

1.2 ENHANCED AND OPTIMIZED GMDH ALGORITHMS

TS-Based GMDH Model and Its application*

He Changzheng, Zhu Bing, Zheng Mingcui

Business School of Sichuan University, Chengdu 610064, P.R. China
hechangzheng@scu.edu.cn, zhubing1866@hotmail.com

Abstract. *In this paper, FRI algorithm which has some deficiencies in feature extraction of market segments groups is improved. By replacing Mamdani fuzzy inference with TS fuzzy inference, a new TS-based GMDH model is built. The algorithm is realized with simple Matlab code. It has been demonstrated in the empirical research that TS-based GMDH model improves the deficiencies of FRI in extracting features of different market groups. This result is further development of the theory and method of GMDH and provides a new approach for inductive modeling.*

Keywords

FRI algorithm, TS model, GMDH, Feature extraction

1. Introduction

The fuzzy model is usually divided into two categories: Mamdani fuzzy model [1] and TS fuzzy model [2]. Self-Organising Fuzzy Rule Induction (FRI) is a modeling technique combining Group Method of Data Handling (GMDH) [3] and Mamdani fuzzy model and FRI can be called GMDH based on Mamdani. FRI can extract features from sample data automatically and form fuzzy model that qualitatively describes system behaviors using a natural language [4]. Although FRI is appropriate to extraction of distinctive features between different segments market, we find that FRI has a low modeling accuracy in extracting features when there are many common features and few distinctive features in different segment markets [5].

The consequent part of Mamdani fuzzy model consists of fuzzy sets defined in the space of output variables while the TS fuzzy model is composed of a linear function of some input variables. Compared with Mamdani fuzzy model, TS fuzzy model has the advantage that approximates complex nonlinear systems with fewer rules and high modeling accuracy [6]. In this paper we attempt to substitute TS fuzzy inference for Mamdani inference used in FRI and divide the sample data into training set and testing set. The data division technique that is not used and can't be used in FRI is the core of the GMDH^[7]. In this way, a GMDH model based on TS fuzzy inference is built, which is called TS-GMDH for short. It has been demonstrated in the empirical research that TS-GMDH, in comparison with FRI, has relatively high accuracy, even when feature differences between customer segments are not obvious.

2. TS Fuzzy Model

The TS fuzzy model was proposed by Takagi and Sugeno in 1985. For multi-input single-output systems, the typical TS model consists of a set of IF-THEN rules and each rule is composed of an antecedent part and a consequent part as follows^[2]:

$$R^l: \text{if } x_1 \text{ is } A_1^{s_1}, \dots, x_m \text{ is } A_m^{s_m} \text{ then } y^l = P_0^l + P_1^l x_1 + \dots + P_m^l x_m, \quad l = 1, 2, \dots, M \quad (1)$$

where R^l denotes the l th rule x_i is the i th input variable, P_j^l represents the j th parameter of the l th rule, $A_i^{s_i}$ represents a fuzzy set defined in the space of i th input variable x_i .

*This work has been supported by National Natural Science Foundation of China (70771067).

The final output y is the weighted average of each rule's output y^l according to following formula:

$$y = \frac{\sum_{l=1}^M G^l y^l}{\sum_{l=1}^M G^l} \quad (2)$$

where G^l is the firing strength of l th rule and calculated as follows:

$$G^l = \prod_{i=1}^m A_i^{S_i}(x_i)$$

where \prod denotes a fuzzy conjunction operator, $A_i^{S_i}(x_i)$ is a membership function corresponding to a fuzzy set $A_i^{S_i}$ of the i th input variable. According to (2), the parameters of $y^l, l=1,2,\dots,M$ could be estimated by Ordinary Least Square (OLS).

3. TS-GMDH modeling algorithm

The modeling of TS-GMDH is very similar to FRI, the only differences are that the min-max fuzzy inference^[4] has become TS fuzzy inference, and the sample data set N is divided to a training set A and a testing set $B(N = A \cup B)$. The main steps of TS-GMDH modeling are as follows:

- 1) Fuzzification of variables. Every input variable $x_i (i = 1, 2, \dots, n)$ is transformed into m fuzzy linguistic variables, there will be mn inputs fuzzy variables in the first layer in the network (see [4] and Fig. 1).
- 2) Forming of the first generation TS models. All the input fuzzy sets are combined in pairs to form the first generation TS models. For example, if two fuzzy sets $A_i^{S_i}$ and $A_j^{S_j}$ defined in the space of input variable x_i and $x_j (i, j = 1, 2, \dots, n)$ are combined, we get following TS model:

$$R_1^l : \begin{cases} \text{if } x_i \text{ is } A_i^{S_i} \text{ then } y_1^1 = a_{10} + a_{1i}x_i \\ \text{if } x_j \text{ is } A_j^{S_j} \text{ then } y_1^2 = b_{10} + b_{1j}x_j \end{cases}$$

$$i, j = 1, 2, \dots, n, i \neq j; l = 1, 2, \dots, m^2 C_n^2$$

where C_n^2 is the number of combination* the parameters in the consequent parts are estimated by OLS in the training set A (See Section 2)

- 3) Selecting Models. F_l best TS models are selected by external criterion in the testing set B (see the Fig. 1). Regularity criterion [7] in following form is used as external criterion of our algorithm:

$$\sum_{i=1}^{|B|} \left(y - \left(\frac{G^1}{G^1 + G^2} \hat{y}_1^1 + \frac{G^2}{G^1 + G^2} \hat{y}_1^2 \right) \right)^2 \quad (5)$$

where y is the sample output, $|B|$ denotes the number of sample data in the set B , \hat{y}_1^1 and \hat{y}_1^2 represent the outputs of the two rules in the model, $G^1 = A_i^{S_i}(x_i)$ and $G^2 = A_j^{S_j}(x_j)$ are firing strength of the two rules.

- 4) Rules fusion. The two rules of R_1^l are merged into one rule which is still denoted by R_1^l :

$$R_1^l : \text{if } x_i \text{ is } A_i^{S_i} \text{ and } x_j \text{ is } A_j^{S_j} \text{ then } y_1^l = a_{10} + a_{1i}x_i + a_{1j}x_j, l = 1, \dots, F_l \quad (6)$$

where fuzzy set $A_i^{S_i}$ and $A_j^{S_j}$ will do the conjunction operation, the coefficients of linear function in the

consequent part remain to be decided.

- 5) Circulation of the algorithm. Step 2), step 3) and step 4) are repeated to create the 2th, 3th...generation TS models until the external criterion (5) begin to increase, F_k fuzzy rules make up of the final TS model, and its parameters in the consequent part are estimated by OLS in the training set A (See Section 2).
- 6) Calculating the final output. The output y is computed using the consequent parts of the F_k rules of the final TS model according formula (2).

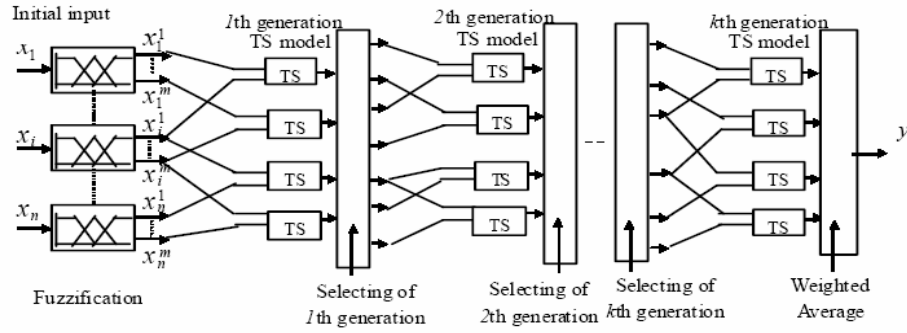


Fig. 1. Network of TS-GMDH modeling

4. Empirical research

In order to verify the effectiveness of TS-GMDH in feature extracting, it is used to extract the feature of different consumer groups of medium-priced cigarette. The sample data comes from the consumer market research of medium-priced cigarette conducted by a cigarette manufacture located in Sichuan province of China. 180 sample data and 60 variables are obtained from the research (including 1 input variable and 59 output variable). The purpose of modeling is to extract features of two smoking group—heavy smoking group and mild smoking group from data analysis. We use FRI and TS-GMDH respectively and randomly choose 150 sample data denoted as set N for modeling and 30 sample data denoted as set M for validating. FRI is realized in software KnowledgeMiner and TS-GMDH with Matlab code. Table 1 shows the results.

The smokers, whether heavy smoking group or mild smoking group, have some common features as follows: males, aged between 30 and 40, low education degree and low income (less than 1000 RMB per month). These conclusion can be drawn from descriptive statistical analysis of sample data. Therefore, use of FRI in feature extraction results in poor modeling accuracy, especially in the validation set M , as Table 1 has shown. On the contrary, TS-GMDH has relatively high modeling accuracy in both modeling data set N and validation data set M . This phenomenon indicates that TS-GMDH improve the performance of FRI in feature extraction when feature difference between customer segments is not obvious.

Tab.1. Comparison for FRI and TS-GMDH

Method	Heavy smoking group		Mild smoking group	
	TV	M	TV	M
FRI	83.33%	70%	83.33%	70%
TS-GMDH	94%	93.33%	89.33%	87.33%

5. Conclusion

In this paper TS model is integrated with the mechanism of inductive modeling, and a new model named TS-GMDH is proposed. The TS-GMDH model improves FRI in two aspects: Mamdani fuzzy inference is replaced by TS fuzzy inference and sample data for modelling is divided into two subset. These improvements make TS-GMDH has higher modeling accuracy and suitable for feature extracting when there are many common features and few distinctive features between different segment markets. It is proved by empirical study that TS-GMDH has make up for FRI's deficiencies in feature extracting. This new method is a beneficial exploration of theory and method of GMDH and offer a more practical tool for the analysis of enterprise market researches data.

References

- [1] Mamdani E.H. Application of fuzzy algorithms for control of simple dynamic plant. *Proceeding of the Institution of Electrical Engineers*, 1974, 121: 1585-1588.
- [2] Takagi T, Sugeno M. Fuzzy Identification of Systems and its Application to Modelling and Control [J]. *IEEE Transaction on Systems, Man, and Cybernetics*, 1985, 15 (1), 116-132.
- [3] Ivakhnenko A.G. The group method of data handling in prediction problems. *Soviet Automatic Control c/c of Avtomatika*, 1976, 9(6):21-30.
- [4] Mueller J.A., Lemke F. *Self-organizing data mining*. Herstellung. Berlin: Libri Books on Demand, 2000.
- [5] Changzheng He. *Self-organizing data mining and economic forecasting*. Beijing: Science Press, 2005:155-162.
- [6] Juang C.F., Lin C.T. An self-constructing neural fuzzy inference network and its applications. *IEEE Trans. Fuzzy Syst.* 1998, 6 (1): 12-31.
- [7] Madala H.R, Ivakhnenko A.G. *Inductive learning algorithms for complex systems modeling*. Boca Raton, London, Tokyo: CRC Press. Inc. 1994.