Information Sciences 217 (2012) 56–64

Contents lists available at SciVerse ScienceDirect

Information Sciences

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

A reduced support vector machine approach for interval regression analysis

Chia-Hui Huang *

Department of Business Administration, National Taipei College of Business, No. 321, Section 1, Jinan Rd., Zhongzheng District, Taipei City 100, Taiwan

ARTICLE INFO

Article history:
Received 26 September 2011
Received in revised form 4 May 2012
Accepted 23 June 2012
Available online 2 July 2012

Keywords:
Interval regression analysis
Fuzzy regression
Reduced support vector machine

ABSTRACT

The support vector machine (SVM) has been shown to be an efficient approach for a variety of classification problems. It has also been widely used in pattern recognition, regression and distribution estimation for separable data. However, there are two problems with using the SVM model: (1) Large-scale: when dealing with large-scale data sets, the solution may be difficult to find when using SVM with nonlinear kernels; (2) Unbalance: the number of samples from one class is much larger than the number of samples from the other classes. It causes the excursion of separation margin. Under these circumstances, developing an efficient method is necessary.

Recently, the use of the reduced support vector machine (RSVM) was proposed as an alternative to the standard SVM. It has been proven more efficient than the traditional SVM in processing large-scaled data. In this paper, we introduce the principle of RSVM to evaluate interval regression analysis. The main idea of the proposed method is to reduce the number of support vectors by randomly selecting a subset of samples.

© 2012 Elsevier Inc. All rights reserved.

1. Introduction

Since Tanaka et al. [38] introduced the fuzzy regression model with symmetric fuzzy parameters, many researchers have studied the properties of fuzzy regression extensively. A collection of recent studies to fuzzy regression analysis can be found in [2,3,5,6,9,10,13–17,35–39].

The fuzzy regression model can be simplified to interval regression analysis, which is considered as the simplest version of possibilistic regression analysis with interval coefficients. Some coefficients in interval linear regression models tend to become separable due to the characteristics of linear programming (LP) [38]. To alleviate the issue of LP, Tanaka and Lee [37] propose an interval regression analysis with a quadratic programming (QP) approach, which gives a more diverse spread of coefficients than by using the LP approach.

The support vector machine (SVM) has been widely used in pattern recognition, regression and distribution estimation [1,4,7,8,11,12,21,23–31,33,40–42]. Arun Kumar et al. [1] incorporated prior knowledge represented by polyhedral sets into a relatively new family of SVM classifiers based on two non-parallel hyperplanes and proposed two formulations termed as knowledge based Twin SVM (KBTWSVM) and knowledge based Least Squares Twin SVM (KBLSTWSVM), both capable of generating non-parallel hyperplanes from given data and prior knowledge. Maldonado et al. [21] proposed an embedded method for feature selection and classification using kernel-penalized SVM (KP-SVM). The KP-SVM simultaneously selects relevant features during classifier construction by penalizing each feature's use in the dual formulation of SVM. Savitha et al. [28] developed a fast learning fully complex-valued extreme learning machine classifier, referred to as Circular Complex-valued Extreme Learning Machine (CC-ELM) for handling real-valued classification problems. CC-ELM classifier uses a single hidden layer network with a non-linear input/hidden layer and a linear output layer. The results show that CC-ELM

* Tel.: +886 2 2322 6402; fax: +886 2 2322 6404.
E-mail address: leohuang@webmail.ntcb.edu.tw

0020-0255/$ - see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ins.2012.06.030

performs better than other existing real-valued and complex-valued classifiers, especially when the data sets are highly unbalanced. Unler et al. [40] introduced a hybrid filter-wrapper feature subset selection algorithm based on particle swarm optimization (PSO) for SVM classification. The filter model is based on the mutual information and is a composite measure of feature relevance and redundancy with respect to the feature subset selected. The experimental results show that the mr^{2}PSO algorithm is competitive in terms of both classification accuracy and computational performance.

Recently, using SVM to solve the interval regression model has become an alternative approach. Hong and Hwang introduced SVM for multivariate fuzzy regression analysis [13] and evaluated interval regression models with quadratic loss SVM [14]. Jeng et al. [16] developed a support vector interval regression networks (SVIRNs) based on both SVM and neural networks. Bisserier et al. [2] proposed a revisited fuzzy regression method where a linear model is identified from Crisp-Inputs Fuzzy-Outputs (CISO) data. D'Urso et al. [10] presented fuzzy clusterwise regression analysis with LR fuzzy response variable and numeric explanatory variables. The suggested model is to allow for linear and non-linear relationship between the output and input variables.

However, there are two problems with using the SVM model:

1. Large-scale: when dealing with large-scale data sets, the solution may be difficult to find when using SVM with nonlinear kernels.
2. Unbalance: the number of samples from one class is much larger than the number of samples from other classes. It causes the excursion of separation margin.

Under these circumstances, developing an efficient method is necessary. The reduced support vector machine (RSVM) has been proven more efficient than the traditional SVM in processing large-scaled data [18–20]. The main purpose of RSVM is to reduce the number of support vectors by randomly selecting a subset of samples. In this report, we introduce the principle of RSVM to evaluate interval regression analysis.

This paper is organized in the following manner. Section 2 reviews interval regression analysis by the QP approach, thus unifying the possibility and necessity models. Section 3 proposes the formulation of RSVM in evaluating the interval regression models. Section 4 provides the numerical examples to illustrate the methods. Finally, Section 5 presents the concluding remarks.

2. Interval regression analysis with the QP approach

In this section we review interval regression analysis by a quadratic programming (QP) approach, thus unifying the possibility and necessity models proposed by Tanaka and Lee [37].

An interval linear regression model is described as

$$Y(\boldsymbol{x}_j) = A_0 + A_1 x_{1j} + \cdots + A_n x_{nj} \tag{1}$$

where $Y(\boldsymbol{x}_j), j = 1, 2, \ldots, q$ is the estimated interval corresponding to the real input vector $\boldsymbol{x}_j = (x_{1j}, x_{2j}, \ldots, x_{nj})^T$. An interval coefficient A_i is defined as $(a_i, c_i), i = 1, 2, \ldots, n$, where a_i is a center and c_i is a radius. Hence, A_i can also be represented as

$$A_i = \{u \mid a_i - c_i \leqslant u \leqslant a_i + c_i\} = [a_i - c_i, a_i + c_i] \tag{2}$$

The interval linear regression model (1) can also be expressed as

$$Y(\boldsymbol{x}_j) = A_0 + A_1 x_{1j} + \cdots + A_n x_{nj} = (a_0, c_0) + (a_1, c_1)x_{1j} + \cdots + (a_n, c_n)x_{nj} = \left(a_0 + \sum_{i=1}^{n} a_i x_{ij}, c_0 + \sum_{i=1}^{n} c_i |x_{ij}|\right) \tag{3}$$

For a data set with non-fuzzy inputs and interval outputs, two interval regression models, the possibility and necessity models, are considered. By assumption, the center coefficients of the possibility regression model and the necessity regression model are the same [37].

Assume that input–output data $(\boldsymbol{x}_j; Y_j)$ are given as

$$(\boldsymbol{x}_j; Y_j) = (1, x_{1j}, \ldots, x_{nj}; Y_j), \ j = 1, 2, \ldots, q \tag{4}$$

where $\boldsymbol{x}_j$ is the jth input vector, Y_j is the corresponding interval output that consists of a center y_j and a radius e_j denoted as $Y_j = (y_j, e_j)$, and q is a data size. For this data set, the possibility and necessity estimation models are defined as

$$Y^*(\boldsymbol{x}_j) = A_0^* + A_1^* x_{1j} + \cdots + A_n^* x_{nj} \tag{5}$$

$$Y_*(\boldsymbol{x}_j) = A_{0^*} + A_{1^*} x_{1j} + \cdots + A_{n^*} x_{nj} \tag{6}$$

where the interval coefficients A_i^* and A_{i*} are defined as $A_i^* = (a_i^*, c_i^*)$ and $A_{i*} = (a_{i*}, c_{i*})$, respectively. The interval $Y^*(\boldsymbol{x}_j)$ estimated by the possibility model must include the observed interval Y_j. The interval $Y_*(\boldsymbol{x}_j)$ estimated by the necessity model must also be included in the observed interval Y_j. The following inclusion relations exist

$$Y_*(\boldsymbol{x}_j) \subseteq Y_j \subseteq Y^*(\boldsymbol{x}_j) \tag{7}$$

The interval coefficients A_i^* and A_i^* can be denoted as

Download English Version:

https://daneshyari.com/en/article/393345

Download Persian Version:

https://daneshyari.com/article/393345

Daneshyari.com