to make use of the information from the reference scenes to improve the modeling effect of the current scene, thereby yielding a model with better generalization capability. Here, transfer learning is an effective solution to the corresponding modeling task because it can enhance the model by leveraging the information available from the reference scenes.

As fuzzy system modeling is an important regression model with extensive applications [21-23, 41-43], it is promising to incorporate transfer learning with the model. To the best of our knowledge, however, the study of transfer learning for fuzzy system modeling has not been reported before. For fuzzy system modeling, transfer learning is very useful in real-world modeling tasks where traditional fuzzy modeling methods may not work very well. For example, trained fuzzy systems have much weaker generalization capability when the training data are insufficient or only partially available [24, 25, 31]. The situation is common in real-world applications in which the sensors and setup for data sampling can be unstable due to noisy environment or other malfunctions, leading to insufficiency of data for the modeling task.

In order to tackle the problems with traditional fuzzy modeling as described above, a feasible remedial strategy is to boost up the performance by taking advantage of the useful information from reference scenes (or related scenes), which can be the data in the scenes, or relevant knowledge like the density distribution and fuzzy rules. The simplest way to obtain the information from reference scenes is to directly use the data collected from the reference scenes, but this approach can cause two major issues. First, due to the necessity of privacy protection in some proprietary applications, such as the aforementioned fermentation process, the data of the reference scenes cannot always be obtained. Under this situation, the knowledge about the reference scenes, e.g. density distribution and model parameters, can be obtained more easily to enhance the modeling of the current scene. Second, drifting phenomenon may exist between the reference scenes and the current scene, which makes it inappropriate to directly use the data from the former in the latter. These two issues should be properly addressed in order to develop an effective transfer regression modeling strategy for fuzzy systems.

In this study, a fuzzy system modeling approach with knowledge-leverage capability on reference scenes is exploited. In view of its popularity, the Takagi-Sugeno-Kang-type fuzzy system (TSK-FS) is chosen to be incorporated with a knowledge-leverage mechanism and hence the knowledge-leverage based TSK-FS (KL-TSK-FS) is proposed. A novel objective criterion is proposed to integrate the model knowledge of the reference scenes and the data of the current scene, and the induced fuzzy rules of the model are learned accordingly. The knowledge of the reference scenes will effectively makes up the deficiency in learning due to the lack of data in the current scene. The contributions of this study include the following aspects.

(1) A novel knowledge-leverage based transfer learning mechanism is developed for the classical intelligent modeling model – the fuzzy system.

(2) A knowledge-leverage based TSK-FS – KL-TSK-FS, and its modeling/learning algorithm is proposed, which is more adaptive to situations where the data are only partially available from the current scene while some useful knowledge exists in the reference scenes.

(3) The proposed method is distinctive in preserving data privacy since it is the knowledge (e.g. the corresponding model parameters), instead of the data of the reference scene, that is used by the algorithm. Thus, the proposed transfer learning regression method possesses good privacy protection capability for the data in the reference scene.

The ability to deal with situations where the data are scarce, missing or protected is a remarkable feature that makes the proposed KL-TSK-FS method very suitable for many real-world applications. In addition to microbiological fermentation process modeling as discussed above, another important application is medical diagnosis modeling for certain diseases. In this modeling task, two critical challenges are often confronted: insufficiency of training samples (i.e. limited patient data) and the requirement of patient privacy protection. The KL-TSK-FS is just the right method to meet the challenges and is promising for various health informatics applications.

The rest of this paper is organized as follows. In section II, the concept and principle of classical TSK-FS systems and the learning algorithms are reviewed. In section III, the concepts of knowledge leverage and knowledge-leverage based fuzzy systems are introduced. A specific knowledge-leverage based fuzzy system, i.e. KL-TSK-FS, based on the ε -insensitive criterion and L2-norm penalty terms, is proposed in section IV. The system is evaluated with the experiments described in section V, along with the results and discussion. The conclusions are given in the final section. A list of main notations used in this paper is provided in Table III to enhance readability.

TABLE III
MAIN NOTATIONS USED IN THIS STUDY

Notations	Explanations
A_j^k	The *j*-th fuzzy subset in the *k*-th fuzzy rule of TSK fuzzy system.
R^d	The *d*-dimension real space.
$f^k(\mathbf{x})$	The output of the *k*-th fuzzy rule of TSK fuzzy system.
$\mu^k(\mathbf{x})$, $\tilde{\mu}^k(\mathbf{x})$	Fuzzy membership and normalized fuzzy membership of the *k*-th fuzzy rule.
c_i^k, δ_i^k	The center and width of Gaussian fuzzy membership function.
$\mathbf{p}_g$	Parameter of linear regression model constructed from TSK fuzzy system.
ε	ε -insensitive parameter
$\boldsymbol{\alpha}^+, \boldsymbol{\alpha}^-$	Lagrangian multiplier vector
ξ^+, ξ^-	Vector of slack variables
J	Generalization performance index for regression

II. CLASSICAL TSK-TYPE FUZZY SYSTEMS

Classical fuzzy system models include the Takagi-Sugeno-Kang (TSK) model [26], Mamdani-Larsen

(ML) model [27] and Generalized Fuzzy Model (GFM) [28]. Among them, the TSK model is the most popular one due to its effectiveness. In this study, the TSK model is adopted to develop the proposed KL-TSK-FS and its learning algorithm. In this section, the concept and principle of classical TSK fuzzy model and the learning algorithms are reviewed.

A. Concept and Principle

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

TSK Fuzzy Rule R^k :

$$\text{IF } x_1 \text{ is } A_1^k \wedge x_2 \text{ is } A_2^k \wedge \cdots \wedge x_d \text{ is } A_d^k \ , \quad (1)$$

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

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

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

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,

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

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

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 [36]. Here, h is a scale parameter and can be adjusted manually.

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

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

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

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

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

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

then Eq. (2.a) can be formulated as the following linear regression problem [14]

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

Thus, the problem of TSK fuzzy model training can be transformed into the learning of the parameters in the corresponding linear regression model [20,30].

B. Classical TSK–FS Learning Algorithms

1) TSK-FS Learning based on Least Square Criterion

Given a training dataset $D_{tr} = \{\mathbf{x}_i, y_i \,|\, \mathbf{x}_i \in R^d, y_i \in R, i = 1, \cdots, N\}$, for fixed antecedents obtained via clustering of the input space (or by other partition techniques), the least square (LS) solution to the consequent parameters is to minimize the following LS criterion function [29], that is,

$$\min_{\mathbf{p}_g} E = \sum_{i=1}^{N} \left(y_i^o - y_i\right)^2 = \sum_{i=1}^{N} \left(\mathbf{p}_g{}^T \mathbf{x}_{gi} - y_i\right)^2 = (\mathbf{y} - \mathbf{X}_g \mathbf{p}_g)^T (\mathbf{y} - \mathbf{X}_g \mathbf{p}_g) , \quad (4)$$

where $\mathbf{X}_g = [\mathbf{x}_{g1}, \cdots, \mathbf{x}_{gN}]^T \in R^{N \times K \cdot (d+1)}$ and $\mathbf{y} = [y_1, \cdots, y_N]^T \in R^N$.

The most popular LS criterion-based TSK fuzzy model training algorithm is the one used in the adaptive-network-based fuzzy inference systems (ANFIS) [29]. For this type of algorithms, a main shortcoming is that they usually have weak robustness for modeling tasks involving noisy and/or small datasets.

2) TSK-FS Learning based on ε-insensitive Criterion and L1-norm Penalty

Besides the least square criterion, another important criterion for TSK fuzzy model training is the ε-insensitive criterion [30]. Given a scalar g and a vector $\mathbf{g} = [g_1, \cdots, g_g]^T$, the corresponding ε-insensitive loss functions take the following forms [30]: $|g|_\varepsilon = g - \varepsilon$ （$g > \varepsilon$） , $|g|_\varepsilon = 0$ （$g \le 0$） and $|\mathbf{g}|_\varepsilon = \sum_{i=1}^{d} |g_i|_\varepsilon$. For the linear regression problem of the TSK fuzzy model in Eq. (3.f), the ε-insensitive loss based criterion function [30] is defined as

$$\min_{\mathbf{p}_g} E = \sum_{i=1}^{N} \left|y_i^o - y_i\right|_\varepsilon = \sum_{i=1}^{N} \left|\mathbf{p}_g^T \mathbf{x}_{gi} - y_i\right|_\varepsilon \quad (5.a)$$

In general, the inequalities $y_i - \mathbf{p}_g^T \mathbf{x}_{gi} < \varepsilon$ and $\mathbf{p}_g^T \mathbf{x}_{gi} - y_i < \varepsilon$ are not satisfied for all data pairs $(\mathbf{x}_{gi}, y_i)$. By introducing the slack variables $\xi_i^+ \ge 0$ and $\xi_i^- \ge 0$, Eq. (5.a) can be equivalently written as

$$\min_{\mathbf{p}_g,\xi_i^+,\xi} E=\sum_{j=1}^{N}\left(\xi_i^+ +\xi_i^-\right) \tag{5.b}$$
$$\text{s.t.}\ \begin{cases} y_i-\mathbf{p}_g^T\mathbf{x}_{gi}<\varepsilon+\xi_i^+ \\ \mathbf{p}_g^T\mathbf{x}_{gi}-y_i<\varepsilon+\xi_i^- \end{cases},\ \xi_i^+\ge 0,\xi_i^-\ge 0\quad \forall\, i\,.$$

Furthermore, by introducing the regularization term [30], Eq. (5.b) is modified to become

$$\min_{\mathbf{p}_g,\xi^+,\xi^+}\quad E=\frac{1}{\tau}\sum_{i=1}^{N}\left(\xi_i^+ +\xi_i^-\right)+\frac{1}{2}\mathbf{p}_g{}^T\mathbf{p}_g \tag{5.c}$$
$$\text{s.t.}\ \begin{cases} y_i-\mathbf{p}_g^T\mathbf{x}_{gi}<\varepsilon+\xi_i^+ \\ \mathbf{p}_g^T\mathbf{x}_{gi}-y_i<\varepsilon+\xi_i^- \end{cases},\ \xi_i^+\ge 0,\xi_i^-\ge 0\quad \forall\, i\,,$$

where $\tau>0$ controls the tradeoff between the complexity of the regression model and the tolerance of the errors. Here, ξ_i^+ and ξ_i^- can be taken as the L1-norm penalty terms and thus Eq. (5.c) is an objective function based on L1-norm penalty terms. This type of TSK training algorithm is referred to as L1-TSK-FS. The dual optimization in Eq. (5.c) is a quadratic programming (QP) problem, which can be expressed as

$$\begin{aligned}\max_{\boldsymbol{\alpha}^+,\boldsymbol{\alpha}^-}\ & -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\alpha_i^+-\alpha_i^-)(\alpha_j^+-\alpha_j^-)\mathbf{x}_{gi}^T\mathbf{x}_{gi} \\ & -\sum_{i=1}^{N}\varepsilon(\alpha_i^++\alpha_i^-)+\sum_{i=1}^{N}y_i(\alpha_i^+-\alpha_i^-)\end{aligned} \tag{5.d}$$
$$\text{s.t.}\ \sum_{i=1}^{N}(\alpha_i^+-\alpha_i^-)=0\,,\ \alpha_j^+,\alpha_j^-\in[0,\tau]\quad\forall\, i\,.$$

Compared to the LS-criterion based algorithms, the ε-insensitive criterion based method has been shown to be more robust against noisy and small datasets.

3) TSK-FS Learning based on ε-insensitive Criterion and L2-norm Penalty

In addition to the L1-norm penalty terms in Eq. (5.c), the L2-norm penalty terms are also employed to develop ε-insensitive criterion based TSK-FS learning method [20], where the insensitive parameter ε is introduced as a penalty term in the objective function. This is similar to the approaches used in other L2-norm penalty-based methods, e.g. L2-norm support vector regression (L2-SVR) [32]. For TSK fuzzy model training, the ε-insensitive objective function based on L2-norm penalty terms is then given by

$$\begin{aligned}\min_{\mathbf{p}_g,\xi^+,\xi^+,\varepsilon}\ & g\left(\mathbf{p}_g,\boldsymbol{\xi}^+,\boldsymbol{\xi}^-,\varepsilon\right)=\frac{1}{\tau}\cdot\frac{1}{N}\sum_{i=1}^{N}\left((\xi_i^+)^2+(\xi_i^-)^2\right) \\ & +\frac{1}{2}\mathbf{p}_g^T\mathbf{p}_g+\frac{2}{\tau}\cdot\varepsilon\end{aligned} \tag{6.a}$$
$$\text{s.t.}\ \begin{cases} y_i-\mathbf{p}_g^T\mathbf{x}_{gi}<\varepsilon+\xi_i^+ \\ \mathbf{p}_g^T\mathbf{x}_{gi}-y_i<\varepsilon+\xi_i^- \end{cases}\quad\forall i\,.$$

Compared with the L1-norm penalty-based ε-insensitive criterion function, the L2-norm penalty-based criterion is advantageous because of the following characteristics: 1) the constraints $\xi_i^+\ge 0$ and $\xi_i^-\ge 0$ in Eq. (5.c) are not required for the optimization; 2) the insensitive parameter ε can be obtained automatically by optimization without the need of manual setting. Similar properties can also be found in other L2-norm penalty-based machine learning algorithms, such as L2-SVR [32]. For convenience, the L2-norm penalty based ε-insensitive TSK fuzzy model training is referred to as L2-TSK-FS in this paper. Based on optimization theory, the dual problem in Eq. (6.a) can be formulated as the following QP problem.

$$\begin{aligned}\max_{\boldsymbol{\alpha}^+,\boldsymbol{\alpha}^-}\ & -\sum_{i=1}^{N}\sum_{j=1}^{N}(\alpha_i^+-\alpha_i^-)(\alpha_j^+-\alpha_j^-)\cdot\mathbf{x}_{gi}^T\mathbf{x}_{gj}-\sum_{i=1}^{N}\frac{N\tau}{2}(\alpha_i^+)^2 \\ & -\sum_{i=1}^{N}\frac{N\tau}{2}(\alpha_i^-)^2+\sum_{i=1}^{N}\alpha_i^+\cdot y_i\cdot\tau-\sum_{i=1}^{N}\alpha_i^-\cdot y_i\cdot\tau\end{aligned} \tag{6.b}$$
$$\text{s.t.}\ \sum_{i=1}^{N}(\alpha_i^++\alpha_i^-)=1\,,\ \alpha_i^+,\alpha_i^-\ge 0\quad\forall\, i\,.$$

Notably, the characteristic of the QP problem in Eq. (6.b) enables the use of core-set based Minimal Enclosing Ball (MEB) approximation technique to solve problems involving very large datasets [32]. The scalable L2-TSK-FS learning algorithm (STSK) has thus been proposed in this regard [20].

III. Knowledge Leverage Based TSK-FS

A. Framework of Classical Data-driven TSK-FS Learning

The discussions in section II indicate that the training methods commonly used for TSK fuzzy systems have become 'data-driven methods'. Typically, these methods obtain the model parameters by optimization techniques based on a certain objective criterion and the data sampled from the scene to be modeled. Almost all data-driven training algorithms developed for fuzzy system modeling only consider the data collected from the current scene, even if the information available from the reference scenes is useful. The situation is shown in Fig. 2.

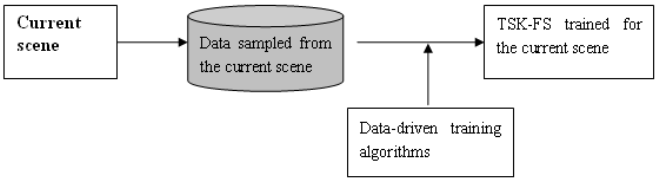

Fig. 2. Traditional data-driven fuzzy system modeling

B. Learning with Knowledge Leverage

As mentioned in the introductory section, when data are insufficient, the existing data-driven fuzzy system training methods will no longer be effective and the generalization performance of the trained systems become inferior. A promising strategy to cope with this issue is to capitalize on other reference scenes which may contain important information that is relevant to the current scene, at least to a certain extent. Two categories of useful information are usually available from the reference scenes, i.e., the *data* and the *knowledge*, as shown in Fig. 3.

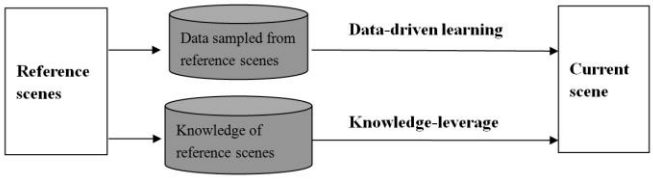

Fig. 3. Leverage on useful information in the reference scenes for the modeling of the current scene

The characteristics of these two categories of information are delineated respectively as follows.

1) *Data* refer to the original information in the references scenes that can be further processed to obtain knowledge. In some situations, however, the data are not always available. For example, many data samples cannot be made open due to privacy protection. Moreover, even if the data of reference scenes are available, it may not be appropriate to directly adopt them for the modeling task in the current scene. This is because drifting may exist between the models of different scenes, so that some of the data from the reference scenes could instead cause negative effect on the system modeling of the current scene.

2) *Knowledge* is the other kind of information obtainable from the reference scenes. The types of knowledge are diverse, including density distribution and model parameters of the reference scenes, for example. Most of them can be obtained by learning procedures. For example, the parameters of a fuzzy system developed for a reference scene can be learned by certain fuzzy system training algorithm based on the data collected from that scene. Despite of the fact that most of the knowledge obtained cannot be inversely mapped to the original data (which is a favorable feature from the perspective of privacy preservation), the knowledge from the reference scenes is important information to improve the modeling of the current scene.

Based on the discussions above, we can see that it is more appropriate to exploit the use of knowledge from the reference scenes, instead of the data from the reference scenes and the data from the current scene, to accomplish the modeling task of the current scene when the data are insufficient. The proposed KL-TSK-FS in the paper is developed using this strategy to train TSK-FS.

C. Framework of KL-TSK-FS Learning

The framework for the construction of KL-TSK-FS can be described with Fig. 4. As shown in the figure, there are two major information sources for the learning of a fuzzy system, 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 of the current scene.

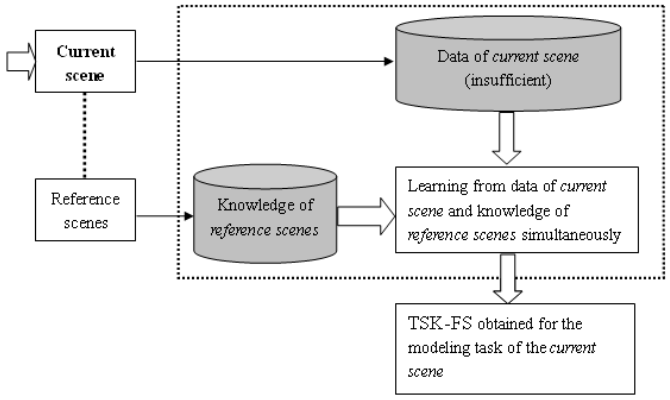

Fig. 4. Framework of KL-TSK-FS learning

IV. KL-TSK-FS Modeling Method

To take advantage of knowledge leverage for TSK-FS, the KL-TSK-FS is proposed by integrating the ε -insensitive criterion and the L2-norm penalty based TSK-FS learning strategy using the corresponding knowledge-leverage mechanism. The goal is to effectively use the knowledge of the reference scenes to remedy the deficiency caused by data insufficiency in the current scene, and to develop an efficient learning algorithm for the fuzzy system.

A. Objective Criterion Integrating the Knowledge from Reference Scenes

For a TSK-FS constructed by the ε -insensitive criterion based technique, the corresponding model parameters obtained in the reference scenes can be regarded as the knowledge. To develop an effective KL-TSK-FS for model learning of the current scene, we propose an optimization criterion which integrates the knowledge of the reference scenes as follows,

$$\min_{\mathbf{p}_g} \ \sum_{i=1}^{N} |\mathbf{p}_g^T \mathbf{x}_{gi} - y_i|_\varepsilon + \lambda(\mathbf{p}_g - \mathbf{p}_{g0})^T(\mathbf{p}_g - \mathbf{p}_{g0}) . \tag{7}$$

The optimization criterion in Eq. (7) contains two terms. The first term refers to the learning of the data of the current scene for the desired TSK-FS. This term is included so that the desired TSK-FS will fit the sampled training data of the current scene as accurate as possible. The second term refers to the knowledge leverage of the reference scene, with $\mathbf{p}_{g0}$ denoting model parameters learned from the reference scenes. The purpose is to estimate the desired parameters by approximating the model obtained from the reference scenes. The parameter λ in Eq. (7) 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. As in L2-TSK-FS [20], we introduce the terms *structure risk* and ε *-insensitive penalty* into Eq. (7) to obtain the following objective criterion,

$$\begin{aligned} \min_{\mathbf{p}_g, \xi^+, \xi^-, \varepsilon} \ & \frac{1}{\tau} \cdot \frac{1}{N} \sum_{i=1}^{N} ((\xi_i^+)^2 + (\xi_i^-)^2) + \frac{1}{2}(\mathbf{p}_g^T \mathbf{p}_g) \\ & + \frac{2}{\tau} \cdot \varepsilon + \lambda(\mathbf{p}_g - \mathbf{p}_{g0})^T(\mathbf{p}_g - \mathbf{p}_{g0}) \end{aligned} \tag{8}$$

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

Note that the first three terms in Eq. (8) are directly inherited from the L2-TSK-FS [20] and the last term is referred to as the *knowledge-leverage term* which is used to learn the knowledge from the reference scenes. Based on the objective criterion in Eq. (8), we can derive the corresponding learning rules for the proposed KL-TSK-FS.

B. Parameter Solution for KL-TSK-FS

Given the optimization problem in Eq. (8), Theorem 1 below is proposed for parameter solution.

Theorem 1: The dual problem of Eq. (8) is a QP problem as shown in Eq. (9).

$$\begin{aligned} \max_{\boldsymbol{\alpha}^-, \boldsymbol{\alpha}^+} \ & \frac{-1}{2(1+2\lambda)} \cdot \sum_{i=1}^{N} \sum_{j=1}^{N} (\alpha_i^+ - \alpha_i^-)(\alpha_j^+ - \alpha_j^-) \mathbf{x}_{gi}^T \mathbf{x}_{gj} \\ & - \frac{N\tau}{4} \cdot \sum_{i=1}^{N} \left((\alpha_i^+)^2 + (\alpha_i^-)^2 \right) - \frac{2\lambda}{1+2\lambda} \cdot \sum_{i=1}^{N} (\alpha_i^+ - \alpha_i^-) \mathbf{p}_{g0}^T \mathbf{x}_{gi} \\ & + \sum_{i=1}^{N} (\alpha_i^+ - \alpha_i^-) y_i \end{aligned}$$

$$\text{s.t. } \sum_{i=1}^{N}\alpha_i^- + \sum_{i=1}^{N}\alpha_i^+ = \frac{2}{\tau},\ \alpha_i^- \ge 0,\ \alpha_i^+ \ge 0. \qquad (9)$$

Proof: By using the Lagrangian optimization theorem, we can obtain the following Lagrangian function for Eq. (8)

$$\begin{aligned} &L(\mathbf{p}_g, \boldsymbol{\xi}^+, \boldsymbol{\xi}^-, \varepsilon, \boldsymbol{\alpha}^+, \boldsymbol{\alpha}^-) \\ &= \frac{1}{\tau}\cdot\frac{1}{N}\sum_{i=1}^{N}((\xi_i^+)^2 + (\xi_i^-)^2) + \frac{1}{2}(\mathbf{p}_g^T\mathbf{p}_g) + \frac{2}{\tau}\cdot\varepsilon + \lambda(\mathbf{p}_g - \mathbf{p}_{g0})^T(\mathbf{p}_g - \mathbf{p}_{g0}) \\ &\quad + \sum_{i=1}^{N}\alpha_i^+(y_i - \mathbf{p}_g^T\mathbf{x}_{gi} - \varepsilon - \xi_i^+) + \sum_{i=1}^{N}\alpha_i^-(\mathbf{p}_g^T\mathbf{x}_{gi} - y_i - \varepsilon - \xi_i^-). \end{aligned} \qquad (10)$$

According to the dual theorem, the minimum of the Lagrangian function in Eq. (10) with respect to $\mathbf{p}_g, \boldsymbol{\xi}^+, \boldsymbol{\xi}^-, \varepsilon$ is equal to the maximum of the function with respect to $\boldsymbol{\alpha}^+, \boldsymbol{\alpha}^-$. Then the following equations can be considered as the necessary conditions of the optimal solution:

$$\frac{\partial L}{\partial \mathbf{p}_g} = \mathbf{p}_g + 2\lambda(\mathbf{p}_g - \mathbf{p}_{g0}) - \sum_{i=1}^{N}(\alpha_i^+ - \alpha_i^-)\mathbf{x}_{gi} = 0, \qquad (11.a)$$

$$\frac{\partial L}{\partial \xi_i^+} = \frac{2}{N\tau}\xi_i^+ - \alpha_i^+ = 0, \qquad (11.b)$$

$$\frac{\partial L}{\partial \xi_i^-} = \frac{2}{N\tau}\xi_i^- - \alpha_i^- = 0, \qquad (11.c)$$

$$\frac{\partial L}{\partial \varepsilon} = \frac{2}{\tau} - \sum_{i=1}^{N}\alpha_i^- - \sum_{i=1}^{N}\alpha_i^+ = 0. \qquad (11.d)$$

From Eqs. (11.a)-(11.d), we have

$$\mathbf{p}_g = \frac{2\lambda\mathbf{p}_{g0} + \sum_{i=1}^{N}(\alpha_i^+ - \alpha_i^-)\mathbf{x}_{gi}}{1+2\lambda}, \qquad (12.a)$$

$$\xi_i^+ = \frac{N\tau\cdot\alpha_i^+}{2}, \qquad (12.b)$$

$$\xi_i^- = \frac{N\tau\cdot\alpha_i^-}{2}, \qquad (12.c)$$

$$\sum_{i=1}^{N}\alpha_i^- + \sum_{i=1}^{N}\alpha_i^+ = \frac{2}{\tau}. \qquad (12.d)$$

Substituting Eqs. (12.a)-(12.d) into Eq. (10), we obtain the dual problem for Eq. (8), i.e.,

$$\begin{aligned} \max_{\boldsymbol{\alpha}^-,\boldsymbol{\alpha}^+}\ & \frac{-1}{2(1+2\lambda)}\cdot\sum_{i=1}^{N}\sum_{j=1}^{N}(\alpha_i^+ - \alpha_i^-)(\alpha_j^+ - \alpha_j^-)\mathbf{x}_{gi}^T\mathbf{x}_{gj} \\ & - \frac{N\tau}{4}\cdot\sum_{i=1}^{N}\left((\alpha_i^+)^2 + (\alpha_i^-)^2\right) - \frac{2\lambda}{1+2\lambda}\cdot\sum_{i=1}^{N}(\alpha_i^+ - \alpha_i^-)\mathbf{p}_{g0}^T\mathbf{x}_{gi} \\ & + \sum_{i=1}^{N}(\alpha_i^+ - \alpha_i^-)y_i + \frac{\lambda}{1+2\lambda}\cdot\mathbf{p}_{g0}^T\mathbf{p}_{g0} \end{aligned}$$

$$\text{s.t. } \sum_{i=1}^{N}\alpha_i^- + \sum_{i=1}^{N}\alpha_i^+ = \frac{2}{\tau},\ \alpha_i^- \ge 0,\ \alpha_i^+ \ge 0,\ \forall i. \qquad (12.e)$$

Since the optimal solution of the dual problem, i.e., $(\boldsymbol{\alpha}^+)^*, (\boldsymbol{\alpha}^-)^*$, is independent of $\lambda/(1+2\lambda)\cdot\mathbf{p}_{g0}^T\mathbf{p}_{g0}$, Eq. (12.e) is equivalent to the following equation,

$$\begin{aligned} \max_{\boldsymbol{\alpha}^-,\boldsymbol{\alpha}^+}\ & \frac{-1}{2(1+2\lambda)}\cdot\sum_{i=1}^{N}\sum_{j=1}^{N}(\alpha_i^+ - \alpha_i^-)(\alpha_j^+ - \alpha_j^-)\mathbf{x}_{gi}^T\mathbf{x}_{gj} \\ & - \frac{N\tau}{4}\cdot\sum_{i=1}^{N}\left((\alpha_i^+)^2 + (\alpha_i^-)^2\right) - \frac{2\lambda}{1+2\lambda}\cdot\sum_{i=1}^{N}(\alpha_i^+ - \alpha_i^-)\mathbf{p}_{g0}^T\mathbf{x}_{gi} \\ & + \sum_{i=1}^{N}(\alpha_i^+ - \alpha_i^-)y_i \end{aligned}$$

$$\text{s.t. } \sum_{i=1}^{N}\alpha_i^- + \sum_{i=1}^{N}\alpha_i^+ = \frac{2}{\tau},\ \alpha_i^- \ge 0,\ \alpha_i^+ \ge 0,\ \forall i. \qquad (12.f)$$

Thus, Theorem 1 is hold.

It is clear from the above results that the optimization problem in Eq. (9) for TSK fuzzy model training can be transformed into a QP problem that can be directly solved by existing QP solutions, such as the working set based QP solution [33].

With the optimal solution $(\boldsymbol{\alpha}^+)^*, (\boldsymbol{\alpha}^-)^*$ of the dual problem in Eq. (9), we can get the optimal solution of the primal problem in Eq. (8) based on the relations presented in Eqs. (12.a)-(12.d). The optimal model parameters of the trained TSK-FS, i.e., $(\mathbf{p}_g)^*$, is then given by

$$(\mathbf{p}_g)^* = \frac{2\lambda\mathbf{p}_{g0} + \sum_{i=1}^{N}\left((\alpha_i^+)^* - (\alpha_i^-)^*\right)\mathbf{x}_{gi}}{1+2\lambda}, \qquad (13.a)$$

which can be further expressed as

$$(\mathbf{p}_g)^* = \gamma\mathbf{p}_{g0} + (1\text{-}\gamma)\ \mathbf{p}_{gc}, \qquad (13.b)$$

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

From Eq. (13.b), we can see that the final optimal parameter $(\mathbf{p}_g)^*$ obtained for the desired TSK-FS contains two parts, i.e. $\gamma\cdot\mathbf{p}_{g0}$ and $(1\text{-}\gamma)\cdot\mathbf{p}_{gc}$. While $(1\text{-}\gamma)\cdot\mathbf{p}_{gc}$ 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. Thus, the final model parameter $(\mathbf{p}_g)^*$ is a balance between these two kinds of knowledge.

C. Learning Algorithm for KL-TSK-FS

Based on the findings in the previous section, the learning algorithm of the proposed KL-TSK-FS is developed and described as follows.

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. (8).
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 Eqs. (9) and (13.a) 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.

The computational complexity of the algorithm is analyzed as follows. The whole algorithm includes two main parts: 1) acquisition of the antecedent parameters of the fuzzy system; and 2) learning of the consequent parameters. For the first part, since the antecedent parameters are inherited directly from the reference scene as the available knowledge, the computational complexity is O(1). For the second, the consequent parameters are obtained by solving the QP problem in Eq. (9) and the complexity is usually $O(N^2)$ for typical QP problems. However,

it can be further reduced to O(N) with some sophisticated algorithms, such as the working set based algorithm [33]. Therefore, the computational complexity of the proposed algorithm is between O(N) and O(N^2), depending on the QP solutions used. In this study, we adopt the working set based QP solution [33] for solving the QP problem concerned.

Remark 1: Although the proposed algorithm has distinctive advantages, it has inherent drawbacks due to the learning mechanism. First, the method assumes that some useful knowledge in the reference scene is available. Otherwise, it is not helpful to use the method since no knowledge can be transferred in the training procedure. Second, it also assumes that the available knowledge is useful for the modeling task of current scene. Thus, if the knowledge from reference scene is irrelevant knowledge, the modeling effect on the current scene may not be improved effectively.

V. Experimental Results

A. Experimental Settings

The proposed learning algorithm for KL-TSK-FS 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 IV. Here, datasets generated from the reference scene and the current scene are denoted by D1 and D2 respectively. The proposed learning algorithm is evaluated from the following two aspects.

(1) *Comparison with traditional L2-TSK-FS*. The performance of KL-TSK-FS is compared comprehensively with that of three L2-TSK-FS methods implemented under different conditions. That is, a total of four TSK-FSs are developed. They are constructed by (i) L2-TSK-FS based on the data in the reference scene, (ii) L2-TSK-FS based on the data in the current scene, (iii) L2-TSK-FS based on the data in both the current scene and the reference scene, and (iv) the proposed KL-TSK-FS, which are denoted respectively by L2-TSK-FS(D1), L2-TSK-FS (D2), L2-TSK-FS (D1+D2) and KL-TSK-FS(D2+Knowledge). With these four fuzzy systems, the testing data, i.e. D2_test, of the current scene are used to evaluate their generalization capability.

(2) *Comparison with regression methods designed for datasets with missing or noisy data*. Three related regression methods are employed to compare with the proposed KL-TSK-FS for performance evaluation. The three methods include: (i) TS-fuzzy system-based support vector regression (TSFS-SVR) [34]; (ii) fuzzy system learned through fuzzy clustering and support vector machine (FS-FCSVM) [35]; and (iii) Bayesian task-level transfer learning for non-linear regression method (HiRBF) [15].

The methods adopted for performance comparison from these two aspects are summarized in Table V. The following generalization performance index *J* is used in the experiments [29],

$$J = \sqrt{\sum_{i=1}^{N}(y_i' - y_i)^2 \Big/ N \Big/ \sum_{i=1}^{N}(y_i - \bar{y})^2 \Big/ N}\,, \tag{14}$$

where *N* is the number of test datasets, 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.

TABLE IV
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 V
THE METHODS ADOPTED FOR PERFORMANCE COMPARISON

L2-norm penalty based TSK-FS modeling methods		Four methods designed for noisy/missing data	
The proposed KL-TSK-FS (D2+knowledge)	L2-TSK-FS(D1) [20]	The proposed KL-TSK-FS	TSFS-SVR [34]
	L2-TSK-FS (D2) [20]		FS-FCSVM [35]
	L2-TSK-FS (D1+D2) [20]		HiRBF [16]

In these experiments, the hyper parameters in the proposed method and the algorithms adopted for performance comparison are determined by using the five-fold cross-validation (CV) strategy with the training datasets. The final model is then trained with the whole training dataset using the parameter values obtained by the CV strategy. 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 it is necessary to satisfy the following requirements: 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 $Y = f(x) = \sin(x) * x, x \in [-10,10]$ is used to describe the reference scene and to generate the dataset D1. On the other hand, the function $y = r * f(x) = r * \sin(x) * x, x \in [-10,10]$ is used to describe the current scene and to generate the dataset D2 and the 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 of missing data into the training set generated for the current scene. The settings for generating the synthetic datasets are described in Table VI. For example, the functions used to simulate the two related scenes, with the relation parameter $r = 0.85$, are shown in Fig. 5(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. 5(b). The figures also show the two data-missing intervals [-6,-3] and [0, 4] introduced into the dataset.

TABLE VI
DETAILS OF THE SYNTHETIC DATASETS

Reference scene	Current scene		
Dataset	Training set		Testing set
Size	Interval with missing data	Size	Size
400	[-6,-3] and [0,4]	144	200
Relation parameter between the two scenes: r =0.9, 0.85, 0.8, 0.75 and 0.7			

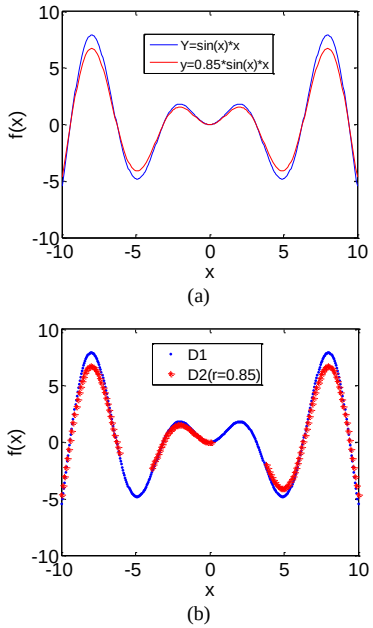

Fig. 5. Functions representing two different scenes with the relation parameter r =0.85 and the corresponding sampled data from these scenes: (a) the functions representing the reference scene (Y) and the current scene (y); (b) the data of the reference scene and the training data of the current scene with missing data in the intervals [-6, -3] and [0, 4].

TABLE VII
GENERALIZATION PERFORMANCE (J) OF THE PROPOSED KL-TSK-FS METHOD AND THE TRADITIONAL L2-TSK-FS METHODS ON THE SYNTHETIC DATASETS

Interval with data missing	Relation parameter (r)	L2-TSK-FS (D1)	L2-TSK-FS (D2)	L2-TSK-FS (D1+D2)	KL-TSK-FS (D2+ Knowledge)
[-6,-3] and [0, 4]	0.9	0.1343	0.2858	0.1012	**0.0501**
	0.85	0.1908	0.2813	0.1434	**0.0516**
	0.8	0.2574	0.2864	0.1983	**0.1094**
	0.75	0.3525	0.2841	0.2627	**0.1534**
	0.7	0.4406	0.2821	0.3432	**0.2388**

TABLE VIII
GENERALIZATION PERFORMANCE (J) OF THE PROPOSED KL-TSK-FS METHOD AND THREE RELATED REGRESSION METHODS ON THE SYNTHETIC DATASETS

Interval with data missing	Relation parameter (r)	TSFS-SVR	FS-FCSVM	HiRBF	KL-TSK-FS
[-6,-3] and [0,4]	0.9	0.2972	0.3161	0.2621	**0.0501**
	0.85	0.2989	0.3179	0.2619	**0.0516**
	0.8	0.2983	0.3170	0.2687	**0.1094**
	0.75	0.2933	0.3167	0.2639	**0.1534**
	0.7	0.2970	0.3185	0.2611	**0.2388**

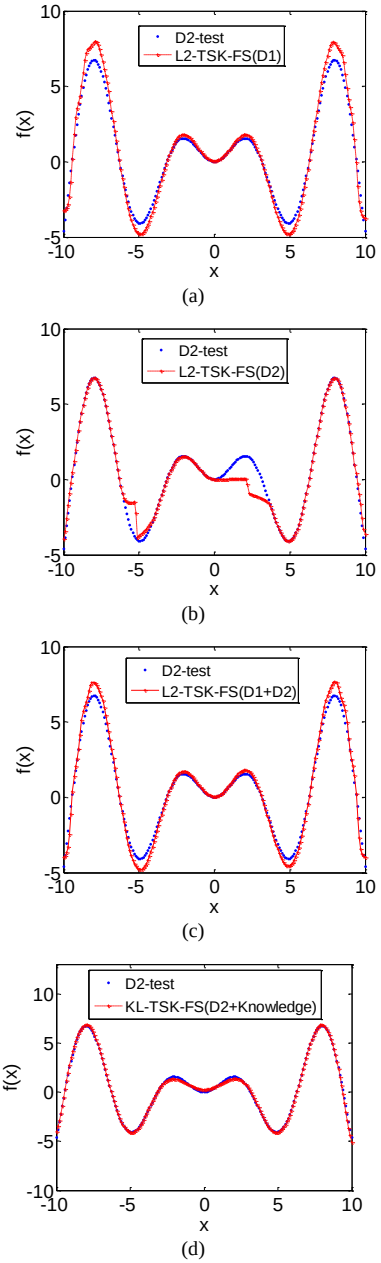

Fig. 6. Modeling results of the proposed KL-TSK-FS method and the three traditional L2-TSK-FS methods using the synthetic datasets shown in Fig. 5(b): (a) L2-TSK-FS on (D1); (b) L2-TSK-FS on (D2); (c) L2-TSK-FS on D1and D2; (d) the proposed KL-TSK-FS on D2 and , (D2+Knowledge).

2) Comparing with the Traditional L2-TSK-FS Modeling Methods

The performance of the proposed KL-TSK-FS and the three traditional L2-norm penalty based TSK-FS modeling methods is evaluated and compared using the synthetic datasets. The experimental results are shown in Table VII and Fig. 6. The following observations can be made from the results:

(1) It can be seen from Table VII that the generalization performance of the proposed KL-TSK-FS is better than that of the traditional L2-TSK-FS methods.

(2) Fig. 6(a) shows the modeling results of the L2-TSK-FS obtained by using the data of the reference scene only. The results indicate that drifting exists between the reference scene and the current scene, as evident from the discrepancies between the two curves in the figure. Hence, the generalization performance of the TSK-FS obtained by L2-TSK-FS from the reference scene is weak for the current scene. The findings show that the use of the data of the reference scene alone is not appropriate for the modeling of the current scene.

(3) Fig. 6(b) shows the modeling results of the L2-TSK-FS obtained by using the data of the current scene only. The results indicate that the generalization performance of the TSK-FS obtained by L2-TSK-FS is even much weaker for the current scene. An obvious reason is that the data in the training set is insufficient, which degrades the generalization capability of the obtained TSK-FS. The prediction performance is especially poor in the intervals with missing data in the training dataset.

(4) Fig. 6(c) shows the modeling results of the L2-TSK-FS obtained by using the data of both the current scene and the reference scene. Although the data of both scenes have been used for training, the generalization performance of the obtained TSK-FS is still not good enough for the current scene. This can be explained with two reasons. First, drifting occurs between the reference and current scenes, i.e., not all data in the reference scene are useful for the modeling task of the current scene. Some of them may even produce negative effect. Second, the size of the reference scene is larger than that of the current scene, which makes the obtained TSK-FS more apt to approximate the reference scene rather than the current scene.

(5) Fig. 6(d) shows the modeling results of the proposed KL-TSK-FS. By inspecting Fig. 6(a) and Fig. 6(d), we can see that the KL-TSK-FS demonstrates better prediction results than the L2-TSK-FS. The latter only uses the data of reference scene. Besides, it is also evident from Fig. 6(b) and Fig. 6(d) that, by introducing the knowledge-leverage mechanism, the proposed KL-TSK-FS has effectively remedied the deficiency of the L2-TSK-FS obtained by the data of the current scene. Furthermore, by comparing Fig. 6(c) and Fig. 6(d), we also find that the KL-TSK-FS has demonstrated better generalization performance than the L2-TSK-FS which employs the data of both the reference and current scenes. It is noteworthy to point out that the KL-TSK-FS also possesses better privacy protection capability than the methods that use the data of the reference scenes directly. In practice, methods requiring data of all the scenes are no longer feasible when the data in the reference scenes are not available due to the necessity of privacy protection, or in situations where only the knowledge is disclosed. Therefore, the proposed KL-TSK-FS is particularly suitable for these situations, illustrating its distinctiveness in privacy protection.

3) Comparing with Regression Methods Designed for Missing or Noisy Data

The performance of the proposed KL-TSK-FS method is evaluated by comparing its performance with that of the three regression methods designed for handling noisy/missing data, i.e., TSFS-SVR, FS-FCSVM and HiRBF. The evaluation is performed on the synthetic datasets. The experimental results are shown in Table VIII and Fig. 7, and the following observations can be obtained:

1) The KL-TSK-FS has demonstrated better generalization performance than the other three methods.

2) The results in Fig. 7(a) and Fig. 7(b) show that the support vector learning based fuzzy modeling methods TSFS-SVR and FS-FCSVM are able to give better generalization performance to some extent. For example, although the data in the interval [-6, -3] are missing, these two methods still demonstrate promising generalization capability at this interval. However, the generalization abilities of these two methods in the other data-missing interval [0, 4] are not satisfactory.

3) Although the transfer learning based method HiRBF has used the data in both the current scene and the reference scene, it is evident form Fig. 7(c) that this method cannot effectively cope with the problem caused by missing data, still exhibiting poor generalization ability in the two data-missing intervals.

4) Fig. 7(d) shows that the proposed KL-TSK-FS is able to offer acceptable generalization capability in the two data-missing intervals, indicating that the method has effectively leveraged the useful knowledge from the reference scene to remedy the generalization abilities in the training procedure.

C. Real-world Datasets

1) The Glutamic Acid Fermentation Process Modeling

To further evaluate the performance of the proposed knowledge-leverage based fuzzy system learning method, an experiment is conducted to apply the method to model a biochemical process with real-world datasets [20]. The datasets 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. 8. 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 IX.

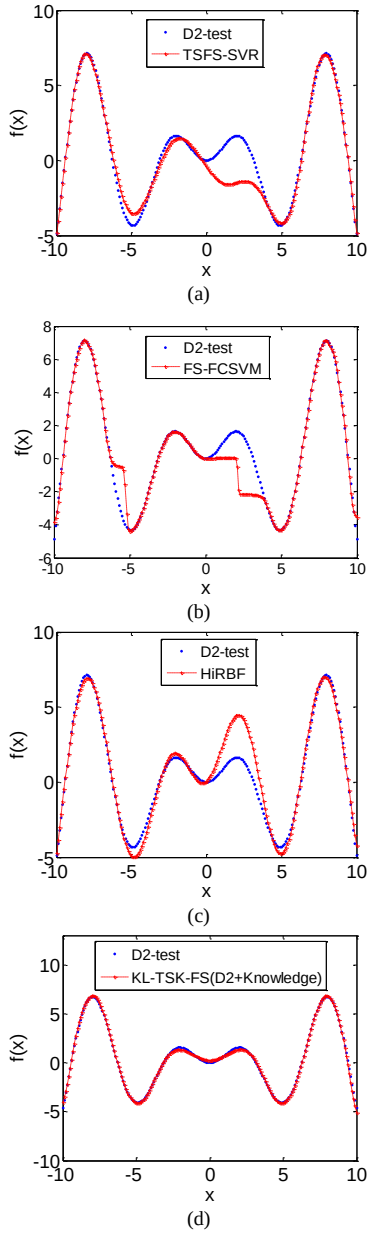

Fig. 7. Modeling results of the proposed KL-TSK-FS method and three related regression methods using the synthetic datasets in Fig. 5(b): (a) TSFS-SVR, (b) FS-FCSVM, (c) HiRBF and (d) KL-TSK-FS.

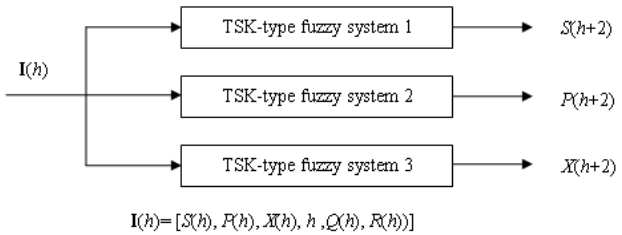

Fig. 8. Illustration of the glutamic acid fermentation process prediction model based on TSK-FS.

TABLE IX
THE FERMENTATION PROCESS MODELING DATASETS

	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.

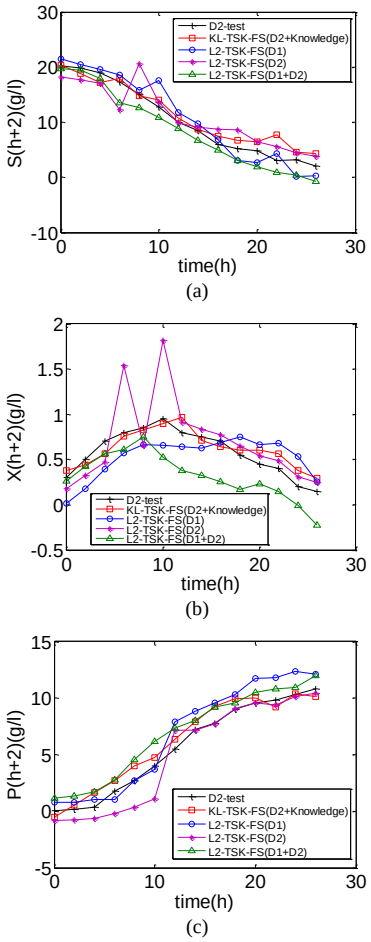

Fig. 9. Performance comparison between the proposed KL-TSK-FS method and three traditional L2-TSK-FS methods in fermentation process modeling: the prediction results of (a) $S(h+2)$ for the 21st batch; (b) $X(h+2)$ for the 21st batch; (c) $P(h+2)$ for the 21st batch.

2) Comparing with the Traditional L2-TSK-FS Modeling Methods

The experimental results of fermentation process modeling using the proposed method KL-TSK-FS and the traditional L2-TSK-FS are given in Table X and Fig. 9, where the modeling effect of the 21st batch is only given in Fig. 9 and that of the 20th batch is omitted for brevity. The findings are similar to those presented in section IV-B for the experiments performed on the synthetic datasets. The modeling results of the KL-TSK-FS are better than that of the three traditional L2-TSK-FS methods.

Especially from Fig. 9, fluctuations are evident from the modeling results of some existing algorithms, which are mainly caused by the lack of information in the learning procedure for the corresponding intervals. As the proposed method can effectively exploit not only the data of the current scene but also

the useful knowledge of the reference scenes, the obtained TSK-FS has demonstrated good adaptive abilities and generated better modeling effect. It can also be seen from the experimental results that, even if the data in the training data of the current scene are missing, the generalization capability of the TSK-FS obtained by the proposed KL-TSK-FS does not degraded significantly. Similar observation can also be obtained from the experimental results in Fig. 10 in the next section. This remarkable feature is very valuable for the biochemical process modeling task as data insufficiency is common problem due to poor sensor sensitivity and noisy environment.

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

Output	L2-TSK-FS (D1)	L2-TSK-FS (D2)	L2-TSK-FS (D1+D2)	KL-TSK-FS (D2+ Knowledge)
$S(h+2)$	0.2792	0.3944	0.2804	**0.1239**
$X(h+2)$	0.8342	1.1203	1.0642	**0.4548**
$P(h+2)$	0.2842	0.3255	0.2533	**0.1482**

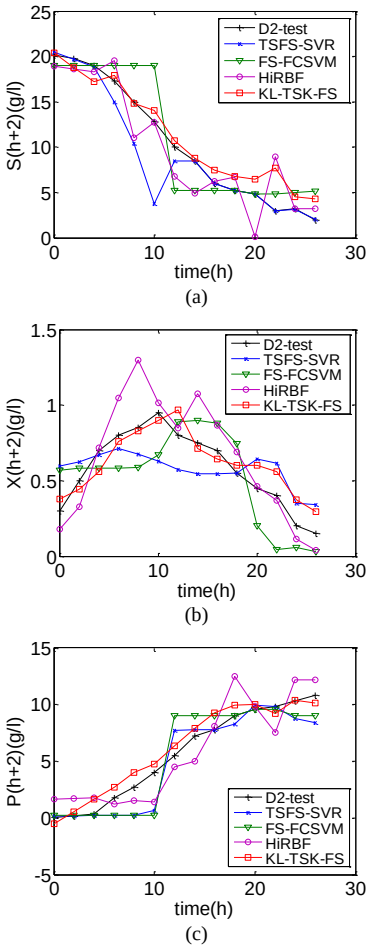

Fig. 10. Performance comparison between the proposed KL-TSK-FS method and three regression methods in fermentation process modeling: the prediction results of (a) $S(h+2)$ for the 21st batch; (b) $X(h+2)$ for the 21st batch; (c) $P(h+2)$ for the 21st batch.

3) Comparing with the Regression Methods Designed for Missing or Noisy Data

The experimental results of fermentation process modeling using the proposed method KL-TSK-FS and three regression methods (i.e., TSFS-SVR, FS-FCSVM and HiRBF) are shown in Table XI and Fig. 10 (as in Fig. 9, only the modeling result of the 21st batch is given). Similar to the findings presented in section V-B-3 for the experiments conducted with the synthetic datasets, in general, the proposed KL-TSK-FS has demonstrated better generalization performance in fermentation process modeling than the other three regression methods. This can again be explained by the fact that the proposed KL-TSK-FS has effectively leveraged the useful knowledge from the reference scene in the training procedure such that the influence of the missing data can be properly reduced.

TABLE XI
GENERALIZATION PERFORMANCE (J) OF THE KL-TSK-FS METHOD AND RELATED REGRESSION METHODS IN FERMENTATION PROCESS MODELING

Output	TSFS-SVR	FS-FCSVM	HiRBF	KL-TSK-FS
$S(h+2)$	0.3452	0.3750	0.3510	**0.1239**
$X(h+2)$	0.7295	0.6118	0.7026	**0.4548**
$P(h+2)$	0.3574	0.4144	0.4117	**0.1482**

VI. CONCLUSIONS

In this study, the concept of knowledge leverage is introduced and used to develop the knowledge-leverage based TSK-type fuzzy system KL-TSK-FS. The corresponding learning algorithm is also presented. The goal is to remedy the deficiency due to data insufficiency in the current scene by leveraging the useful knowledge available from the reference scenes. The proposed algorithm can learn from not only the data of the current scene but also the knowledge of the reference scenes. Moreover, since all that is needed for the training of the fuzzy system of the current scene is the knowledge of the reference scenes (not the data), the learning algorithm is able to preserve data confidentiality in the reference scenes.

The experimental results have demonstrated the attractiveness and effectiveness of the proposed method when compared with the existing methods. Despite the promising performance, there are still rooms for further improvement. For example, for the proposed knowledge-leverage based TSK-type fuzzy system, reduction of training time is needed in order to make possible online applications. Development of more advanced transfer learning mechanisms for fuzzy system learning is also important to deal with scenes with inadequate data. Future work will be devoted to these issues.

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] 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.

[3] 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.

[4] S. Bickel, M. Brückner, and T. Scheffer, "Discriminative learning for differing training and test distributions," *Proc. 24th Int. Conf. Machine Learning, pp. 81-88*, 2007.

[5] M. Sugiyama, S. Nakajima, H. Kashima, P.V. Buenau, and M. Kawanabe, "Direct importance estimation with model selection and its

application to covariate shift adaptation," *Proc. 20th Ann. Conf. Neural Information Processing Systems*, Dec. 2008.

[6] N.D. Lawrence and J.C. Platt, "Learning to learn with the informative vector machine," *Proc. 21st Int. Conf. Machine Learning*, July 2004.

[7] A. Schwaighofer, V. Tresp, and K. Yu, "Learning Gaussian process kernels via hierarchical Bayes," *Proc. 17th Ann. Conf. Neural Information Processing Systems*, pp. 1209-1216, 2005.

[8] J. Gao, W. Fan, J. Jiang, and J. Han, "Knowledge transfer via multiple model local structure mapping," *Proc. 14th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, pp. 283-291, Aug. 2008.

[9] L. Mihalkova, T. Huynh, and R.J. Mooney, "Mapping and revising markov logic networks for transfer learning," *Proc. 22nd Assoc. for the Advancement of Artificial Intelligence (AAAI) Conf. Artificial Intelligence*, pp. 608-614, July 2007.

[10] L. Mihalkova and R.J. Mooney, "Transfer learning by mapping with minimal target data," *Proc. Assoc. for the Advancement of Artificial Intelligence (AAAI '08) Workshop Transfer Learning for Complex Tasks*, July 2008.

[11] J. Davis and P. Domingos, "Deep transfer via second-order Markov logic," *Proc. Assoc. for the Advancement of Artificial Intelligence (AAAI '08) Workshop Transfer Learning for Complex Tasks*, July 2008.

[12] S.J. Pan, I.W. Tsang, J.T. Kwok, and Q. Yang, "Domain adaptation via transfer component analysis," *IEEE Trans. Neural Networks*, vol. 22, no.2, pp.199-210, 2011.

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

[14] Z. Wang, Y. Song, and C. Zhang, "Transferred dimensionality reduction," *Proc. European Conf. Machine Learning and Knowledge Discovery in Databases (ECML/PKDD '08)*, pp. 550-565, Sept. 2008.

[15] 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.

[16] L. Borzemski and G. Starczewski, "Application of transfer regression to TCP throughput prediction," *Proc. First Asian Conference on Intelligent Information and Database Systems*, pp. 28-33, 2009.

[17] W. Mao, G. Yan, J. Bai, and H. Li, "Regression transfer learning based on principal curve," *Lecture Note on Computer Science 6063*, pp. 365–372, 2010.

[18] J. Liu, Y. Chen, and Y. Zhang, "Transfer regression model for indoor 3D location estimation," *Lecture Note on Computer Science 5916*, pp. 603–613, 2010.

[19] D. Pardoe and P. Stone, "Boosting for regression transfer," *Proc. Int. Conf. Machine Learning*, pp. 863-870, 2010.

[20] 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 System*, vol. 19, no.2, pp.210-226, 2011.

[21] J. M. Mendel, *Uncertain rule-based fuzzy logic systems: Introduction and new directions*. Upper Saddle River, NJ: Prentice-Hall, 2001.

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

[23] S.T. Wang, *Neural-fuzzy systems and their application*. Beijing: Beijing Aeronautical University Press, 1998.

[24] L.X. Wang, *Adaptive fuzzy systems and control: Design and stability analysis*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1994.

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

[26] T. Takagi and M. Sugeno, "Fuzzy identification of systems and its application to modeling and control," *IEEE Trans. Systems Man and Cybernetics*, vol.15, no.1, pp. 116-132, 1985.

[27] E.H. Mamdani, "Application of fuzzy logic to approximate reasoning using linguistic synthesis," *IEEE Trans. Computer*, vol. C-26 (12), pp. 1182- 1191, 1977.

[28] M.F. Azeem, M. Hanmandlu, and N. Ahmad, "Generalization of adaptive neural-fuzzy inference systems," *IEEE Trans. Neural Networks*, vol. 11, no. 6, pp. 1332- 1346, 2000.

[29] J.-S. R. Jang, "ANFIS: Adaptive-network-based fuzzy inference systems," *IEEE Trans. on Systems, Man, and Cybernetics*, vol.23, no.3, pp. 665-685, May 1993.

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

[31] J. Yen, L. Wang, and C. W. Gillespie, "Improving the interpretability of TSK fuzzy models by combining global learning and local learning," *IEEE Trans. Fuzzy Systems*, vol. 6, no. 4, pp. 530-537, Aug. 1998.

[32] I.W. Tsang, J.T. Kwok, and J.M. Zurada, "Generalized core vector machines," *IEEE Trans. Neural Networks*, vol.17, no.5, pp.1126- 1140, 2006.

[33] R.E. Fan, P.H. Chen, C.J. Lin, "Working Set Selection Using Second Order Information for Training Support Vector Machines," *Journal of Machine Learning Research*, vol. 6, pp. 1889-1918, 2005.

[34] C.F. Juang, S.H. Chiu, and S.J. Shiu, "Fuzzy system learned through fuzzy clustering and support vector machine for human skin color segmentation," *IEEE Trans. Systems Man and Cybernetics*, vol. 37, no.6, pp.1077-1087, 2007.

[35] C.F. Juang and C.D. Hsieh, "TS-fuzzy system-based support vector regression," *Fuzzy Sets and Systems*, vol.160, no.17, pp.2486-2504, 2009.

[36] Z.H. Deng, K.S. Choi, F.L. Chung, S.T. Wang, "Enhanced soft subspace clustering integrating within-cluster and between-cluster information," *Pattern Recognition*, vol. 43, no. 3, pp. 767-781, 2010.

[37] 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.

[38] L.X. Duan, I.W. Tsang, D. Xu, "Domain Transfer Multiple Kernel Learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, *vol. 34*, no.3, pp. 465-479, 2012.

[39] 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.

[40] 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.

[41] H. C. Huang, Y. H. Chen, and G. Y. Lin, "Fuzzy-based Bacterial Foraging for Watermarking Applications," *Proc. Ninth International Conference on Hybrid Intelligent Systems*, pp.214-217, Shenyang, China, 2009.

[42] K. M. Singh, "Fuzzy Rule Based Median Filter for Gray-scale Images," *Journal of Information Hiding and Multimedia Signal Processing*, vol. 2, no. 2, pp. 108-122, April 2011.

[43] H. M. Feng, J. H. Horng, and S. M. Jou, "Bacterial Foraging Particle Swarm Optimization Algorithm Based Fuzzy-VQ Compression Systems," *Journal of Information Hiding and Multimedia Signal Processing*, vol. 3, no. 3, pp. 227-239, July 2012.

[44] X. Cao, Z. Wang, P. Yan, and X. Li, "Transfer learning for pedestrian detection," Neurocomputing, vol. 100, no. 1, pp. 51–57, 2013.

[45] X. Gao, X. Wang, X. Li, and D. Tao, "Transfer latent variable model based on divergence analysis," *Pattern Recognition*, vol. 44, no. 10, pp. 2358–2366, 2011.

[46] J. Liu, M. Shah, B. Kuipers, and S. Savarese, "Cross-view action recognition via view knowledge transfer," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 3209–3216.

[47] H. Wang, F. Nie, H. Huang, and C. Ding, "Dyadic transfer learning for cross-domain image classification," in *Computer Vision (ICCV), 2011 IEEE International Conference on. IEEE*, 2011, pp. 551–556.

[48] I.H. Jhuo, D. Liu, DT Lee, and S.F. Chang, "Robust visual domain adaptation with low-rank reconstruction," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 2168–2175.

[49] L. Duan, D. Xu, and S.F. Chang, "Exploiting web images for event recognition in consumer videos: A multiple source domain adaptation approach," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. IEEE*, 2012, pp. 1338–1345.

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

[51] R. Gopalan, R. Li, and R. Chellappa, "Domain adaptation for object recognition: An unsupervised approach," in *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011, pp. 999–1006.