Computational Statistics and Data Analysis 54 (2010) 333–347

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Constrained linear regression models for symbolic interval-valued variables

Eufrásio de A. Lima Neto [a], Francisco de A.T. de Carvalho [b,*]

[a] Departamento de Estatística, Universidade Federal da Paraíba, Cidade Universitária s/n, CEP 58051-900, João Pessoa (PB), Brazil

[b] Centro de Informática, Universidade Federal de Pernambuco, Av. Prof. Luiz Freire, s/n, Cidade Universitária, CEP 50740-540, Recife (PE), Brazil

ARTICLE INFO

Article history:
Received 21 June 2007
Received in revised form 24 October 2008
Accepted 20 August 2009
Available online 28 August 2009

ABSTRACT

This paper introduces an approach to fitting a constrained linear regression model to interval-valued data. Each example of the learning set is described by a feature vector for which each feature value is an interval. The new approach fits a constrained linear regression model on the midpoints and range of the interval values assumed by the variables in the learning set. The prediction of the lower and upper boundaries of the interval value of the dependent variable is accomplished from its midpoint and range, which are estimated from the fitted linear regression models applied to the midpoint and range of each interval value of the independent variables. This new method shows the importance of range information in prediction performance as well as the use of inequality constraints to ensure mathematical coherence between the predicted values of the lower ($\hat{y}_{Li}$) and upper ($\hat{y}_{Ui}$) boundaries of the interval. The authors also propose an expression for the goodness-of-fit measure denominated *determination coefficient*. The assessment of the proposed prediction method is based on the estimation of the average behavior of the *root-mean-square error* and *square of the correlation coefficient* in the framework of a Monte Carlo experiment with different data set configurations. Among other aspects, the synthetic data sets take into account the dependence, or lack thereof, between the midpoint and range of the intervals. The bias produced by the use of inequality constraints over the vector of parameters is also examined in terms of the mean-square error of the parameter estimates. Finally, the approaches proposed in this paper are applied to a real data set and performances are compared.

1. Introduction

Due to the explosive growth in the use of databases, new approaches have been proposed for discovering regularities and summarizing information stored in large data sets. The development of robust, efficient machine learning algorithms for processing this data and the falling cost of computational power enable the use of computationally intensive methods for data analysis. *Symbolic Data Analysis* (SDA – (Bock and Diday, 2000)) has been introduced as a new domain related to multivariate analysis, pattern recognition and artificial intelligence for extending classical exploratory data analysis and statistical methods to symbolic data. Symbolic data allows multiple (sometimes weighted) values for each variable and new variable types (interval, categorical multi-valued and modal variables) have been introduced. These new variables make it possible to take into account the variability and/or uncertainty in the data.

The prediction of the values of a dependent variable from other (independent) variables that are presumed to explain the variability of the former is a common task in pattern recognition and data analysis fields. The classical regression model for

* Corresponding author. Tel.: +55 81 21268430; fax: +55 81 21268438.
E-mail addresses: eufrasio@de.ufpb.br (E.d.A. Lima Neto), fatc@cin.ufpe.br, francisco.carvalho@pq.cnpq.br (F.d.A.T. de Carvalho).

0167-9473/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2009.08.010

usual quantitative data is used in order to predict the behavior of a dependent variable Y as a function of other independent variables that are responsible for the variability of variable Y. However, to fit this model to the data, it is necessary to estimate a vector $\boldsymbol{\beta}$ of parameters from the data vector $\mathbf{Y}$ and the model matrix $\mathbf{X}$, supposed with complete rank p. Estimations using the *method of least squares* do not require any probabilistic hypothesis on the variable Y. This method consists of minimizing the sum of the square of errors. A detailed study on linear regression models for classical data can be found in Scheffé (1959), Draper and Smith (1981) and Montgomery and Peck (1982), among others.

In regression analysis of quantitative data, the items are usually represented as a vector of quantitative measures. However, due to recent advances in information technologies, it is now common to record interval data. In the framework of SDA, interval data appear when the observed values of the variables are intervals from the set of real numbers $\Re$. Moreover, interval data arise in practical situations, such as recording monthly interval temperatures at meteorological stations, daily interval stock prices, etc. Another source of interval data is the aggregation of huge databases into a reduced number of groups, the properties of which are described by symbolic interval variables. Therefore, tools for interval-valued data analysis are very much required.

Different approaches have been introduced for analyzing symbolic interval data. Bertrand and Goupil (0000) and Billard and Diday (2003) introduced central tendency and dispersion measures suitable for interval-valued data. De Carvalho (1995) proposed histograms for interval-valued data. Concerning factorial methods, (Cazes et al., 1997; Lauro and Palumbo, 2000) and, more recently, (Billard et al., 2007) presented principal component analysis methods suitable for interval-valued data. Palumbo and Verde (2000) and Lauro et al. (2000) generalized factorial discriminant analysis (FDA) to interval-valued data. Groenen et al. (2006) introduced a multidimensional scaling method for managing interval dissimilarities. Regarding supervised classification methods, (Ichino et al., 1996) introduced a symbolic classifier as a region-oriented approach for interval-valued data. Rasson and Lissoir (2000) presented a symbolic kernel classifier based on dissimilarity functions suitable for interval-valued data. Périnel and Lechevallier (2000) proposed a tree-growing algorithm for classifying interval-valued data. Concerning interval-valued time series, Maia et al. (2008) have introduced approaches to interval-valued time series forecasting.

SDA provides a number of clustering methods for symbolic data. These methods differ with regard to the type of symbolic data considered, their cluster structures and/or the clustering criteria considered. With hierarchical clustering methods, an agglomerative approach has been introduced that forms composite symbolic objects using a join operator whenever mutual pairs of symbolic objects are selected for agglomeration based on minimum dissimilarity (Gowda and Diday, 1991) or maximum similarity (Gowda and Diday, 1992). Ichino and Yaguchi (1994) defined generalized Minkowski metrics for mixed feature variables and presented dendrograms obtained from the application of standard linkage methods for data sets containing numeric and symbolic feature values. Chavent (1998) proposed a divisive clustering method for symbolic data that simultaneously furnishes a hierarchy of the symbolic data set and a monothetic characterisation of each cluster in the hierarchy. Guru et al. (2004) and Guru and Kiranagi (2005) introduced agglomerative clustering algorithms based, respectively, on similarity and dissimilarity functions that are multi-valued and non-symmetric.

Concerning partitioning (fuzzy and hard) clustering algorithms for interval-valued data, Bock (2002) proposed several clustering algorithms for symbolic data described by interval variables and presented a sequential clustering and updating strategy for constructing a Self-Organising Map (SOM) to visualize interval-valued data. Chavent and Lechevallier (2002) proposed a dynamic clustering algorithm for interval-valued data, in which the class representatives are defined by an optimality criterion based on a modified Hausdorff distance. Souza and De Carvalho (2004) presented partitioning clustering methods for interval-valued data based on (adaptive and non-adaptive) city-block distances. De Carvalho et al. (2006) proposed an algorithm using an adequacy criterion based on adaptive Hausdorff distances. More recently, De Carvalho (2007) introduced adaptive and non-adaptive fuzzy c-means clustering methods for partitioning interval-valued data as well as (fuzzy) cluster and partition interpretation tools.

In the framework of *Symbolic Data Analysis*, Billard and Diday (2000) presented the first approach to fitting a linear regression model to an interval-valued data set. Their approach consists of fitting a linear regression model to the midpoint of the interval values assumed by the variables in the learning set and applies this model to the lower and upper boundaries of the interval values of the independent variables to predict, respectively, the lower and upper boundaries of the interval value of the dependent variable. Lima Neto and De Carvalho (2008) improved this approach by presenting a new method based on two linear regression models – the first regression model on the midpoints of the intervals and the second one on the ranges – which reconstruct the boundaries of the interval values of the dependent variable in a more efficient manner than the Billard and Diday method.

However, neither method ensures that the predicted values of the lower boundaries ($\hat{y}_{Li}$) will be lower than or equal to the predicted values of the upper boundaries ($\hat{y}_{Ui}$). Judge and Takayama (1966) addressed the use of constraints in regression models for usual data in order to ensure the positiveness of the dependent variable Y. In this paper, we introduce a constrained linear regression model for interval-valued data that ensures this mathematical coherence between the predicted values $\hat{y}_{Li}$ and $\hat{y}_{Ui}$. The probabilistic assumptions that involve the linear regression model theory for classical data will not be considered in the case of symbolic data (interval variables), since this is still an open research topic. Thus, the problem will be investigated as an optimization problem, in which we wish to fit the best hyper plane that minimizes a predefined criterion. Moreover, we illustrate the importance of the use of restrictions in these linear regression models by analyzing the number of times that $\hat{y}_{Li} \geq \hat{y}_{Ui}$ in the former linear regression models *without* constraints and we present expressions for a goodness-of-fit measure denominated *determination coefficient*.