International Journal of Innovative
Computing, Information and Control
Volume 4, Number 9, September 2008

ICIC International ©2008 ISSN 1349-4198
pp. 2325–2332

SUPPORT VECTOR REGRESSION WITH INPUT DATA UNCERTAINTY

Ping Zhong and Laisheng Wang

College of Science
China Agricultural University
Beijing 100083, P. R. China
pingsunshine@yahoo.com.cn; wanglaish@sina.com

Received July 2007; revised December 2007

Abstract. *Support vector regression is an important and ongoing research subject in machine learning. Conventional support vector regression implicitly assumes that the parameters in the optimization problems are known exactly. However, in practice, the parameters have perturbations since they are estimated from the training data that are usually subject to noise. In this article, we propose a second-order cone programming formulation for designing regression functions which are robust to input uncertainty. In addition, we give the intuitive geometric interpretation of the proposed formulation and extend it to the nonlinear case. The preliminary numerical experiments confirm the robustness of the proposed formulation.*

Keywords: Machine learning, Support vector regression, Second-order cone program, Robust estimation

1. **Introduction.** Given a training data set $\mathcal{T} = \{(x_1, y_1), \cdots, (x_l, y_l)\} \subseteq \mathcal{X} \times R$ with the input data $x_i \in \mathcal{X}$ and the observation $y_i \in R$, the main goal of regression problem is to find a function $f : \mathcal{X} \to R$ that can correctly predict the observation values of new input data points by learning from the given training data set $\mathcal{T}$. Geometrically this corresponds to a linear or nonlinear surface fitting the given points. The best-known solution to this problem is that proposed independently by Gauss and Legendre of choosing the line that minimizes the sum of the squares of the distances from the training points. This technique is known as least squares. However, ignorance of those errors that fall within some tolerance, say ε, may lead to a better generalization ability. At the same time, applying the idea of support vector machines (SVMs), the function $f(x)$ is made as flat as possible in fitting the training data. This problem is called ε-support vector regression (ε-SVR) [1-3]. Conventionally, ε-SVR is formulated as a convex quadratic programming problem. Regression functions are sought under the implicit assumption that noise is confined to the observations $\{y_i\}_{i=1}^l$ while the input data $\{x_i\}_{i=1}^l$ are not corrupted with noise. However, in practice, sampling errors, modeling errors and instrument errors may preclude the possibility of knowing the input data exactly. So the parameters in ε-SVR formulation have perturbations since they are estimated from the training data. As pointed out by Goldfarb and Iyengar [4], the solutions to the optimization problems are sensitive to parameter perturbations. When the input data contain noise, unaware of this situation, the conventional learning process tries to fit these unwanted data and may lead to incorrect prediction. So it will be useful to find robust estimators which do not change significantly when input data are only known approximately subject to some uncertainty. This problem is studied by Pannagadatta et al. in [5]. They regard each x_i as a random