Neural Networks 57 (2014) 1–11

Contents lists available at ScienceDirect

Neural Networks

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

Noise model based ν-support vector regression with its application to short-term wind speed forecasting

Qinghua Hu[a,b,*], Shiguang Zhang[a,c], Zongxia Xie[d], Jusheng Mi[a,e], Jie Wan[f]

[a] College of Mathematics and Information Science, Hebei Normal University, Shijiazhuang, Hebei, 050024, China
[b] School of Computer Science and Technology, Tianjin University, Tianjin, 300072, China
[c] College of Mathematics and Computer Science, Hengshui University, Hengshui, Hebei, 053000, China
[d] School of Computer Software, Tianjin University, Tianjin, 300072, China
[e] Hebei Key Laboratory of Computational Mathematics and Applications, Shijiazhuang, Hebei, 050024, China
[f] School of Energy Science and Engineering, Harbin Institute of Technology, Harbin, 150001, China

ARTICLE INFO

Article history:
Received 17 September 2013
Received in revised form 9 March 2014
Accepted 1 May 2014
Available online 13 May 2014

Keywords:
Support vector regression
Noise model
Loss function
Inequality constraints
Wind speed forecasting

ABSTRACT

Support vector regression (SVR) techniques are aimed at discovering a linear or nonlinear structure hidden in sample data. Most existing regression techniques take the assumption that the error distribution is Gaussian. However, it was observed that the noise in some real-world applications, such as wind power forecasting and direction of the arrival estimation problem, does not satisfy Gaussian distribution, but a beta distribution, Laplacian distribution, or other models. In these cases the current regression techniques are not optimal. According to the Bayesian approach, we derive a general loss function and develop a technique of the uniform model of ν-support vector regression for the general noise model (N-SVR). The Augmented Lagrange Multiplier method is introduced to solve N-SVR. Numerical experiments on artificial data sets, UCI data and short-term wind speed prediction are conducted. The results show the effectiveness of the proposed technique.

1. Introduction

Regression is an old topic in the domain of learning functions from a set of samples (Hastie, Tibshirani, & Friedman, 2009). It provides researchers and engineers with a powerful tool to extract hidden rules of data. The trained model is used to predict future events with the information of past or present events. Regression analysis is now successfully applied in nearly all fields of science and technology, including the social sciences, economics, finance, wind power prediction for grid operation. However this domain is still attracting much attention from research and application domains.

Generally speaking, there are three important issues in designing a regression algorithm: model structures, objective functions and optimization strategies. The model structures include linear or nonlinear functions (Park & Lee, 2005), neural networks (Spech, 1990), decision trees (Esposito, Malerba, & Semeraro, 1997), and so on; optimization objectives include ϵ-insensitive loss (Cortes & Vapnik, 1995; Vapnik, 1995; Vapnik, Golowich, & Smola, 1996), squared loss (Suykens, Lukas, & Vandewalle, 2000; Wu, 2010; Wu & Law, 2011), robust Huber loss (Olvi & David, 2000) etc. According to the formulation of optimization functions, a collection of optimization algorithms (Ma, 2010) have been developed. In this work, we focus on the problem which optimal formulation should be considered with respect to different error models.

Suppose we are given a set of training data

$$D_l = \{(x_1, y_1), (x_2, y_2), \ldots, (x_l, y_l)\}, \tag{1}$$

where $x_i \in R^L, y_i \in R, \ i = 1, 2, \ldots, l$. Take a multivariate linear regression task f as an example. The form is

$$f(x) = \omega^T \cdot x + b, \tag{2}$$

where $\omega \in R^L, b \in R, \ i = 1, 2, \ldots, l$. The task is to learn the parameter vectors ω and parameter b, by minimizing the objective function

$$g_{LR} = \sum_{i=1}^{l} (y_i - \omega^T \cdot x_i - b)^2. \tag{3}$$

* Corresponding author at: College of Mathematics and Information Science, Hebei Normal University, Shijiazhuang, Hebei, 050024, China. Tel.: +86 22 27401839.
E-mail address: huqinghua@tju.edu.cn (Q. Hu).

http://dx.doi.org/10.1016/j.neunet.2014.05.003
0893-6080/

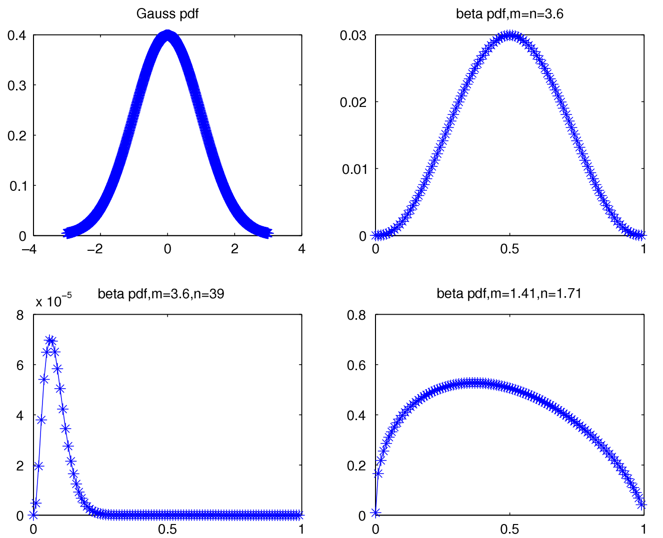

Fig. 1. Gaussian PDF and beta PDF of parameters.

The objective function of the sum-of-squares error is usually used in regression. The trained model is optimal, if the samples have been corrupted by independent and identical probability distributions (i.i.d.). Noise satisfying Gaussian distribution with zeros mean and variance σ^2, i.e., $y_i = f(x_i) + \xi_i,\ i = 1, \ldots, l,\ \xi_i \sim N(0, \sigma^2)$.

In the recent years, the support vector regressor (SVR) is growing up as a popular technique (Cortes & Vapnik, 1995; Cristianini & Shawe, 2000; Smola & Schölkopf, 2004; Vapnik, 1995, 1998, 1999; Vapnik et al., 1996; Wu, 2010; Wu & Law, 2011). It is a universal regression machine based on the V–C dimension theory. This technique is developed with the Structural Risk Minimization (SRM) principle, which has shown its effectiveness in applications. The classical SVR is optimized by minimizing Vapnik's ϵ-insensitive loss function of residuals and has achieved good performance in a variety of practical applications (Bayro-Corrochano & Arana-Daniel, 2010; Duan, Xu, & Tsang, 2012; Huang, Song, Wu, & You, 2012; Kwok & Tsang, 2003; Lopez & Dorronsoro, 2012; Yang & Ong, 2011).

In 1995, ϵ-SVR was proposed by Vapnik and his research team (Cortes & Vapnik, 1995; Vapnik, 1995; Vapnik et al., 1996). In 2000, ν-SVR was introduced by Schölkopf, Smola, Williamson, and Bartlett (2000), which automatically computes ϵ. Suykens et al. (2000) constructed least squares support vector regression with Gaussian noise (LS-SVR). Wu (Wu, 2010; Wu & Law, 2011) and Pontil, Mukherjee, and Girosi (1998) constructed ν-support vector regression with Gaussian noise (GN-SVR). If the noise obeys the Gaussian distribution, the outputs of the models are optimal. However, it was found that the noise in some real-world applications, just like wind power forecast and direction-of-arrival estimation problem, does not satisfy Gaussian distribution, but a beta distribution or Laplace distribution, respectively. In these cases these regression techniques are not optimal.

The principle of ν-support vector regression (ν-SVR) can be written as (Chalimourda, Schölkopf, & Smola, 2004; Chih-Chung & Chih-Jen, 2002; Schölkopf et al., 2000):

$$\begin{aligned}
&\min\left\{g_{P_{\nu\text{-SVR}}} = \frac{1}{2}\|\omega\|^2 + C\cdot\left(\nu\epsilon + \frac{1}{l}\sum_{i=1}^{l}(\xi_i + \xi_i^*)\right)\right\}\\
&\text{Subject to}:\\
&\quad \omega^T\cdot x_i + b - y_i \le \epsilon + \xi_i\\
&\quad y_i - \omega^T\cdot x_i - b \le \epsilon + \xi_i^*\\
&\quad \xi_i, \xi_i^* \ge 0, \quad i = 1, 2, \ldots, l,\ \epsilon \ge 0,
\end{aligned} \tag{4}$$

where ξ_i, ξ_i^* are two slack variables. The constant $C > 0$ determines the trade-off between the flatness of f and the amount up to which deviations larger than ϵ are tolerated. $\nu \in (0, 1]$ is a constant which controls the number of support vectors. In the ν-SVR the size of ϵ is not given a priori but a variable. Its value is traded off against the model complexity and slack variables via a constant ν (Chalimourda et al., 2004). This corresponds to dealing with a so-called ϵ-insensitive loss function (Cortes & Vapnik, 1995; Vapnik, 1995) described by

$$c_\epsilon(\xi) = |\xi|_\epsilon = \begin{cases} 0, & \text{if } |\xi| \le \epsilon,\\ |\xi| - \epsilon, & \text{otherwise.}\end{cases} \tag{5}$$

In 2002, ϵ-SVR for a general noise model was proposed in Schölkopf and Smola (2002):

$$\begin{aligned}
&\min\left\{g_{\epsilon\text{-SVR}} = \frac{1}{2}\|\omega\|^2 + C\cdot\left(\sum_{i=1}^{l}\widetilde{c}(\xi_i) + \widetilde{c}(\xi_i^*)\right)\right\}\\
&\text{Subject to}:\\
&\quad \omega^T\cdot x_i + b - y_i \le \epsilon + \xi_i\\
&\quad y_i - \omega^T\cdot x_i - b \le \epsilon + \xi_i^*\\
&\quad \xi_i, \xi_i^* \ge 0, \quad i = 1, 2, \ldots, l,
\end{aligned} \tag{6}$$

where $c(x, y, f(x)) = \widetilde{c}(|y - f(x)|_\epsilon)$ is a general convex loss function in the sample point (x_i, y_i) of D_l. $|y - f(x)|_\epsilon$ in (5) is Vapnik's ϵ-insensitive loss function.

Using Lagrange multiplier techniques (Cortes & Vapnik, 1995; Vapnik, 1995), Problem (4) can be transformed to a convex optimization problem with a global minimum. At the optimum, the regression estimate takes the form $f(x) = \sum_{i=1}^{l}(\alpha_i^* - \alpha_i)(x_i\cdot x) + b$, where $(x_i \cdot x)$ is the inner product.

In 2002, Bofinger, Luig, and Beyer (2002) found that the output of wind turbine systems is limited between zero and the maximum power and the error statistics do not follow a normal distribution. In 2005, Fabbri, Román, Abbad, and Quezada (2005) believed that the normalized produced power p must be within the interval $[0, 1]$ and the beta function is more appropriate to fit the error than the standard normal distribution function. Bludszuweit, Antonio, and Llombart (2008) showed the advantages of using the beta probability distribution function (PDF), instead of the Gaussian PDF, for approximating the forecast error distribution. The error ϵ between the predicted values x_p and the measured values x_m obeys the beta distribution in the forecast of wind power, and the PDF of ϵ is $f(\epsilon) = \epsilon^{m-1}\cdot(1-\epsilon)^{n-1}\cdot h$, $\epsilon \in (0, 1)$, the parameters m and n are often called hyperparameters because they control the distribution of the variable ϵ ($m > 1, n > 1$), h is the normalization factor and parameters m and n are determined by the values of the mean (which is the predicted power) and the standard deviation (Bishop, 2006; Canavos, 1984). Fig. 1 shows plots of Gaussian distribution and the beta distribution for different values of hyperparameters. In 2007, Zhang, Wan, Zhao, and Yang (2007) and Randazzo, Abou-Khousa, Pastorino, and Zoughi (2007) presented the estimation results under a Laplacian noise environment in the direction-of-arrival of coherent electromagnetic waves impinging estimation problem. Laplace distribution is frequently encountered in various machine learning areas, e.g., the over-complete wavelet transform coefficients of images, processing in Natural images, etc. (Eltoft, Kim, & Lee, 2006; Park & Lee, 2005).

Based on the above analysis, we know that the error distributions do not satisfy Gaussian distribution in some real-world applications. We try to study the optimal loss functions for different error models.

It is not suitable to apply the GN-SVR to fit functions from data with non-Gaussian noise. In order to solve the above problems,