IGI PUBLISHING

ITB15227

701 E. Chocolate Avenue, Suite 200, Hershey PA 17033-1240, USA
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.igi-global.com

This paper appears in the publication, **Data Mining and Knowledge Discovery Technologies** edited by D. Taniar © 2008, IGI Global

Chapter X

Model Free Data Mining[1]

Can Yang, Zhejiang University, Hangzhou, P. R. China
Jun Meng, Zhejiang University, Hangzhou, P. R. China
Shanan Zhu, Zhejiang University, Hangzhou, P. R. China
Mingwei Dai, Xi'an Jiao Tong University, Xi'an, P. R. China

Abstract

Input selection is a crucial step for nonlinear regression modeling problem, which contributes to build an interpretable model with less computation. Most of the available methods are model-based, and few of them are model-free. Model-based methods often make use of prediction error or sensitivity analysis for input selection and model-free methods exploit consistency. In this chapter, we show the underlying relationship between sensitivity analysis and consistency analysis for input selection, and then derive an efficient model-free method from our common sense, and then formulate this common sense by fuzzy logic, thus it can be called fuzzy consistency analysis (FCA). In contrast to available methods, FCA has the following desirable properties: (1) it is a model-free method so that it will not be biased on a specific *model, exploiting "what the data say" rather than "what the model say," which is the essential point of data mining—input selection should not be biased on a specific model, (2) it is implemented as efficiently as classical model-free methods, but more flexible than them, and (3) it can be directly applied to a data set with mix continuous and discrete inputs without doing rotation. Four benchmark problems study indicates that the proposed method works effectively for nonlinear problems. With the input selection procedure, the underlying reasons, which affect the prediction are work out, which helps to gain an insight into a specific problem and servers the purpose of data mining very well.*

Introduction

For real-world problems such as data mining or system identification, it is quite common to have tens of potential inputs to the model under construction. The excessive inputs not only increase the complexity of the computation necessary for building the model (even degrade the performance of the model, which is the curse of dimensionality) (Bellman, 1961; Hastie, Tibshirani, & Friedman, 2001), but also impair the transparency of the underlying model. Therefore, a natural choice of the solution is the number of inputs actually used for modeling should be reduced to the necessary minimum, especially when the model is nonlinear and contains many parameters. Input selection is thus a crucial step for the purposes of (1) removing noises or irrelevant inputs that do not have any contribution to the output; (2) removing inputs that depend on other inputs; and (3) making the underlying model more concise and transparent. However, figuring out which ones to keep and which ones to drop is a daunting task. Large arrays of feature selection methods, like the principal component analysis (PCA), have been introduced in linear regression problems. However, they usually fail to discover the significant inputs in real-world applications, which often involve nonlinear modeling problems.

Input selection thus has drawn great attention in recent years, and some methods have been presented which can be categorized into two groups:

1. Model-based methods, which use a particular model in order to find the significant inputs. In general, model-based methods do input selection mainly by (a) trial and error or (b) sensitivity analysis. They need to try different combinations of inputs to find a good subset for our model to use.
   a. **Trial and error:** This kind of method usually builds up a specified model first with training data, and then checks its prediction accuracy by checking data. Wang (2003) proposed his method for variable importance ranking based on mathematic analysis of approximation accuracy. A relatively