Expert Systems with Applications 42 (2015) 2502–2509

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Set-valued samples based support vector regression and its applications

Jiqiang Chen [a], Witold Pedrycz [b,c,d], Minghu Ha [a,*], Litao Ma [a]

[a] School of Science, Hebei University of Engineering, Handan 056038, China
[b] Department of Electrical & Computer Engineering, University of Alberta, Edmonton T6R 2V4 AB, Canada
[c] Department of Electrical and Computer Engineering, Faculty of Engineering, King Abdulaziz University, Jeddah 21589, Saudi Arabia
[d] Systems Research Institute, Polish Academy of Sciences, Warsaw, Poland

ARTICLE INFO

Article history:
Available online 30 September 2014

Keywords:
Support vector machine
Regression problems
Set-valued sample
Wind speed
Peak particle velocity
Prediction

ABSTRACT

In this study, we address the regression problem on set-valued samples that appear in applications. To solve this problem, we propose a support vector regression approach for set-valued samples that generalizes the classical ε-support vector regression. First, an initial representative point (or an element) for every set-valued sample is selected, and a weighted distance between the initial representative point and other points is determined. Second, based on the classification consistency principle, a search algorithm to determine the best representative point for every set-valued datum is designed. Thus, the set-valued samples are converted into numeric samples. Finally, a support vector regression that is based on set-valued data is constructed, and the regression results of the set-valued samples can be approximated using the method used for the numeric samples. Furthermore, the feasibility and efficiency of the proposed method is demonstrated using experiments with real-world examples concerning wind speed prediction and the prediction of peak particle velocity.

1. Introduction

Support vector machines (SVMs) were proposed by Drucker, Burges, and Kaufman (1997) and Vapnik, Golowich, and Smola (1997) in the 1990s. The SVM, including the support vector classification algorithm (SVC) and support vector regression algorithm (SVR), was one of the first statistical learning algorithms to use the kernel function theory in the field of machine learning. The SVM is specifically tailored to the small sample case by solving a convex quadratic optimization problem to obtain a globally optimal solution with existing information. In this process, the SVM solves the local minimization problem a neural network algorithm cannot avoid. The SVM uses the kernel function as a nonlinear transformation to enable samples to be mapped to a high-dimensional feature space. In this manner, we can solve the original problem in the high-dimensional space. SVM solves the problem of a high-dimensional disaster successfully. Currently, support vector machines are the subject of extensive attention and are attracting a growing number of researchers studying them from different viewpoints (Chapelle, Sindhwani, & Keerthi, 2008; Deng & Tian, 2009; Ha, Wang, & Zhang, 2010; Ji, Pang, & Qiu, 2010; Karasuyama & Takeuchi, 2010; Kavousi-Fard, Samet, & Marzbani, 2014; Li & Fang, 2004; Narwaria & Lin, 2010; Tsang, Zhang, & Chawla, 2009;Yang & Liu, 2007).

By considering the insensitive loss function ε introduced by Vapnik, the support vector classification algorithm is extended to the support vector regression algorithm. The support vector regression algorithm obtains a linear regression function in high-dimensional feature space. One obvious drawback of standard SVR is that the prior knowledge of specific issues cannot be incorporated into the learning process. Considering that prior knowledge is useful to the performance of the algorithm, Chuang (2007) proposed a fuzzy weighted SVR with a fuzzy partition to address the problem of boundary effects. Li, Mersereau, and Simske (2007) introduced a new algorithm for the restoration of a noisy blurred image based on support vector regression applied to blind image deconvolution. Lauer and Bloch (2008) explored the addition of constraints to the linear programming formulation of the support vector regression problem for the incorporation of prior knowledge, and they proposed a new method for the simultaneous approximation of multiple outputs linked by prior knowledge. Juang and Cheng (2009) proposed the Takagi–Sugeno fuzzy system-based support vector regression (TSFS-SVR), which is motivated by the TS-type fuzzy rules and fuzzy clustering. The capabilities of TSFS-SVR were demonstrated by conducting simulations in clean and noisy function approximations and signal prediction. Farooq, Guergachi, and Krishnan (2010) presented a novel prior knowledge-based Green's kernel for support vector regression after

* Corresponding author.
E-mail addresses: jiqiang516@163.com (J. Chen), wpedrycz@ualberta.ca (W. Pedrycz), mhha@hbu.edu.cn (M. Ha), ltma1821@163.com (L. Ma).

http://dx.doi.org/10.1016/j.eswa.2014.09.038
0957-4174/

reviewing the correspondence between support vector kernels that are used in support vector machines and regularization operators that are used in regularization networks. Seo, Yim, and Kim (2011) investigated empirical modeling of the superconductor-triggered type fault current limiter (STFCL) using principal component-based and fuzzy support vector regression for the prediction and detection of faults in the STFCL. Yeh, Huang, and Lee (2011) developed a two-stage multiple-kernel learning algorithm by incorporating sequential minimal optimization and the gradient projection method to address the problem of manually tuning the hyper parameters of the kernel functions in stock market forecasting problems. Liu and Xue (2012) designed a class of kernels by linearly combining the kernels that correspond to each rule via fuzzy entropies for all the fuzzy rules and constructed a new support vector regression based on fuzzy a priori information. Additionally, there are many research studies (Chen, He, & Wang, 2010; Chen, Bo, & Liu, 2011; Chen, Xue, & Ha, 2014; Gordini, 2014; Ha, Wang, & Chen, 2013; Harris, 2013; Kang & Cho, 2014; Li, Tax, & Duin, 2013; Manivannan, Aggarwal, & Devabhaktuni, 2012; Saito, Rezende, & Falcao, 2014; Zhou & Chellappa, 2006) that focus on improving the classical support vector machine and that incorporate a priori knowledge.

However, for all the above-described studies, the training set is specified as

$$S = \{(x_1, y_1), (x_2, y_2), \ldots, (x_l, y_l)\}, \quad y_i \in R, \ i = 1, \ldots, l,$$

where $x_i \in R^n$ is a numeric vector. In applications, there are many regression problems where the data are not numeric vectors but set-valued samples, represented as

$$S = \{(A_1, y_1), (A_2, y_2), \ldots, (A_l, y_l)\}, \quad y_i \in R,$$

or

$$S = \{(x_1, A_1), (x_2, A_2), \ldots, (x_l, A_l)\}, \quad x_i \in R, \ i = 1, \ldots, l, \tag{1}$$

where $A_i \subset R^n$ is a set. For example, in a wind farm, wind speed is monitored ten times $v_j (j = 1,2,\ldots,10)$ every ten minutes, and the mean value $\overline{v} = \sum_{j=1}^{10} v_j$ is output as the monitored value every ten minutes. The wind speed for the next ten minutes is predicted according to the mean value, which is still a numerical number, resulting in a much larger prediction error. Therefore, in order to increase the prediction accuracy, the data used in prediction should be all of the values $v_j (j = 1,2,\ldots,10)$ that can be described by a set $A = \{v_1, v_2, \ldots, v_{10}\}$, not the mean value $\overline{v}$. Then, the problem of wind speed prediction is based on the data set

$$S = \{(t_1, A_1), (t_2, A_2), \ldots, (t_l, A_l)\}, \quad t_i \in [0, +\infty), \ i = 1, \ldots, l, \tag{2}$$

where $A_i = \{v_{i1}, v_{i2}, \ldots, v_{i10}\}$ are sets, not vectors. Thus, the problem is regarded as a regression problem for set-valued samples, which cannot be addressed by the classical SVRs (which are designed for regression problems based on numerical numbers). Therefore, it is necessary to focus on constructing the regression method for set-valued samples.

In the prediction problem of wind speed, values v_i of wind speed every ten minutes are obtained from sets A_i $(i = 1,\ldots,l)$. In other words, v_i can be seen as the representative point (or element) of the corresponding set A_i $(i = 1,\ldots,l)$. Inspired by this idea, and in accordance with the classification consistency principle (i.e., the number of misclassified points is controlled within the allowable range as far as possible), we design an algorithm to search for the best representative point (see the stars in Fig. 1) for every set-valued sample (see the circles in Fig. 1). Then, the set-valued training data are converted into the classical training data, and we can use classical support vector regression to solve the regression problems of the set-valued data.

This paper is organized as follows. In Section 2, a weighted distance between the representative point and other points (or elements) is defined. In Section 3, an algorithm is designed to search for the best representative point of every set-valued datum, and the set-valued data are converted to classical data. In Section 4, a set-valued samples based support vector regression (SSVR) is constructed, which can give us the regression results of the set-valued samples. In Section 5, experiments for set-valued data are provided to illustrate the proposed method, and the results show that the proposed method is effective and feasible. In Section 6, we draw conclusions.

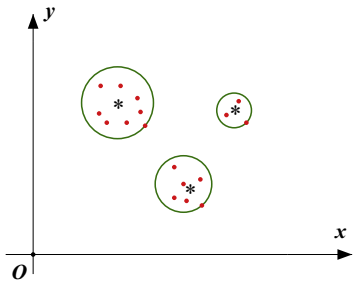

Fig. 1. Set-valued data and their representative points.

2. Weighted distance between the representative point and the other points

Here, it is noted that the importance of every component of point x in set A is different from each other. For example, suppose that x_1 and x_2 are two students; the four features are shown in Table 1.

Therefore, the feature vectors of x_1 and x_2 are $V_{x_1} = (1, 18, 180, 80)$ and $V_{x_2} = (1, 18, 170, 75)$, respectively. Because x_1 and x_2 have the same "Gender" and "Age", we cannot identify them from the two features. In other words, "Gender" and "Age" do not have effect in the identification. Moreover, because x_1 and x_2 have a "Body Height" of 180 and 170, respectively, they can be identified from "Body Height" only. Therefore, the four features play different roles in the identification of x_1 and x_2. For this reason, we introduce the concept of the index classification weight and weighted distance.

2.1. Definition of the index classification weight

Let l be the number of set-valued data, let $\{A_i | i = 1,2,\ldots,l\}$ be the collection of the set-valued data, let N_i be the number of elements x in set A_i, and let $N = \sum_{i=1}^{l} N_i$.

Let m_i $(i = 1,2,\ldots,l)$ be the (initial) representative point of set A_i and $j = 1,2,\ldots,n$. Let us use the following notation

$$m_i = (m_{i1}, m_{i2}, \ldots, m_{in}) \tag{3}$$

$$\overline{m}_j = \frac{1}{l}\sum_{i=1}^{l} m_{ij} \tag{4}$$

$$\sigma_j^2 = \frac{1}{l}\sum_{i=1}^{l} (m_{ij} - \overline{m}_j)^2 \tag{5}$$

$$\lambda_j = \sigma_j^2 / \sum_{t=1}^{n} \sigma_t^2 \tag{6}$$

Table 1
Four features of students x_1 and x_2.

Students	Gender (male = 1, female = 0)	Age (years)	Body Height (cm)	Body Weight (kg)
x_1	1	18	180	80
x_2	1	18	170	75