Pattern Recognition ∎ (∎∎∎∎) ∎∎∎–∎∎∎

Contents lists available at ScienceDirect

Pattern Recognition

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

Learning kernel logistic regression in the presence of class label noise

Jakramate Bootkrajang*, Ata Kabán

School of Computer Science, The University of Birmingham, Birmingham B15 2TT, UK

ARTICLE INFO

Article history:
Received 20 October 2012
Received in revised form
8 May 2014
Accepted 11 May 2014

Keywords:
Classification
Label noise
Model selection
Multiple Kernel Learning

ABSTRACT

The classical machinery of supervised learning machines relies on a correct set of training labels. Unfortunately, there is no guarantee that all of the labels are correct. Labelling errors are increasingly noticeable in today's classification tasks, as the scale and difficulty of these tasks increases so much that perfect label assignment becomes nearly impossible. Several algorithms have been proposed to alleviate the problem of which a robust Kernel Fisher Discriminant is a successful example. However, for classification, discriminative models are of primary interest, and rather curiously, the very few existing label-robust discriminative classifiers are limited to linear problems.

In this paper, we build on the widely used and successful kernelising technique to introduce a label-noise robust Kernel Logistic Regression classifier. The main difficulty that we need to bypass is how to determine the model complexity parameters when no trusted validation set is available. We propose to adapt the Multiple Kernel Learning approach for this new purpose, together with a Bayesian regularisation scheme. Empirical results on 13 benchmark data sets and two real-world applications demonstrate the success of our approach.

1. Introduction

Traditional supervised learning machines rely on a correct set of class labels. There is however no guarantee that all the labels will be correct in practice, either due to the scale of the labelling task, the lack of information available to determine the class labels or the subjectivity of the labelling experts.

The presence of class label noise inherent in training samples has been reported to deteriorate the performance of the existing classifiers in a broad range of classification problems including biomedical data analysis [20,30] and image classification [24,47]. More recently, class label noise emerges as a side effect of crowdsourcing practices where annotators of different backgrounds are asked to perform labelling tasks. For example Amazon's Mechanical Turk, Citizen science, Galaxy Zoo to name just a few. Although, the problem posed by the presence of class label noise is acknowledged in the literature, it is often naively ignored in practice. Part of the reason for this may be that uniform/symmetric label noise is *relatively* harmless [21,22,12,27].

There is an increasing research literature that aims to address the issues related to learning from samples with noisy class label assignments. The seemingly straightforward approach is by means of data preprocessing where any suspect samples are removed or relabelled [7,1,29,37,31,18]. However, these approaches hold the risk of removing useful data too, which is detrimental to the classification performance, especially when the number of training examples is limited (e.g. in biomedical domains). Most previous approaches try to detect mislabelled instances based on various heuristics, and very few take a principled modelling approach with the notable exceptions of [32,24,25,36].

Lawrence and Scholkopf [24] incorporated a probabilistic model of random label flipping into their robust Kernel Fisher Discriminant (rKFD) for binary classification. Based on the same model, Li et al. [25] conducted extensive experiments on more complex data sets, which convincingly demonstrated the value of explicit modelling. The rKFD was later extended to multi-class setting by [3] and this has further motivated the recent development of a label noise-tolerant Hidden Markov Model to improve segmentation [15].

While all these works demonstrate the great potential and flexibility of a model based approach, most existing work falls in the category of generative methods. For classification problems, discriminative methods are of interest, and similar algorithmic

Abbreviation: LR, Logistic Regression; rLR, robust Logistic Regression; KLR, Kernel Logistic Regression; rKLR, robust Kernel Logistic Regression; rMKLR, robust Multiple Kernel Logistic Regression

* Corresponding author. Present address: Department of Computer Science, Chiang Mai University, Chiang Mai 50200, Thailand.

E-mail addresses: jakramate.b@cmu.ac.th (J. Bootkrajang), A.Kaban@cs.bham.ac.uk (A. Kabán).

http://dx.doi.org/10.1016/j.patcog.2014.05.007
0031-3203/© 2014 Elsevier Ltd. All rights reserved.

Nomenclature

$\mathcal{D}$	a data set
y	true label
$\hat{y}$	predicted label
m	dimensionality of data
κ	a kernel
$\mathbf{w}$	logistic regression parameter vector
$\boldsymbol{\eta}$	kernel combination coefficient vector
Ω	label flipping probability matrix
$\mathbf{x}$	a data point
$\tilde{y}$	observed label
N	number of data points
K	number of classes
S	number of kernels
ζ	regularisation on $\mathbf{w}$
ξ	regularisation on $\boldsymbol{\eta}$
ω_{jk}	element of Ω

developments for discriminative classifiers are still limited. For example, Madger et al. [28] studied logistic regression with known label flip probabilities and they reckon problems when these probabilities are unknown. Hausman et al. [17] have given a foundation of a statistical model for the binary classification problem but provide no algorithmic solution to the learning of label noise parameters.

Recently Raykar et al. [36] proposed an EM algorithm to learn a latent variable model extension of logistic regression, for data with multiple sets of noisy labels. Our initial work [4] suggested a more efficient gradient-based algorithm to optimise a similar latent variable model for problems where only a single set of labels is available. A sparse extension of the model has also been developed in [4]. However all of these developments are limited to linear problems. In this paper we focus on non-linear classification with labelling errors which is not as trivial as it might look at first.

Since the introduction of the kernel trick, many linear classifiers have been harnessed with an ability to solve non-linear problems, whereby their usage extends to a wider range of applications. Generally, deploying a kernel machine also involves determining good kernel parameters, and Cross-Validation (CV) has long been an established standard approach. However, when class label noise is present, it becomes unclear why would CV be a good approach since then all candidate models will be validated against noisy class labels. The issue has also been briefly discussed in [24,6]. In [24], the authors resort to using a 'trusted validation set' to select optimal kernel parameters. The trusted set must be labelled carefully, which seriously restricts the applicability of the method. For example in crowdsourcing it would be very difficult (if not impossible) to construct such a trusted set.

We start by straightforwardly formulating a robust Kernel Logistic Regression (rKLR) as an extension of the robust Logistic Regression (rLR). We present a simple yet effective algorithm to learn the classifier and investigate whether or not CV is a reasonable approach for model selection in the presence of labelling errors. As we shall see, we find that performing CV in noisy environments gives rise to a slightly under-fitted model. We then propose a robust Multiple Kernel Logistic Regression algorithm (rMKLR) based on the so-called Multiple Kernel Learning (MKL) framework (an extensive survey in recent advances of MKL is given in [16]) and the Bayesian regularisation technique [9] to automate the model selection step without using any cross-validation. From this we obtain improvements in both generalisation performance and learning speed. The genealogy of the proposed methods is summarised in Fig. 1, which serves as a roadmap for the next section.

Throughout this work, similar to the related work above, we will focus on label noise occurring at random – the flipping of labels is assumed to be independent of the contents of the data features. The reason for this is simplicity and generic applicability. Alternative models of label noise are discussed after the Experiments section.

2. Robust kernel logistic regression

Consider a set of training samples $\mathcal{D}=\{(\mathbf{x}_n,\tilde{y}_n)\}_{n=1}^N$, where $\mathbf{x}_n\in\mathbb{R}^m$ and $\tilde{y}_n\in\{0,1\}$ denotes the observed (possibly noisy) label of $\mathbf{x}_n$. Kernel logistic regression produces a non-linear decision boundary, $f(\mathbf{x})$, by forming a linear decision boundary in the space of the non-linearly transformed input vectors. By the representer theorem [19], the optimal $f(\mathbf{x})$ has the form

$$f(\mathbf{x})=\sum_{n=1}^{N} w_n\kappa(\cdot,\mathbf{x}_n) \tag{1}$$

where $\kappa(\cdot,\cdot)$ is a positive definite reproducing kernel that gives an inner product in the transformed space.

Denoting by $\mathbf{w}$ the parameter vector with entries $w_n, n=1,\dots,N$, we define the probability of an observed label $\tilde{y}_n$ as a linear combination of the probabilities that the true label of a point is 0 or 1:

$$\begin{aligned} p(\tilde{y}=k|\kappa(\cdot,\mathbf{x}_n),\mathbf{w}) &= \sum_{j=0}^{1} p(\tilde{y}=k|y=j)p(y=j|\kappa(\cdot,\mathbf{x}_n),\mathbf{w}) \\ &= \sum_{j=0}^{1} \omega_{jk}p(y=j|\kappa(\cdot,\mathbf{x}_n),\mathbf{w}) \end{aligned} \tag{2}$$

Here, $p(\tilde{y}=k|y=j)=\omega_{jk}$ are probabilistic factors representing the probability that the true label j flips into the observed label k. These parameters form a label transition table, Ω, that we will refer to as the *flip matrix*. The full set of parameters for this robust model will be denoted as $\Theta=\{\mathbf{w},\Omega\}$. Now, fitting the robust kernel logistic regression is equivalent to maximising the following log-likelihood:

$$\mathcal{L}(\Theta)=\sum_{n=1}^{N}\sum_{k=0}^{1}\mathbb{1}(\tilde{y}_n=k)\log\, p(\tilde{y}_n=k|\kappa(\cdot,\mathbf{x}_n),\Theta)-\zeta\sum_{n=1}^{N}w_n^2 \tag{3}$$

where $\mathbb{1}(\cdot)$ is the Kronecker delta function. We also included an L 2 regularisation term to express our preference for a smooth (and non-sparse) model.

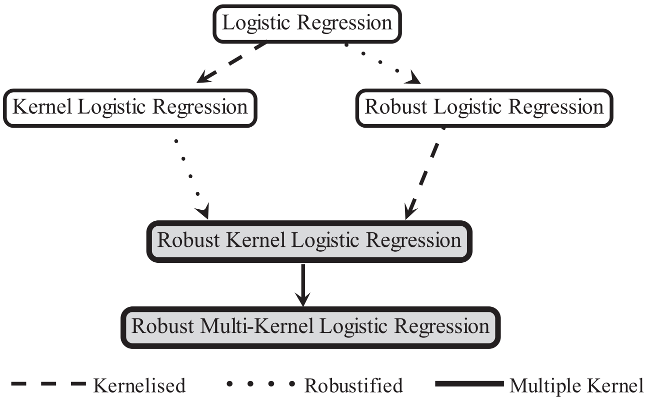
Logistic Regression
Kernel Logistic Regression
Robust Logistic Regression
Robust Kernel Logistic Regression
Robust Multi-Kernel Logistic Regression
Kernelised
Robustified
Multiple Kernel

Fig. 1. Genealogy of the robust Kernel Logistic Regression and the robust Multi-Kernel Logistic Regression methods. The highlighted boxes are the classifiers proposed in this paper. Note that there are two paths to arrive at the robust Kernel Logistic Regression.

Please cite this article as: J. Bootkrajang, A. Kabán, Learning kernel logistic regression in the presence of class label noise, Pattern Recognition (2014), http://dx.doi.org/10.1016/j.patcog.2014.05.007