Theta}^{\frac{1}{2}}\mathbf{T}^{-\frac{1}{2}}(x - m) + \mu. \tag{8}$$

Here, the matrix power $\mathbf{A}^p$ of a symmetric matrix $\mathbf{A}$ is defined through its eigendecomposition $\mathbf{R}\mathbf{\Lambda}\mathbf{R}^{\mathrm{t}}$, with $\mathbf{R}$ a rotation matrix and where the power acts on the diagonal of the eigenmatrix only:

$$\mathbf{A}^p = \mathbf{R}\mathbf{\Lambda}^p\mathbf{R}^{\mathrm{t}}. \tag{9}$$

By and large, the transformation sees to it that the overall mean and covariance statistics of the labeled data match the respective statistics as measured on all data. That is, on the transformed data, the corresponding m and $\mathbf{T}$ equal μ and $\mathbf{\Theta}$, respectively. The next step is to simply train a regular LDA on this transformed training data, providing the semi-supervised estimates for m_i' and $\mathbf{W}'$. By means of Eq. (7), the corresponding $\mathbf{B}'$ in the transformed space can be determined and, by construction, one has

$$\mu = \sum_{i=1}^{K} p_i m_i' \tag{10}$$

and

$$\mathbf{\Theta} = \mathbf{W}' + \mathbf{B}'. \tag{11}$$

As the transformation applied is affine, we can actually directly estimate the m_i's and the $\mathbf{W}'$ from the original space. Given the class means m_i and the within-class covariance matrix $\mathbf{W}$ determined on the *untransformed* labeled data only, the following equations hold:

$$m_i' = \mathbf{\Theta}^{\frac{1}{2}}\mathbf{T}^{-\frac{1}{2}}(m_i - m) + \mu \tag{12}$$

and

$$\mathbf{W}' = \mathbf{\Theta}^{\frac{1}{2}}\mathbf{T}^{-\frac{1}{2}}\mathbf{W}\mathbf{T}^{-\frac{1}{2}}\mathbf{\Theta}^{\frac{1}{2}}. \tag{13}$$

This expresses the m_i's and $\mathbf{W}'$ in terms of first and second order moment statistics in the original space.

3.2. Constrained linear discriminant analysis: the invariant cases

Now, what happens if the data gets transformed by a nonsingular affine transformation A? For regular LDA it is easily checked that the class means and within-class covariance matrix change accordingly. The new means become $A(m_i)$ and the new average within-class covariance matrix becomes $\mathbf{A}^{\mathrm{t}}\mathbf{W}\mathbf{A}$, in which the matrix $\mathbf{A}$ describes the linear part of the affine transformation A. As a result the classification outcome of LDA is invariant to such affine transforms in the sense discussed in Section 1 (cf. Fukunaga, 1990; McLachlan, 1992). That this does not hold for the semi-supervised setting, however, can for instance be concluded directly from the fact that for an arbitrary nonsingular matrix $\mathbf{A}$, one will find that

$$\mathbf{A}^{\mathrm{t}}\mathbf{W}'\mathbf{A} \neq (\mathbf{A}^{\mathrm{t}}\mathbf{\Theta}\mathbf{A})^{\frac{1}{2}}(\mathbf{A}^{\mathrm{t}}\mathbf{T}\mathbf{A})^{-\frac{1}{2}}\mathbf{W}(\mathbf{A}^{\mathrm{t}}\mathbf{T}\mathbf{A})^{-\frac{1}{2}}(\mathbf{A}^{\mathrm{t}}\mathbf{\Theta}\mathbf{A})^{\frac{1}{2}}. \tag{14}$$

To find a semi-supervised approach with the same level of invariance as original LDA, a trick inspired by Loog and Duin (2004) is applied. The idea is simply that all data are somehow standardized prior to the semi-supervised parameters being

determined using Eqs. (12) and (13). Once the relevant parameters are determined in this transformed space, they can be transformed back to the original space, where they can be used to construct the semi-supervised version of LDA.

The particular standardizing transform employed here first carries out a centering of the feature vectors such that the overall data mean μ coincides with the origin. Following this, a simultaneous diagonalization of the total covariance matrices $\mathbf{T}$ and $\boldsymbol{\Theta}$ is performed. The latter can be achieved by first determining the matrix $\mathbf{E}$ of eigenvectors coming from the generalized eigenvalue problem

$$\mathbf{TE} = \boldsymbol{\Theta}\mathbf{E}\boldsymbol{\Lambda}. \tag{15}$$

Subsequently, all data is transformed with $\mathbf{E}$, leading to two diagonal total covariance matrices $\mathbf{D}_T = \mathbf{E}^t\mathbf{T}\mathbf{E}$ and $\mathbf{D}_\Theta = \mathbf{E}^t\boldsymbol{\Theta}\mathbf{E}$. For this standardized feature space, the semi-supervised within covariance and class means are determined. Finally, by the inverse transformation, the sought-after parameters in the original space are found. All in all, we find the following expressions for m_i' and $\mathbf{W}'$:

$$m_i' = \mathbf{E}^{-t}\mathbf{D}_\Theta^{\frac{1}{2}}\mathbf{D}_T^{-\frac{1}{2}}\mathbf{E}^t(m_i - m) + \mu \tag{16}$$

and

$$\mathbf{W}' = \mathbf{E}^{-t}\mathbf{D}_\Theta^{\frac{1}{2}}\mathbf{D}_T^{-\frac{1}{2}}\mathbf{E}^t\mathbf{W}\mathbf{E}\mathbf{D}_T^{-\frac{1}{2}}\mathbf{D}_\Theta^{\frac{1}{2}}\mathbf{E}^{-1}. \tag{17}$$

To reduce notational clutter, we use the notation $-t$ for the inverse of the transpose (as well as the transpose of the inverse). That is, for an invertible matrix $\mathbf{A}$, we define $\mathbf{A}^{-t} = (\mathbf{A}^{-1})^t = (\mathbf{A}^t)^{-1}$.

A final remark is that the generalized eigenvalue decomposition is not unique. Given a nonsingular diagonal matrix $\mathbf{D}$ and a permutation matrix $\mathbf{P}$, the matrix $\mathbf{F} = \mathbf{EDP}$ is also a solution to the generalized eigenvalue problem in Eq. (15). In the same way as for $\mathbf{E}$, applying $\mathbf{F}$ to $\mathbf{T}$ and $\boldsymbol{\Theta}$ diagonalizes them. It may lead to different diagonal matrices though. Still, any solution to the above generalized eigenvalue decomposition will lead to exactly the same expressions for m_i' and $\mathbf{W}'$. To see this, first write out $\mathbf{F}^t\mathbf{T}\mathbf{F}$ to show that it equals $\mathbf{P}^t\mathbf{D}^t\mathbf{D}_T\mathbf{D}\mathbf{P}$. A similar expression holds for $\mathbf{F}^t\boldsymbol{\Theta}\mathbf{F}$. Secondly, because $\mathbf{F}^t\mathbf{T}\mathbf{F}$ is diagonal, $\mathbf{P}^t(\mathbf{D}^t\mathbf{D}_T\mathbf{D})\mathbf{P}$ is in fact its eigendecomposition. Therefore, $(\mathbf{F}^t\mathbf{T}\mathbf{F})^{-\frac{1}{2}}$ equals $\mathbf{P}^t(\mathbf{D}^t\mathbf{D}_T\mathbf{D})^{-\frac{1}{2}}\mathbf{P}$. Again a similar expression holds for $(\mathbf{F}^t\boldsymbol{\Theta}\mathbf{F})^{\frac{1}{2}}$. Finally, we can plug these expansions into Eqs. (16) and (17) and simplify them to find exactly the same m_i' and $\mathbf{W}'$ back. For instance, for the first equation we have

$$\begin{aligned}
&(\mathbf{EDP})^{-t}\mathbf{P}^t(\mathbf{D}^t\mathbf{D}_\Theta\mathbf{D})^{\frac{1}{2}}\mathbf{P}\mathbf{P}^t(\mathbf{D}^t\mathbf{D}_T\mathbf{D})^{-\frac{1}{2}}\mathbf{P}(\mathbf{EDP})^t(m_i - m) + \mu \\
&= \mathbf{E}^{-t}\mathbf{D}^{-t}\mathbf{P}^{-t}\mathbf{P}^t\mathbf{D}_\Theta^{\frac{1}{2}}\mathbf{D}_T^{-\frac{1}{2}}\mathbf{P}\mathbf{P}^t\mathbf{D}^t\mathbf{E}^t(m_i - m) + \mu \\
&= \mathbf{E}^{-t}\mathbf{D}^{-t}\mathbf{D}_\Theta^{\frac{1}{2}}\mathbf{D}_T^{-\frac{1}{2}}\mathbf{D}^t\mathbf{E}^t(m_i - m) + \mu = \mathbf{E}^{-t}\mathbf{D}_\Theta^{\frac{1}{2}}\mathbf{D}_T^{-\frac{1}{2}}\mathbf{E}^t(m_i - m) + \mu \\
&= m_i',
\end{aligned}$$

where at various places we used the facts that diagonal matrices commute and that $\mathbf{P}$ is a projection matrix and so $\mathbf{P}^t = \mathbf{P}^{-1}$. The same can now easily be checked for $\mathbf{W}'$.

4. Experimental setup and results

For the experiments, nine real-world data sets, all having two classes, are taken from the UCI Machine Learning Repository (Asuncion and Newman, 2007). In addition, seven benchmark data sets as discussed in Chapelle et al. (2006, Chapter 21) are considered. Of the latter selection, six are two-class problems, while one is six-class. The data sets used, together with some basic specifications, can be found in Table 1.

4.1. UCI data sets

The possibility to improve LDA by semi-supervised learning has our main interest. A comparison is therefore made to the standard, supervised setting and LDA trained by means of self-learning (McLachlan, 1975; Yarowsky, 1995) (comparable results would be obtained by any EM approach (Cohen et al., 2004; Cozman and Cohen, 2006; Nigam et al., 1998)). Nevertheless, to put the performance of LDA on these data sets in perspective, we also considered the efficiency of logistic regression and nearest neighbor classification. We chose these classifiers because of their very different nature (linear vs. highly nonlinear) and the absence of any parameters to tune. The latter is important as we will, among others, be dealing with small samples. The logistic regression and nearest neighbor (1NN) are regular supervised methods. In the next section, a comparison is made to two state-of-the-art semi-supervised linear classifiers.

Table 1
Top part: basic properties of the nine two-class, real-world, UCI data sets considered (Asuncion and Newman, 2007). Bottom part: same numbers for the used benchmark data sets from (Chapelle et al., 2006). The `coil` data set is the only classification problem with six classes. All others have merely two.

Data set	# Objects	Dim.	Smallest prior
haberman	306	3	0.26
ionosphere	351	33	0.36
parkinsons	195	22	0.25
pima	768	8	0.35
sonar	208	60	0.47
spectf	267	44	0.21
transfusion	748	3	0.24
wdbc	569	30	0.37
g24lc	1500	241	0.50
g24ld	1500	241	0.50
digit1	1500	241	0.49
usps	1500	241	0.20
coil	1500	241	0.17
bci	400	117	0.50
text	1500	11,900	0.50

On the UCI data sets, experiments were carried out with 10 and 100 labeled training objects in total. In all cases, we made sure every class has at least one training sample. We consider learning curves as a function of the number of *unlabeled* instances. This setting should disclose both the sensitivity of the self-learning to an abundance of unlabeled data and the improvements that may generally be obtained given various quantities of unlabeled data, also small ones. The number of unlabeled objects considered are 2, 8, 32, 128, 512, 2048, and 8192. For every combination of number of unlabeled objects and labeled objects, 1000 experiments were carried out. In every iteration, a new data set was generated randomly. In order to be able to do so on the limited amount of samples in the UCI data sets, instances were drawn with replacement. It basically assumes that the empirical distribution of every data set is its true distributions and allows us to measure the error rates on the full data set. This approach enabled us to properly study the influence of the constraint estimation on real-world data without having to deal with the extra variation due to cross validation and the like. For rather flexible classifiers this might give unacceptably optimistically biased results. In our case, however, it is mainly about the different versions of LDA, which we believe can be compared in a fair way based on these experiments.

The experiments occasionally suffer from a singular average within-class covariance matrix, whose inverse is needed to apply LDA. In this case, instead of LDA, the use of a NMC in the subspace spanned by the singular components of the within covariance can be justified.

Figs. 1 and 2 provide averaged learning curves for, respectively, 10 and 100 labeled samples for eight of the nine data sets. For reasons of readability, `spect` has been excluded from this figure but its behavior resembles that of `spectf`. The lighter bands around the learning curves sketch the standard deviations of the mean

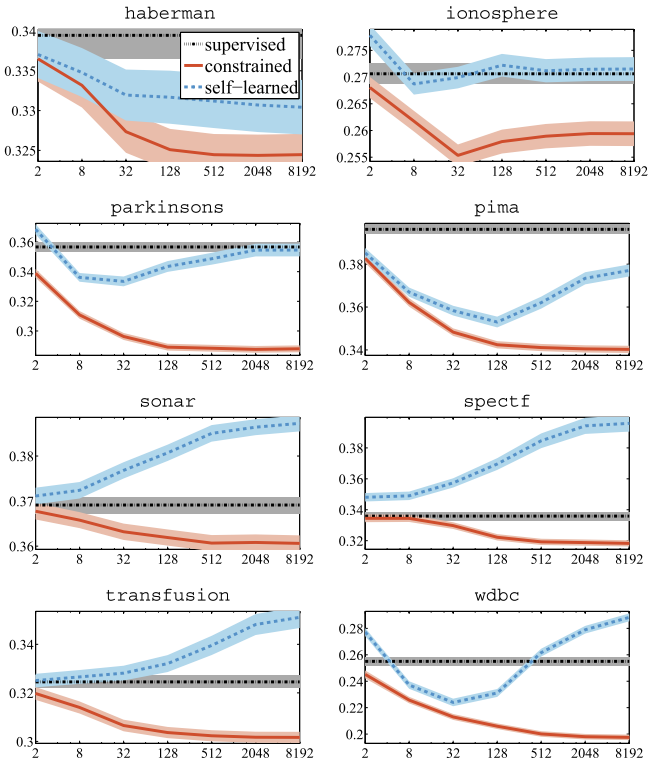

Fig. 1. Error rates against number of unlabeled data points for the supervised, constrained semi-supervised, and self-learned classifiers on eight of the nine real-world UCI data sets for various unlabeled sample sizes and a total of 10 labeled training samples. Lighter bands around the learning curves show the standard deviations of the mean error rates.

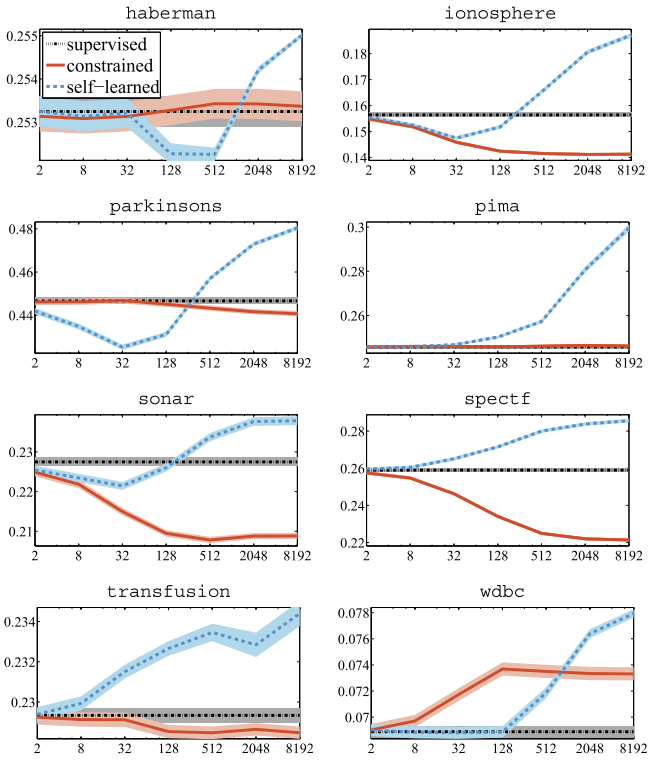

Fig. 2. Error rates against number of unlabeled data points for the supervised, constrained semi-supervised, and self-learned classifiers on eight of the nine real-world UCI data sets for various unlabeled sample sizes and a total of 100 labeled training samples. Lighter bands around the learning curves show the standard deviations of the mean error rates.

error rates. Firstly, note that in most of the experiments the constrained LDA performs best of the three LDA schemes compared. Self-learning leads to deteriorated performance with increasing unlabeled data sizes in many a case. Another interesting observation is that adding only a moderate amount of unlabeled objects often allows our semi-supervised constrained approach to already outperform regular supervised LDA. The improvements are, however, equally moderate and apparent mainly in the experiments with 10 labeled instances. In some cases our approach seems to perform worse, notably for `haberman` and `wdbc` with 100 labeled samples. In those cases, however, the difference with supervised LDA is in the third decimal only and therefore practically not significant. Still, the difference on `wdbc` is statistically significant.

Qualitatively the results are very convincing as it is easily appreciated from the figures that the curves of the constrained approach are below the other two curves in most cases. For a more quantitative impression, Tables 2 and 3 show error rates for all nine data sets for 10 and 100 labeled samples, respectively. Next to the results obtained with regular LDA and our constrained approach, the results for supervised logistic regression and 1NN are reported as well. The most interesting observation that can be made is that even though LDA is among the best *supervised* classifiers in only 5 out of 18 cases, constrained semi-supervised LDA is among the best in 9 cases. Especially in the small sample case of 10 labeled samples the benefits of semi-supervised learning for LDA are absolutely evident: in 6 out of 9 cases constrained LDA performs the best. For the larger sample size of 100, 1NN is dominant with 5 out of 9 lowest error rates. Considering the nonparametric nature of this classifier, the relative low dimensionality of many of the data sets, and the optimistic bias due to the sampling discussed in the beginning of this section, this outcome is not surprising. It is, in fact, interesting to see that in many cases LDA still performs rather well compared to logistic regression and 1NN also in this setting.

Table 2
Error rates for supervised LDA, constrained semi-supervised LDA, logistic regression, and 1NN on all nine UCI data sets. The semi-supervised approach is trained with an unlabeled sample size of 8192 and a total of 10 labeled training samples. The numbers corresponds to the point completely on the right of the plot in Fig. 1. Underlining indicates the best performing method. Boldfaced fonts indicate results do not differing significantly from the best result (based on a Wilcoxon signed-rank test).

Data set	LDA	Constr. LDA	Logistic	1NN
`haberman`	0.339	**0.324**	0.354	0.351
`ionosphere`	0.271	**0.259**	0.268	0.269
`parkinsons`	0.357	0.288	0.326	**0.282**
`pima`	0.396	**0.340**	0.406	0.375
`sonar`	0.369	0.361	**0.344**	0.399
`spect`	0.335	**0.295**	0.333	0.318
`spectf`	0.336	**0.318**	0.340	**0.320**
`transfusion`	0.325	**0.302**	0.334	0.348
`wdbc`	0.255	0.198	0.183	**0.136**

4.2. Benchmark data sets from Chapelle et al.

Additional experiments were carried on seven of the benchmark data sets from Chapelle et al. (2006) (recall Table 1). In this, constrained LDA was compared to the two linear, state-of-the art, semi-supervised classification procedures discussed in the same book. The one is an entropy regularized logistic classifier (Grandvalet and Bengio, 2005) and the other is a transductive linear SVM (TSVM, Joachims, 1999). Both approaches model the additional assumption of low density separation. The former does this

Table 3
Error rates for supervised LDA, constrained semi-supervised LDA, logistic regression, and 1NN on all nine UCI data sets. The semi-supervised approach is trained with an unlabeled sample size of 8192 and a total of 100 labeled training samples. These results corresponds to the points completely on the right of the plot in Fig. 2. Underlining indicates the best performing method. Boldfaced fonts indicate results do not differing significantly from the best result (based on a Wilcoxon signed-rank test).

Data set	LDA	Constr. LDA	Logistic	1NN
haberman	0.253	0.253	0.254	**0.237**
ionosphere	0.156	0.141	0.160	**0.127**
parkinsons	0.447	0.441	0.187	**0.108**
pima	**0.246**	**0.246**	0.247	0.297
sonar	0.227	0.209	0.229	**0.141**
spect	0.197	**0.187**	0.196	0.189
spectf	0.259	0.221	0.237	**0.196**
transfusion	0.229	**0.228**	0.229	0.259
wdbc	**0.069**	0.073	0.094	0.078

through a regularization term that measures class overlap based on conditional entropy, while the latter employs an additional constraints that explicitly tries to maximize the margin on the test samples as well. In both cases, it leads to an objective function that is not convex (upwards or downwards) and can be difficult to optimize. Chapelle et al. (2006) also provides the protocol to carry out the experiments and therefore we can compare our results to the previous two methods in a relatively fair way. The results for entropy regularization and TSVM are therefore taken directly from Chapelle et al. (2006).

Tables 4 and 5 give the results for the two training set sizes of 10 and 100 samples, respectively. The right most column provides the results of our semi-supervised LDA. Note that the underline and boldface annotations mean something different from the ones in Tables 2 and 3: an underline indicates that constrained LDA is better than entropy regularization, while boldface means that the same LDA approach performs better than TSVM. These experimental outcomes are based on ten-fold cross validation, though no significance tests could be carried out as we only have the data from the constrained LDA at our disposal. Finally, we note that for the six-class coil data set, we trained the multiclass classifier in a one-against-all fashion.

Tables 4 and 5 show that, particularly in the small sample setting of 10 labeled training instances, the performance of constrained LDA is rather good: on six out of seven sets, it outperforms at least one of the two competitors and it is the overall best in two cases. With 100 labeled samples, the two discriminative approaches become markedly better. An explanation for this might be that generative approaches would typically attain a higher-asymptotic error rate when comparable classifier models (or hypothesis spaces) are considered (cf. Ng and Jordan, 2002).

Table 4
Error rates entropy regularization, transductive SVM, and constrained semi-supervised LDA, on seven data sets taken from Chapelle et al. (2006) for 10 labeled samples. The results for entropy regularization and TSVM are taken from Chapelle et al. (2006). Underlining of the constrained LDA result indicates that it performs better than entropy regularization. Boldfaced fonts indicate constrained LDA performs better than TSVM.

Data set	Entropy reg.	TSVM	Constrained LDA
g241c	0.474	0.209	0.403
g241d	0.458	0.464	**0.411**
digit1	0.244	0.206	0.236
usps	0.203	0.307	**0.235**
coil	0.665	0.500	0.690
bci	0.477	–	**0.462**
text	0.421	0.286	0.393

Table 5
Error rates entropy regularized logistic regression, transductive SVM, and constrained semi-supervised LDA, on seven data sets taken from Chapelle et al. (2006) for 100 labeled samples. The results for entropy regularization and TSVM are taken from Chapelle et al. (2006). Underlining of the constrained LDA result indicates that it performs better than entropy regularization. Boldfaced fonts indicate constrained LDA performs better than TSVM.

Data set	Entropy reg.	TSVM	Constrained LDA
g241c	0.210	0.182	0.271
g241d	0.254	0.238	0.283
digit1	0.073	0.180	0.182
usps	0.122	0.211	**0.163**
coil	0.295	0.427	**0.352**
bci	0.289	–	**0.284**
text	0.249	0.223	0.267

5. Discussion and conclusions

In the vein of Loog (2010), we proposed to perform semi-supervised linear discriminant analysis (LDA) by making use of known constraints that link label-independent and label-dependent parameters. This is a rather different way of thinking about semi-supervised learners and should be of interest in its own respect. In particular, we revisited the work from Loog (2012), which first presented a form of moment-constrained semi-supervised LDA, and corrected and extended this work. One of the key changes to the original work is that the revised approach is affine invariant, which is a property that standard LDA also enjoys.

An advantage of our semi-supervised LDA is that it is practically as easy to train as regular LDA. There is no need for regularization schemes or iterative procedures as in Hartley and Rao (1968); McLachlan (1975, 1977); Nigam et al. (1998). We only need fairly standard operations on matrices, including a generalized eigenvalue decomposition, which are relatively easy to employ. These operations, might be unstable however, especially when the number of labeled feature vectors is small compared to the dimensionality of the feature space. In our code, inversion and power operations are implemented through a singular value decomposition, which provides relatively stable solutions. Nonetheless, one should realize that numerics could play a negative role at times and might need to be designed with more care when applying our semi-supervised LDA to very high-dimensional data sets. In this respect, regularization may definitely be a way of improving the method as well.

The proposed semi-supervised approach to LDA is a significant advance form the earlier proposed semi-supervised nearest mean classifier (Loog, 2010). Our approach allows to take into account important constraints on second-order moments and not only simple constraints on sample means. The other important improvement, already mentioned, is that our novel semi-supervised LDA is constructed to be properly affine invariant. One remark we should make here, which also relates to those about numerics, is that invariance cannot be enforced anymore as soon as the total covariance matrix $\mathbf{T}$ of the labeled data becomes singular. The most obvious way to proceed in this case is to initially reduce the feature space to the subspace spanned by the eigenvectors of $\mathbf{T}$ and perform semi-supervised LDA in that subspace (cf. Loog, 2007).

The experiments demonstrate that often significant improvements can be obtained by enforcing labeled-unlabeled constraints on the parameters. This is in contrast with the results obtained by self-learned LDA that often times performs dramatically worse. Even to the performance of regular, supervised LDA, the self-learner often loses out. Also when compared to other linear classifiers

like supervised and semi-supervised logistics regression or linear transductive SVMs our constrained approach often provides improved classification performance. Moreover, the results also demonstrate that, especially in the small labeled sample setting, a flexible classifier like supervised 1NN is often outperformed by the constrained LDA (even though the sampling used in the experiments favors the 1NN classifier!). Irrespective of the convincing results obtained by the constrained LDA, from a conceptual point of view, we are not there yet. Also our current version of the semi-supervised constrained LDA occasionally leads to deteriorations in performance with more unlabeled data, however small. Two recent findings pertain to this discussion and are elaborated on in the following two paragraphs.

In Loog and Duin (2012), the counterintuitive observation was made that there are problems on which particular classifiers attain, in expectation, their optimum error rate at a training set size which is *finite*. That is, for the classifier to be trained optimally for these classification tasks, one should not necessarily use all the labeled data available; using less might lead to an improvement. For this unexpected phenomenon to potentially arise, it seems two ingredients are needed. Firstly, it is most noticeable in case the classification model is clearly misaligned with the actual problem considered. Secondly, the learning algorithm should not minimize the actual classification error but some surrogate loss. The latter requirement is virtually always fulfilled. The level to which the former is met depends, among others, on the flexibility of the classification method. As an example, Loog and Duin (2012) present a simple, artificial problem on which NMC, LDA, Fisher's discriminant, linear SVM, logistic regression, and probably most other linear classifiers "dip". It seems plausible that similarly counterintuitive behavior may occasionally arise in the case of semi-supervised learning when increasing the numbers of unlabeled data points, even for models that do not rely on additional assumptions, such as the constrained LDA.

Loog and Jensen (submitted for publication) have demonstrated that the original semi-supervised NMC approach from Loog (2010) can be cast into a constrained log-likelihood formulation. As also the regular supervised version of NMC can be understood in terms of a model optimized under log-likelihood, this suggests that we can not only compare the supervised and semi-supervised NMC based on error rates but also in terms of the logarithmic loss on the test set. Following up on this, Loog and Jensen (submitted for publication) show empirically that measuring the performance of this classifier in terms of logarithmic loss, one does get the monotonic learning curve one expects, even when the corresponding curve for the error rate shows very irregular behavior. Clearly, one cannot compare classifiers optimizing different losses in this way. A supervised and a semi-supervised version of one and the same classifier, however, could be readily compared. Regarding our semi-supervised constrained LDA, should we try to understand what objective function is actually being optimized? Is it anyway perfectly fine to compare the constrained LDA with regular LDA through the log-likelihood? Or should we take this insight as a reason to abandon the ad hoc constrained approach advocated in this paper and proceed to investigate different approaches (cf. Loog and Jensen, 2012)?

On a rather different note, it is of interest to consider the earlier-mentioned idea of Fan et al. (2009) from the perspective of our constrained LDA. Although the idea from Fan et al. (2009) is not directly applicable in the classifier setting, it is possible to directly employ our constrained estimates of the within and between class covariance matrices to form a semi-supervised dimensionality reduction technique based on the Fisher criterion. It is of interest to compare our proposal to the one from Fan et al. (2009) and investigate to what extent they lead to different subspaces and what the consequence of this is for subsequent classification schemes.

Winding up, the one important general point we conveyed is that it is possible to perform semi-supervised LDA without making additional, extrinsic assumptions on the characteristics of the data distribution but by exploiting certain intrinsic regularities of the classification scheme considered. The principle of linking label-dependent and label-independent estimates is of course more broadly applicable. The problem, however, is that it is not directly clear which constraints to apply when dealing with many of the other decision rules. One of the primary open issues therefore is if there is a general principle of constructing and applying constraints leading to intrinsically motivated semi-supervised learning for a much broader class of classifiers. One obvious direction to start in would be to study kernelized and flexibilized LDA (Hastie et al., 2001) for which the current work seems of direct relevance.

Acknowledgements

The reviewers are sincerely acknowledged for their critical appraisal of this work. Their comments, remarks, and suggestions not only made this paper considerably longer but also significantly more interesting and convincing. I also would like to warmly thank Are Jensen with whom I had some rather insightful discussions on semi-supervision over the past few years. He really helped me shape my thoughts on this subject, some of which can be found back in this article.

References

Abney, S., 2004. Understanding the Yarowsky algorithm. Computational Linguistics 30 (3), 365–395.
Asuncion, A., Newman, D., 2007. UCI machine learning repository. <http://www.ics.uci.edu/mlearn/MLRepository.html>.
Basu, S., Banerjee, A., Mooney, R., 2002. Semi-supervised clustering by seeding. In: Proceedings of the Nineteenth International Conference on Machine Learning. pp. 19–26.
Ben-David, S., Lu, T., Pál, D., 2008. Does unlabeled data provably help? Worst-case analysis of the sample complexity of semi-supervised learning. In: COLT 2008. pp. 33–44.
Bengio, Y., Delalleau, O., Le Roux, N., 2006. Label propagation and quadratic criterion. In: Semi-Supervised Learning. MIT Press (Chapter 11).
Chapelle, O., Schölkopf, B., Zien, A., 2006. Semi-Supervised Learning. MIT Press, Cambridge, MA.
Cohen, I., Cozman, F., Sebe, N., Cirelo, M., Huang, T., 2004. Semisupervised learning of classifiers: theory, algorithms, and their application to human-computer interaction. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1553–1567.
Cozman, F., Cohen, I., 2006. Risks of semi-supervised learning. In: Semi-Supervised Learning. MIT Press (Chapter 4).
Culp, M., Michailidis, G., 2008. An iterative algorithm for extending learners to a semi-supervised setting. Journal of Computational and Graphical Statistics 17 (3), 545–571.
Dempster, A., Laird, N., Rubin, D., 1977. Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society: Series B (Methodological) 39 (1), 1–38.
Duda, R., Hart, P., 1973. Pattern Classification and Scene Analysis. John Wiley & Sons.
Duin, R., Loog, M., Pekalska, E., Tax, D., 2010. Feature-based dissimilarity space classification. LNCS, vol. 6388. Springer, pp. 46–55.
Efron, B., 2001. [Statistical modeling: the two cultures]: Comment. Statistical Science, 218–219.
Fan, B., Lei, Z., Li, S., 2009. Normalized LDA for semi-supervised learning. In: 8th IEEE International Conference on Automatic Face & Gesture Recognition. IEEE, pp. 1–6.
Fukunaga, K., 1990. Introduction to Statistical Pattern Recognition. Academic Press.
Grandvalet, Y., Bengio, Y., 2005. Semi-supervised learning by entropy minimization. In: Advances in Neural Information Processing Systems, vol. 17. MIT Press, pp. 529–536.
Hand, D., 2006. Classifier technology and the illusion of progress. Statistical Science 21 (1), 1–14.
Hartley, H., Rao, J., 1968. Classification and estimation in analysis of variance problems. Review of the International Statistical Institute 36 (2), 141–147.
Hastie, T., Tibshirani, R., Friedman, J., 2001. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer Verlag.
Joachims, T., 1999. Transductive inference for text classification using support vector machines. In: Proceedings of the Sixteenth International Conference on Machine Learning. pp. 200–209.

Liu, Q., Sung, A., Chen, Z., Liu, J., Huang, X., Deng, Y., 2009a. Feature selection and classification of MAQC-II breast cancer and multiple myeloma microarray gene expression data. PLoS ONE 4 (12), e8250.

Liu, W., Laitinen, S., Khan, S., Vihinen, M., Kowalski, J., Yu, G., Chen, L., Ewing, C., Eisenberger, M., Carducci, M., Nelson, W., Yegnasubramanian, S., Luo, J., Wang, Y., Xu, J., Isaacs, W., Visakorpi, T., Bova, G., 2009b. Copy number analysis indicates monoclonal origin of lethal metastatic prostate cancer. Nature Medicine 15 (5), 559–565.

Loog, M., 2007. A complete characterization of a family of solutions to a generalized fisher criterion. Journal of Machine Learning Research 8 (9), 2121–2123.

Loog, M., 2010. Constrained parameter estimation for semi-supervised learning: the case of the nearest mean classifier. In: Proceedings of ECML PKDD. LNAI. Springer, pp. 291–304.

Loog, M., 2012. Semi-supervised linear discriminant analysis using moment constraints. In: Partially Supervised Learning. LNAI, vol. 7081. Springer, pp. 32–41.

Loog, M., Duin, R., 2004. Linear dimensionality reduction via a heteroscedastic extension of LDA: the Chernoff criterion. IEEE Transactions on Pattern Analysis and Machine Intelligence 26 (6), 732–739.

Loog, M., Duin, R., 2012. The dipping phenomenon. In: Structural, Syntactic, and Statistical Pattern Recognition. LNCS, vol. 7626. Springer, pp. 310–317.

Loog, M., Jensen, A., 2012. Constrained log-likelihood-based semi-supervised linear discriminant analysis. In: Structural, Syntactic, and Statistical Pattern Recognition. LNCS, vol. 7626. Springer, pp. 327–335.

Loog, M., Jensen, A., submitted for publication. Semi-supervised nearest mean classification through a constrained log-likelihood.

Mann, G., McCallum, A., 2010. Generalized expectation criteria for semi-supervised learning with weakly labeled data. The Journal of Machine Learning Research 11, 955–984.

McLachlan, G., 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.

McLachlan, G., 1977. Estimating the linear discriminant function from initial samples containing a small number of unclassified observations. Journal of the American Statistical Association 72 (358), 403–406.

McLachlan, G., 1992. Discriminant Analysis and Statistical Pattern Recognition. John Wiley & Sons.

McLachlan, G., Ganesalingam, S., 1982. Updating a discriminant function on the basis of unclassified data. Communications in Statistics – Simulation and Computation 11 (6), 753–767.

Mira, A., Isella, C., Renzulli, T., Cantarella, D., Martelli, M., Medico, E., 2009. The GAB2 signaling scaffold promotes anchorage independence and drives a transcriptional response associated with metastatic progression of breast cancer. Oncogene 28 (50), 4444–4455.

Ng, A., Jordan, M., 2002. On discriminative vs. generative classifiers: a comparison of logistic regression and naive Bayes. In: Advances in Neural Information Processing Systems, vol. 14. MIT Press, pp. 841–848.

Nigam, K., McCallum, A., Thrun, S., Mitchell, T., 1998. Learning to classify text from labeled and unlabeled documents. In: Proceedings of the Fifteenth National Conference on Artificial Intelligence. pp. 792–799.

Ripley, B., 1996. Pattern Recognition and Neural Networks. Cambridge Univiversity Press.

Roepman, P., Jassem, J., Smit, E., Muley, T., Niklinski, J., van de Velde, T., Witteveen, A., Rzyman, W., Floore, A., Burgers, S., Giaccone, G., Meister, M., Dienemann, H., Skrzypski, M., Kozlowski, M., Mooi, W., van Zandwijk, N., 2009. An immune response enriched 72-gene prognostic profile for early-stage non-small-cell lung cancer. Clinical Cancer Research 15 (1), 284.

Salazar, R., Roepman, P., Capella, G., Moreno, V., Simon, I., Dreezen, C., Lopez-Doriga, A., Santos, C., Marijnen, C., Westerga, J., et al., 2011. Gene expression signature to improve prognosis prediction of stage II and III colorectal cancer. Journal of Clinical Oncology 29 (1), 17–24.

Singh, A., Nowak, R., Zhu, X., 2009. Unlabeled data: now it helps, now it doesn't. In: Advances in Neural Information Processing Systems, vol. 21. MIT Press, pp. 1513–1520.

Szummer, M., Jaakkola, T., 2002. Partially labeled classification with Markov random walks. In: Advances in Neural Information Processing Systems, vol. 14. MIT Press, pp. 945–952.

Titterington, D., 1976. Updating a diagnostic system using unconfirmed cases. Journal of the Royal Statistical Society: Series C (Applied Statistics) 25 (3), 238–247.

Vittaut, J., Amini, M., Gallinari, P., 2002. Learning classification with both labeled and unlabeled data. In: Machine Learning: ECML 2002. pp. 69–78.

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.

Zhou, Z., Li, M., 2010. Semi-supervised learning by disagreement. Knowledge and Information Systems 24 (3), 415–439.

Zhu, X., Ghahramani, Z., 2002. Learning from labeled and unlabeled data with label propagation. Technical Report CMU-CALD-02-107, Carnegie Mellon University.

Zhu, X., Goldberg, A., 2009. Introduction to Semi-Supervised Learning. Morgan & Claypool Publishers.