Neural Networks 84 (2016) 57–66

Contents lists available at ScienceDirect

Neural Networks

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

Semi-supervised learning for ordinal Kernel Discriminant Analysis

M. Pérez-Ortiz [a,*], P.A. Gutiérrez [b], M. Carbonero-Ruz [a], C. Hervás-Martínez [b]

[a] Department of Quantitative Methods, Universidad Loyola Andalucía, 14004 - Córdoba, Spain

[b] Department of Computer Science and Numerical Analysis, University of Córdoba, 14070 - Córdoba, Spain

ARTICLE INFO

Article history:
Received 2 February 2016
Received in revised form 9 August 2016
Accepted 15 August 2016
Available online 25 August 2016

Keywords:
Ordinal regression
Discriminant analysis
Semi-supervised learning
Classification
Kernel learning

ABSTRACT

Ordinal classification considers those classification problems where the labels of the variable to predict follow a given order. Naturally, labelled data is scarce or difficult to obtain in this type of problems because, in many cases, ordinal labels are given by a user or expert (e.g. in recommendation systems). Firstly, this paper develops a new strategy for ordinal classification where both labelled and unlabelled data are used in the model construction step (a scheme which is referred to as semi-supervised learning). More specifically, the ordinal version of kernel discriminant learning is extended for this setting considering the neighbourhood information of unlabelled data, which is proposed to be computed in the feature space induced by the kernel function. Secondly, a new method for semi-supervised kernel learning is devised in the context of ordinal classification, which is combined with our developed classification strategy to optimise the kernel parameters. The experiments conducted compare 6 different approaches for semi-supervised learning in the context of ordinal classification in a battery of 30 datasets, showing (1) the good synergy of the ordinal version of discriminant analysis and the use of unlabelled data and (2) the advantage of computing distances in the feature space induced by the kernel function.

1. Introduction

With the advent of the big data era and the increased popularity of machine learning, the number of scientific data-driven applications is growing at an abrupt pace. Because of this increased necessity, new related research avenues are explored every year. In this sense, the recently coined term weak supervision (Hernández-González, Inza, & Lozano, 2016) refers to those classification machine learning problems where the labelling information is not as accessible as in the fully-supervised problem (where a label is associated to each pattern). The problem of semi-supervised learning (i.e. learning from both labelled and unlabelled observations) is an example that has been the focus of many machine learning researchers in the past years. In many real-world applications, obtaining labelled patterns could be a challenging task, however, unlabelled examples might be available with little or no cost. The main idea behind semi-supervised learning is to take advantage from unlabelled data when constructing the machine classifier (and this is done using different assumptions on the unlabelled data: smoothness, clustering or manifold assumptions (Chapelle, Schölkopf, & Zien, 2010; Wang, Shen, & Pan, 2009; Zhu, 2005)). These learning approaches have been empirically and theoretically studied in the literature and represent a suitable solution for such circumstances, where the use of unlabelled data has been seen to improve the performance of the model and stabilise it. Semi-supervised learning has been mainly studied for binary classification (Cai, He, & Han, 2007; Ortigosa-Hernández, Inza, & Lozano, in press) and regression (Zhu, 2005), although recently the main focus has shifted to multi-class problems (Ortigosa-Hernández et al., in press; Soares, Chen, & Yao, 2012; Xu, Anagnostopoulos, & Wunsch, 2007) (and even multi-dimensional ones (Ortigosa-Hernández et al., 2012)). This paper tackles the use of unlabelled data in the context of ordinal classification (Gutiérrez, Pérez-Ortiz, Sánchez-Monedero, Fernández-Navarro, & Hervás-Martínez, 2016), a learning paradigm which shares properties of both classification and regression.

Ordinal regression (also known as ordinal classification) can be defined as a relatively new learning paradigm whose aim is to learn a prediction rule for ordered categories. In contrast to multinomial classification, there exists some ordering among the elements of $\mathcal{Y}$ (the labelling space) and both standard classifiers and the zero–one loss function do not capture and reflect this ordering appropriately (Gutiérrez et al., 2016) (leading to worse models in terms of errors in the ordinal scale). Concerning regression, $\mathcal{Y}$ is a non-metric space.

* Corresponding author.
E-mail addresses: i82perom@uco.es, mariaperez@uloyola.es (M. Pérez-Ortiz), pagutierrez@uco.es (P.A. Gutiérrez), mcarbonero@uloyola.es (M. Carbonero-Ruz), chervas@uco.es (C. Hervás-Martínez).

http://dx.doi.org/10.1016/j.neunet.2016.08.004
0893-6080/

An explanatory example of order among categories is the Likert scale (Likert, 0000), a well-known methodology used for questionnaires, where the categories correspond to the level of agreement or disagreement for a series of statements. The scheme of a typical five-point Likert scale could be: {*Strongly disagree, Disagree, Neither agree or disagree, Agree, Strongly Agree*}, where the natural order among categories can be appreciated. The major problem within this kind of classification is that the misclassification errors should not be treated equally, e.g., misclassifying a *Strongly disagree* pattern as *Strongly agree* should be more penalised than a misclassification with the *Disagree* category.

Several issues must be highlighted when developing new ordinal classifiers in order to exploit the presence of this order among categories. Firstly, this implicit data structure should be learned by the classifier in order to minimise the different ordinal classification errors (Gutiérrez et al., 2016), and, secondly, different evaluation measures or metrics should be developed in this context. The most popular approach for this type of problems is threshold models (Chu & Ghahramani, 2005; McCullagh & Nelder, 1989; Shashua & Levin, 2003; Sun, Li, Wu, Zhang, & Li, 2010). These methods are based on the idea that, to model ordinal ranking problems from a regression perspective, one can assume that some underlying real-valued outcomes exist (also known as latent variable), which are, in practice, unobservable.

Recently, a version of the well-known Kernel Discriminant Analysis algorithm has been proposed for ordinal regression (Sun et al., 2010), showing different advantages with respect to other ordinal classification methods, i.e. a lower computational complexity and the ability to capture the associated class distributions. In essence, the formulation seeks for the projection that allows the greater separation for the classes, but maintaining the classes ordered in the projection (to avoid serious misclassification errors). This algorithm, Kernel Discriminant Learning for Ordinal Regression (KDLOR), has shown great potential and competitiveness against other specially designed ordinal classifiers.

However, supervised ordinal regression approaches present limitations when there are few data (Srijith, Shevade, & Sundararajan, 2013; Wu, Sun, Liang, Tang, & Cai, 2015), which is a common situation in this setting, where most ordinal classification problems are labelled by a user or expert (a process that could be expensive or time-consuming), and the number of classes is usually relatively high (which hinders the class discrimination to a great extent). Consider, for example, the case of a film recommendation system, where most users might not have interest in labelling data, therefore unlabelled data exist and are easily available. In this sense, the paradigm of semi-supervised learning would use the unlabelled data along with the labelled data to learn more precise models. The development and analysis of semi-supervised ordinal regression algorithms is, therefore, of great interest. However, the number of works in the literature approaching this problem is very low (Liu, Liu, Zhong, & Chan, 2011; Seah, Tsang, & Ong, 2012; Srijith et al., 2013; Wu et al., 2015), where only two of them focus on developing ordinal and semi-supervised classifiers (Srijith et al., 2013; Wu et al., 2015) (the remainder focuses on related frameworks, such as the transductive problem (Liu et al., 2011; Seah et al., 2012) or clustering (Xiao, Liu, & Hao, 2016), which are out of the scope of this paper).

We propose and test different approaches to deal with semi-supervised ordinal classification problems. Firstly, we extend the KDLOR algorithm to make use of unlabelled data via the smoothness and manifold assumptions, (i.e. (1) points nearby are likely to share the same label, and (2) the projection should not only match the classification task but also respect the geometric structure inferred from labelled and unlabelled data points). Secondly, this paper proposes to compute the graph Laplacian (used for the previous objective) in the feature space induced by the kernel function, as opposed to computing it in the input space. Since the final objective function is computed in the feature space, this is a crucial consideration for the proposed technique. Finally, we also propose a new method for semi-supervised kernel learning based on kernel-target alignment to use in conjunction with (ordinal) kernel methods. Kernel learning techniques are a common choice to optimise the kernel parameters and adequately fit the data using a kernel function (Cortes, Mohri, & Rostamizadeh, 2012; Cristianini, Kandola, Elisseeff, & Shawe-Taylor, 2002). We test our proposals in a set of 30 ordinal classification datasets and compare them to other strategies, the results showing the good synergy of combining labelled and unlabelled data in the context of ordinal regression.

The rest of the paper is organised as follows: Section 2 shows a description of previous concepts; Section 3 presents the proposal of this work; Section 4 describes the specific characteristics of the datasets and the experimental study analyses the results obtained; and finally, Section 5 outlines some conclusions and future work.

2. Previous notions

This section introduces some of the previous work in the area of the paper.

Consider a training sample $D = \{\mathbf{x}_i, y_i\}_{i=1}^N \subseteq \mathcal{X} \times \mathcal{Y}$ generated i.i.d. from a (unknown) joint distribution $P(\mathbf{x}, y)$, where $\mathcal{X} \subseteq \mathbb{R}^d$ and $\mathcal{Y} = \{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_Q\}$. In the ordinal regression setup, the labelling space is ordered due to the data ranking structure ($\mathcal{C}_1 \prec \mathcal{C}_2 \prec \cdots \prec \mathcal{C}_Q$, where $\prec$ denotes this order information). Let N be the number of patterns in the training sample, N_q the number of samples for the qth class and $\mathbf{X}_q$ the set of patterns belonging to class $\mathcal{C}_q$.

Furthermore, let $\mathcal{H}$ denote a high-dimensional Hilbert space. Then, for any mapping of patterns $\Phi : \mathcal{X} \to \mathcal{H}$, the inner product $k(\mathbf{x}_i, \mathbf{x}_j) = \langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j) \rangle_{\mathcal{H}}$ of the mapped inputs is known as a kernel function, giving rise to a positive semidefinite (PSD) matrix $\mathbf{K}$ for a given input set $\{\mathbf{x}_i\}_{i=1}^N$.

2.1. Discriminant learning

This learning paradigm is one of the pioneers and leading techniques in the machine learning area, being currently used for supervised dimensionality reduction and classification. The main goal of this technique can be described as finding the optimal linear projection for the data (from which different classes can be well separated). To do so, the algorithm analyses two objectives: the maximisation of the between-class distance and the minimisation of the within-class distance, by using variance–covariance matrices ($\mathbf{S}_b$ and $\mathbf{S}_w$, respectively) and the so-called Rayleigh coefficient ($J(\mathbf{w}) = \frac{\mathbf{w}^{\mathrm{T}}\mathbf{S}_{\mathrm{b}}\mathbf{w}}{\mathbf{w}^{\mathrm{T}}\mathbf{S}_{\mathrm{w}}\mathbf{w}}$, where $\mathbf{w}$ is the projection for the data). To achieve these objectives, the $Q - 1$ eigenvectors associated to the highest eigenvalues of $\mathbf{S}_{\mathrm{w}}^{-1} \cdot \mathbf{S}_{\mathrm{b}}$ are computed.

The between-class and within-class scatter matrices ($\mathbf{S}_b$ and $\mathbf{S}_w$, respectively) are defined as follows (when considering the kernel version):

$$\mathbf{S}_{\mathrm{w}} = \frac{1}{N}\sum_{q=1}^{Q}\sum_{\mathbf{x}_i \in \mathbf{X}_q}(\Phi(\mathbf{x}_i) - \mathbf{M}_q^{\Phi})(\Phi(\mathbf{x}_i) - \mathbf{M}_q^{\Phi})^{\mathrm{T}}, \tag{1}$$

$$\mathbf{S}_{\mathrm{b}} = \frac{1}{N}\sum_{q=1}^{Q} N_q(\mathbf{M}_q^{\Phi} - \mathbf{M}^{\Phi})(\mathbf{M}_q^{\Phi} - \mathbf{M}^{\Phi})^{\mathrm{T}}, \tag{2}$$

where $\mathbf{M}_q^{\Phi} = \frac{1}{N_q}\sum_{\mathbf{x}_i \in \mathbf{X}_q}\Phi(\mathbf{x}_i)$, and $\mathbf{M}^{\Phi} = \frac{1}{N}\sum_{i=1}^{N}\Phi(\mathbf{x}_i)$. The objectives presented can be achieved by the maximisation of the