2014 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)
July 6-11, 2014, Beijing, China

Knowledge-Leverage Based TSK Fuzzy System with Improved Knowledge Transfer

Zhaohong Deng, *IEEE Senior Member*, Yizhang Jiang, Longbing Cao, *IEEE Senior Member*, Shitong Wang

***Abstract*—In this study, the improved knowledge-leverage based TSK fuzzy system modeling method is proposed in order to overcome the weaknesses of the knowledge-leverage based TSK fuzzy system (TSK-FS) modeling method. In particular, two improved knowledge-leverage strategies have been introduced for the parameter learning of the antecedents and consequents of the TSK-FS constructed in the current scene by transfer learning from the reference scene, respectively. With the improved knowledge-leverage learning abilities, the proposed method has shown the more adaptive modeling effect compared with traditional TSK fuzzy modeling methods and some related methods on the synthetic and real world datasets.**

***Index Terms*—Improved KL-TSK-FS, Fuzzy systems, Knowledge leverage, Missing data, Fuzzy modeling, Transfer learning**

I. INTRODUCTION

What is transfer learning? Most modeling methods require sufficient data to be collected for model learning. On one hand, in many real world applications, the available data may be insufficient since the data is scare or very noisy. In this situation, many tradition modeling methods become unfeasible. On the other hand, for a current scene, usually some reference scenes exist along with amounts of useful information. Although these reference scenes are different from the current scene, they are similar to it to a certain extent. How to use useful information in the reference scene to improve the modeling effect for the current scene is becoming a significant area of research. Transfer learning is just the technique to address this topic [1-11]. Recently, transfer learning has been studied extensively for different learning tasks, such as supervised learning[2-8] and unsurprised learning[9-11]. In this study, our focus is transfer learning for fuzzy systems. As a kind of classical intelligent models, fuzzy systems have been extensively applied in many fields [12]. Thus, for the situation where transfer learning is required, it is very significant that the fuzzy system modeling methods with the transferring abilities are available. For example, the modeling of fermentation process [6, 13] is one example where the transfer learning is required.

How to make the classical fuzzy modeling methods to have the ability of transfer learning? In [5], a kernel density estimation based transfer learning mechanism is introduced to develop the modeling method with the transfer learning abilities for the Mamdani-Larsen fuzzy system (ML-FS), i.e., the knowledge-leverage based ML fuzzy system (KL-ML-FS) modeling method. In [6], one kind of transfer learning mechanism is proposed for the development of the transfer learning TSK-FS modeling method, i.e., the knowledge-leverage based TSK-FS (KL-TSK-FS) modeling method. In the above transfer learning fuzzy system modeling methods, the KL-TSK-FS modeling method has shown the stronger learning abilities than the KL-ML-FS modeling method [5,6], but there are still many rooms to improve it due to the following weaknesses for KL-TSK-FS. (1) The antecedent parameters of the TSK-FS constructed by KL-TSK-FS algorithm is directly inherited from the model obtained in the reference scene, which make the obtained model not very appropriate to the modeling task of the current scene. (2) The knowledge-leverage mechanism used for the parameter learning of the consequents is still much weaker. Thus, more advanced knowledge-leverage transfer learning mechanism can be expected.

How to further improve the ability of transfer learning for KL-TSK-FS? In this study, a novel KL-TSK-FS with improved knowledge-transfer (KL-TSK-FS-IKT) is proposed, and we have addressed the KL-TSK-FS-IKT from two aspects as follows.

1) The transfer fuzzy c-mean clustering technique is proposed to realize the knowledge-leverage for the antecedents, which can make the parameter learning in the antecedents from the available data in the current scene and the knowledge of the reference scene simultaneously.

2) The improved knowledge-leverage mechanism is also introduced for the parameter learning in the consequents. Besides the knowledge-leverage term in the original KL-TSK-FS modeling method, an additional knowledge-leverage term has been introduced, which will make the obtained model parameters in the consequents to absorb more knowledge from the reference scene in the learning procedure.

The rest of this paper is organized as follows. In section II, the concept and principle of classical TSK-FS systems, and the KL-TSK-FS modeling method are reviewed, respectively. In

Z.H. Deng is with the School of Digital Media, Jiangnan University, Wuxi 214122, China and the Department of Biomedical Engineering, University of California Davis, Davis, 95616, USA (e-mail: dzh666828@aliyun.com and zhdeng@ucdavis.edu).

Y. Z Jiang and S.T. Wang are with the School of Digital Media, Jiangnan University, Wuxi 214122, China (e-mail: s101914015@vip.jiangnan.edu.cn, wxwangst@aliyun.com).

L. B. Cao is with the Advanced Analytics Institute, University of Technology Sydney, Australia (e-mail: longbing.cao@uts.edu.au)

This work was supported in part by the National Natural Science Foundation of China (61170122, 61272210), the Ministry of education program for New Century Excellent Talents (NCET-120882), the Fundamental Research Funds for the Central Universities (JUDCF13030) and 2013 Postgraduate Student's Creative Research Fund of Jiangsu Province (CXZZ13_0760), Australian Research Council Discovery Grants (DP130102691) and Linkage Grants (LP120100566).

Section III, the existing weaknesses of the KL-TSK-FS modeling method are discussed. In section IV, the KL-TSK-FS-IKT is proposed. The proposed method is evaluated with the experiments described in section V, along with the results and discussion. The conclusions are given in the final section.

II. Classical TSK-type Fuzzy Systems

The TSK model is the most popular fuzzy system model due to its effectiveness. In this section, the concept and principle of classical TSK-FS are first reviewed briefly. And then the KL-TSK-FS is introduced with the more details.

A. TSK Fuzzy Systems

For TSK fuzzy systems, the most commonly used fuzzy inference rules are defined as follows.

TS K Fuzzy Rule R^k :

IF x_1 is A_1^k $\wedge x_2$ is A_2^k $\wedge\cdots\wedge x_d$ is A_d^k (1)

Then $f^k(\mathbf{x}) = p_0^k + p_1^k x_1 + \cdots + p_d^k x_d$, $k = 1,\cdots,K$.

In Eq. (1), A_i^k is a fuzzy subset subscribed by the input variable x_i for the *k*-th rule; *K* is the number of fuzzy rules, and $\wedge$ is a fuzzy conjunction operator. Each rule is premised on the input vector $\mathbf{x} = [x_1, x_2, \cdots, x_d]^T$, and maps the fuzzy sets in the input space $A^k \subset R^d$ to a varying singleton denoted by $f^k(\mathbf{x})$. When *multiplicative conjunction* is employed as the conjunction operator, *multiplicative implication* as the implication operator, and *additive disjunction* as the disjunction operator, the output of the TSK fuzzy model can be formulated as

$$y^0 = \sum_{k=1}^{K} \frac{\mu^k(\mathbf{x}) f^k(\mathbf{x})}{\sum_{k'=1}^{K} \mu^{k'}(\mathbf{x})} = \sum_{k=1}^{K} \tilde{\mu}^k(\mathbf{x}) f^k(\mathbf{x}) , \quad (2.a)$$

where $\mu^k(\mathbf{x})$ and $\tilde{\mu}^k(\mathbf{x})$ denote the fuzzy membership function and the normalized fuzzy membership associated with the fuzzy set A^k . These two functions can be calculated by using

$$\mu^k(\mathbf{x}) = \prod_{i=1}^{d} \mu_{A_i^k}(x_i) \text{ and} \quad (2.b)$$

$$\tilde{\mu}^k(\mathbf{x}) = \mu^k(\mathbf{x}) \Big/ \sum_{k'=1}^{K} \mu^{k'}(\mathbf{x}) . \quad (2.c)$$

A commonly used fuzzy membership function is the Gaussian membership function which can be expressed by

$$\mu_{A_i^k}(x_i) = \exp\left(\frac{-(x_i - c_i^k)^2}{2\delta_i^k}\right) , \quad (2.d)$$

where the parameters c_i^k, δ_i^k can be estimated by clustering techniques or other partition methods. For example, with fuzzy c-means (FCM) clustering, c_i^k, δ_i^k can be estimated as follows,

$$c_i^k = \sum_{j=1}^{N} u_{jk} x_{ji} \Big/ \sum_{j=1}^{N} u_{jk} , \quad (2.e)$$

$$\delta_i^k = h \cdot \sum_{j=1}^{N} u_{jk} (x_{ji} - c_i^k)^2 \Big/ \sum_{j=1}^{N} u_{jk} , \quad (2.f)$$

where u_{jk} denotes the fuzzy membership of the *j*-th input data $\mathbf{x}_j = (x_{j1}, \cdots, x_{jd})^T$, belonging to the *k*-th cluster obtained by FCM clustering [14]. Here, h is a scale parameter and can be adjusted manually.

When the premise of the TSK fuzzy model is determined, let

$$\mathbf{x}_e = (1, \mathbf{x}^T)^T , \quad (3.a)$$

$$\tilde{\mathbf{x}}^k = \tilde{\mu}^k(\mathbf{x}) \mathbf{x}_e , \quad (3.b)$$

$$\mathbf{x}_g = ((\tilde{\mathbf{x}}^1)^T, (\tilde{\mathbf{x}}^2)^T, \cdots, (\tilde{\mathbf{x}}^K)^T)^T , \quad (3.c)$$

$$\mathbf{p}^k = (p_0^k, p_1^k, \cdots, p_d^k)^T \quad (3.d)$$

and

$$\mathbf{p}_g = ((\mathbf{p}^1)^T, (\mathbf{p}^2)^T, \cdots, (\mathbf{p}^K)^T)^T , \quad (3.e)$$

then Eq. (2.a) can be formulated as the following linear regression model [15]

$$y^o = \mathbf{p}_g^T \mathbf{x}_g . \quad (3.f)$$

Thus, the problem of TSK fuzzy model training can be transformed into the learning of the parameters in the corresponding linear regression model [6,13,15]. And the ε-insensitive criterion and L2-norm penalty based learning algorithm of its consequent parameters is proposed in [13].

B. KL-TSK-FS

In order to make the TSK-FS have the ability of transfer learning. In [6], a novel knowledge-leverage based TSK-FS is proposed, i.e., KL-TSK-FS. The basic idea and its algorithm are briefly described in this subsection.

1) Framework of KL-TSK-FS Learning

The framework for the construction of the KL-TSK-FS is described in Fig.1. As shown in the figure, there are two major information sources for the learning of a TSK-FS, namely, data of the current scene and knowledge of the reference scenes. With these two categories of information, parameter learning is carried out and the fuzzy system is obtained for the modeling task in the current scene accordingly.

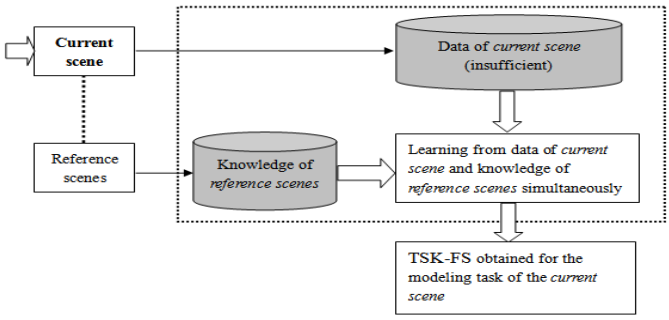


Fig.1 Framework of the KL-TSK-FS

2) Objective Criterion and Parameter Solution

The objective function of the KL-TSK-FS is constructed as follows [6].

$$\min_{\mathbf{p}_g, \xi^+, \xi^-, \varepsilon} f_{L2-TSK-FS} + \lambda(\mathbf{p}_g - \mathbf{p}_{g0})^T(\mathbf{p}_g - \mathbf{p}_{g0})$$

$$f_{L2-TSK-FS} = \frac{1}{N\tau}\sum_{i=1}^{N}((\xi_i^+)^2 + (\xi_i^-)^2) + \frac{1}{2}(\mathbf{p}_g^T \mathbf{p}_g) + \frac{2}{\tau}\cdot\varepsilon$$

$$s.t. \begin{cases} y_i - \mathbf{p}_g^T \mathbf{x}_{gi} < \varepsilon + \xi_i^+ \\ \mathbf{p}_g^T \mathbf{x}_{gi} - y_i < \varepsilon + \xi_i^- \end{cases}, \forall i . \quad (4)$$

The optimization criterion in Eq. (4) contains two terms. The first term refers to learning from the data of the current scene, which is directly inherited from the L2-TSK-FS [13]. This term is included so that the desired TSK-FS will fit the sampled training data of the current scene as accurately as possible. The second term refers to knowledge-leverage from the reference scene, with $\mathbf{p}_{g0}$ denoting the knowledge of the model parameters learned in the reference scene. The purpose of this term is to estimate the desired parameters of the consequents by approximating the model parameters in the reference scene. The parameter λ in Eq. (4) is used to balance the influence of these two terms and the optimal value can be determined by using the commonly used cross-validation strategy in machine learning.

Based on the objective criterion in Eq. (4), the dual problem of the corresponding parameter learning of KL-TSK-FS can be found in [6].

III. Existing Weaknesses of KL-TSK-FS

In this section, we discuss the existing weaknesses of the KL-TSK-FS modeling method. For convenience, the learning algorithm of the KL-TSK-FS is firstly given below [6].

Learning algorithm for KL-TSK-FS

Step 1	Introduce the knowledge of the reference scenes, i.e., the model parameter.
Step 2	Set the balance parameters τ, λ in Eq. (4).
Step 3	Use Eqs. (2.d)-(3.e) and the antecedent parameters of the fuzzy model obtained from the reference scenes to construct the dataset $\mathbf{x}_{gi}$ for the model to be trained, i.e., the linear regression model in Eq. (3.f), which is associated with the fuzzy system to be constructed for the current scene.
Step 4	Use Eq. (4) to obtain the final consequent parameters $(\mathbf{p}_g)^*$ of the desired TSK-FS in the current scene.
Step 5	Use the antecedent parameters inherited from the reference scenes and the consequent parameters obtained in step 4 to generate the fuzzy system for the current scene.

From the above learning algorithm of the KL-TSK-FS, we give the following analysis of the weaknesses of KL-TSK-FS.

(1) First, we can see that the antecedent parameters of the TSK-FS constructed in the current scene are directly inherited from the model obtained in the reference scene. This strategy results in the antecedent parameters being not particularly appropriate for the modeling task in the current scene since they cannot be learned from any information, such as the training data, in the current scene.

(2) Second, the consequent parameters can only absorb knowledge from the reference scene by the introduced term $(\mathbf{p}_g - \mathbf{p}_{g0})^T(\mathbf{p}_g - \mathbf{p}_{g0})$ as shown in Eq.(4). Thus, it seems that the knowledge-leverage from the reference scene is still not enough. It can be expected that more knowledge-leveraged terms can be introduced to improve the learning abilities for the consequents of the current scene.

With the above analysis, we know that it is a very important work to investigate the improved KL-TSK-FS modeling method. In the sequent sections, an improved KL-TSK-FS modeling method will be proposed for this purpose.

IV. KL-TSK-FS-IKT

A. Framework of KL-TSK-FS-IKT

The framework for the proposed KL-TSK-FS-IKT modeling method can be described with Fig. 2(b). As shown in the figure, there are two following aspects addressed to improve the knowledge-leverage abilities: (1) Transfer clustering based knowledge-leverage in the antecedent; and (2) Improved knowledge-leverage mechanism in the consequents. For a comparison, Fig. 2(a) shows the knowledge-leverage mechanism of the KL-TSK-FS modeling method in [6]. In the sequent subsections, the proposed knowledge-leverage mechanisms for the antecedents and consequents are described in detail, respectively. Then the algorithm of the KL-TSK-FS-IKT modeling method is presented.

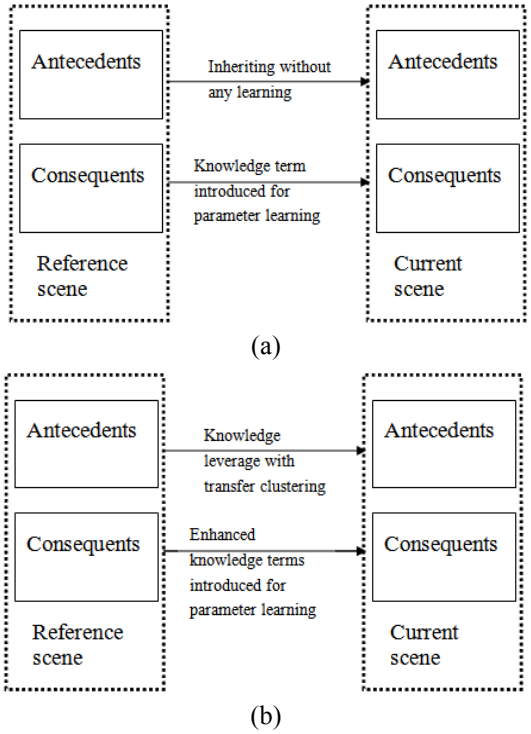


Fig. 2 Knowledge-leverage mechanism in the KL-TSK-FS modeling method and the proposed KL-TSK-FS-IKT modeling method. (a) KL-TSK-FS; (b) KL-TSK-FS-IKT.

B. Improved Learning for The Antecedents with Transfer Clustering

From Eq.(2.d), we can see that the commonly used Gaussian membership function in the antecedents of TSK-FS includes two types of the parameters, c_i^k and δ_i^k, $k = 1, \cdots, K$, $i = 1, \cdots, d$. For parameter vectors $\mathbf{c}^k = [c_i^k, \cdots, c_d^k]$, they can be taken as the cluster centers obtained by a certain clustering method on the input data of the training dataset. In KL-TSK-FS, the parameters $\mathbf{c}^k = [c_i^k, \cdots, c_d^k]$ of the TSK-FS trained in the reference scene are assumed as the available knowledge and are directly used for the TSK-FS constructed in the current scene. Thus, these parameters are not very proper

for the current scene since no information in the current scene is used for the learning of these parameters. In order to overcome this weakness, the following transfer fuzzy c-mean (TFCM) clustering technique is proposed for the learning of the antecedent parameters.

First, we take the parameters $\mathbf{c}_r^k=[c_{i,r}^k,\cdots,c_{d,r}^k]$ as the available knowledge in the reference scene, which represent the K centers of the fuzzy partitions in the input space of the reference scene. Then, we propose a TFCM clustering method to obtain the K centers of the fuzzy partitions in the input space of the current scene with the following objective function.

$$\min_{\mathbf{U},\mathbf{C}_c} J_{TFCM}=\sum_{k=1}^{K}\sum_{j=1}^{N}u_{kj}^m\left\|\mathbf{x}_j-\mathbf{c}_c^k\right\|^2+\lambda_a\cdot\sum_{k=1}^{K}\left(\sum_{j=1}^{N}u_{kj}^m\right)\left\|\mathbf{c}_c^k-\mathbf{c}_r^k\right\|^2$$
$$\text{s.t. } u_{kj}\in[0,1],\ \sum_{k=1}^{K}u_{kj}=1,\ 1\le j\le N. \quad (5)$$

In Eq.(5), the $\mathbf{x}_j$ are the available input data for model training in the current scene; $\mathbf{c}_c^k$ represent the K centers of the fuzzy partitions in the input space of the current scene; u_{ij} denotes the membership of data $\mathbf{x}_j$ belonging to the k -th cluster; $\mathbf{U}=[u_{ij}]_{K\times N}$ and $\mathbf{C}=[c_{i,c}^k]_{K\times d}$ denote the fuzzy partition matrix and the center matrix, respectively; λ_a is a balance parameter to control the influence of different terms in the objective function.

In particular, we can see that the first term in Eq. (5) is directly inherited from the classical FCM algorithm, which is used to learn the fuzzy partition matrix and the cluster center matrix based on the available data $\mathbf{x}_j$ in the current scene. The second term in Eq. (5) is a knowledge-leverage term, which can be used to learn the cluster centers of the current scene from the knowledge of the reference scene.

With Eq. (5) and the optimization technique used in FCM, we can easily obtain the following learning rules for the fuzzy partition matrix and the cluster center matrix.

$$\mathbf{c}_c^k=\frac{\sum_{j=1}^{N}u_{kj}^m\mathbf{x}_j+\lambda_a\sum_{j=1}^{N}u_{kj}^m\mathbf{c}_r^k}{(1+\lambda_a)\sum_{j=1}^{N}u_{kj}^m}=\frac{1}{(1+\lambda_a)}\frac{\sum_{j=1}^{N}u_{kj}^m\mathbf{x}_j}{\sum_{j=1}^{N}u_{kj}^m}+\frac{\lambda_a}{(1+\lambda_a)}\mathbf{c}_r^k, \quad (6)$$

$$u_{kj}=\frac{\left[1\Big/\left(\left\|\mathbf{x}_j-\mathbf{c}_c^k\right\|^2+\lambda_a\left\|\mathbf{c}_c^k-\mathbf{c}_f^k\right\|^2\right)\right]^{1/(m-1)}}{\left[\sum_{k'=1}^{C}\left(1\Big/\left(\left\|\mathbf{x}_j-\mathbf{c}_c^{k'}\right\|^2+\lambda_a\left\|\mathbf{c}_c^{k'}-\mathbf{c}_f^{k'}\right\|^2\right)\right)\right]^{1/(m-1)}} \quad (7)$$

From Eq.(6), we can see that the obtained cluster centers of the current scene contains two parts, i.e., $(1/1+\lambda_a)\sum_{j=1}^{N}u_{kj}^m\mathbf{x}_j/\sum_{j=1}^{N}u_{kj}^m$ and $(\lambda_a/1+\lambda_a)\mathbf{c}_r^k$, which can taken the knowledge learned from the data in the current scene and the knowledge in the related scene, respectively. With the strategy in Eq. (2.f), we can easily evaluate the parameters δ_i^k for the current scene based on the clustering results obtained by TFCM.

C. Improved Learning for The Consequents

In this subsection, we will investigate the improved learning mechanism to improve the knowledge-leverage abilities of the KL-TSK-FS for the learning of the consequent parameters. In particular, the following objective function is proposed for this purpose.

$$\min_{\mathbf{p}_g,\xi^+,\xi^-,\varepsilon} f_{L2-TSK-FS}+\lambda(\mathbf{p}_g-\mathbf{p}_{g0})^T(\mathbf{p}_g-\mathbf{p}_{g0})+\frac{1}{N\tilde{\tau}}\sum_{i=1}^{N}\left|\mathbf{p}_g^T\mathbf{x}_{gi}-\mathbf{p}_{g0}^T\mathbf{x}_{gi}\right|_{\tilde{\varepsilon}}+\frac{2}{\tau}\cdot\tilde{\varepsilon}$$
$$f_{L2-TSK-FS}=\frac{1}{N\tau}\sum_{i=1}^{N}((\xi_i^+)^2+(\xi_i^-)^2)+\frac{1}{2}(\mathbf{p}_g^T\mathbf{p}_g)+\frac{2}{\tau}\cdot\varepsilon \quad (8.a)$$
$$\text{s.t. }\begin{cases}y_i-\mathbf{p}_g^T\mathbf{x}_{gi}<\varepsilon+\xi_i^+\\ \mathbf{p}_g^T\mathbf{x}_{gi}-y_i<\varepsilon+\xi_i^-\end{cases}$$

Comparing Eq.(8.a) with Eq.(4), it can be seen that additional terms $(1/N\tilde{\tau})\sum_{i=1}^{N}\left|\mathbf{p}_g^T\mathbf{x}_{gi}-\mathbf{p}_{g0}^T\mathbf{x}_{gi}\right|_{\tilde{\varepsilon}}$ and $(2/\tau)\cdot\tilde{\varepsilon}$ have been introduced. The first one of these two terms also contains the knowledge of the related scene and is used for the consequent learning in the current scene. This term implicates that the trained model in the current scene is apt to obtain the consistent decision result with that obtained by the model in the reference scene if the knowledge of the related scene is useful. For the added term $(2/\tau)\cdot\tilde{\varepsilon}$, it is the penalty term of the $\tilde{\varepsilon}$-insensitive parameter. As stated in the L2-TSK-FS [13], this insensitive parameter can be learned automatically when the penalty term $(2/\tau)\cdot\tilde{\varepsilon}$ is added in the objective function.

For easy optimization, by introducing the slack variables and the L2-norm penalty terms, Eq.(8.a) can be equivalent expressed as follows.

$$\min_{\mathbf{p}_g,\xi^+,\xi^-,\varepsilon} f_{L2-TSK-FS}+\lambda(\mathbf{p}_g-\mathbf{p}_{g0})^T(\mathbf{p}_g-\mathbf{p}_{g0})+\frac{1}{N\tilde{\tau}}\sum_{i=1}^{N}((\eta_i^+)^2+(\eta_i^-)^2)+\frac{2}{\tau}\cdot\tilde{\varepsilon} \quad (8.b)$$
$$f_{L2-TSK-FS}=\frac{1}{N\tau}\sum_{i=1}^{N}((\xi_i^+)^2+(\xi_i^-)^2)+\frac{1}{2}(\mathbf{p}_g^T\mathbf{p}_g)+\frac{2}{\tau}\cdot\varepsilon$$
$$\text{s.t. }\begin{cases}y_i-\mathbf{p}_g^T\mathbf{x}_{gi}<\varepsilon+\xi_i^+\\ \mathbf{p}_g^T\mathbf{x}_{gi}-y_i<\varepsilon+\xi_i^-\\ \mathbf{p}_g^T\mathbf{x}_{gi}-\mathbf{p}_{g0}^T\mathbf{x}_{gi}<\tilde{\varepsilon}+\eta_i^+\\ \mathbf{p}_g^T\mathbf{x}_{gi}-\mathbf{p}_{g0}^T\mathbf{x}_{gi}<\tilde{\varepsilon}+\eta_i^-\end{cases}$$

For the objective criterion in Eq.(8.b), its dual problem can be formulated as the following QP problem.

$$\max_{\boldsymbol{\alpha}^+,\boldsymbol{\alpha}^-,\boldsymbol{\beta}^+,\boldsymbol{\beta}^-} -\frac{1}{2}\left\{\frac{N\tau}{2}\sum_{i=1}^{N}\left[(\alpha_i^+)^2+(\alpha_i^-)^2\right]+\frac{N\tilde{\tau}}{2}\sum_{i=1}^{N}\left[(\beta_i^+)^2+(\beta_i^-)^2\right]\right.$$
$$+\frac{1}{1+2\lambda}\cdot\sum_{i=1}^{N}\sum_{j=1}^{N}(\alpha_i^+-\alpha_i^-)(\alpha_j^+-\alpha_j^-)\mathbf{x}_{gi}^T\mathbf{x}_{gj}+\frac{1}{1+2\lambda}\cdot\sum_{i=1}^{N}\sum_{j=1}^{N}(\beta_i^+-\beta_i^-)(\beta_j^+-\beta_j^-)\mathbf{x}_{gi}^T\mathbf{x}_{gj} \quad (8.c)$$
$$+\frac{2}{1+2\lambda}\cdot\sum_{i=1}^{N}\sum_{j=1}^{N}(\beta_i^+-\beta_i^-)(\alpha_j^+-\alpha_j^-)\mathbf{x}_{gi}^T\mathbf{x}_{gj}+\sum_{i=1}^{N}(\alpha_i^+-\alpha_i^-)(y_i-\frac{2\lambda}{1+2\lambda}\mathbf{p}_{g0}^T\mathbf{x}_{gi})$$
$$\left.+\frac{1}{1+2\lambda}\sum_{i=1}^{N}(\beta_i^+-\beta_i^-)\mathbf{p}_{g0}^T\mathbf{x}_{gi}\right\}$$

$$\text{s.t. }\alpha_i^+\ge0,\alpha_i^-\ge0,\ \sum_{i=1}^{N}\left(\alpha_i^++\alpha_i^-\right)=\frac{2}{\tau},\ \beta_i^+\ge0,\beta_i^-\ge0,\ \sum_{i=1}^{N}\left(\beta_i^++\beta_i^-\right)=\frac{2}{\tilde{\tau}}.$$

where $\boldsymbol{\alpha}^+,\boldsymbol{\alpha}^-,\boldsymbol{\beta}^+,\boldsymbol{\beta}^-$ are the Lagrangian multiplier vector, i.e., the solution variables of the dual problem of Eq. (8.b). For save the space, the derivation of Eq. (8.c) is omitted.

According to the KTT optimal theory, the optimal consequent parameters of the trained TSK-FS, i.e., $(\mathbf{p}_g)^*$, can be finally given by

$$\mathbf{p}_g^*=\frac{2\lambda}{1+2\lambda}\mathbf{p}_{g0}+\frac{1}{1+2\lambda}\left(\sum_{i=1}^{N}((\alpha_i^+)^*-(\alpha_i^-)^*)\mathbf{x}_{g,t}+\sum_{i=1}^{N}((\beta_i^+)^*-(\beta_i^-)^*)\mathbf{x}_{gj}\right) \quad (9.a)$$

where $(\alpha_i^+)^*,(\alpha_i^-)^*,(\beta_i^+)^*,(\beta_i^-)^*$ are the optimal solutions of

the dual problem in Eq. (8.c). Furthermore, Eq. (9.a) can be written as follows.

$$(\mathbf{p}_g)^* = \gamma \mathbf{p}_{g0} + (1-\gamma)\ \mathbf{p}_{gc}, \quad (9.b)$$

with $\gamma = \frac{2\lambda}{1+2\lambda}$, $\mathbf{p}_{gc} = \sum_{i=1}^{N}((\alpha_i^+)^* - (\alpha_i^-)^*)\mathbf{x}_{gj,t} + \sum_{i=1}^{N}((\beta_i^+)^* - (\beta_i^-)^*)\mathbf{x}_{gj}$.

The final optimal parameter $(\mathbf{p}_g)^*$ in Eq. (9.b) shows that the consequent parameters of the trained TSK-FS still contains the knowledge of two parts, i.e. $\gamma \cdot \mathbf{p}_{g0}$ and $(1-\gamma)\cdot \mathbf{p}_{gc}$. While $(1-\gamma)\cdot \mathbf{p}_{gc}$ can be regarded as the knowledge learned from the data of the current scene, $\gamma \cdot \mathbf{p}_{g0}$ can be taken as the knowledge inherited from the reference scenes. Please note that in essence the $(1-\gamma)\cdot \mathbf{p}_{gc}$ term is related with both scenes since the involved $\boldsymbol{\alpha}^+, \boldsymbol{\alpha}^-, \boldsymbol{\beta}^+, \boldsymbol{\beta}^-$ parameters in this term are influenced by the information of two scenes simultaneously in the learning procedure as shown in Eq. (8.c). Now, comparing the result in Eq.(9.b) and the corresponding Eq.(13.a) in [6], we can find that the proposed improved objective function has given much stronger knowledge-leverage abilities than that used in [6].

D. Algorithm of KL-TSK-FS-IKT

The learning algorithm of the proposed KL-TSK-FS-IKT is described in detail below.

Algorithm for KL-TSK-FS-IKT

Step 1	Introduce the knowledge of the reference scenes, i.e., the model parameter.
Step 2	Set the balance parameters λ_a in Eq. (5) and $\tau, \tilde{\tau}, \lambda$ in Eq. (8.a), respectively.
Step 3	Use Eqs. (5)-(7) and (2.f) to learn the antecedent parameters of the TSK-FS in the current scene.
Step 4	Use Eqs. (2.d)-(3.e) and the antecedent parameters of the TSK-FS obtained from the Step3 to construct the dataset $\mathbf{x}_{gi}$ for the linear model to be trained, i.e., the linear regression model in Eq. (3.f), which is associated with the TSK-FS to be constructed for the current scene.
Step 5	Use Eqs.(8.c) to obtain the consequent parameters of the TSK-FS in the current scene.

V. Experimental Results

A. Experimental Settings

The proposed learning algorithm for KL-TSK-FS-IKT is evaluated by using both synthetic and real-world datasets. Details about the evaluation will be described in detail in section V-B and V-C respectively. For clarity, the notations of the datasets and their definitions are listed in Table I. Here, datasets generated from the reference scene and the current scene are denoted by D1 and D2 respectively.

TABLE I NOTATIONS OF THE ADOPTED DATASETS AND THEIR DEFINITIONS

Notation	Definitions
D1	Dataset generated from the reference scene
D2	Dataset generated from the current scene for training
D2_test	Dataset generated from the current scene for testing
r	Relation parameter between the reference scene and the current scene, which is used to construct the synthetic datasets.

TABLE II. THE METHODS ADOPTED FOR PERFORMANCE COMPARISON

Algorithm	Descriptions
L2-TSK-FS (D1) [13]	L2-TSK-FS based on the data in the reference scene
L2-TSK-FS (D2) [13]	L2-TSK-FS based on the data in the current scene
L2-TSK-FS (D1+D2) [13]	L2-TSK-FS based on the data in both the current scene and the reference scene
HiRBF [3]	Bayesian task-level transfer learning for non-linear regression method
KL-TSK-FS (D2+Knowledge) [6]	Knowledge-leverage based TSK fuzzy system modeling method
KL-TSK-FS-IKT (D2+Knowledge)	The proposed improved KL-TSK-FS modeling method

The following generalization performance index J is used in the experiments [16],

$$J = \sqrt{\sum_{i=1}^{N}(y_i' - y_i)^2 \Big/ N \Big/ \sum_{i=1}^{N}(y_i - \bar{y})^2 \Big/ N}, \quad (10)$$

where N is the number of test data, y_i is the output for the i-th test input, y_i' is the fuzzy model output for the i-th test input, and $\bar{y} = \sum_{i=1}^{N} y_i \Big/ N$. The smaller the value of J, the better the generalization performance.

In the experiments, the hyper parameters of all the methods adopted for the comparison are determined by using the five folds cross-validation strategy with the training datasets. All the algorithms are implemented using MATLAB on a computer with Intel Core 2 Duo P8600 2.4 GHz CPU and 2GB RAM.

B. Synthetic Datasets

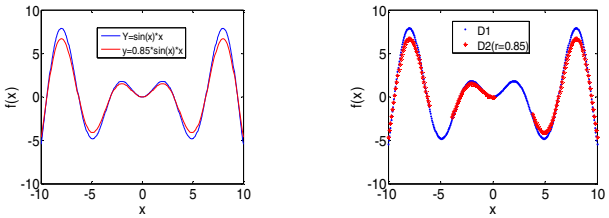
1) Generation of Synthetic Datasets: Synthetic datasets are generated to simulate the scenes in the study and the following requirements need to be satisfied: 1) the reference scene should be related to the current scene, i.e., the reference and current scenes are different but related; 2) some of the data of the current scene are not available or missing. In other words, the data available from the current scene are insufficient.

Based on the above requirements, the function $Y = f(x) = x * \sin(x), x \in [-10,10]$ is used to describe the reference scene and to generate dataset D1. On the other hand, the function $y = r * f(x) = r * x * \sin(x), x \in [-10,10]$ is used to describe the current scene and to generate dataset D2 and testing dataset D2_test for the current scene. Here, r is a relation parameter between the reference scene and the current scene. The parameter is used to control the degree of similarity/difference between these two scenes. When r =1, the two scenes are identical. On the other hand, the lack of information (data insufficiency) is simulated by introducing intervals with missing data into the training set generated for the current scene. The settings for generating the synthetic datasets are described in Table III. For example, the two functions used to simulate the two related scenes, with the relation parameter $r = 0.85$, are shown in Fig. 3(a). The datasets of the reference scene and the training sets of the

current scene, generated with the same relation parameter (i.e. $r = 0.85$), are shown in Fig. 3(b). The figures also show the two data-missing intervals [-6, -3] and [0, 4] introduced into the dataset.

TABLE III DETAILS OF THE SYNTHETIC DATASETS

Reference scene	Current scene		
Dataset	Training set		Testing set
Size of dataset	Interval with data missing	Size of dataset	Size of dataset
400	[-6,-3] and [0,4]	144	200
Relation parameter between the two scenes: r =0.9, 0.85 and 0.75			



(a) (b)

Fig.3 Functions representing two different scenes with the relation parameter $r = 0.85$ and the corresponding sampled data from these scenes: (a) the functions representing the reference scene (Y) and the current scene (y); (b) the data of the reference scene and the training data of the current scene with missing data in the intervals [-6,-3] and [0,4]

2) Performance Comparison: The modeling results of all the algorithms on the synthetic datasets are reported in Table IV. Fig. 4 shows the modeling effect on the dataset with $r = 0.85$. Since the modeling results on the other synthetic datasets are similar to those given in Fig. 4, they are not presented here due to space limitations. The following observations can be made from the reported results.

(1) It can be seen from Table IV that the generalization performance of the proposed KL-TSK-FS-IKT is better than that of several related methods adopted in the experiment.

(2) Fig. 4(a)-(c) show the modeling results of the L2-TSK-FS obtained by using different datasets. The results show that no matter how the datasets constitute, the L2-TSK-FS do not have the best performance due to its absence of the transfer ability.

(4) The result of the HiRBF algorithm is shown in Fig.4 (d). Although the transfer learning-based method HiRBF has used the data in both the current scene and the reference scene in the training, it is evident from Fig. 4 (d) that this method cannot effectively cope with the problem caused by the missing data with its transfer learning mechanism, still exhibiting poor generalization ability in the two data-missing intervals.

(5) Fig. 4(e) and Fig. 4(f) show the modeling effect of the KL-TSK-FS and the proposed KL-TSK-FS-IKT. Both algorithms have knowledge-leveraged abilities from the related scene and show a promising modeling effect. In particular, these methods are able to give an acceptable generalization capability in the two data-missing intervals, indicating that they have effectively leveraged useful knowledge from the reference scene and remedied the generalization abilities in the training procedure.

(6) By comparing the KL-TSK-FS with the KL-TSK-FS-IKT, we can see that the KL-TSK-FS-IKT demonstrates stronger knowledge-leverage abilities from the performance index in Table IV and the visual effect in Fig. 4. The experimental results confirm that the proposed improved knowledge-leverage mechanism in the KL-TSK-FS-IKT is advantageous to that used in the KL-TSK-FS.

TABLE IV GENERALIZATION PERFORMANCE (J) OF THE ADOPTED METHOD ON THE SYNTHETIC DATASETS

Interval with data missing	Relation parameter (r)	L2-TSK-FS (D1)	L2-TSK-FS (D2)	L2-TSK-FS (D1+D2)
[-6,-3] and [0, 4]	0.9	0.1343	0.2858	0.1012
	0.85	0.1908	0.2813	0.1434
	0.75	0.3525	0.2841	0.2627
Interval with data missing	Relation Parameter (r)	HiRBF	KL-TSK-FS (D2+Knowledge)	KL-TSK-FS-IKT (D2+Knowledge)
[-6,-3] and [0,4]	0.9	0.2621	0.0501	**0.0283**
	0.85	0.2619	0.0516	**0.0332**
	0.75	0.2639	0.1534	**0.1278**

C. *Real-world Datasets*

1) The Glutamic Acid Fermentation Process Modeling: To further evaluate the performance of the proposed knowledge-leverage based TSK-FS modeling method, an experiment is conducted to apply the method to model a biochemical process with a real-world dataset [6, 8]. The dataset adopted originates from the glutamic acid fermentation process, which is a multiple-input-multiple-output system. The input variables of the dataset include the fermentation time h, glucose concentration $S(h)$, thalli concentration $X(h)$, glutamic acid concentration $P(h)$, stirring speed $R(h)$, and ventilation $Q(h)$, where $h = 0, 2, \cdots, 28$. The output variables are glucose concentration $S(h+2)$, thalli concentration $X(h+2)$, and glutamic acid concentration $P(h+2)$ at a future time $h+2$. The TSK-FS based biochemical process prediction model is illustrated in Fig. 5. The data in this experiment were collected from 21 batches of fermentation processes, with each batch containing 14 effective data samples. In this experiment, in order to match the situation discussed in this study, the data are divided into two scenes, i.e., the reference scene and the current scene, as described in Table V.

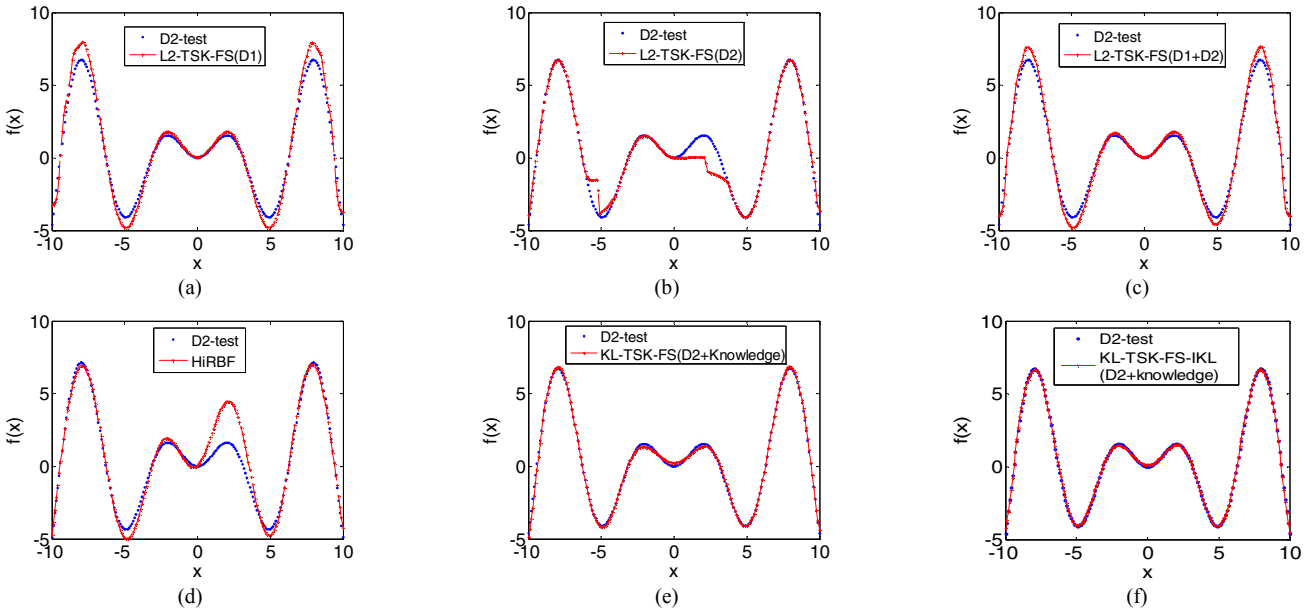


Fig.4 Modeling effect of different methods by using the synthetic datasets shown in Fig.4(b). (a) L2-TSK-FS(D1); (b) L2-TSK-FS (D2); (c) L2-TSK-FS (D1+D2); (d) HiRBF; (e) KL-TSK-FS (D2+Knowledge) and (f) KL-TSK-FS-IKT (D2+Knowledge).

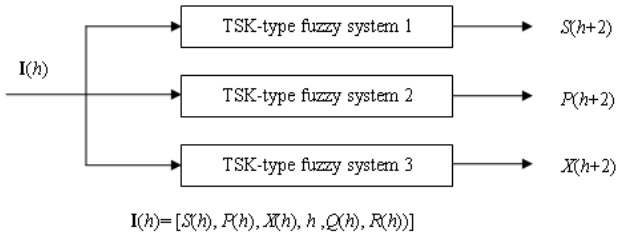


Fig. 5 Illustration of the glutamic acid fermentation process prediction model based on the TSK-FSs.

TABLE V. THE FERMENTATION PROCESS MODELING DATASET

	Data of reference scene (D1)	Data of current scene	
		Training set (D2)*	Testing set (D2_test)
Batches	1-16	17-19	20-21
Size of dataset	224	30	28

*For training set of the current scene, information is missing at time $h = 6, 8, 10, 12$.

2) Performance Comparison: The experiment results of fermentation process modeling using different methods are given in Table VI.

Table VI shows that the modeling results of the KL-TSK-FS-IKT method are better than the results of the other methods. This can be explained again by the fact that the proposed method can effectively exploit not only the data of the current scene but also useful knowledge of the reference scene in the training procedure for the current scene. It can be seen from the experiment results that, even if the data in the training data of the current scene are missing, the generalization capability of the TSK-FSs obtained by KL-TSK-FS and the proposed KL-TSK-FS-IKT is not degraded significantly. This remarkable feature is very valuable for biochemical process modeling since the lack of sampled data is common due to the poor sensitivity of sensors in the noisy environment.

From the results in Table VI, we can see that although the KL-TSK-FS method also has knowledge-leverage abilities, due to the insufficient knowledge-leverage learning, its generalization abilities are much weaker than the proposed KL-TSK-FS-IKT method. Thus, the KL-TSK-FS-IKT will be more promising than KL-TSK-FS in the practical application of fermentation process modeling.

TABLE VI. GENERALIZATION PERFORMANCE (J) OF THE PROPOSED KL-TSK-FS METHOD AND THE TRADITIONAL L2-TSK-FS METHODS IN FERMENTATION PROCESS MODELING

Output	L2-TSK-FS(D1)	L2-TSK-FS(D2)	L2-TSK-FS(D1+D2)
$S(h+2)$	0.2792	0.3944	0.2804
$X(h+2)$	0.8342	1.1203	1.0642
$P(h+2)$	0.2842	0.3255	0.2533
Output	HiRBF	KL-TSK-FS	KL-TSK-FS-IKT
$S(h+2)$	0.3510	0.1239	**0.1108**
$X(h+2)$	0.7026	0.4548	**0.3578**
$P(h+2)$	0.4117	0.1482	**0.1069**

VI. CONCLUSIONS

In this study, the improved knowledge-leverage based TSK fuzzy system modeling method is proposed in order to overcome the existing weaknesses of the existing knowledge-leverage based TSK fuzzy system modeling. In particular, two improved knowledge-leverage strategies have been introduced for the learning of the antecedent parameters and consequent parameters, respectively. With the improved knowledge-leverage learning abilities, the proposed method has shown a better modeling effect compared with traditional TSK fuzzy modeling methods and other related methods on synthetic and real world datasets. Despite the promising performance of the proposed method, there is still room for further improvement. For example, more advanced transfer learning mechanisms can be expected for TSK fuzzy system modeling. In near future, we will have a further study in this

issue in depth.

REFERENCES

[1]. S.J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowledge Data Engineering,* vol. 22, no. 10, pp. 1345–1359, Oct. 2010.

[2]. L.X. Duan, D. Xu, I. W. Tsang, "Domain adaptation from multiple sources: a domain-dependent regularization approach," *IEEE Trans. Neural Netw. Learning Syst., vol. 23*, no. 3, pp. 504-518, 2012.

[3]. P. Yang, Q. Tan, and Y. Ding, "Bayesian task-level transfer learning for non-linear regression," *Proc. Int. Conf. on Computer Science and Software Engineering,* pp. 62-65, 2008.

[4]. J.W. Tao, K.F.L. Chung, S.T. Wang, "On minimum distribution discrepancy support vector machine for domain adaptation." Pattern Recognition, vol. 45, no.11, pp. 3962-3984, 2012.

[5]. Z.H. Deng, Y.Z. Jiang, F.L. Chung, H. Ishibuchi, S.T. Wang, "Knowledge-Leverage Based Fuzzy System and Its Modeling." *IEEE Trans. Fuzzy Systems*, vol. 21, no. 4, pp. 597-609, 2013.

[6]. Z.H. Deng, Y.Z. Jiang, K.S. Choi, F.L. Chung, S.T. Wang, "Knowledge-Leverage-Based TSK Fuzzy System Modeling." *IEEE Trans. Neural Networks and Learning Systems*, vol. 24, no. 8, pp. 1200-1212, 2013.

[7]. X. Liao, Y. Xue, and L. Carin, "Logistic regression with an auxiliary data source," *Proc. 21st Int. Conf. Machine Learning*, pp. 505-512, Aug. 2005.

[8]. J. Huang, A. Smola, A. Gretton, K.M. Borgwardt, and B. Schölkopf, "Correcting sample selection bias by unlabeled data," *Proc. 19th Ann. Conf. Neural Information Processing Systems*, 2007.

[9]. W.H. Jiang and F.L. Chung, "Transfer Spectral Clustering," *Proc. European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD)*, Bristol, UK, 24-28 Sept. 2012.

[10]. B. Gong, Y. Shi, F. Sha, and K. Grauman, "Geodesic flow kernel for unsupervised domain adaptation," *Proc. 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2012, pp. 2066–2073.

[11]. W. Dai, Q. Yang, G. Xue, and Y. Yu, "Self-taught clustering," *Proc. 25th Int. Conf. Machine Learning*, pp. 200-207, July 2008.

[12]. J.S.R. Jang, C.T. Sun, and E. Mizutani, *Neuro-fuzzy and soft-computing*. Upper Saddle River, NJ: Prentice-Hall, 1997.

[13]. Z.H. Deng, K.S. Choi, F.L. Chung, S.T. Wang, "Scalable TSK fuzzy modeling for very large datasets using minimal-enclosing-ball approximation," *IEEE Trans. Fuzzy Systems*, vol. 19, no.2, pp.210-226, 2011.

[14]. J.C. Bezdek, J. Keller, and R. Krishnapuram, *Fuzzy models and algorithms for pattern recognition and image processing*. San Francisco: Kluwer Academic Publishers, 1999.

[15]. J. Leski, "TSK-fuzzy modeling based on ε-insensitive learning," *IEEE Trans. Fuzzy Systems*, vol. 13, no.2, pp. 181-193, 2005

[16]. J.S.R. Jang, "ANFIS: adaptive-network-based fuzzy inference system," *IEEE Trans. Systems, Man and Cybernetics*, 23, no. 3, pp. 665-685, 1993.