Chapter 2.34
Cluster-Based Input Selection for Transparent Fuzzy Modeling[1]

Can Yang
Zhejiang University, China

Jun Meng
Zhejiang University, China

Shanan Zhu
Zhejiang University, China

ABSTRACT

Input selection is an important step in nonlinear regression modeling. By input selection, an interpretable model can be built with less computational cost. Input selection thus has drawn great attention in recent years. However, most available input selection methods are model-based. In this case, the input data selection is insensitive to changes. In this article, an effective model-free method is proposed for the input selection. This method is based on sensitivity analysis using Minimum Cluster Volume (MCV) algorithm. The advantage of our proposed method is that with no specific model needed to be built in advance for checking possible input combinations, the computational cost is reduced, and changes of data patterns can be captured automatically. The effectiveness of the proposed method is evaluated by using three well-known benchmark problems that show that the proposed method works effectively with small and medium-sized data collections. With an input selection procedure, a concise fuzzy model is constructed with high accuracy of prediction and better interpretation of data, which serves well the purpose of patterns discovery in data mining.

INTRODUCTION

Fuzzy Inference System (FIS) is one of the most important applications in fuzzy logic and fuzzy set theory (Zadeh, 1973). FIS is useful for description, prediction and control in many fields, such as data mining, diagnosis, decision support, system identification and control. The strength of FIS comes from two aspects: (1) it can handle linguistic concepts; (2) it is universal approxima-

tions (Wang & Mendel, 1992). Although expert knowledge can be easily incorporated in building a FIS, it has been seen that the expert knowledge would lead to an insufficient accuracy of FIS in modeling complex systems. So in recent years, many researchers have attempted to generate FIS from observed data (Wang & Mendel, 1992; Jang, 1993; Babuska, 1998; Sugeno & Yasukawa, 1993; Kosko, 1997).

In real-world applications such as system identification problems, it is common to have tens of potential inputs to a model under construction. The excessive inputs not only impair transparency of the underlying model, but also increase computational complexity in building the model. Therefore, the number of inputs actually used in modeling should be reduced to a sufficient minimum, especially when the model is nonlinear and has high dimensionality. Input selection is thus becoming 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; (3) making the underlying model more concise and transparent. A large array of input selection methods, like analysis of correlation, the principal component analysis (PCA) and the least squares method have been introduced in linear regression problems. However, they usually fail to discover significant inputs in real-world applications which often involve nonlinear modeling. Relatively a few methods are available for input selection in nonlinear modeling. These methods found in literature can generally be divided into two categories:

1. **Model-based methods** that use a specific model to search for significant inputs. Finding an optimal solution of input selection often requires examining different models for all possible combinations of inputs, which becomes computationally intractable even for a reasonable number of input attributes. In order to avoid this, heuristic criteria often were introduced. A relatively simple and fast method was proposed in Jang (1996) by using ANFIS. This method is based on an assumption that the ANFIS model with the smallest root mean squared error (RMSE) after one epoch of training has a greater potential to achieve a lower RMSE when given more epochs of training. The heuristic method, which generates ANFIS sequentially by involving increased number of inputs, is called forward selection. Although this method was developed for ANFIS, the same idea could be used for other types of FISs. Methods presented in Tanaka, Sano, and Watanabe (1995) and Chiu (1996) used a different heuristic method; namely, backward selection. Lin and Cunningham (1994) proposed a method based on fuzzy curves that represent the sensitivity of an output with respect to other inputs. Some other methods, based on criteria like individual discrimination power and entropy variation index, were proposed in Hong and Chen (1999) and Pal (1999). However, these methods are restricted to deal with classification problems and assumed that the input variables are independent. This assumption usually cannot be satisfied in real-world problems. Another algorithm was proposed in Wang (2003), based on mathematic analysis of approximation accuracy.
2. **Model-free methods** that do not need to develop models for measuring relevant inputs. For instance, the method proposed in He and Asada (1993) exploits the continuity property of nonlinear functions. The so-called Lipschitz coefficients are computed in order to find the optimal order of an input-output model. Emami, Turksen, and Goldenberg (1998) proposed a method based on geometric criterion, in which a non-significant index was developed in order to find the most important inputs. A