Neural Networks 59 (2014) 51–60

Contents lists available at ScienceDirect

Neural Networks

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

Ordinal regression neural networks based on concentric hyperspheres

Pedro Antonio Gutiérrez [a,*], Peter Tiňo [b], César Hervás-Martínez [a]

[a] University of Córdoba, Department of Computer Science and Numerical Analysis, Rabanales Campus, Albert Einstein building, 14071 - Córdoba, Spain

[b] School of Computer Science, The University of Birmingham, Birmingham B15 2TT, United Kingdom

ARTICLE INFO

Article history:
Received 7 February 2014
Received in revised form 15 May 2014
Accepted 7 July 2014
Available online 11 July 2014

Keywords:
Ordinal regression
Ordinal classification
Neural networks
Latent variable

ABSTRACT

Threshold models are one of the most common approaches for ordinal regression, based on projecting patterns to the real line and dividing this real line in consecutive intervals, one interval for each class. However, finding such one-dimensional projection can be too harsh an imposition for some datasets. This paper proposes a multidimensional latent space representation with the purpose of relaxing this projection, where the different classes are arranged based on concentric hyperspheres, each class containing the previous classes in the ordinal scale. The proposal is implemented through a neural network model, each dimension being a linear combination of a common set of basis functions. The model is compared to a nominal neural network, a neural network based on the proportional odds model and to other state-of-the-art ordinal regression methods for a total of 12 datasets. The proposed latent space shows an improvement on the two performance metrics considered, and the model based on the three-dimensional latent space obtains competitive performance when compared to the other methods.

1. Introduction

When we face an ordinal regression (OR) problem, the objective is to predict the label y_i of an input vector $\mathbf{x}_i$, where $\mathbf{x}_i \in \mathcal{X} \subseteq \mathbb{R}^k$ and $y_i \in \mathcal{Y} \in \{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_Q\}$. This is done by estimating a classification rule or function $F : \mathcal{X} \to \mathcal{Y}$ to predict the labels of new samples. In a supervised setting, we are given a training set of N points, $D = \{(\mathbf{x}_i, y_i), 1 \leq i \leq N\}$. All these considerations can be also found in standard nominal classification, but, for OR, a natural label ordering is included, which is given by $\mathcal{C}_1 \prec \mathcal{C}_2 \prec \cdots \prec \mathcal{C}_Q$. The symbol $\prec$ is an order relation representing the nature of the classification problem and expressing that a label is before another in the ordinal scale.

OR problems are very common in real settings, although the machine learning community has often treated them from a standard (nominal) perspective, ignoring the order relationship, $\prec$, between classes. Some examples of application fields where OR is found are credit rating (Dikkers & Rothkrantz, 2005), econometric modelling (Mathieson, 1996), medical research (Cardoso, da Costa, & Cardoso, 2005) or face recognition (Kim & Pavlovic, 2010), to name a few. Considering the order relationship between classes can result in two significant benefits: (1) minimisation of specific classification errors, and (2) incorporation of the ordering into the classifier. With respect to the first benefit, it is clear that one should focus on predicting categories as close as possible to the real one when tackling an OR problem. Hence, OR methods are aimed to minimise those errors that involve large category gaps in the ordinal scale. As an example, consider a tumour classification problem where the categories are {*benign, dangerous, malign*}. Misclassification of *malign* tumours as *dangerous* is preferred to assign the label *benign* to a *malign* tumour and OR methods will generally minimise this second type of errors. The second benefit comes from the fact that label order is usually present, in a direct way in the input space or through a latent space representation (Sánchez-Monedero, Gutiérrez, Tino, & Hervás-Martínez, 2013). Imbuing a classifier with this ordering will generally improve generalisation performance, as the classifier is better representing the nature of the task.

The field of OR has experienced significant development in the last decade, with many new methods adapted from traditional machine learning methodologies, from support vector machine (SVM) formulations (Chu & Keerthi, 2007) to Gaussian processes (Chu & Ghahramani, 2005) or discriminant learning (Sun, Li, Wu, Zhang, & Li, 2010). For all these methods, although classifier construction is motivated and undertaken from different points of view, the final models share a common structure or nature. They exploit the fact that it is natural to assume that an unobserved continuous variable underlies the ordinal response variable (e.g. the actual age of the person appearing in a picture

* Corresponding author. Tel.: +34 957218349; fax: +34 957218630.
E-mail addresses: pagutierrez@uco.es, zamarck@gmail.com (P.A. Gutiérrez).

http://dx.doi.org/10.1016/j.neunet.2014.07.001
0893-6080/

for an age classification problem). This variable is called latent variable, and methods based on that assumption are known as threshold models (Verwaeren, Waegeman, & De Baets, 2012). Indeed, this structure can be found in one of the first models for OR, the proportional odds model (POM) (McCullagh, 1980), which is a probabilistic model estimating the cumulative probabilities of the different ordered categories and leading to linear decision boundaries. Threshold models methodologies estimate:

- A function $f(\mathbf{x})$ that tries to predict the values of the latent variable.
- A set of thresholds $\mathbf{b} = (b_1, b_2, \ldots, b_{Q-1}) \in \mathbb{R}^{Q-1}$ to represent intervals in the range of $f(\mathbf{x})$, which must satisfy the constraints $b_1 \leq b_2 \leq \cdots \leq b_{Q-1}$.

From a practical perspective, threshold models are basically trying to find a one-dimensional projection ($f(\mathbf{x})$) where patterns are ordered according to the class labels. Finding this projection can be a problem for real world datasets. If we consider linear models for $f(\mathbf{x})$, the chances that the patterns exhibit a linear ordering relationship are certainly very low. If we consider nonlinear models, the pressure to find this nonlinear projection can result in unnatural or too complex projections leading to poorer generalisation performance. This paper proposes to relax this pressure by allowing a higher dimensional representation of the latent space. This is done by ordering patterns in an L-dimensional space, where each class region is limited by concentric hyperspheres (centred in the origin). The ordering of the classes is imposed by assuring that the radii of the hyperspheres are also ordered.

Another popular way to tackle OR problems is to decompose the original task into several binary tasks, where each binary task consists of predicting if the patterns belong to a category higher (in the ordinal scale) than a given label $\mathcal{C}_q$. One model is estimated for each class in the ordinal scale. The approach is presented in the work of Frank and Hall (2001), also proposing a way to fuse probabilities given for all binary tasks. Later on, there have been two different lines of research where binary classification and OR were linked in a more direct way (Cardoso & da Costa, 2007; Li & Lin, 2007; Lin & Li, 2012). Instead of learning Q different binary classifiers, a single binary classifier is learnt, where the category examined is included as an additional feature and training patterns are replicated and weighted. The framework in Li and Lin (2007) and Lin and Li (2012) is more generic, in the sense that it can be applied to different cost matrices.

The Error-Correcting Output Codes (ECOC) methodology is a popular and effective coding method to learn complex class targets, which can be used also for OR. The main idea is to associate each class $\mathcal{C}_q$ with a column of a binary coding matrix $\mathbf{M}_{R\times Q}$, where each entry of the matrix $\mathbf{M}(j, q) \in \{-1, +1\}$, Q is the number of classes, R is the number of binary classifiers, $1 \leq q \leq Q$ and $1 \leq j \leq R$. After training the binary classifiers, prediction is then accomplished by choosing the column of $\mathbf{M}$ closest to the set of decision values, where the distance function should be selected according to the error function minimised during learning (Allwein, Schapire, & Singer, 2001; Dietterich & Bakiri, 1995).

In the field of neural networks, there have been some proposals for OR problems. The first one dates back to 1996, when Mathieson proposed a non-linear version of the POM (Mathieson, 1996, 1999) by setting the projection $f(\mathbf{x})$ to be the output of a neural network. Although the results were quite promising, the method was evaluated for a very specific dataset. A more extensive battery of experiments should be done to further validate the proposal.

Costa (1996) derived another neural network architecture to exploit the ordinal nature of the data. It was based on a "partitive approach", where probabilities are assigned to the joint prediction of constrained concurrent events.

Other approach (Cheng, Wang, & Pollastri, 2008) applies the coding scheme of Frank and Hall and a decision rule based on examining output nodes with an order and selecting the first one whose output is higher than a predefined threshold T. The problem of this method is that inconsistencies can be found in the predictions (i.e. a sigmoid with value higher than T after the index selected).

The ordinal neural network (oNN) of Cardoso and da Costa (2007) adapts the previously discussed data replication method to neural networks (a single model for binary decomposition using an extended and replicated version of the dataset), allowing the derivation of nonlinear decision boundaries.

Additionally, extreme learning machines (ELMs) have been used as a very fast method to fit single layer neural networks, where the hidden neurons weights are random, and the output weights are analytically obtained (Huang, Zhou, Ding, & Zhang, 2012). They have been adapted to OR (Deng, Zheng, Lian, Chen, & Wang, 2010), considering again the Frank and Hall coding scheme and a prediction based on the ECOC loss-based decoding approach (Allwein et al., 2001), i.e. the chosen label is that which minimises the exponential loss. Another recent paper by Riccardi, Fernandez-Navarro, and Carloni (in press) introduces a cost-sensitive approach for adapting the stagewise additive modelling using a multiclass exponential boosting algorithm (SAMME, which is the multiclass version of the well-known AdaBoost) to OR problems. They consider ELMs as the base classifier and they introduce three different loss functions, affecting the update rule of the error estimation and/or of the pattern weights (Riccardi et al., in press). From the three variants introduced in the paper, the third one (which adapts the update rule of both the error estimation and the pattern weights) obtains the best results. The OR model proposed in Fernandez-Navarro, Riccardi, and Carloni (in press) adapts ELM to OR problems by imposing monotonicity constraints in the weights connecting the hidden layer with the output layer. The optimum of the inequality constrained least squares problem is determined analytically according to the closed-form solution estimated from the Karush–Kuhn–Tucker conditions.

A conceptually different methodology is proposed by da Costa, Alonso, and Cardoso (2008) and da Costa and Cardoso (2005) for training OR models, with a special attention to neural networks. They assume that the random variable class of a pattern should follow a unimodal distribution. Two possible implementations are considered: a parametric one, where a specific discrete distribution is assumed and the associated free parameters are estimated by a neural network; and a non-parametric one, where no distribution is assumed but the error function is modified to avoid errors from distant classes. Finally, the approach in Dobrska, Wang, and Blackburn (2012) is a distribution-independent methodology for OR based on pairwise preferences. The strength of dependency between two data instances (continuous preferences) is shown to improve algorithmic performance, obtaining competitive results.

In this paper, we extend the proposal of Mathieson (1996, 1999), deriving a nonlinear version of the POM based on neural networks. We present a common learning framework to fit the parameters of a nominal neural network (NNN) and the neural network based on the POM (POMNN). The framework is then used to fit an extended version of the POMNN, where, as previously discussed, the latent space is assumed to be L-dimensional and the patterns are ordered by considering $Q-1$ concentric hyperspheres. The underlying motivation is to relax the imposition of projecting all patterns in a real line.

With regard to the relationship between ECOC and the proposal of this paper, one single model is used for learning the ordinal target, and the problem is not decomposed into several binary ones. In this way, the latent space structure relates each pattern to the posterior probabilities without learning multiple binary classifiers.

This paper is organised as follows: Section 2 is devoted to a brief analysis of the POM model, closely related to the models proposed;