Neurocomputing 128 (2014) 371–388

Contents lists available at ScienceDirect

Neurocomputing

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

# Interval kernel regression

Roberta A.A. Fagundes [a], Renata M.C.R. de Souza [a,*], Francisco José A. Cysneiros [b]

[a] Centro de Informática, Univeridade Federal de Pernambuco, Av. Jornalista Anibal Fernandes, s/n - Cidade Universitária, CEP 50740-560 Recife (PE), Brazil
[b] Departamento de Estatística, Centro de Ciências Exatas, Universidade Federal de Pernambuco, Av. Prof. Luiz Freire, s/n - Cidade Universitária, CEP 50740-540 Recife (PE), Brazil

ARTICLE INFO

*Article history:*
Received 9 January 2013
Received in revised form
3 July 2013
Accepted 28 August 2013
Communicated by W.S. Hong
Available online 30 October 2013

*Keywords:*
Kernel regression
Symbolic data analysis
Interval-valued data
Noises

ABSTRACT

Kernel regression is more attractive when it is not possible to determine explicit parametric form of the model and moreover, it does not depend on probabilistic distribution. This paper introduces kernel regression in which the input data set is described by interval-value variables. Two model families are considered. The first family estimates the bounds of the intervals regarding either a smooth function for center variables of the intervals (first model) or two smooth functions for range and center variables, respectively (second model). The second family performs the estimates of the intervals based on regression mixtures. These mixtures assume either a smooth function for center variables and a linear function based on least squares for range variables (third model) or a smooth function for range variables and a linear function for center variables (fourth model). The predictions of the lower and upper bounds of new intervals are computed and two different simulation studies are carried out to validate these predictions. Five real-life interval data sets are also considered. The prediction quality is assessed by a mean magnitude of relative error calculated from a test data set.



## 1. Introduction

Kernel regression methods can be found in a range of application domains, and continue to grow in popularity. Most kernel methods adaptively balance the model complexity against its precision at fitting the training data and achieving remarkable generalization performance. Besides, the kernel methods show great ability to deal with high-dimensional problems. In this sense, the kernel methods provide attractive solutions for many practical problems.

The aim of a kernel regression analysis is to produce a reasonable analysis to the unknown nonparametric regression function $m$, where for $n$ data points $(Y_i, X_i)$ and observation errors $(\varepsilon_i)$, the relationship can be modeled as

$$Y_i = m(X_i) + \varepsilon_i, \quad i = 1, \ldots, n. \tag{1}$$

Kernel regression method, in fact, is a non-parametric fitting computing which has been widely used in many science and engineering areas such as pattern discrimination and intelligent computing. The kernel regression only depends on local data to determine the model structure without any parameter which refer mostly to distribution-free methods [11]. It has three main purposes:

1. It provides a versatile method of exploring a general relationship between variables.
2. It gives predictions of observations yet to be made without reference to a fixed parametric form.
3. It provides a tool for finding spurious observations by studying the influence of isolated points. The flexibility of the method is extremely helpful in a preliminary and exploratory statistical analysis. The nonparametric analysis could help in suggesting simple parametric formulations of the regression relationship.

Although kernel regression shows good purposes, it fails at some points:

1. It makes less efficient use of data than other least squares methods. It requires fairly large, densely sampled data sets in order to produce good models. This is because kernel regression relies on the local data structure when performing the local fitting.

* Corresponding author. Tel.: +55 8121268430; fax: +55 8121268438.
*E-mail address:* rmcrs@cin.ufpe.br (R.M.C.R. de Souza).

0925-2312/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2013.08.029

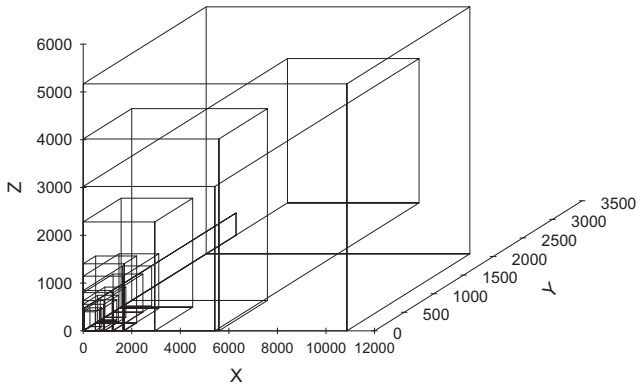


**Fig. 1.** 3D scatter plot: number of operators ($X$), number of operands ($Z$) and number of lines ($Y$).

2. The cost of predicting new examples can be high (since most of the computation takes place at this stage).
3. It is very sensitive to the curse of dimensionality.

The statistical treatment of interval data has been considered in the context of *Symbolic Data Analysis* (*SDA*) which is a knowledge discovery and data management field related to multivariate analysis, pattern recognition and artificial intelligence. An extensive coverage of earlier symbolic data analysis methods can be found in [8]. *SDA* focuses on the analysis of data sets where individuals are described by variables that can represent internal variation and/or structure. Symbolic data values can be intervals, histograms, distributions, lists of values, taxonomies, etc. The term symbolic is used to stress the fact that the values are of a different nature.

Some data sets naturally consist of symbolic interval data as for example, the data set of minimum and maximum temperatures naturally represented by intervals, while many other interval symbolic data sets result from the aggregation of large classical data sets. For example, in a database of software project modules, we are surely more interested in describing the behavior of the project rather than each module by itself.

The analysis requires then that the software data for each module be somehow aggregated to obtain the information of interest. It quickly becomes apparent that variation and structure must be taken into account when investigating this data set. Here the observed variability for each project is of utmost importance, and cannot be kept by summary statistics like an average, a median or a mode – leading to serious loss of important information.

Fig. 1 presents 13 projects of the NASA described by 3 interval variables in order to estimate software size: number of operators ($X$), number of operands ($Z$) and number of lines ($Y$).

Symbolic Data Analysis provides a framework where the variability observed may effectively be considered in the data representation, and methods developed that take it into account [17]. This work introduces kernel regression models for data described by interval-value variables. Interval kernel regression (IKR) contributes to advance of the SDA literature in the regression framework since nonparametric regression for interval data is an open topic. IKR has the same advantages and disadvantages of the kernel regression for classic data. However, IKR is able to model relationship between response and predictors that allows to take the intrinsic variability.

The rest of this paper is organized as follows, Section 2 discusses the related works of the *SDA* literature. Section 3 describes four kernel regression models for interval data proposed in this paper. Section 4 presents a performance analysis of these models using five real-valued interval data sets. Section 5 discusses two simulation studies with different scenarios of interval data in order to evaluate the performance in more details. Finally, Section 5 gives the concluding remarks.

## 2. Related works

In *SDA*, the most of the interval regression models consider that their parameters are estimated by the minimization squared error criterion. Billard and Diday [1] presented an approach to extend the classic linear regression model to symbolic interval data by fitting the method of least squares to the centers of the interval valued data assumed by the interval variables. Billard and Diday [2] proposed another approach that fits two independent on the lower and upper bounds of the intervals. Billard and Diday [3] also included explanatory variables as well as hierarchical variable structure into symbolic regression framework.

Maia and De Carvalho [15] showed a least absolute deviation regression model suitable for manage interval-valued data and modeling based on regression $L_1$. Lima Neto and De Carvalho [12] proposed the center and range method for fitting interval valued data as an improvement in comparison with the methods presented in Billard and Diday [1,2]. Lima Neto and De Carvalho [13] proposed an approach that fits a constrained linear regression model on the center and range of the interval values in order to ensure mathematical coherence between the predicted values of the lower and upper boundaries of the interval. In previous work, Fagundes et al. [10] presented a robust prediction method for symbolic interval data based on the simple linear robust regression methodology.

In the context of regression models for interval data that assume probability distributions for the errors, Domingues [9] proposed a methodology of analysis for interval data based on a symmetrical linear regression. In this model the prediction of the lower and upper bounds of the intervals is not damaged in the presence of interval outliers defined by center outliers. Lima Neto et al. [14] introduce the bivariate symbolic regression model for interval data based on generalized linear model theory. Souza et al. [19] introduced multi-class logistic linear regression models for the lower and upper bounds of the intervals conjointly and separately. Although all these regression methods of the literature of *SDA* can be applied to solve regression problems in different domains, they are not able to model well smooth relationships between response and predictor of the interval variables. So, nonparametric regression methods for interval data are need.