Pattern Recognition 63 (2017) 115–126

Contents lists available at ScienceDirect

Pattern Recognition

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

Robust semi-supervised least squares classification by implicit constraints

Jesse H. Krijthe [a,b,*], Marco Loog [a,c]

[a] Pattern Recognition Laboratory, Delft University of Technology, Mekelweg 4, 2628CD Delft, The Netherlands
[b] Department of Molecular Epidemiology, Leiden University Medical Center, Einthovenweg 20, 2333ZC Leiden, The Netherlands
[c] Image Group, University of Copenhagen, Universitetsparken 5, DK-2100 Copenhagen, Denmark

ARTICLE INFO

Article history:
Received 14 December 2015
Received in revised form
12 July 2016
Accepted 18 September 2016
Available online 20 September 2016

Keywords:
Semi-supervised learning
Robust
Least squares classification

ABSTRACT

We introduce the implicitly constrained least squares (ICLS) classifier, a novel semi-supervised version of the least squares classifier. This classifier minimizes the squared loss on the labeled data among the set of parameters implied by all possible labelings of the unlabeled data. Unlike other discriminative semi-supervised methods, this approach does not introduce explicit additional assumptions into the objective function, but leverages implicit assumptions already present in the choice of the supervised least squares classifier. This method can be formulated as a quadratic programming problem and its solution can be found using a simple gradient descent procedure. We prove that, in a limited 1-dimensional setting, this approach never leads to performance worse than the supervised classifier. Experimental results show that also in the general multidimensional case performance improvements can be expected, both in terms of the squared loss that is intrinsic to the classifier and in terms of the expected classification error.

1. Introduction

We consider the problem of semi-supervised learning of binary classification functions. As in the supervised paradigm, the goal in semi-supervised learning is to construct a classification rule that maps objects in some input space to a target outcome, such that future objects map to correct target outcomes as well as possible. In the supervised paradigm this mapping is learned using a set of L training objects and their corresponding outputs. In the semi-supervised scenario we are given an additional and often large set of U unlabeled objects. The challenge of semi-supervised learning is to incorporate this additional information to improve the classification rule.

The goal of this work is to build a semi-supervised version of the least squares classifier that is robust against deterioration in performance meaning that, at least in expectation, its performance is not worse than supervised least squares classification. While it may seem like an obvious requirement for any semi-supervised method, current approaches to semi-supervised learning do not have this property. In fact, performance can significantly degrade as more unlabeled data is added, as has been shown in [1,2], among others. This makes it difficult to apply these methods in practice, especially when there is a small amount of labeled data to identify possible reduction in performance. A useful property of any semi-supervised learning procedure would therefore be that its performance does not degrade as we add more unlabeled data. Additionally, many semi-supervised learning procedures are formulated as hard-to-optimize, non-convex objective functions. A more satisfactory state of affairs for semi-supervised classification would therefore be methods that are easier to train and that, on average, do not lead to worse classification performance than their supervised alternatives.

We present a novel approach to semi-supervised learning for the least squares classifier that we will refer to as implicitly constrained least squares classification (ICLS). ICLS leverages implicit assumptions present in the supervised least squares classifier to construct a semi-supervised version. This is done by minimizing the supervised loss function subject to the constraint that the solution has to correspond to the solution of the least squares classifier for some labeling of the unlabeled objects.

As this work is specifically concerned with least squares classification, we note several reasons why this is a particularly interesting classifier to study: first of all, the least squares classifier is a discriminative classifier. Some have claimed semi-supervised learning without additional assumptions is impossible for discriminative classifiers [3,4]. Our results show that this does not strictly hold.

Secondly, the closed-form solution for the supervised least squares classifier allows us to study its theoretical properties. In

* Corresponding author at: Pattern Recognition Laboratory, Delft University of Technology, Mekelweg 4, 2628CD Delft, The Netherlands.
E-mail addresses: jkrijthe@gmail.com (J.H. Krijthe), m.loog@tudelft.nl (M. Loog).

http://dx.doi.org/10.1016/j.patcog.2016.09.009
0031-3203/

particular, in the univariate setting without intercept and assuming perfect knowledge of P_X, the distribution of the feature, we show that this procedure *never* gives worse performance in terms of the squared loss criterion compared to the supervised least squares classifier. Moreover, using the closed-form solution we can rewrite our semi-supervised approach as a quadratic programming problem, which can be solved through a simple gradient descent with boundary constraints.

Lastly, least squares classification is a useful and adaptable classification technique allowing for straightforward use of, for instance, regularization, sparsity penalties or kernelization [5–9]. Using these formulations, it has been shown to be competitive with state-of-the-art methods based on loss functions other than the squared loss [7] as well as computationally efficient on large datasets [10].

This work builds on [11] and offers a more complete exposition: we show ICLS can be formulated as a quadratic programming problem, we extend the experimental results section by including an alternative semi-supervised procedure, adding additional datasets and discussing the 'peaking' phenomenon. Moreover, we extend the theoretical result with conditions when one is likely to see improvement of the proposed approach over the supervised classifier.

The main contributions of this paper are

- A novel convex formulation for robust semi-supervised learning using squared loss (Eq. (5)).
- A proof that this procedure never reduces performance in terms of the squared loss for the 1-dimensional case without intercept (Theorem 1).
- An empirical evaluation of the properties of this classifier (Section 6).

The rest of this paper is organized as follows. Section 2 gives an overview of related work on semi-supervised learning. Section 3 gives a high level overview of the method while Section 4 introduces our semi-supervised version of the least squares classifier in more detail. We then derive a quadratic programming formulation and present a simple way to solve this problem through bounded gradient descent. Section 5 contains a proof of the improvement of the ICLS classifier over the supervised alternative. This proof is specific to classification with a single feature, without including an intercept in the model. For the multivariate case, we present an empirical evaluation of the proposed approach on benchmark datasets in Section 6 to study its properties. The final sections discuss the results and conclude.

2. Related work

Many diverse approaches to semi-supervised learning have been proposed [12,13]. While semi-supervised techniques have shown promise in some applications, such as document classification [14], peptide identification [15] and cancer recurrence prediction [16], it has also been observed that these techniques may give performance worse than their supervised counterparts. See for instance [1,2], for an analysis of this problem, and [17] for a practical example in part-of-speech tagging. In these cases, disregarding the unlabeled data would lead to better performance.

Some [18,19] have argued that *agnostic* semi-supervised learning, which [18] defines as semi-supervised learning that is at least no worse than supervised learning, can be achieved by cross-validation on the limited labeled data. Agnostic semi-supervised learning follows if we only use semi-supervised methods when their estimated cross-validation error is significantly lower than those of the supervised alternatives. As the results of [18] indicate, this criterion may be too conservative: given the small amount of labeled data, a semi-supervised method will only be preferred if the difference in performance is very large. If the difference is less distinct, the supervised learner will always be preferred and we potentially ignore useful information from the unlabeled objects. Moreover, this cross-validation approach can be computationally demanding.

2.1. Self-learning

A simple approach to semi-supervised learning is offered by the self-learning procedure [20] also known as Yarowsky's algorithm [21,22] or retagging [17]. Taking any classifier, we first estimate its parameters on only the labeled data. Using this trained classifier we label the unlabeled objects and add them, or potentially only those we are most confident about, with their predicted labels to the labeled training set. The classifier parameters are re-estimated using these labeled objects to get a new classifier. One iteratively applies this procedure until the predicted labels of the unlabeled data no longer change.

One of the advantages of this procedure is that it can be applied to any supervised classifier. It has also shown practical success in some application domains, particularly document classification [14,22]. Unfortunately, the process of self-training can also lead to severely decreased performance, compared to the supervised solution [1,2]. One can imagine that once an object is incorrectly labeled and added to the training set, its incorrect label may be reinforced, leading the solution away from the optimum. Self-learning is closely related to expectation maximization (EM) based approaches [21]. Indeed, expectation maximization suffers from the same issues as self-learning [13]. In Section 6 we compare the proposed approach to self-learning for the least squares classifier.

2.2. Additional assumptions

Some semi-supervised methods leverage the unlabeled data by introducing assumptions that link properties of the features alone to properties of the label of an object given its features. Commonly used assumptions are the smoothness assumption: objects that are close in the feature space likely share the same label; the cluster assumption: objects in the same cluster share a label; and the low density assumption enforcing that the decision boundary should be in a region of low data density.

The low-density assumption is used in entropy regularization [23] as well as for support vector classification in the transductive support vector machine (TSVM) [24] and closely related semi-supervised SVM (S^3VM) [25,26]. In these approaches an additional term is added to the objective function to push the decision boundary away from regions of high density. Several approaches have been put forth to minimize the resulting non-convex objective function, such as the convex concave procedure [27] and difference convex programming [26,28].

In all these approaches to semi-supervised learning, a parameter controls the importance of the unlabeled points. When the parameter is correctly set, it is clear, as [19] claims, that TSVM is always no worse than supervised SVM. It is, however, non-trivial to choose this parameter, given that semi-supervised learning is most interesting in cases where we have limited labeled objects, making a choice using cross-validation very unstable. In practice, therefore, TSVM can also lead to performance worse than the supervised support vector machine, as well will also see in Section 6.3.

2.3. Safe semi-supervised learning

Refs. [29,30] attempt to guard against the possibility of deterioration in performance by not introducing additional assumptions,