Applied Soft Computing 31 (2015) 405–411

Contents lists available at ScienceDirect

Applied Soft Computing

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

A novel support vector regression for data set with outliers

Jie Hu*, Kai Zheng

Institute of Knowledge Based Engineering, School of Mechanical Engineering, Shanghai Jiao Tong University, 800 Dongchuan Road, Shanghai 200240, PR China

ARTICLE INFO

Article history:
Received 3 March 2014
Received in revised form 26 January 2015
Accepted 15 March 2015
Available online 27 March 2015

Keywords:
Support vector regression
Outlier
Fuzzification theory
Inconsistency matrix
Neighbors match operator

ABSTRACT

Support vector machine (SVM) is sensitive to the outliers, which reduces its generalization ability. This paper presents a novel support vector regression (SVR) together with fuzzification theory, inconsistency matrix and neighbors match operator to address this critical issue. Fuzzification method is exploited to assign similarities on the input space and on the output response to each pair of training samples respectively. The inconsistency matrix is used to calculate the weights of input variables, followed by searching outliers through a novel neighborhood matching algorithm and then eliminating them. Finally, the processed data is sent to the original SVR, and the prediction results are acquired. A simulation example and three real-world applications demonstrate the proposed method for data set with outliers.

1. Introduction

The support vector machine (SVM) initially proposed by Cortes and Vapnik [1,2] is drawing close attention due to its high generalization in solving practical problems such as nonlinearity, small samples and over-fitting. The SVM is a learning machine based on the structural risk minimization (SRM) inductive principle to achieve the generalized performance. Unlike some traditional approaches which attempt to minimize the empirical risk, the SVM also considers the minimization of Vapnik–Chervonenkis (VC) dimension. The main idea of the SVM is to compute a linear regression function in a higher dimensional feature space mapped from the original input space. This function is established with a portion of the training data, which are called support vectors. The SVM has been successfully applied to various fields – classification [3,4], time prediction [5–7] and regression [8,9].

The SVM falls into two categories, one is the support vector classification (SVC) and the other is the support vector regression (SVR). When the SVM is exploited in the process of time prediction or regression estimation, the approaches are defined as the SVR. The model generated by the SVR only depends on the support vectors, not all of the training samples. An early overview of the fundamental ideas underlying the SVR has been given by Smola and Schölkopf [8], which also includes some popular algorithms for training the SVM, as well as some modifications and extensions.

The approximation scheme using the SVR depends only on the support vectors, rather than all the training samples. It is an important advantage, while simultaneously it makes the SVR too sensitive to outliers and increases the risk of over-fitting [10]. If the collected data set contains outliers, the learning process may not recognize such a situation and then try to fit those abnormal data, thus resulting in an erroneous approximation function [6,11].

Generally for real-world applications, data sets often contain multiple variables as well as noise or outliers that are inconsistent with the other data. Outliers may occur for a variety of reasons, such as environment changes or erroneous measurements. Developing methods for reducing the influence of outliers in the SVM have attracted considerable researchers to study on it. Lin and Wang [12] proposed the fuzzy SVM by assigning different fuzzy memberships to different training samples. Jin et al. [13] transformed the given data into a higher feature space through a fuzzy system, and exploited genetic algorithms to improve the fuzzy feature transformation. However, it is difficult to define the membership to the training samples especially when there is not any prior knowledge. Williams [14] proposed a new version called scaled SVM based on the extreme value theory and by computing the mean and variance of the generalization errors for reducing generalization error. Li [15] introduced an idea of separating outliers through the K-nearest neighbor algorithm to guarantee high generalization. Zhan and Shen [16] proposed a new training method for the SVM, in which an adaptive penalty term in the objective function is designed to suppress the influences of outliers, while the method is relatively complicated. Zhang and Wang [17] proposed a rough margin based SVM based on the rough set theory, in which more training samples

* Corresponding author. Tel.: +86 21 3420 6552; fax: +86 21 3420 6313.
E-mail address: hujie@sjtu.edu.cn (J. Hu).

http://dx.doi.org/10.1016/j.asoc.2015.03.033
1568-4946/© 2015 Elsevier B.V. All rights reserved.

can be adaptively considered with different penalty depending on their positions.

Nevertheless, each of the above approaches was proposed mainly for improving the efficiency and generalization of the SVC, rather than the SVR. To our knowledge, there have not yet been quantities of researches in reducing the influence of outliers for SVM regression. Chuang et al. [18] proposed a robust SVR network, in which the traditional robust statistics is exploited to improve the acquired regression model, while this method needs extensive computation and additional parameters. Suykens et al. [11] proposed a weighted least squares SVM (LS-SVM) to reduce the effects of outliers, while the final result was seriously influenced by the selection of the extra parameters.

In terms of outlier detection, several of the most popular techniques are distance-based approach [19], distribution-based approach [20], depth-based approach [21]. However, it is well known that these algorithms are subjected to the dimension curse, and some of them also require the weights of multiple variables. In addition to these approaches above, there are also density-based approach [22] and rough set theory-based approach [23], but the former is sensitive to the parameters defining the neighborhood, and the latter applies to discrete data rather continuous data. Moreover, SVR is used for outlier detection with nonlinear functions and multidimensional input [24], but it is difficult to apply because of adjusting parameters and high computational cost.

Drove by the above analysis, a novel support vector regression method for data set with outliers is devised in this paper. In this approach, fuzzy similarity is introduced to define the similarities between each pair of two training samples. The inconsistency matrix is exploited to compute the weights of the input variables. A new neighborhood matching algorithm is used to judge whether the training samples are outliers. Then these outliers are eliminated and the data set without outliers is sent to the SVR.

The remainder of this paper is arranged as follows: Section 2 reviews a general background of the SVM and its limitations when outliers exist, and the basic theory of fuzzy rough set theory. The proposed SVR for data set with outliers is presented in Section 3, as well as the introduction of several key steps of the proposed method, including fuzzy similarity (FS) calculation, weight calculation, neighborhood matching degree calculation and modeling for the SVR. A simulation example and three real-world applications are used to illustrate the validity of the method in Sections 4 and 5 respectively. Conclusion follows in Section 6.

2. Methodological background

2.1. SVM for regression

A regression problem can be defined as to determine a function for approximating the output from a set of training data $X = \{(x_1, y_1), (x_2, y_2), \ldots (x_l, y_l)\}$, where $x_i \in R^l$ (l is the number of the training samples) denotes the input space, and $y_i \in R$ denotes its corresponding output value for $i = 1, 2, \ldots, l$. As mentioned above, the SVR is to approximate the given observations by a linear function and a nonlinear transformation from R^l to a high-dimensional feature space F. The general function for SVR takes the form as follows:

$$f(x) = (w \cdot \varphi(x)) + b \tag{1}$$

where $w \in R^l$ denotes a weight vector, $b \in R$ denotes a threshold, and $\varphi(\cdot)$ is the nonlinear transformation from R^l to the feature space F. Based on the SVR theory, the value of w and b should be determined by minimizing the structural risk:

$$R_{SR}(f) = \frac{1}{2}||w^2|| + C\sum_{i=1}^{l} L_\varepsilon(y) \tag{2}$$

where $||w||^2$ indicates the complexity of the regression function, which should be minimized in the approximation process, and w can be written by the form of Eq. (3) with Lagrange multiplier α_i and α_i^*; $C > 0$ is a regular constant determining the penalties to the empirical errors, a large value of which tends to minimize the error in the regression process and get lower generalization, while a small value of which allows the error but tends to get higher generalization; $L_\varepsilon(y)$ denotes the loss function determined by the insensitive parameter ε and the error between the real output value y and the estimated one $f(x)$ as Eq. (4).

$$w = \sum_{i=1}^{l} (\alpha_i - \alpha_i^*)\varphi(x_i) \tag{3}$$

$$L_\varepsilon(y) = \begin{cases} 0, & \left|f(x) - y\right| < \varepsilon \\ \left|f(x) - y\right| - \varepsilon, & \text{otherwise} \end{cases} \tag{4}$$

The general function can be rewritten by substituting Eq. (3) into Eq. (4) as follows:

$$f(x) = \sum_{i=1}^{l} (\alpha_i - \alpha_i^*)(\varphi(x_i) \cdot \varphi(x)) + b = \sum_{i=1}^{l} (\alpha_i - \alpha_i^*)k(x_i, x) + b \tag{5}$$

where the $(\varphi(x_i) \cdot \varphi(x))$ can be replaced with kernel function $k(x_i, x)$, which maps the input space into a higher dimensional feature space and reflects the prior knowledge on data. The details of several main kernel functions can be referred to Ref. [25], and the selection of them should be undertaken by the user.

The dual problem corresponding to the original optimization problem is to maximize

$$\sum_{i=1}^{l} y_i(\alpha_i^* - \varepsilon) - \varepsilon\sum_{i=1}^{l} (\alpha_i + \alpha_i^*) - \frac{1}{2}\sum_{i,j=1}^{l} (\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j)k(x_i, x_j) \tag{6}$$

which subjects to

$$\sum_{i=1}^{l} (\alpha_i^* - \alpha_i) = 0 \tag{7}$$

$$\alpha_i, \alpha_i^* \in [0, C], \quad i = 1, 2, \ldots l$$

As mentioned above, only a portion of training samples are support vectors, which have the nonzero values of the corresponding Lagrange multipliers in Eq. (5). If the requirement $\left|f(x) - y\right| < \varepsilon$ is met, the corresponding training samples will not contribute to the regression as their Lagrange multipliers equaling to zero; while $\left|f(x) - y\right| \geq \varepsilon$, the corresponding training samples may become support vectors with their nonzero Lagrange multipliers.

After the value of w is determined in Eq. (3), the variable b can be computed according to the Karush–Kuhn–Tucker (KKT) condition as follows:

$$\begin{aligned} &b = y_i - (w \cdot x_i) - \varepsilon \quad \text{if} \quad \alpha_i \in [0, C] \\ \text{or} \quad &b = y_i - (w \cdot x_i) - \varepsilon \quad \text{if} \quad \alpha_i^* \in [0, C] \end{aligned} \tag{8}$$

Then the construction of the SVR has been completed.

It can be seen that if over-fitting phenomena occurs, some inaccurate information like outlier may also be modeled into the regression function, thus making the function unsmooth. During the regression process, SVR tends to minimize the empirical error to some extent, which in turn, lowers the generalization of the acquired regression function. Therefore, these outliers should be removed before the implementation of the SVR.