Computational Statistics and Data Analysis 82 (2015) 199–206

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Kernel multilogit algorithm for multiclass classification

Oscar Dalmau [a,*], Teresa E. Alarcón [b], Graciela González [a]

[a] Centro de Investigación en Matemáticas, A.C. Jalisco S/N Col. Valenciana, 36240, Guanajuato, Gto, Mexico

[b] Centro Universitario de los Valles, Universidad de Guadalajara, Carretera Guadalajara - Ameca Km. 45.5 C.P. 46600, Ameca, Jalisco, Mexico

ARTICLE INFO

Article history:
Received 19 April 2014
Received in revised form 2 September 2014
Accepted 2 September 2014
Available online 18 September 2014

Keywords:
Classification
Multilogit function
Linear model
Kernel

ABSTRACT

An algorithm for multi-class classification is proposed. The soft classification problem is considered, where the target variable is a multivariate random variable. The proposed algorithm transforms the original target variable into a new space using the multilogit function. Assuming Gaussian noise on this transformation and using a standard Bayesian approach the model yields a quadratic functional whose global minimum can easily be obtained by solving a set of linear system of equations. In order to obtain the classification, the inverse multilogit-based transformation should be applied and the obtained result can be interpreted as a 'soft' or probabilistic classification. Then, the final classification is obtained by using the 'Winner takes all' strategy. A Kernel-based formulation is presented in order to consider the non-linearities associated with the feature space of the data. The proposed algorithm is applied on real data, using databases available online. The experimental study shows that the algorithm is competitive with respect to other classical algorithms for multiclass classification.

1. Introduction

The classification problem is a machine learning and statistic task (Bishop, 2006; Hastie et al., 2003; Duda et al., 2001) that consists in inferring a function from a labeled training dataset. This function, also called classifier, allows one to identify or predict to which category a new observation belongs. The classification problem is an active area in machine learning for both theoretical and practical problems. Some well-known classification algorithms are: Linear classifiers (Fan et al., 2008), Quadratic classifiers (Hastie et al., 2003), Logistic regression (Ryali et al., 2010; Hastie et al., 2003), Naive Bayes classifier (Chandra and Gupta, 2011; Lu et al., 2006; Lopez-Cruz et al., 2013; Taheri et al., 2011), Support vector machines (Cuingnet et al., 2010, 2011) and Decision trees (Gray and Fan, 2008; Kim and Loh, 2001) among others.

In this work we present an algorithm for multiclass classification. The algorithm borrows concepts already used in some classification algorithms, such as Linear classifiers and Multilogistic regression (MLR), and puts together to achieve a good classifier. The main idea of the algorithm is to transform not only the feature vector but also the target or response variable. There is a recent work where the authors also transform the response variable, see Balasubramanian and Lebanon (2012), although in this case the aim is to select response variables. This approach is more similar to the classical variable selection problem (or feature selection). In particular, they propose a sparse linear model, which is estimated through an optimization problem. A similar formulation for variable selection and regression was proposed by Tibshirani in Tibshirani (1994). Our approach is more related to the 'kernel trick' technique in which a particular function (polynomial, Gaussian, … etc.) is

* Corresponding author. Tel.: +52 473 732 7155; fax: +52 473 732 5749.
E-mail addresses: osdalmau@gmail.com, dalmau@cimat.mx (O. Dalmau).

http://dx.doi.org/10.1016/j.csda.2014.09.007
0167-9473/

applied to the predictor variable. Our algorithm employs the *direct-multilogit* function to convert the response variable into a new space, unlike the MLR model that utilizes the *inverse-multilogit* function to model the class-conditional density function.

Our proposal tries to overcome the principal drawbacks of the MLR and of the Multiple Output Model (MOM) while taking advantages of the good characteristics of both methods. The main difference of the MLR model with respect to our proposal is that MLR directly models the original discrete target variable using the Maximum Likelihood Estimator approach what produces a non-linear optimization problem whereas our algorithm models a transformed target variable through a linear model, for some advantages of linear models see Hastie et al. (Hastie et al., 2003, Chapter 3). On the other hand, one of the principal limitation of the MOM model is that it can under or overestimate the response variable, i.e., the estimation could produce predictions outside the range [0, 1], (Caudill, 1988) which, according to the model, does not have sense, i.e., the estimation loses the probabilistic interpretation. In our case, the components of the transformed target variable are real numbers so our method does not have the same problematic and when we apply the inverse-multilogit transformation, the components of the estimated target variable belongs to [0, 1] and the sum is equal to 1 what provides us a probabilistic interpretation of the estimation.

The proposed algorithm assumes that for each feature vector one has a multivariate response variable whose components can be interpreted as the confidence of the data to belong to a class of a set of categories, i.e., a soft classification. A particular case of this assumption is taken in the classic classification problem, i.e., the 'hard classification', in which a data belongs to exactly one category. The algorithm consists of 3 steps. First, we transform the original target variable by using an invertible transformation in order to obtain a transformed target variable in a new space. Here we use the *multilogit* function although the algorithm accepts other transformations. In the second step, we model the transformed target variable given the original or a transformed feature vector. For modeling the transformed target variable, we propose a linear model what resembles the Multiple Output Model for classification (Bishop, 2006, Chapter 4), (Hastie et al., 2003, Chapter 3), (Duda et al., 2001, Chapter 5). The previous modeling allows us to predict the transformed variable for a new data. In a third step, the algorithm estimates the target variable in the original space.

The new algorithm is presented through a quadratic functional that should be minimized with respect to the parameters of the multilogit function. The global minimum of this functional can easily be obtained by solving a set of linear systems. In the present work, we show experimentally that the proposed algorithm is competitive with other classical algorithms for multiclass classification.

This paper is organized as follows. In Section 2 we present an overview of the *multilogit function* and some notation. Section 3 describes the formulation of the proposed algorithm. In the first subsection we present the algorithm based on a linear model, and in a second subsection we present the final version of the algorithm as a kernel-based extension of the previous linear algorithm. Section 4 presents a comparison of the proposed algorithm with other classification algorithms using simulated and real data. Finally, in the last section, we present our conclusions.

2. Multilogit function and overview of related models

In this section we present an overview of the Multilogit function, the Multilogit Regression Model and the Multiple Output Model. Additionally, we present the notation used in this manuscript.

2.1. Multilogit function: notation

It is well-known that multilogit function (Bishop, 2006, p. 209) allows us to model a multinomial response variable $\boldsymbol{y} \in \mathbb{R}^C$ (class labels) in terms of a set of explanatory variables $\boldsymbol{x} \in \mathbb{R}^d$ (feature vector), see Bohning (1992). In our case, we write the categorical target variable $\boldsymbol{y}$ using the 1-of-C coding[1] (Bishop, 2006, p. 424), i.e. $y_c \in \{0, 1\}$, $\boldsymbol{y}^T\mathbf{1} = 1$. Let us also define the set of labels $\mathcal{C} = \{1, \dots, C\}$. If we denote the *conditional probability* $Pr(\boldsymbol{y} = \boldsymbol{e}_c|\boldsymbol{x})$ as

$$p_c(\boldsymbol{x}) \stackrel{def}{=} Pr(\boldsymbol{y} = \boldsymbol{e}_c|\boldsymbol{x}), \tag{1}$$

then, the multinomial logit-model is given by

$$p_c(\boldsymbol{x}) = \frac{e^{f(\boldsymbol{x};\boldsymbol{\theta}_c)}}{1 + \sum_{i=1}^{C-1} e^{f(\boldsymbol{x};\boldsymbol{\theta}_i)}}, \quad c = 1, 2, \dots, C-1; \tag{2}$$

$$p_C(\boldsymbol{x}) = \frac{1}{1 + \sum_{i=1}^{C-1} e^{f(\boldsymbol{x};\boldsymbol{\theta}_i)}}. \tag{3}$$

[1] Note that the possible values of $\boldsymbol{y}$ are the elements of the canonical basis $\{\boldsymbol{e}_c : 1 \le c \le C\}$ of $\mathbb{R}^C$.