

SUPPORT VECTOR REGRESSION WITH INTERVAL-INPUT INTERVAL-OUTPUT

Wensen An^{1,2†}, Cecilio Angulo², Yanguang Sun¹

¹ R&D Center, Automation Research and Design Institute of Metallurgical Industry, 100071 Beijing, China

² Knowledge Engineering Research Group, Universitat Politècnica de Catalunya, 08800 Vilanova i la Geltrú, Barcelona, Spain

Received: 18-04-2008

Revised: 02-06-2008

Support vector machines (classification and regression) are powerful machine learning techniques for crisp data. In this paper, the problem is considered for interval data. Two methods to deal with the problem using support vector regression are proposed and two new methods for evaluating performance for estimating prediction interval are presented as well.

Keywords: Support vector regression, Interval-input interval-output, Uncertainties, Quadratic programming

1. Introduction

In machine learning, we usually hope that the obtained sample data are as precise as possible in order to construct an exact model. For example, the system model often guides real applications in complex industrial process. And the success of the model-based applications generally depends on how exactly models reproduce the behavior of the real industrial process that they represent¹.

However, the sample data is usually imprecise and may be in some case missing, due to the various limitations. For example, even under the same experimental conditions with the same measuring tools, the obtained data usually differ from one to another because the measuring tools are never perfectly precise and environmental condition changes in industrial measurements². Moreover, discretization is utilized to describe qualitatively the sample data in order to speed up learning process when existing large-scale training dataset^{3, 4}. This is to say, the data often has the feature of uncertainty in real life. Generally, interval is a common tool to describe the data with uncertainty^{5, 6}.

Currently, little work has been done to study support vector machines (SVM)^{7, 8} with interval information. Interval-based support vector classification problems are discussed in literatures^{4, 9, 10}, while Chuang and Hwang^{11, 12, 13, 14, 15} paid attention to regression problems with interval output. However, the methods of

Chuang and Hwang were not completely based on interval because the input data they used were still crisp, not interval, that is, their methods were based on crisp-input interval-output (CIO) in their research.

In this paper, the interval-input interval-output (IIIO) problems based on support vector regression (SVR) are discussed. Moreover, we give two definitions of prediction accuracy in order to evaluate the differences between predicted outputs and real outputs.

The rest of the paper is organized as follows. In section 2, we discuss the decomposed approach to deal with IIIO problems based on SVR technique. First, we give the basic idea, and then the simulation results are demonstrated. Finally, two definitions of prediction accuracy to evaluate the differences between predicted outputs and real outputs are presented. Following that section 3 studies quadratic programming (QP) based SVR method to solve IIIO problems. Section 4 concludes the paper with a brief summary.

2. Decomposed approach

Let the dataset of intervals $'D$ be:

$$'D = \{('X_i, 'Y_i) | i = 1, \dots, n\} \in I^m \times I. \quad (1)$$

with interval input $'X_i$ and associated targets interval output $'Y_i$, respectively.

where

[†] Corresponding author. Tel: +86 10-63812255, Fax: +86 10-63851580. E-mail: anwensen@yahoo.com.cn

W.S. An, C. Angulo and Y.G. Sun

I^m denotes m-dimensional closed interval;

$$\begin{aligned} {}^l X_i &= ({}^l X_{i1}, {}^l X_{i2}, \dots, {}^l X_{im})^T; \\ {}^l X_{ij} &= (x_{ij}; s_{ij}), {}^l Y_i = (y_i; r_i); \end{aligned} \quad (2)$$

x_{ij}, s_{ij} denotes center and radius of interval ${}^l X_{ij}$;

y_i, r_i denotes center and radius of interval ${}^l Y_i$;

$x_{ij}, s_{ij}, y_i, r_i \in \square$;

$s_{ij}, r_i \geq 0$

Suppose the function, which approximates the relations between the interval input and interval output, is expressed as:

$${}^l Y = f({}^l X) \quad (3)$$

Moreover, by denoting c_Y, r_Y center and radius of the interval output ${}^l Y$, i.e. ${}^l Y = (c_Y; r_Y)$, we can obtain:

$${}^l Y = f({}^l X) = c_Y \pm r_Y = g(x_{ij}; s_{ij}) \pm h(x_{ij}; s_{ij}) \quad (4)$$

This means that standard SVR technique can be utilized to solve regression problems with interval by decomposing interval to its center and radius.

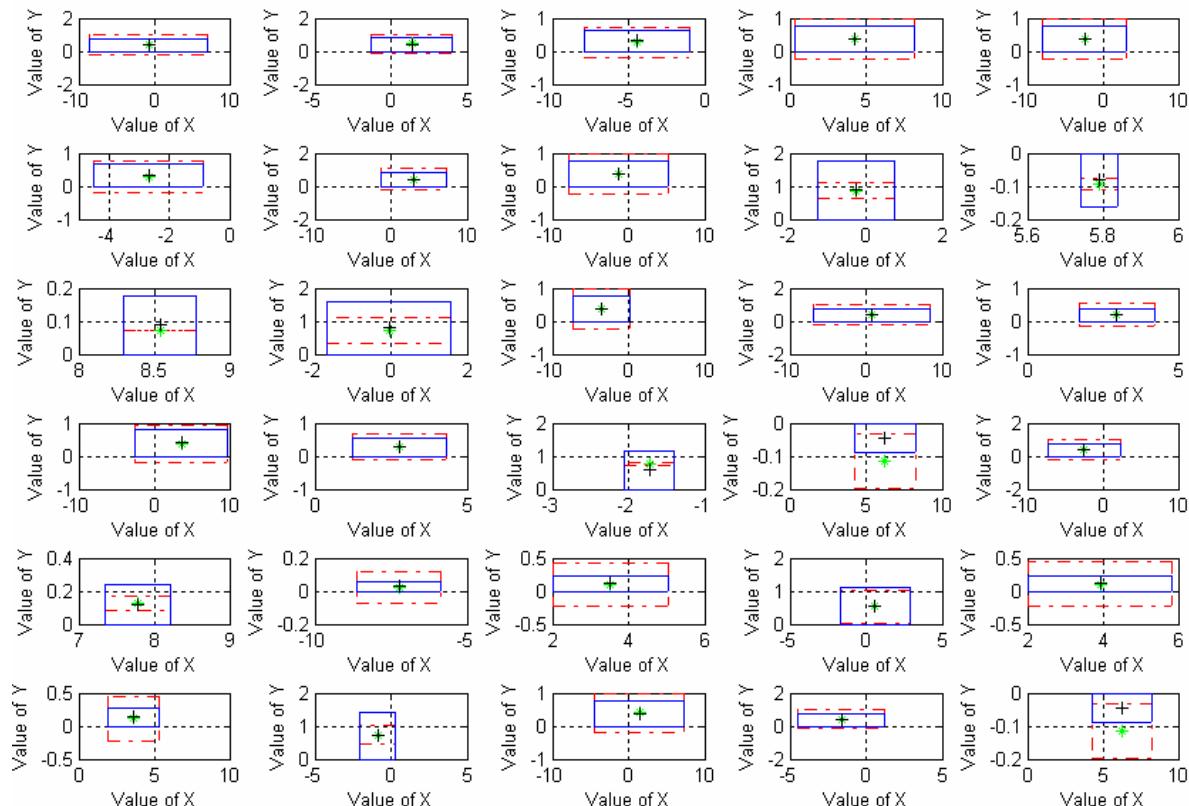


Fig.2. This is comparisons between predicted intervals and real ones used by *Sinc* function.

Now we demonstrate the results of the method through two experiments. In order to get some sample data, we first generate 150 points from *Sinc* function, whose expression is as $f(x) = \text{Sinc}(x) = \sin x/x$, and 2-dimensional Gabor function¹⁶, whose expression is as $f(x) = 0.5\pi * \exp(-2x^2)\cos(2\pi x^2)$, respectively; then create 150 pairs of interval by selecting two points randomly among them each time. Finally, we randomly select one of the 120 pairs of interval as the training data, the rest as the test data. In the experiments, LIBSVM¹⁷ is utilized and Guassian RBF is selected as the kernel of SVR. Moreover, 10-fold cross-validation method is used to obtain the parameters of kernel functions.

The *Sinc* function and generated sample data from it are shown in figure 1, while figure 2 shows the prediction results between predicted intervals and real interval. In figure 2, the dotted line denotes the real intervals and the real line denotes the predicted intervals. The symbols of “plus” and “star” denote the centers of real intervals and predicted intervals, respectively.

The 2-dimensional *Gabor* function and generated sample data from it are shown in figure 3, while figure 4

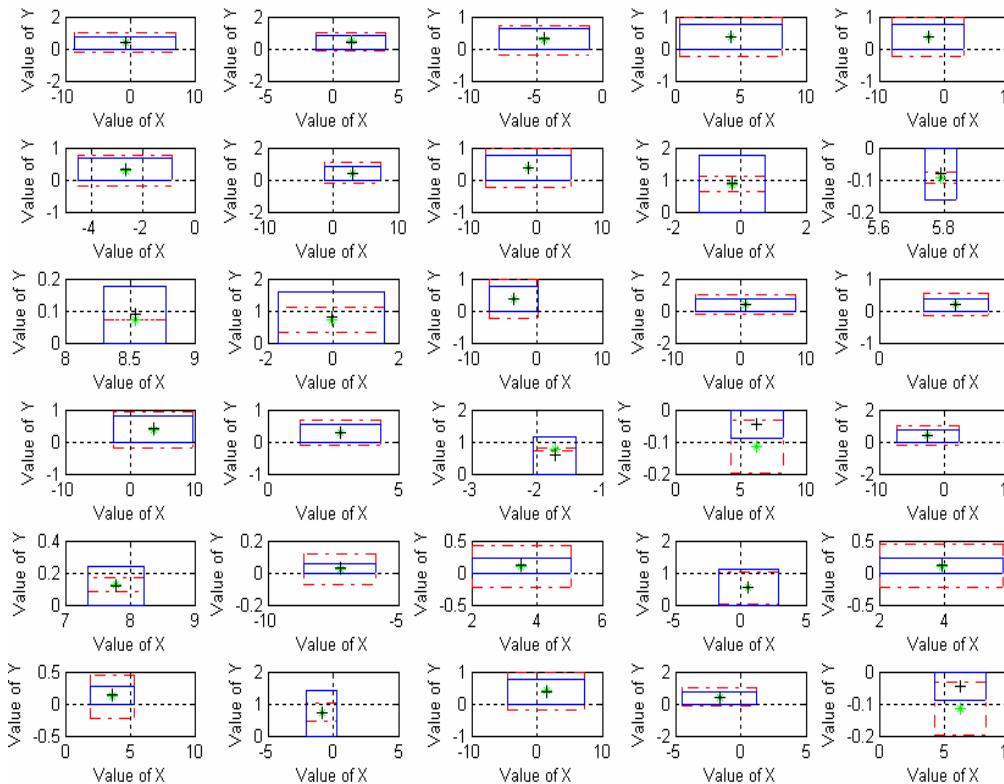


Fig.4 This is comparisons between predicted intervals and real ones used by *Gabor* function.

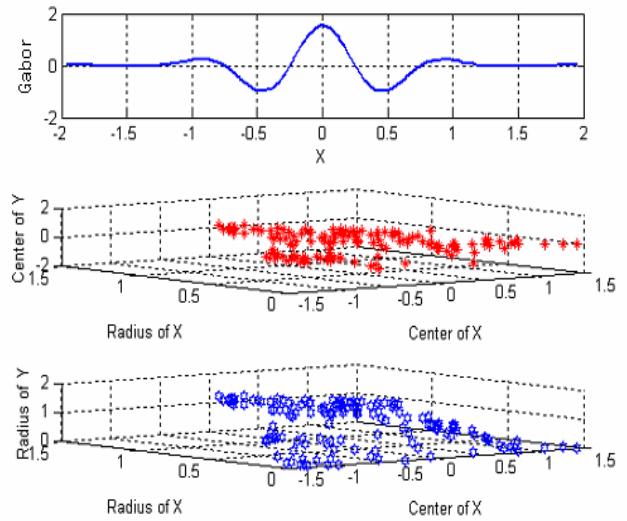


Fig. 3. This is *Gabor* function and generated sample data.

W.S. An, C. Angulo and Y.G. Sun

shows the prediction results between predicted intervals and real interval. In figure 4, the dotted line denotes the real intervals and the real line denotes the predicted intervals. The symbols of “plus” and “star” denote the centers of real intervals and predicted intervals, respectively.

In order to evaluate the differences between predicted intervals and real intervals, we define two expressions of prediction accuracy as follows.

Suppose real interval output and predicted interval output corresponding with the same interval input $'X_i$ are denoted by $'Y_i$ and $'\tilde{Y}_i$. The area of rectangle by interval $'X_i$ and $'Y_i$ is denoted by A_i^{real} , the area of rectangle constructed by interval $'X_i$ and $'\tilde{Y}_i$ is denoted by A_i^{pred} , and the overlapped area between the above rectangles is denoted by A_i^{lap} . Then we have the following definitions:

$$\begin{aligned} Accu_1 &= \frac{1}{N} \sum_{i=1}^N \frac{2A_i^{lap}}{A_i^{real} + A_i^{pred}} \\ Accu_2 &= \frac{2 \sum_{i=1}^N A_i^{lap}}{\sum_i^N A_i^{real} + \sum_i^N A_i^{pred}} \end{aligned} \quad (5)$$

According to the expression (5), the accuracy rates on the two experiments are as follows: i) for the sample data generated from *Sinc* function, $Accu_1 = 0.81221$ and $Accu_2 = 0.96736$; ii) for the sample data generated from 2-diminsional *Gabor* function, $Accu_1 = 0.88778$, $Accu_2 = 0.97547$. From the results of the experiments, the accuracy rate in the second form is higher than the first one, and this fact is easy to understand because that, when the overlapped rectangle constructed by predicted interval output is big on the big interval input and small on the small interval input, the big overlapped rectangles will “weaken” the influences of the small ones according to the second formulation of prediction accuracy.

3. QP-based approach

As we know, the standard algorithm of SVR can be realized via solving a quadratic programming (QP) problem, and then utilizes the kernel trick. So now a problem comes across, i.e. can we construct an interval-based SVR model by solving a QP problem like standard SVR?

Suppose we have the datasets of intervals like expression (1) and expression (2), and the non-linear regression model can be described as follows:

$$'Y_i = \psi('X_i) = a_0 'X_{i0} + a_1 'X_{i1} + \cdots + a_m 'X_{im} \quad (6)$$

where $X_{i0} = (x_{i0}; s_{i0}) = (1; 1)$, and $a_0, a_1, \dots, a_m \in \mathbb{Q}$ are regression coefficients.

According to the interval analysis^{5, 6}, the expression (6) can be formulated:

$$\begin{aligned} 'Y_i &= a_0(x_{i0}; s_{i0}) + a_1(x_{i1}; s_{i1}) + \cdots + a_m(x_{im}; s_{im}) \\ &= \left(\sum_{j=0}^m a_j x_{ij}; \sum_{j=0}^m |a_j| s_{ij} \right) \\ &= (\mathbf{a}^T \mathbf{x}_i; |\mathbf{a}|^T \mathbf{s}_i) \end{aligned} \quad (7)$$

where $\mathbf{a} = (a_0, a_1, \dots, a_m)^T$, $\mathbf{x}_i = (x_{i0}, x_{i1}, \dots, x_{im})^T$, $\mathbf{s}_i = (s_{i0}, s_{i1}, \dots, s_{im})^T$.

In a similar way to build the standard quadratic loss function based SVR algorithm, the QP-based interval SVR model can be constructed as follows:

$$\begin{aligned} \min \quad & \frac{1}{2} \|\mathbf{a}\|^2 + \frac{C}{2} \left[\sum_{i=1}^n \xi_{1i}^2 + \sum_{i=1}^n (\xi_{2i}^2 + \xi_{2i}^{*2}) + \sum_{i=1}^n (\xi_{3i}^2 + \xi_{3i}^{*2}) \right] \\ \text{s.t.} \quad & \begin{cases} \mathbf{a}^T \mathbf{s}_i \leq \xi_{1i} \\ y_i - \mathbf{a}^T \mathbf{x}_i \leq \xi_{2i} \\ \mathbf{a}^T \mathbf{x}_i - y_i \leq \xi_{2i}^* \\ \mathbf{a}^T \mathbf{x}_i + |\mathbf{a}|^T \mathbf{s}_i \leq y_i - r_i + \xi_{3i}, \quad i = 1, 2, \dots, n \\ \mathbf{a}^T \mathbf{x}_i - |\mathbf{a}|^T \mathbf{s}_i \geq y_i - r_i - \xi_{3i}^* \\ \xi_{1i}, \xi_{2i}, \xi_{2i}^*, \xi_{3i}, \xi_{3i}^* \geq 0 \end{cases} \end{aligned} \quad (8)$$

However, the formulation (8) is the primal model of interval SVR. How to efficiently solve its dual problem and make good use of kernel trick, or realize it directly in the primal problem¹⁸ is still a further research topic.

4. Conclusions

We have presented the approaches to deal with problem with interval-input interval-output using support vector regression algorithm in this paper. We have also proposed two ways to evaluate the prediction accuracy, that is, the differences between predicted values and real values. In future work, we intend to solve the proposed QP-based interval SVR model and employ them to the industrial applications.

Acknowledgements

Wensen An acknowledges the helpful discussions with Prof. Andreu Català and Dr. Xavier Parra. The authors would like to thank anonymous reviewers for their

Support Vector Regression with Interval-Input Interval-Output

valuable comments that have helped to improve the quality of this paper.

References

1. M. Rebolledo, Rough Intervals--Enhancing Intervals for Qualitative Modeling of Technical Systems, *Artif. Intell.*, **170**(8) (2006): 667–685.
2. P. Nivlet, F. Fournier and J. Royer, Interval Discriminant Analysis: An Efficient Method to Integrate Errors in Supervised Pattern Recognition, in *2nd Int. Symp. Imprecise Probabilities and Their Applications*, eds. G. de Cooman, T. Fine and T. Seidenfeld, (Ithaca, New York, 2001), pp.284–292.
3. H. H. Bock and E. Diday, Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data, Series: *Studies in Classification, Data Analysis, and Knowledge Organization* (Springer-Verlag, NY, 2000).
4. C. Angulo, D. Anguita and L. Gonzalez-Abril, Support Vector Machines for Interval Discriminant Analysis, *Neurocomputing*, **71**(7-9) (2008): 1220–1229.
5. R. E. Moore, *Interval Analysis*, (Prentice-Hall, Englewood, NJ, 1966).
6. L. Jaulin, M. Kieffer, O. Didrit and E. Walter, *Applied Interval Analysis* (Springer-Verlag, Berlin, 2001).
7. V. N. Vapnik, *The Nature of Statistical Learning Theory* (Springer, NY, 1999).
8. V. N. Vapnik, *Statistical Learning Theory* (Wiley, NY, 1998).
9. P. Lingras and C. J. Butz, Interval Set Classifiers Using Support Vector Machines, Proc. 2004 Conf. North American Fuzzy Information Processing Society (NAFIPS04), **2**: 707–710, 2004.
10. Y. G. Zhao, Q. M. He and Q. Chen, An Interval Set Classification Based on Support Vector Machines, Proc. the Joint Int. Conf. Autonomic and Autonomous Systems and Int. Conf. Networking and Services (ICAS/ICNS 2005), 2005.
11. J. T. Jeng, C. C. Chuang and S. F. Su, Support Vector Interval Regression Networks for Interval Regression Analysis, *Fuzzy Sets Syst.*, **138**(2) (2003): 283–300.
12. D. H. Hong and C. H. Hwang, Support Vector Machine for Interval Regression, Proc. the Autumn Conf., Korean Statistical Society, 67–72, 2004.
13. D. H. Hong and C. H. Hwang, Interval Regression Analysis Using Quadratic Loss Support Vector Machine, *IEEE T. Fuzzy Syst.*, **13**(2)(2005): 229–237.
14. C. H. Hwang, D. H. Hong, E. Na, H. Park and J. Shim, Interval Regression Analysis Using Support Vector Machine and Quantile Regression, *Springer Lect. Notes Artif. Int.*, **3613** (2005): 100–109.
15. C. H. Hwang, D. H. Hong and K. H. Seok, Support Vector Interval Regression Machine for Crisp Input and Output Data, *Fuzzy Sets Syst.*, **157** (2006): 1114–1125.
16. Z. H. Zhou, J. X. Wu and W. Tang, Ensembling Neural Networks: Many Could Be Better Than All, *Artif. Intell.*, **137**(1-2) (2002): 239–263.
17. C. C. Chang, C. J. Lin, LIBSVM: A Library for Support Vector Machines (Version 2.3), Department of Computer Science, National Taiwan University, 2001, <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.
18. O. Chapelle, Training A Support Vector Machine in The Primal, Max Planck Institute for Biological Cybernetics Technical Report, No.TR-147, April 2006, [http://www.kyb.mpg.de/publications/attachments/tr_3597\[1\].pdf](http://www.kyb.mpg.de/publications/attachments/tr_3597[1].pdf).