https://doi.org/10.1016/j.ins.2016.03.050
This is the Pre-Published Version.

A Novel Multi-task TSK Fuzzy Classifier and Its Enhanced Version for Labeling-Risk-Aware Multi-task Classification

Yizhang Jiang[1,2], Zhaohong Deng [1,3*], Kup-Sze Choi[3], Fu-Lai Chung[2], Shitong Wang[1,2]
[1] School of Digital Media, Jiangnan University, Wuxi, Jiangsu, P.R. China
[2] Department of Computing, The Hong Kong Polytechnic University, Hong Kong
[3] Center of Smart Heath, The Hong Kong Polytechnic University, Hong Kong

*Corresponding author: dzh666828@aliyun.com

Abstract:

While Takagi-Sugeno-Kang (TSK) fuzzy system has been extensively applied for regression, the paper aims to unveil its potential for classification, of multiple tasks in particular. First, a novel TSK fuzzy classifier (TSK-FC) is presented for pattern classification by integrating the large margin criterion into the objective function. When multiple tasks are concerned, it has been shown that learning of the tasks simultaneously yields better performance than learning independently. In this regard, the ability of TSK-FC is exploited for multi-task learning, where a multi-task TSK fuzzy classifier called MT-TSK-FC is proposed by using a mechanism that does not only use the independent sample information of each task, but also the inter-task correlation information to enhance the classification performance. However, as the number of tasks increases, the learning process is prone to labeling risk, which can lead to considerable degradation in the performance of pattern classification. To reduce the risk, a labeling-risk-aware mechanism is proposed to enhance the performance of the MT-TSK-FC, and the labeling-risk-aware multi-task TSK fuzzy classifier called LRA-MT-TSK-FC is thus developed. Since the three proposed fuzzy classifiers – TSK-FC, MT-TSK-FC and LRA-MT-TSK-FC – can all be implemented by solving the corresponding QP problems,

© 2016. This manuscript version is made available under the CC-BY-NC-ND 4.0 license https://creativecommons.org/licenses/by-nc-nd/4.0/.

global optimal solutions are guaranteed. Experiments on multi-task synthetic and real image datasets are conducted to demonstrate comprehensively the effectiveness of the classifiers.

Keywords**:** TSK fuzzy system, Classification, Large margin, Multi-task learning, Labeling-risk, Labeling-risk-aware mechanism

1. Introduction

There are many fuzzy system based intelligent models that have been proposed for different tasks, such as clustering, regression and classification. Some classical method are summarized in Table 1. Compared with the most of the existing intelligence models, fuzzy system has shown its distinctive advantage in the interpretation [37, 39] and modeling abilities of uncertainty. It has been diversely applied to industrial process control, robot control, finance prediction, complex system control, image processing, medical diagnosis, and so on [8,17-19,33,37,39,40]. Among these fuzzy system models, the TSK fuzzy system is more popular and has been extensively studied in many tasks, including larger-scale data modeling [8], transfer learning modeling [9] and type-2 fuzzy modeling [26,27,35], due to its simplicity and effectiveness. In contrast to abundant amount of its studies on the regression, its studies on classification, especially on multi-task classification, still keeps comparatively scarce. In this work, we try to focus TSK fuzzy system on this aspect.

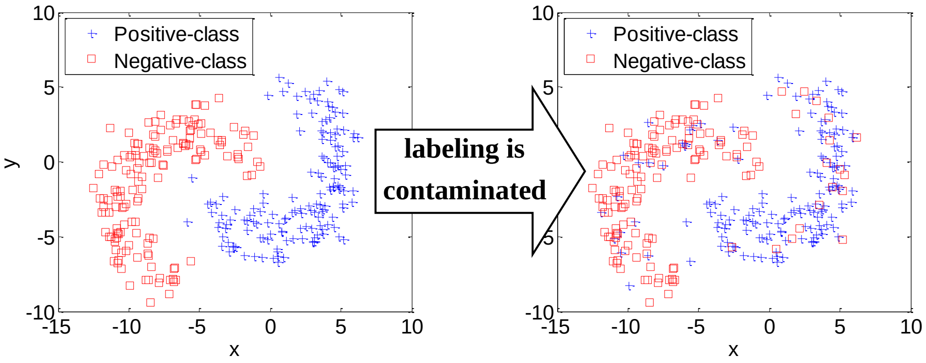
Table 1 Some classical fuzzy systems based methods in pattern recognition.

Author(s) (pub. year)	Ref. No.	Type-1	Type-2	Domain of the problem
Deng et al. (2011 and 2013)	[8,9]	√		**Regression**
Juang et al. (2007 and 2009)	[17, 18]	√		**Regression, Classification**
Leski (2005)	[23]	√		**Regression**
Mikut et al. (2008)	[25]	√		**Classification**
Qin et al. (2008)	[31]	√		**Classification**
Sanchez et al. (2014 and 2015)	[41, 43]		√	**Clustering, Classification**
Melin et al. (2014)	[42]		√	**Clustering and Classification (*Survey*)**
Castillo et al. (2014)	[44]		√	**Regression**
Deng et al. (2014)	[45]	√		**Classification and Regression**
Jiang et al. (2015)	[46]	√		**Classification**
Elkano et al. (2014)	[47]	√		**Classification**
Qun et al. (2006)	[48]		√	**Clustering**
Zheng et al. (2010)	[49]		√	**Classification**
Alcalá-Fdez et al. (2011)	[50]	√		**Classification**
Fazzolari et al. (2014)	[51]	√		**Classification**

Most of fuzzy classifiers are trained by BP-like training algorithms [4,13,22,26,27] and GA-like algorithms[15,20,29,34], which make training usually very slow on large scale data. In addition, most of existing methods train the model using the objective function of minimizing the empirical risk that usually results in the over fitting on the small data set. In this study, we will propose a novel TSK fuzzy classifier (TSK-FC) in which the large margin and structural risk minimization is used to construct its objective function. The proposed TSK-FC has the following characteristics: First, the training of TSK-FC can be equivalently transformed as a classical convex QP problem. Hence, its computational complexity is between $O(N)$ and $O(N^2)$ [12], depending on the QP solver adopted. Compared with GA-like and BP-like training methods, QP based TSK-FC training algorithm has the faster training speed. In addition, the large margin and structural risk minimization based criterion can make the TSK-FC have the better generalization performance than traditional training methods.

Like most existing fuzzy classifiers, the proposed TSK-FC is still a single-task classifier, which is not available for multi-task classification that are becoming more and more common in real-word applications [6]. For multi-task classification problems, in order to get satisfactory classification performance, we should keep in mind that we should not individually apply TSK-FC to each task, due to the fact that multitask learning or learning multiple related tasks simultaneously has better performance than learning these tasks independently [6,21,30,32,38]. Therefore, in this study, we further develop the proposed classifier TSK-FC into its multi-task version called multi-task TSK-FC (MT-TSK-FC) by using the proposed multi-task learning mechanism, which not only takes the advantage of independent sample information for each task, but also effectively uses the inter-task correlation information to enhance the classification performance.

Furthermore, the proposed MT-TSK-FC is extended for the labeling-risk scenarios since the labeling-risk scenarios are common in many applications. For example, a typical labeling-risk scene for single-task classification is shown in Fig.1. In Fig.1(a), a dataset that can be well classified by using a traditional classification algorithm such as SVM [7] or the proposed TSK-FC. However, if the dataset is mislabeled with some samples, as shown in Fig.1(b), the classification algorithms including SVM or TSK-FC cannot work well due to labeling-risk.



(a) The dataset without labeling-risk (b) The dataset with labeling-risk

Fig.1 A typical labeling-risk scene for a single-task classification dataset

In this study, to address the labeling-risk problem, we first propose a new multi-task labeling-risk-control mechanism for labeling-risk classification and then extend MT-TSK-FC into its enhanced version, i.e., labeling-risk-aware multi-task TSK fuzzy classifier (LRA-MT-TSK-FC). Based on the proposed multi-task labeling-risk-control mechanism, the LRA-MT-TSK-FC will become a more adaptive multi-task fuzzy classifier. It can well work on not only the traditional multi-task classification scene, but also the labeling-risk multi-task classification scene.

The contributions of this work can be highlighted as follows.

(1) A novel TSK fuzzy classifier (TSK-FC) based on the large margin criterion and structural risk minimization is presented. Although the proposed TSK-FC is similar to the large margin criterion based SVM from the viewpoint of objective criterion, it has distinctive characteristics compared with SVM. For example, such as TSK-FC has the higher interpretability than SVM, which is very useful for many practical applications. In addition, since the training of TSK-FC is a classical QP problem and computational complexity is between $O(N)$ and $O(N^2)$, it is very faster than many classical fuzzy system construction

algorithms, such as GA-like algorithms and BP-like algorithms.

(2) The proposed single-task TSK-FC is extended into a multi-task version, i.e., MT-TSK-FC. With respect to the multi-task learning framework, we construct a new objective function based on multi-task learning mechanism, which can effectively integrate task independence and inter-task correlation. As we know the proposed MT-TSK-FC is the first multi-task fuzzy classifier. We also show that the training of MT-TSK-FC can also be transformed as a classical QP problem, and its computational complexity keeps the same order as that of TSK-FC.

(3) Since labeling-risk problems are becoming common, to address labeling-risk multi-task classification problems, we further extend MT-TSK-FC into its labeling-risk-aware version LRA-MT-TSK-FC by introducing a new multi-task labeling-risk-control mechanism. We will prove that LRA-MT-FC's training also can still be transformed as a classical QP problem, and hence it can share the same computational complexity as MT-TSK-FC.

(4) The proposed TSK-FC, MT-TSK-FC and LRA-MT-TSK-FC have not only better generalization ability but also more interpretability than many black-box-like single task and/or multi-task classifiers, such as SVM and neural networks.

(5) Extensive experiments on synthetic and real image classification datasets demonstrate that the proposed fuzzy classifiers outperforms or is at least comparable to several existing benchmarking and state-of-the-art methods.

The rest of this paper is organized as follows. In section 2, the concept and principle of classical TSK fuzzy systems are briefly reviewed and TSK-FC is then proposed. In section 3, according to the multi-task learning framework, the multi-task TSK fuzzy classifier

MT-TSK-FC is presented. In section 4, a novel labeling-risk-aware mechanism is proposed for labeling-risk multi-task classification scenarios and then the labeling-risk-aware multi-task TSK fuzzy classifier called LRA-MT-TSK-FC is presented. The experimental results on synthetic and real image classification datasets are reported in Section 5. Finally, conclusions and the potentials of the proposed methods are given in the last section. Appendix A , Appendix B and Appendix C are provided to enhance readability.

2. Single-Task TSK Fuzzy Classifier

In this section, the classical TSK fuzzy system is briefly reviewed. Then, a TSK based fuzzy classifier (TSK-FC) is presented for classification tasks. The characteristics of the proposed classifier is also analyzed.

2.1. Concept and Principle of TSK Fuzzy Systems

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

$$\begin{aligned}&\text{TSK Fuzzy Rule } R^m:\\&\text{IF } x_1 \text{ is } A_1^m \wedge x_2 \text{ is } A_2^m \wedge \cdots \wedge x_d \text{ is } A_d^m\\&\text{Then } f^m(\mathbf{x}) = p_0^m + p_1^m x_1 + \cdots + p_d^m x_d \qquad m = 1, \cdots, M\end{aligned} \tag{1}$$

In Eq. (1), A_i^m is a fuzzy subset subscribed by the input variable x_i for the *m*-th rule; *M* 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 $A^m \subset R^d$ to a varying singleton denoted by $f^m(\mathbf{x})$. 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

$$y^0 = \sum_{m=1}^{M} \frac{\mu^m(\mathbf{x})}{\sum_{m'=1}^{M} \mu^{m'}(\mathbf{x})} \cdot f^m(\mathbf{x}) = \sum_{m=1}^{M} \tilde{\mu}^m(\mathbf{x}) \cdot f^m(\mathbf{x}), \tag{2.a}$$

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

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

$$\tilde{\mu}^m(\mathbf{x}) = \mu^m(\mathbf{x}) \Big/ \sum_{m'=1}^{M} \mu^{m'}(\mathbf{x}). \tag{2.c}$$

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

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

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

$$c_i^m = \sum_{j=1}^{N} u_{jm} x_{ji} \Big/ \sum_{j=1}^{N} u_{jm}, \tag{2.e}$$

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

where u_{jm} denotes the fuzzy membership of the j-th input data $\mathbf{x}_j = (x_{j1}, \cdots, x_{jd})^T$, belonging to the m-th cluster obtained by FCM clustering [3] or other partition methods. Here h is a scalar parameter and can be adjusted manually.

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

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

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

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

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

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

then Eq. (2.a) can be formulated as the following linear regression problem [23]

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

Thus, the training problem of the above TSK model can be transformed into the learning of the parameters in the corresponding linear regression model [8,9,23].

2.2. Classification Strategy

Given a binary training dataset $D = \{\mathbf{x}_i, y_i \mid \mathbf{x}_i \in R^d,\ y_i \in \{\,1\,, -1\,\}, i = 1, \cdots, N\}$, we obtain a trained TSK fuzzy system, whose output can be expressed as Eq. (3.f). Given a testing data point $\mathbf{x}$, its label can be determined by the following decision rule:

$$y^o = \begin{cases} 1 & f(\mathbf{x}) = \mathbf{p}_g^T \mathbf{x}_g > 0 \\ -1 & f(\mathbf{x}) = \mathbf{p}_g^T \mathbf{x}_g < 0 \end{cases}, \text{ i.e., } \ y^o = \begin{cases} \mathbf{p}_g^T \mathbf{x}_g > 0 & y_i > 0 \\ \mathbf{p}_g^T \mathbf{x}_g < 0 & y_i < 0 \end{cases}$$

2.3. Margin Maximization Based Optimization Criterion

For any data point $\{\mathbf{x}_i, y_i\}$ in the given training dataset, with the aim of classification, the margin maximization solution of the consequent parameters is to maximize the following criterion function:

$$\begin{aligned} &\max \quad \varepsilon \\ &s.t. \ \begin{cases} \mathbf{p}_g^T \mathbf{x}_{gi} > \varepsilon & y_i > 0 \\ \mathbf{p}_g^T \mathbf{x}_{gi} < -\varepsilon & y_i < 0 \end{cases} \end{aligned} \tag{4.a}$$

According to the constrained conditions, $y_i \cdot f(\mathbf{x}_i) = y_i \cdot \mathbf{p}_g{}^T \mathbf{x}_{gi} > \varepsilon$ ($i = 1, \cdots, N$) for the output of TSK fuzzy system are expected. Using the above constrained conditions, the criterion in (4.a) can be equivalently written as:

$$\begin{aligned} &\max \quad \varepsilon \\ &\text{s.t.} \quad y_i \cdot (\mathbf{p}_g^T \mathbf{x}_{gi}) > \varepsilon \end{aligned} \tag{4.b}$$

where ε denotes the margin. Since the above conditions cannot always hold for all data points $\mathbf{x}_{gi}$ ($i = 1, \cdots, N$), the following constraints can be adopted by introducing slack variables $\xi_i \geq 0$ $(i=1,\cdots,N)$

$$y_i \cdot f(\mathbf{x}_i) = y_i \cdot \mathbf{p}_g{}^T \mathbf{x}_{gi} > \varepsilon - \xi_i \tag{4.c}$$

Based on the above Eqs.(4.b) and (4.c), we further introduce the similar learning mechanism in SVM, i.e., the large margin criterion and structural risk minimization, to construct the optimization objective function for the proposed fuzzy classifier as follows.

$$\begin{aligned} &\min_{\mathbf{p}_g} \quad -\varepsilon + \tau \cdot \frac{1}{2}(\mathbf{p}_g^T \mathbf{p}_g) + \frac{1}{N}\sum_{i=1}^{N} \xi_i \\ &\text{s.t.} \quad y_i \cdot (\mathbf{p}_g^T \mathbf{x}_{gi}) > \varepsilon - \xi_i,\ \xi_i > 0,\ \ i=1,\cdots,N \end{aligned} \tag{5.a}$$

where $\frac{1}{2}\mathbf{p}_g^T \mathbf{p}_g$ is the regularization term. Eq.(5.a) indicates that the obtained TSK fuzzy classifier will maximize the margin ε and simultaneously minimize the empirical error terms ξ_i . The regularization term may effectively enhance the generalization ability of the TSK fuzzy system for classification. Furthermore, Eq.(5.a) can be reformulated as following optimization problem:

$$\begin{aligned} &\min_{\mathbf{p}_g} \quad \frac{1}{2}(\mathbf{p}_g^T \mathbf{p}_g) + \frac{1}{\tau N}\sum_{i=1}^{N} \xi_i - \frac{1}{\tau}\varepsilon \\ &\text{s.t.} \quad y_i \cdot (\mathbf{p}_g^T \mathbf{x}_{gi}) > \varepsilon - \xi_i,\ \xi_i > 0, \varepsilon > 0,\ \ i=1,\cdots,N \end{aligned} \tag{5.b}$$

According to Eq. (5.b), one may note that the proposed *L1*-norm penalty-based criterion has the following characteristics: 1) The constraints $\xi_i > 0$ are not needed for optimization in (5.b); 2) the margin ε can be obtained automatically by optimization, i.e. without the need of manual setting. Note here that although the proposed TSK-FC has adopted the similar objective criterion to SVM, there are also obvious differences between them: 1) While the obtained hyperplane by SVM is in the original feature space (by linear SVM) or in a kernelized space (by kernelized SVM), the optimal classification hyperplane for TSK-FC is obtained in a distinctive feature space that , is mapped from the original feature space by fuzzy rules. 2) TSK-FC does not involve the kernelization, which make it not need to optimize the kernel parameters, the key parameters in SVM. 3) The classification hyperplane obtained by TSK-FC can be transformed into the fuzzy rules of fuzzy system, which is thus more interpretable.

Based on optimization theory, the dual problem of Eq. (5.b) can be obtained by Lagrange optimization as

$$\begin{aligned} &\max_{\boldsymbol{\lambda}} \quad -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\lambda_i\lambda_j y_i y_j \mathbf{x}_{gi}^T\mathbf{x}_{gj} \\ &\text{s.t.} \qquad \lambda \in [0, \frac{1}{N\tau}] \qquad \sum_{i=1}^{N}\lambda_i = \frac{1}{\tau} \qquad \forall\ i \end{aligned} \tag{5.c}$$

and its matrix form can be expressed as

$$\begin{aligned} &\max_{\boldsymbol{\upsilon}} \quad -\frac{1}{2}\boldsymbol{\upsilon}^T\mathbf{K}\boldsymbol{\upsilon} \\ &\text{s.t.}\ \boldsymbol{\upsilon}^T\mathbf{1} = \frac{1}{\tau}, \quad 0 \le \lambda_i \le \frac{1}{N\tau} \quad \forall\ i \end{aligned} \tag{5.d}$$

where $\boldsymbol{\upsilon} = (\lambda_1, \lambda_2, \cdots, \lambda_N)^T$, $\mathbf{K} = [K_{ij}]_{N\times N}$, $K_{ij} = y_i y_j \mathbf{x}_{gi}{}^T\mathbf{x}_{gj}$. Then, with the dual theory and the optimal solution of $\boldsymbol{\upsilon}$, we can get the optimal $\mathbf{p}_g$ as

$$\mathbf{p}_{\mathrm{g}}^* = \sum_{i=1}^{N}\lambda_i^* y_i \mathbf{x}_{\mathrm{g}i} \tag{6}$$

Please refer to Appendix A for the derivations of Eq.(5-c) and (5-d).

Once $\mathbf{p}_{\mathrm{g}}^*$ is determined, in terms of Eqs. (3.a)-(3.f) and the above classification strategy, the corresponding TSK-FC classifier is directly built. In summary, the training of the proposed fuzzy classifier TSK-FC is still a quadratic programming (QP) optimization problem. Thus, the computational complexity of the proposed method mainly comes from learning the consequent parameters. The consequent parameters of TSK-FC can be obtained by solving the QP problem in Eq.(5.c) 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 [12]. Therefore, the computational complexity of the proposed fuzzy classifier TSK-FC is between $O(N)$ and $O(N^2)$. In this study, we adopt the working set-based QP solution [12] for solving the QP problem concerned.

2.4. Algorithm

Based on the analysis above, we summarized the proposed single-task fuzzy classifier TSK-FC as follows.

Algorithm 1: *TSK-FC*

Stage 1: Constructing the input dataset	
Step 1:	Set the number of fuzzy rules M and the regularization parameter τ .
Step 2:	Determine the antecedents of TSK fuzzy system by using clustering or other partition techniques to partition the dataset in the input space.
Step 3:	Construct the new dataset $\tilde{D}=\left\{\mathbf{x}_{gi}, y_i\right\}$ by using Eqs.(3.a)-3(.c).
Stage 2: Optimizing the objective function of TSK-FC	
Step 4:	Use QP optimizer to solve the objective function in Eq.(5.c) or (5.d)
Stage 3: Obtaining the decision function of TSK-FC	
Step 5:	Obtain the parameters of TSK-FC by using Eqs.(5.d) and (3.d)-(3-e) and get the decision function (2.a) or (3.f) of TSK-FC.

3. Multi-task TSK Fuzzy Classifier

In this section, the proposed TSK-FC is extended for learning in multi-task learning pattern and a multi-task TSK-FC (MT-TSK-FC) is proposed, whose framework is shown in Fig. 2. It can be seen that each fuzzy classifier is trained in a multi-task learning manner by multi-task training datasets, which reserves the independent sample information and takes full use of the inter-task correlation. In the following sub-section, a specific multi-task TSK fuzzy classifier and its training method based on large margin criterion and L1-norm penalty term will be elaborated.

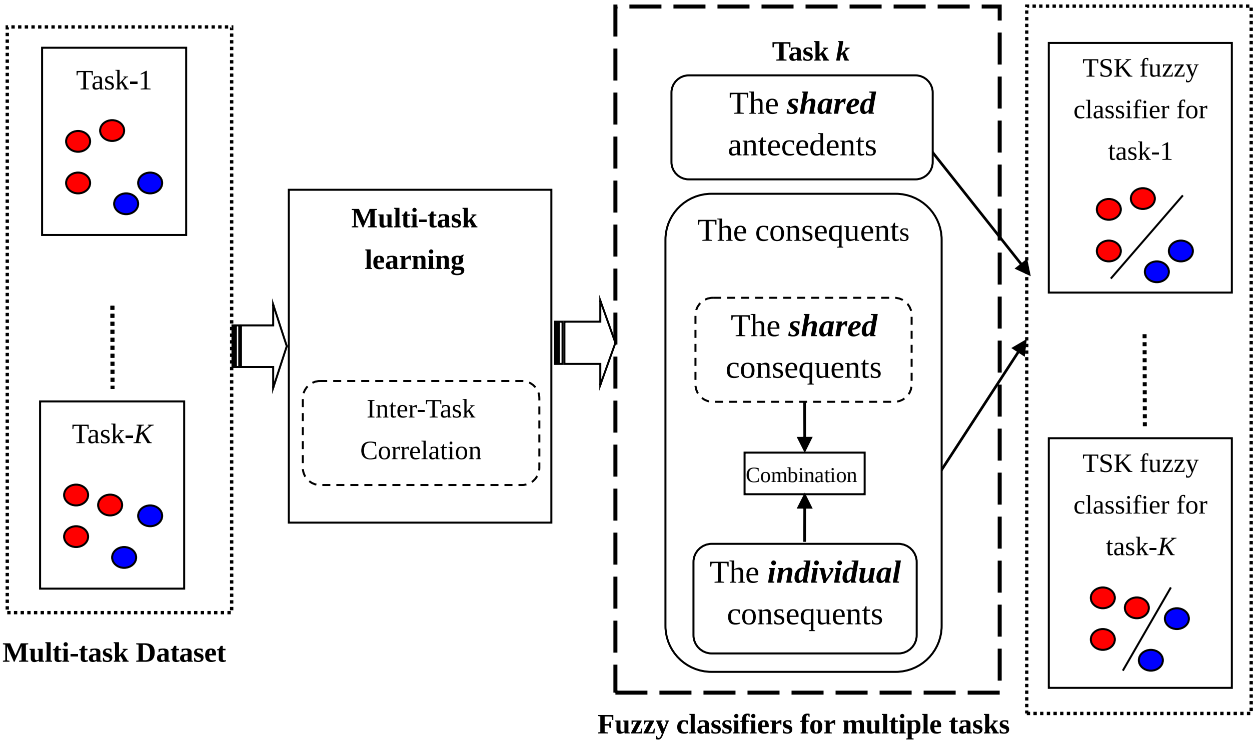


Fig.2 The framework of the proposed learning method for MT-TSK-FC

3.1. Objective Function

When we design the objective function of the multi-task TSK fuzzy system based on the classic ε-insensitive criterion and L1-norm penalty terms, we should consider how to maintain the balance between the unique characteristics of different tasks of data samples (*independence information*) and correlation information (*inter-task hidden correlation*), and how to generalize the independence and correlation information. In order to make TSK fuzzy systems empowered with multi-task learning ability, the following objective function for our proposed MT-TSK-FC which incorporates the concept of multi-task learning is proposed:

$$\begin{aligned} &\min_{\mathbf{p}_{g_0},\boldsymbol{\theta},\xi,\varepsilon} \ \Psi_{\mathrm{S}}\left(\mathbf{p}_{g_0}\right)+\sum_{k=1}^{K} g_{\mathrm{k}}\left(\boldsymbol{\theta}_k,\xi_{\mathrm{k}},\varepsilon_{\mathrm{k}}\right) \\ &\text{s.t.} \ \ y_{i,k}\cdot((\mathbf{p}_{g_0}+\boldsymbol{\theta}_k)^T \mathbf{x}_{gi,k})>\varepsilon_k-\xi_{i,k}, \quad \varepsilon_k>0, \quad \xi_{i,k}>0 \quad \forall i,k \end{aligned} \tag{7}$$

where

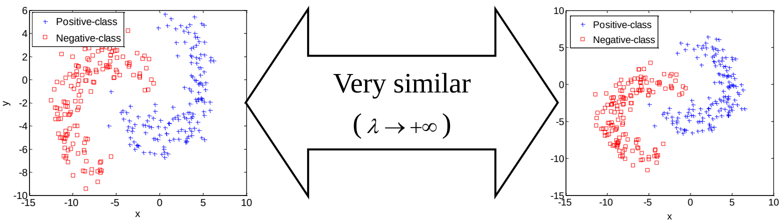
$$\Psi_{\mathrm{S}}\left(\mathbf{p}_{g_0}\right)=\frac{1}{2}\mathbf{p}_{g_0}^T\mathbf{p}_{g_0} \tag{7.a}$$

$$g_k\left(\boldsymbol{\theta}_k,\xi_k,\varepsilon_k\right)=\frac{\lambda}{K}\frac{1}{2}\boldsymbol{\theta}_k^T\boldsymbol{\theta}_k+\frac{1}{N_k\tau_k}\sum_{i=1}^{N_k}\xi_{i,k}-\frac{1}{\tau_k}\varepsilon_k \tag{7.b}$$

After observing Eq.(7), we can find that Eqs.(7.a) and (7.b) play different roles in Eq.(7), i.e., Eq.(7.a) representing the *correlation information* for different tasks and Eq.(7.b) representing the *independence information* of different tasks. Specifically, in order to represent the independence information and the correlation information, we assume the corresponding model parameter $\tilde{\mathbf{p}}_{g,k}$ for task-*k* can be written as $\tilde{\boldsymbol{p}}_{g,k}=\boldsymbol{p}_{g_0}+\boldsymbol{\theta}_k$, where the vector $\boldsymbol{\theta}_k$ tends to zero when different tasks are similar to each other, otherwise the mean vector $\boldsymbol{p}_{g_0}$ tends to zero. Namely, the vector $\boldsymbol{p}_{g_0}$ carries the *correlation information* while the vector $\boldsymbol{\theta}_k$ represents the *independence information*. Note here that the balance parameter λ is very important, it has an impact on $\boldsymbol{\theta}_k$ and control the balance between *independence information* and *correlation information*. Their values can be manually set and can also be taken by cross-validation strategy [16]. In additional, for Eq. (7), having the same advantage as the TSK-FC, its constraints $\xi_{i,k}>0$ for each task are not needed for optimization, and its margins ε_k can also be automatically obtained.

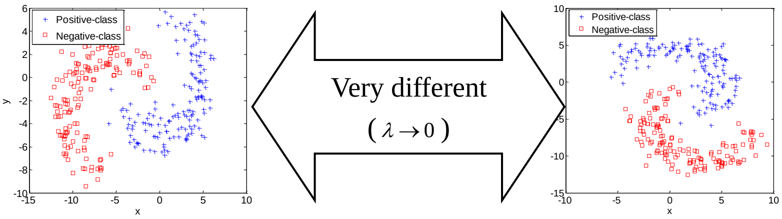
Here, we give an example as shown in Fig.3 to further show how to balance effect of the *independence information* and the *correlation information* by using the balance parameter λ. In Fig. 3, two multi-task scenes are designed, i.e., scene 1 and scene 2. In the scene 1, two tasks are very similar, which indicates that there exists strong correlation between two tasks and weak independence for each task. Namely, the correlation information $\boldsymbol{p}_{g_0}$ is more useful than independence information $\boldsymbol{\theta}_k$ in this scene. Thus, λ should trend to $+\infty$, i.e., each $\boldsymbol{\theta}_k\to 0$, and then $\boldsymbol{p}_{g_0}$ will play a main role in the final model parameter $\tilde{\mathbf{p}}_{g,k}$ ($\tilde{\boldsymbol{p}}_{g,k}=\boldsymbol{p}_{g_0}+\boldsymbol{\theta}_k$) in this scene. Instead, in the scene 2, two tasks are very different, which means there exist strong independence for each task and weak correlation between two tasks. In this scene, the independence information $\boldsymbol{\theta}_k$ should play a main role in the final model parameter $\tilde{\mathbf{p}}_{g,k}$. Thus, λ should trend to 0, i.e., each $\boldsymbol{\theta}_k \to +\infty$. Overall, according to different multi-task scenes, we can adjust the parameter λ to balance the effect of the *independence information* and the *correlation information*. For this purpose, the cross-validation strategy can be used.

***<u>Scene 1</u>*: Task 1 and Task 2 are very similar**



(a) The original two-moon dataset for task 1 (b) Rotated by 10° for task 2

***<u>Scene 2</u>*: Task 1 and Task 2 are very different**



(c) The original two-moon dataset for task 1 (d) Rotated by 90° for task 2

Fig.3 An example of multi-task scenes that there are different extent of independence information and the correlation information between two tasks.

3.2. Parameter Solution

Given the optimization problem in Eq.(7), the dual of Eq.(7) is given as follows.

$$\begin{aligned} &\max_{\boldsymbol{\lambda}_1,\ldots,\boldsymbol{\lambda}_K} L(\boldsymbol{\lambda}_1,\ldots,\boldsymbol{\lambda}_K) \\ &= -\frac{1}{2}\sum_{k=1}^{K}\sum_{l=1}^{K}\sum_{i=1}^{N_k}\sum_{j=1}^{N_l}\lambda_{j,l}\lambda_{i,k}y_{j,l}y_{i,k}\mathbf{x}_{gj,l}^T\mathbf{x}_{gi,k} - \frac{K}{2\lambda}\sum_{k=1}^{K}\sum_{i=1}^{N_k}\sum_{j=1}^{N_k}\lambda_{i,k}\lambda_{j,k}y_{i,k}y_{j,k}\mathbf{x}_{gj,k}^T\mathbf{x}_{gi,k} \\ &\text{s.t.} \quad \lambda_{i,k}\in[0,\frac{1}{N_k\tau_k}] \quad \sum_{i=1}^{N_k}\lambda_{i,k}\ge\frac{1}{\tau_k} \quad \forall k \quad k=1\ldots K \end{aligned} \tag{8}$$

where the constraint $\sum_{i=1}^{N_k}\lambda_{i,k}\ge\frac{1}{\tau_k}$ can be equivalently expressed as $\sum_{i=1}^{N_k}\lambda_{i,k}=\frac{1}{\tau_k}$. In Eq.(8), $\boldsymbol{\lambda}_1,\ldots,\boldsymbol{\lambda}_K$ are the Largangian multiplier vectors, i.e., the solution variables of the dual problem of Eq.(7). The derivation of Eq.(8) can be seen in the Appendix B.

According to the KKT optimal theory, the optimal consequent parameters of the trained MT-TSK-FC for each task, i.e., $\tilde{\boldsymbol{p}}_{g,k}^*$ can be finally given by

$$\boldsymbol{p}_{g_0}^* = \sum_{k=1}^{K}\sum_{i=1}^{N_k}\lambda_{i,k}^* y_{i,k}\boldsymbol{x}_{gi,k} \tag{9.a}$$

$$\boldsymbol{\theta}_k^* = \frac{K}{\lambda}\sum_{i=1}^{N_k}\lambda_{i,k}^* y_{i,k}\boldsymbol{x}_{gi,k} \tag{9.b}$$

$$\tilde{\boldsymbol{p}}_{g,k}^* = \boldsymbol{p}_{g_0}^* + \boldsymbol{\theta}_k^* = \sum_{k=1}^{K}\sum_{i=1}^{N_k}\lambda_{i,k}^* y_{i,k}\boldsymbol{x}_{gi,k} + \frac{K}{\lambda}\sum_{i=1}^{N_k}\lambda_{i,k}^* y_{i,k}\boldsymbol{x}_{gi,k} \tag{9.c}$$

where $\lambda_{i,k}^*$ are the optimal solutions of the dual problem for task k in Eq.(8). The derivation of Eqs.(9.a)-(9.b) can also be seen in the Appendix B.

For Eq.(8), we can give a more compact form as follows. Eq. (8) can be formulated as

$$\begin{aligned} &\arg\max_{\boldsymbol{\upsilon}} \ -\frac{1}{2}\boldsymbol{\upsilon}^T\boldsymbol{K}\boldsymbol{\upsilon} \\ &\text{s.t.}\begin{cases} \boldsymbol{\upsilon}_k^T\mathbf{1} = \dfrac{1}{\tau_k} \\ \boldsymbol{\upsilon}_{i,k}\in[0,\dfrac{1}{N_k\tau_k}] \end{cases} \quad \forall i,k \end{aligned} \tag{10}$$

where

$$\begin{aligned} \boldsymbol{\upsilon} &= (\underbrace{\tilde{\lambda}_{1,1},\ldots,\tilde{\lambda}_{N_1,1}}_{N_1},\underbrace{\tilde{\lambda}_{1,2},\ldots,\tilde{\lambda}_{N_2,2}}_{N_2},\ldots,\underbrace{\tilde{\lambda}_{1,K},\ldots,\tilde{\lambda}_{N_K,K}}_{N_K})^T \\ &= \left((\boldsymbol{\lambda}_1)^T,(\boldsymbol{\lambda}_2)^T,\ldots,(\boldsymbol{\lambda}_K)^T\right)^T \end{aligned} \tag{11.a}$$

$$\tilde{\boldsymbol{K}}_k = [\tilde{k}_{ij}]_{N_k\times N_k}, \tilde{k}_{ij} = \frac{K}{\lambda}y_{i,k}y_{j,k}\boldsymbol{x}_{gj,k}^{\mathrm{T}}\boldsymbol{x}_{gi,k} \tag{11.b}$$

$$\hat{\boldsymbol{K}}_{k,l} = [\tilde{k}_{ij}]_{N_l \times N_k}, \tilde{k}_{ij} = y_{i,l} y_{j,k} \mathbf{x}_{gj,l}^{\mathrm{T}} \mathbf{x}_{gi,k} \tag{11.c}$$

$$\boldsymbol{K} = \begin{pmatrix} \tilde{\boldsymbol{K}}_1 + \hat{\boldsymbol{K}}_{1,1} & \hat{\boldsymbol{K}}_{2,1} & \cdots & \hat{\boldsymbol{K}}_{K,1} \\ \hat{\boldsymbol{K}}_{1,2} & \tilde{\boldsymbol{K}}_2 + \hat{\boldsymbol{K}}_{2,2} & \cdots & \hat{\boldsymbol{K}}_{K,2} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\boldsymbol{K}}_{1,K} & \hat{\boldsymbol{K}}_{2,K} & \cdots & \tilde{\boldsymbol{K}}_K + \hat{\boldsymbol{K}}_{K,K} \end{pmatrix} \tag{11.d}$$

According to Eqs. (8) or (10), it is still a QP problem. We can find the optimal model parameters $\tilde{\boldsymbol{p}}_{g,k}^{*}$ to construct the corresponding MT-TSK-FC for each task. Consequently, a novel decision function can be expressed as follows.

$$y = sign(f(\mathbf{x}_{g,k})) = \left(\tilde{\mathbf{p}}_{g,k}^{*}\right)^{*} \mathbf{x}_{g,k} = (\mathbf{p}_{g_0}^{*} + \boldsymbol{\theta}_k^{*})\mathbf{x}_{g,k} = \begin{cases} 1 & \text{if } f(\mathbf{x}_{g,k}) > 0 \\ -1 & \text{otherwise} \end{cases} \tag{12}$$

3.3. Algorithm

Based on the derivations above, we can summarize the proposed MT-TSK-FC as follows.

Algorithm 2: *MT-TSK-FC: The proposed multi-task fuzzy classifier.*

Stage 1: Constructing multi-task input dataset	
Step 1:	Set the numbers of fuzzy rules M_k
Step 2:	Determine the antecedents of TSK fuzzy system by using clustering or other partition techniques to partition the multi-task dataset in different input spaces.
Step 3:	Construct the new multi-task dataset $\tilde{D}_k = \left\{\mathbf{x}_{gi,k}, y_{i,k}\right\}$ by using Eqs.(3.a)-(3.c).
Stage 2: Optimizing the objective function of MT-TSK-FC	
Step 4:	Set the regularization parameter τ_k and the balance parameter λ.
Step 5:	Use a QP solver to optimize the objective function in Eq.(8) or (10)
Stage 3: Obtaining the decision function of MT-TSK-FC for each task	
Step 6:	Obtain the parameters of MT-TSK-FC by using Eqs.(9.c) and (3.d)-(3-e) and get the decision function (12) of MT-TSK-FC for each task.

4. Labeling-risk-aware MT-TSK-FC

As our analysis in the introduction, we focused our attention on a typical kind of labeling-risk problem, i.e., the label is mislabelled or contaminated. The performance of the trained classifier can not obtain its ideal classification accuracy due to the labeling-risk. To address this problem, in this section, we will first propose a novel labeling-risk-aware mechanism for labeling-risk classification scenarios, and then develop MT-TSK-FC into its enhanced version LRA-MT-TSK-FC. The LRA-MT-TSK-FC classifier has better classification performance and robustness under labeling-risk multi-task classification scenarios.

4.1. Labeling-risk-aware mechanism

Labeling-risk can be explicitly modeled by assuming that the labels in the multi-task training dataset, we have the training dataset $\tilde{D}_k = \left\{ \mathbf{x}_{gi,k}, y_{i,k} \right\}$ for task *k*, where $y_{i,k}$ can be mislabelled or contaminated, i.e., the value of $y_{i,k}$ is changed from +1 to -1 or -1 to +1. Focused on this scene, we introduce a set of random variables $\varsigma_{i,k} \in \{0,1\}, i = 1, \ldots, N_k$ for task *k*, which represent whether the corresponding label $y_{i,k}$ is changed or not, if the value is changed $\varsigma_{i,k} = 1$, if not, $\varsigma_{i,k} = 0$. Accordingly, a novel labeling-risk-control mechanism is proposed as follows.

$$\tilde{y}_{i,k} = y_{i,k}(1 - 2\varsigma_{i,k}) = \begin{cases} -y_{i,k} & \varsigma_{i,k} = 1,\ y_{i,k} \text{ with labeling-risk} \\ y_{i,k} & \varsigma_{i,k} = 0,\ y_{i,k} \text{ without labeling-risk} \end{cases} \tag{13}$$

For Eq.(13), if $\tilde{y}_{i,k}$ with labeling-risk, i.e., $\varsigma_{i,k} = 1$,then $\tilde{y}_{i,k} = -y_{i,k}$, while $\tilde{y}_{i,k} = y_{i,k}$ otherwise.

4.2. LRA-MT-TSK-FC

Observe the dual problem of MT-TSK-FC, i.e., Eq.(8) or Eq.(10), the class labels solely affect two parts, i.e., Eq.(11.b) and Eq.(11.c) under a multi-task scene. In particular, taking labeling-risk into account, we can rewrite the above equations, based on the labeling-risk-aware mechanism, into the following equations.

$$\tilde{\boldsymbol{K}}_k = [\tilde{k}_{ij}]_{N_k \times N_k}, \tilde{k}_{ij} = y_{i,k}(1-2\varsigma_{i,k})y_{j,k}(1-2\varsigma_{j,k})\frac{K}{\lambda}\boldsymbol{x}_{\mathrm{g}j,k}^{\mathrm{T}}\boldsymbol{x}_{\mathrm{g}i,k} \tag{14.a}$$

$$\hat{\boldsymbol{K}}_{k,l} = [\tilde{k}_{ij}]_{N_l \times N_k}, \tilde{k}_{ij} = y_{i,l}(1-2\varsigma_{i,l})y_{j,k}(1-2\varsigma_{j,k})\boldsymbol{x}_{\mathrm{g}j,l}^{\mathrm{T}}\boldsymbol{x}_{\mathrm{g}i,k} \tag{14.b}$$

Note that, in the absence of labeling-risk $\varsigma_{i,k} = 0$, $i = 1,\ldots,N_k$ for each task, Eq.(14.a) and Eq.(14.b) are equivalent to Eq.(11.b) and Eq.(11.c), respectively, i.e., the proposed classifier MT-TSK-FC, while for the label with labeling-risk, i.e., $\varsigma_{i,k} = 1$, the proposed classifier MT-TSK-FC will become a novel labeling-risk-aware MT-TSK-FC (LRA-MT-TSK-FC).

If we assume that every label is independently changed with the same probability for each task, then for the task k, $\varsigma_{i,k}$ is independent and identically distributed. Boolean random variables, whose mean $\mu_k (0 \le \mu_k \le 1)$ is simply the probability of $\varsigma_{i,k} = 1$. Within this assumption, we can compute the expected value of Eq.(14.a) and Eq.(14.b), which are given by

$$E(\tilde{\boldsymbol{K}}_k) = E([\tilde{k}_{ij}]_{N_k \times N_k}),\ E(\tilde{k}_{ij}) = \begin{cases} y_{i,k}y_{j,k}\dfrac{K}{\lambda}\boldsymbol{x}_{\mathrm{g}j,k}^{\mathrm{T}}\boldsymbol{x}_{\mathrm{g}i,k} & i = j \\ y_{i,k}y_{j,k}\dfrac{K}{\lambda}\boldsymbol{x}_{\mathrm{g}j,k}^{\mathrm{T}}\boldsymbol{x}_{\mathrm{g}i,k}\left(1-4\mu_k(1-\mu_k)\right) & i \ne j \end{cases} \tag{15.a}$$

$$E(\hat{\boldsymbol{K}}_{k,l}) = E([\tilde{k}_{ij}]_{N_l \times N_k})\ ,E(\tilde{k}_{ij}) = \begin{cases} y_{i,k}y_{j,l}\boldsymbol{x}_{\mathrm{g}j,k}^{\mathrm{T}}\boldsymbol{x}_{\mathrm{g}i,l} & k = l, i = j \\ y_{i,k}y_{j,k}\boldsymbol{x}_{\mathrm{g}j,k}^{\mathrm{T}}\boldsymbol{x}_{\mathrm{g}i,k}\left(1-4\mu_k(1-\mu_k)\right) & k = l, i \ne j \\ y_{i,l}y_{j,k}\boldsymbol{x}_{\mathrm{g}j,l}^{\mathrm{T}}\boldsymbol{x}_{\mathrm{g}i,k}\left(1-2(\mu_l+\mu_k)+4\mu_l\mu_k\right) & k \ne l, \forall i, j \end{cases} \tag{15.b}$$

The derivation of Eqs.(15.a)-(15.b) can be seen in the Appendix C.

Now, we can use the expected value of Eq.(14.a) and Eq.(14.b), i.e., Eq.(15.a) and Eq.(15.b) to reconstruct the training algorithm of MT-TSK-FC, and a novel training algorithm is accordingly proposed for LRA-MT-TSK-FC as follows.

$$\begin{aligned} &\arg\max_{\boldsymbol{\upsilon}}\ -\frac{1}{2}\boldsymbol{\upsilon}^T\boldsymbol{K}\boldsymbol{\upsilon} \\ &\text{s.t.}\begin{cases} \boldsymbol{\upsilon}_k^T\mathbf{1} = \dfrac{1}{\tau_{\mathrm{k}}} \\ \boldsymbol{\upsilon}_{i,k} \in [0, \dfrac{1}{N_k\tau_k}] \end{cases} \quad \forall i, k \end{aligned} \tag{16}$$

where

$$\boldsymbol{\upsilon}=(\underbrace{\tilde{\lambda}_{1,1},...,\tilde{\lambda}_{N_1,1}}_{N_1},\underbrace{\tilde{\lambda}_{1,2},...,\tilde{\lambda}_{N_2,2}}_{N_2},\ldots,\underbrace{\tilde{\lambda}_{1,K},...,\tilde{\lambda}_{N_K,K}}_{N_K})^T=\left(\left(\boldsymbol{\lambda}_1\right)^T,\left(\boldsymbol{\lambda}_2\right)^T,\ldots,\left(\boldsymbol{\lambda}_K\right)^T\right)^T \tag{17.a}$$

$$\tilde{\boldsymbol{K}}_k=[\tilde{k}_{ij}]_{N_k\times N_k},\tilde{k}_{ij}=\begin{cases} y_{i,k}y_{j,k}\dfrac{K}{\lambda}\boldsymbol{x}_{gj,k}^{\mathrm{T}}\boldsymbol{x}_{gi,k} & i=j\\ y_{i,k}y_{j,k}\dfrac{K}{\lambda}\boldsymbol{x}_{gj,k}^{\mathrm{T}}\boldsymbol{x}_{gi,k}\left(1-4\mu_k(1-\mu_k)\right) & i\neq j\end{cases} \tag{17.b}$$

$$\hat{\boldsymbol{K}}_{k,l}=[\tilde{k}_{ij}]_{N_l\times N_k},\tilde{k}_{ij}=\begin{cases} y_{i,k}y_{j,l}\boldsymbol{x}_{gj,k}^{\mathrm{T}}\boldsymbol{x}_{gi,l} & k=l,i=j\\ y_{i,k}y_{j,k}\boldsymbol{x}_{gj,k}^{\mathrm{T}}\boldsymbol{x}_{gi,k}\left(1-4\mu_k(1-\mu_k)\right) & k=l,i\neq j\\ y_{i,l}y_{j,k}\boldsymbol{x}_{gj,l}^{\mathrm{T}}\boldsymbol{x}_{gi,k}\left(1-2(\mu_l+\mu_k)+4\mu_l\mu_k\right) & k\neq l,\forall i,j\end{cases} \tag{17.c}$$

$$\boldsymbol{K}=\begin{pmatrix} \tilde{\boldsymbol{K}}_1+\hat{\boldsymbol{K}}_{1,1} & \hat{\boldsymbol{K}}_{2,1} & \cdots & \hat{\boldsymbol{K}}_{K,1}\\ \hat{\boldsymbol{K}}_{1,2} & \tilde{\boldsymbol{K}}_2+\hat{\boldsymbol{K}}_{2,2} & \cdots & \hat{\boldsymbol{K}}_{K,2}\\ \vdots & \vdots & \ddots & \vdots\\ \hat{\boldsymbol{K}}_{1,K} & \hat{\boldsymbol{K}}_{2,K} & \cdots & \tilde{\boldsymbol{K}}_K+\hat{\boldsymbol{K}}_{K,K}\end{pmatrix} \tag{17.d}$$

According to Eq. (16), we can find the optimal model parameters $\tilde{\boldsymbol{p}}_{g,k}^{*}$ to construct the corresponding LRA-MT-TSK-FC for each task.

4.3. Algorithm

The learning algorithm of the proposed classifier LRA-TSK-FC is described in detail below.

Algorithm 3: *LRA-MT-TSK-FC: The proposed labeling-risk-aware multi-task fuzzy classifier.*

Stage 1: Constructing multi-task input dataset	
Step 1:	Set the numbers of fuzzy rules M_k .
Step 2:	Determine the antecedents of TSK fuzzy system by using clustering or other partition techniques to partition the multi-task dataset in different input spaces.
Step 3:	Construct the new multi-task dataset $\tilde{D}_k=\left\{\mathbf{x}_{gi,k},y_{i,k}\right\}$ by using Eqs.(3.a)-(3.c).
Stage 2: Optimizing the objective function of LRA-MT-TSK-FC	
Step 4:	Set the regularization parameter τ_k , the balance parameter λ and the mean of labeling-risk μ_k .
Step 5:	Use a QP solver to optimize the objective function in Eq.(16)
Stage 3: Obtaining the decision function of LRA-MT-TSK-FC for each task	
Step 6:	Obtain the parameters of MT-TSK-FC by using Eqs.(9.c) and (3.d)-(3-e) and get the decision function Eq.(12) of LRA-MT-TSK-FC for each task.

***Remark*:** Compared with MT-TSK-FC proposed in the section III, the above training algorithm should reasonably improve the robustness of the trained MT-TSK-FC under a labeling-risk scene. The proposed method only yields a kernel matrix of the dual problem correction, and does not modify the multi-task TSK fuzzy classifier. Please note, it is an heuristic method and it is thus not guaranteed to fulfill any optimality criterion.

5. Experimental Results

5.1. Setup

In order to validate and assess the classification performance of the proposed classifiers TSK-FC, MT-TSK-FC and LRA-MT-TSK-FC, we conduct experiments on a synthetic multi-task dataset [36] and an application of image classification with labeling-risk [28] and report the obtained results in this section. A detailed description of these datasets are given in subsections 5.2 and 5.3. In all experiments, two-third of samples are taken as the training set, and the remaining one-third of samples are used for testing. In this study, we focus our main attentions on the labeling-risk problem. In order to simulate the situation of labeling-risk scenes on multi-task learning, we design two scenes, i.e., single-task risk (single-task labeling-risk scene) which means there just exists labeling-risk on one task for multi-task dataset, and the other is multi-task risk (multi-task labeling-risk scene) which means there exists labeling-risk on all tasks for multi-task dataset. Each scene is constructed with four labeling-risk situations, i.e., 5%-labeling-risk, 10%-labeling-risk, 20%-labeling-risk and 30%-labeling-risk. It should be noted that 5%-labeling-risk represents there exists 5% error labels on training set.

In our experiments, we compare the proposed three fuzzy classifiers with three classical single-task classifiers, i.e., SVM [7], Naïve Bayesian [14] and KNN [10], and one multi-task classifier, i.e., multi-task learning algorithm MT-SVM [11]. For the seven classifiers involved, besides reporting their performances on multi-task scenes, we will focus the robustness of the above seven classifiers under the multi-task labeling-risk scene. For each task, the labeling-risk means that parameter μ_k in four different labeling-risk situations will be fixed on the following three values for the proposed classifier LRA-MT-TSK-FC, i.e., $\mu_k = 0.1$, $\mu_k = 0.3$ and $\mu_k = 0.5$. Under these three different μ_k, the robustness of the above methods will be further observed and discussed. In addition, we will also evaluate the experimental results reasonably by using two traditional evaluation indices, i.e., *Accuracy* and *F1-measure* (or acc and F1 for simplicity, respectively) [24]. In our experiments, the hyperparameters are determined on a training set by five-fold cross-validation strategy within the given grids of the parameter values. All classifiers are implemented using MATLAB on a computer with Intel Core 2 Duo P8600 2.4 GHz CPU and 2GB RAM. For clarity, the detail experimental settings are summarized in Table 2.

Table 2 Settings of the Experiments

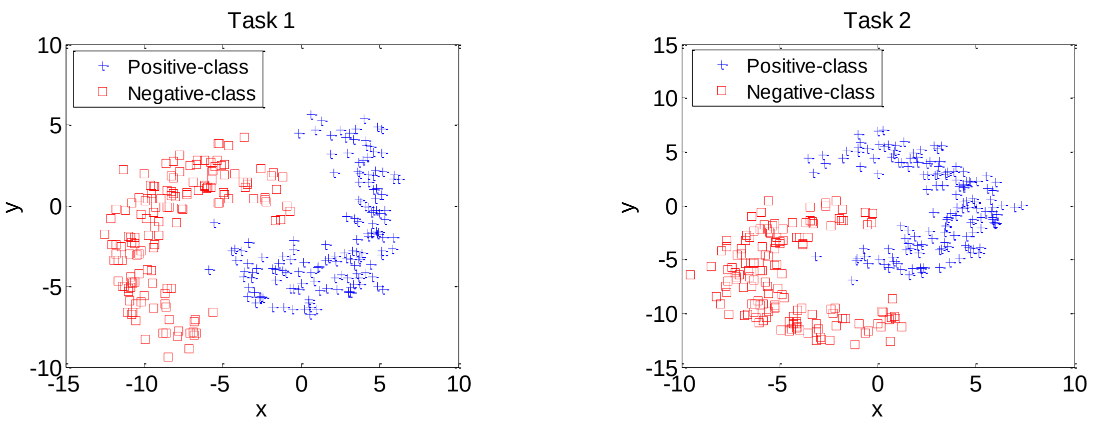
Model training methods	Single-task classification methods	Multi-task classification methods
	1. SVM [7] 2. Naive Bayes [14] 3. KNN [10] **4. TSK-FC**	1. MT-SVM [11] **2. MT-TSK-FC** **3. LRA-MT-TSK-FC**
Performance evaluation approaches	1. Five-fold cross validation strategy is adopted on training set. 2. Accuracy: The proportion of number of testing data predicted correctly to the number of the total testing data. 3. F1-measure: The harmonic mean of precisions and recalls.	
Method-specific settings	1. For KNN, the nearest points number is determined within the parameter set $K = \{1, 2, \cdots, 9, 10\}$ by five-fold cross-validation. 2. For SVM, the Gaussion kernel function $K(\mathbf{x},\mathbf{y}) = e^{-\|\mathbf{x}-\mathbf{y}\|^2/\sigma^2}$ is chosen, and the kernel parameter σ is	1. For MT-SVM, the Gaussion kernel function $K(\mathbf{x},\mathbf{y}) = e^{-\|\mathbf{x}-\mathbf{y}\|^2/\sigma^2}$ is chosen, and the kernel parameter σ is determined within the parameter set $\{2^{-6},2^{-5},\ldots,2^{5},2^{6}\}$ and The regularization parameter C^{A}, C^{B} and D are determined within

determined within the parameter set $\{2^{-6},2^{-5},\ldots,2^{5},2^{6}\}$ and the regularization parameter C is determined within the parameter set $\{2^{-6},2^{-5},\ldots,2^{5},2^{6}\}$ by five-fold cross-validation. 3. For the proposed classifier TSK-FC, the number of fuzzy rules was determined within parameter set $\{5,10,15,20,25,30,40,50,80,100\}$, and the regularization parameter τ was determined within the parameter set $\{2^{-6},2^{-5},\ldots,2^{5},2^{6}\}$ by five-fold cross-validation.	the parameter set $\{2^{-6},2^{-5},\ldots,2^{5},2^{6}\}$ by five-fold cross-validation. 2. For the proposed classifier MT-TSK-FC and LRA-MT-TSK-FC, for each task, the number of fuzzy rules was determined within parameter set $\{5,10,15,20,25,30,40,50,80,100\}$, and the regularization parameter τ_k was determined within the parameter set $\{2^{-6},2^{-5},\ldots,2^{5},2^{6}\}$ by five-fold cross-validation.

5.2. Synthetic Dataset

5.2.1 Two moon dataset

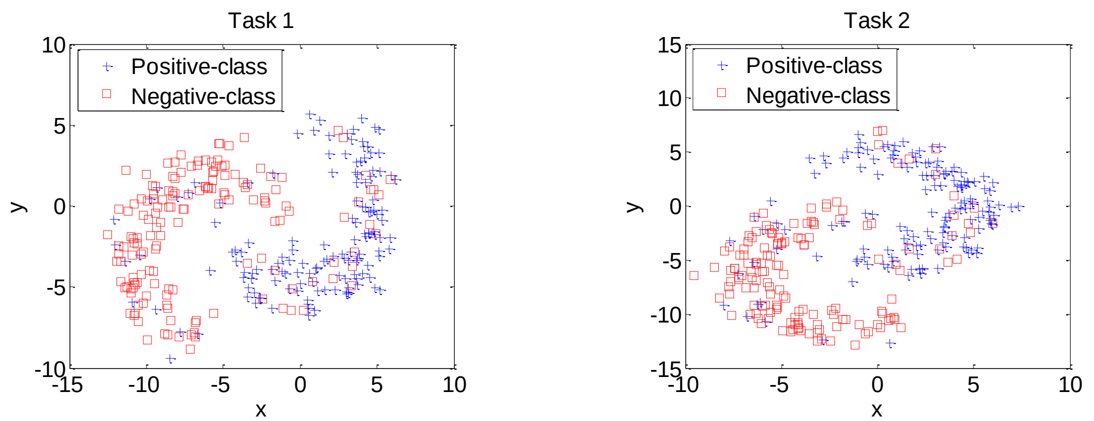
In this subsection, we construct a multi-task synthetic dataset (two-moon dataset) [36] to study the performance of the proposed classifiers, i.e., TSK-FC, MT-TSK-FC and LRA-MT-TSK-FC. The classification performances of the above three proposed classifiers and other benchmarking classifiers are compared using this multi-task synthetic dataset. We consider as first task data a synthetic data set composed of 600 samples generated according to a bi-dimensional pattern of two intertwining moons associated with two specific information classes (300 samples each), as shown in Fig.4(a). The data of another task were generated by rotating anticlockwise the data of first task by 45 degree. Due to rotation, first task and second task data exhibit different distributions, but they still have the structural features. Each task has only two classes of labeled samples (1 positive, -1 negative) as shown in Fig. 4 by “+” and “□” respectively.



(a) The original two-moon dataset for task 1 (b) Rotated by 45° for task 2

Fig.4 Synthetic multi-task dataset: (a) the original two-moon dataset; (b) rotated by 45°

In order to test the robustness of the proposed classifiers under the labeling-risk scene, two labeling-risk scenes are generated by the following situations: 1) Labeling-risk for single-task (single-task labeling-risk scene), i.e., task 1 without labeling-risk and task 2 with four different degrees of labeling-risk situations as described in section 5.1; 2) Labeling-risk for all tasks (multi-task labeling-risk scene), i.e., all the tasks with four different degrees of labeling-risk. An example for a multi-task labeling-risk scene with 30%-labeling-risk is shown in Fig.5.



(a) Task 1 with 30%-labeling-risk (b) Task 2 with 30%-labeling-risk

Fig.5 An example for a multi-task labeling-risk scene with 30%-labeling-risk

5.2.2 Comparative Analysis

According to the experimental results on the synthetic dataset shown in Table 3, Table 4 and Fig.6, one may obtain the following observations:

i) The proposed single-task fuzzy classifier TSK-FC has better or at least comparable accuracy and F1-measures than the other single-task classifiers. The results show that TSK-FC inherits the good performance and distinctive characteristics of fuzzy systems.

ii) Although most single-view classifiers achieve pretty high accuracies and F1-measure for each class, the proposed multi-task TSK fuzzy classifier MT-TSK-FC and MT-SVM classifier obtain consistently higher or at least comparable accuracies and better F1-measures, in particular on the multi-task labeling-risk scene, and the results explained the multi-task classifiers can make use of the *correlation information* among all tasks to enhance its accuracy.

iii) The proposed classifier MT-TSK-FC has comparable performance with MT-SVM in this dataset. But the proposed fuzzy classifiers possess very good interpretability originated from TSK fuzzy systems.

iv) For a labeling-risk scene, we can observe two results: 1) On the single-task labeling-risk scene, both single-task classifier and traditional multi-task classifier will get an good performance on one task (the task without labeling-risk). As shown in Table 3, the single-task classifier gets a better performance on task 1 (Task 1 without any labeling-risk on this scene), and the multi-task classifiers MT-TSK-FC and MT-SVM get better performance on task 2. But the proposed classifier LRA-MT-TSK-FC gets comparable performance on task 1 and a best performance on task 2. 2) On a multi-task labeling-risk scene, similar results can be observed. Pease note, under this scene, the multi-task classifiers MT-TSK-FC and MT-SVM get better performance than other single-task classifiers, it actually indicates the multi-task learning mechanism has the robustness of a labeling-risk scene to a certain extent, but the developed performance is still not very obvious and with the development of the labeling-risk rate the performances of MT-TSK-FC and MT-SVM were getting worse. But the proposed classifier LRA-MT-TSK-FC gets the best performance than other classifiers among all tasks due to labeling-risk-aware mechanism.

In summary, the experimental results illustrate that the proposed single-task TSK-FC , multi-task MT-TSK-FC and multi-task labeling-risk-aware LRA-MT-TSK-FC have distinctive performance in this synthetic dataset when compared with the corresponding counterparts under a multi-task labeling-risk scene.

Table 3. Performances of TSK-FC, MT-TSK-FC, LRA-MT-TSK-FC and the benchmarking classifiers on synthetic dataset under the single-task labeling-risk scene with 30%-labeling-risk

	Classifiers		**Synthetic datasets**					
			Task 1			Task 2		
			Acc	Positive F1	Negative F1	Acc	Positive F1	Negative F1
Single-task	SVM	*Mean*	1	1	1	0.6222	0.5750	0.6600
		Std.	0	0	0	1.17e-16	1.17e-16	1.17e-16
	Naive Bayes	*Mean*	0.9333	0.9302	0.9362	0.6333	0.6348	0.6318
		Std.	8.97e-16	1.12e-15	7.85e-16	0	0	0
	KNN	*Mean*	1	1	1	0.6849	0.6549	0.7096
		Std.	0	0	0	0.0362	0.0360	0.0376
	TSK-FC	*Mean*	1	1	1	0.6978	0.6760	0.7166
		Std.	0	0	0	0.0165	0.0210	0.0155
Multi-task	MT-SVM	*Mean*	0.9556	0.9535	0.9535	0.6830	0.6707	0.6952
		Std.	0	0	0	0	0	0
	MT-TSK-FC	*Mean*	0.9689	0.9686	0.9692	0.7111	0.6891	0.7299
		Std.	0.0192	0.0197	0.0188	0.0208	0.0186	0.0248
	LRA-MT-TSK-FC ($\mu_1 = 0, \mu_2 = 0.1$)	*Mean*	0.9822	0.9818	0.9827	0.8822	0.8835	0.8808
		Std.	0.0230	0.0233	0.0228	0.0531	0.0548	0.0514
	LRA-MT-TSK-FC ($\mu_1 = 0, \mu_2 = 0.3$)	*Mean*	0.9933	0.9931	0.9935	0.9533	0.9524	0.9542
		Std.	0.0149	0.0154	0.0144	0.0093	0.0091	0.0095
	LRA-MT-TSK-FC ($\mu_1 = 0, \mu_2 = 0.5$)	*Mean*	0.9933	0.9930	0.9936	0.9889	0.9883	0.9894
		Std.	0.0099	0.0104	0.0095	0.0079	0.0082	0.0075

Table 4. Performance of TSK-FC, MT-TSK-FC, LRA-MT-TSK-FC and the benchmarking classifiers on a synthetic dataset under the multi-task labeling-risk scene with 30%-labeling-risk

	Classifiers		Synthetic datasets					
			Task 1			Task 2		
			Acc	Positive F1	Negative F1	Acc	Positive F1	Negative F1
Single-task	SVM	*Mean*	0.7000	0.7158	0.6824	0.6667	0.6250	0.7000
		Std.	1.17e-016	1.17e-016	1.17e-016	1.17e-016	0	0
	Naive Bayes	*Mean*	0.6778	0.6791	0.6764	0.6889	0.6718	0.7020
		Std.	2.24e-16	5.61e-16	0	0	0	0
	KNN	*Mean*	0.7118	0.7386	0.6781	0.6860	0.6277	0.7280
		Std.	0.0323	0.0300	0.0381	0.0345	0.0414	0.0316
	TSK-FC	*Mean*	0.7133	0.7298	0.6929	0.6822	0.6167	0.7285
		Std.	0.0480	0.0559	0.0423	0.0127	0.0095	0.0137
Multi-task	MT-SVM	*Mean*	0.7222	0.7573	0.6753	0.7644	0.7985	0.7165
		Std.	0	0	0	0.0050	0.0034	0.0077
	MT-TSK-FC	*Mean*	0.7378	0.7650	0.7032	0.7867	0.8126	0.7499
		Std.	0.0348	0.0304	0.0414	0.0355	0.0379	0.0382
	LRA-MT-TSK-FC	*Mean*	0.9244	0.9322	0.9145	0.8978	0.9139	0.8743
	($\mu_1=0, \mu_2=0.1$)	*Std.*	0.0277	0.0225	0.0354	0.0093	0.0076	0.0121
	LRA-MT-TSK-FC	*Mean*	0.9356	0.9394	0.9311	0.9311	0.9406	0.9179
	($\mu_1=0, \mu_2=0.3$)	*Std.*	0.0093	0.0093	0.0098	0.0145	0.0117	0.0187
	LRA-MT-TSK-FC	*Mean*	0.9733	0.9747	0.9718	0.9800	0.9821	0.9773
	($\mu_1=0, \mu_2=0.5$)	*Std.*	0.0290	0.0276	0.0305	0.0145	0.0129	0.0165

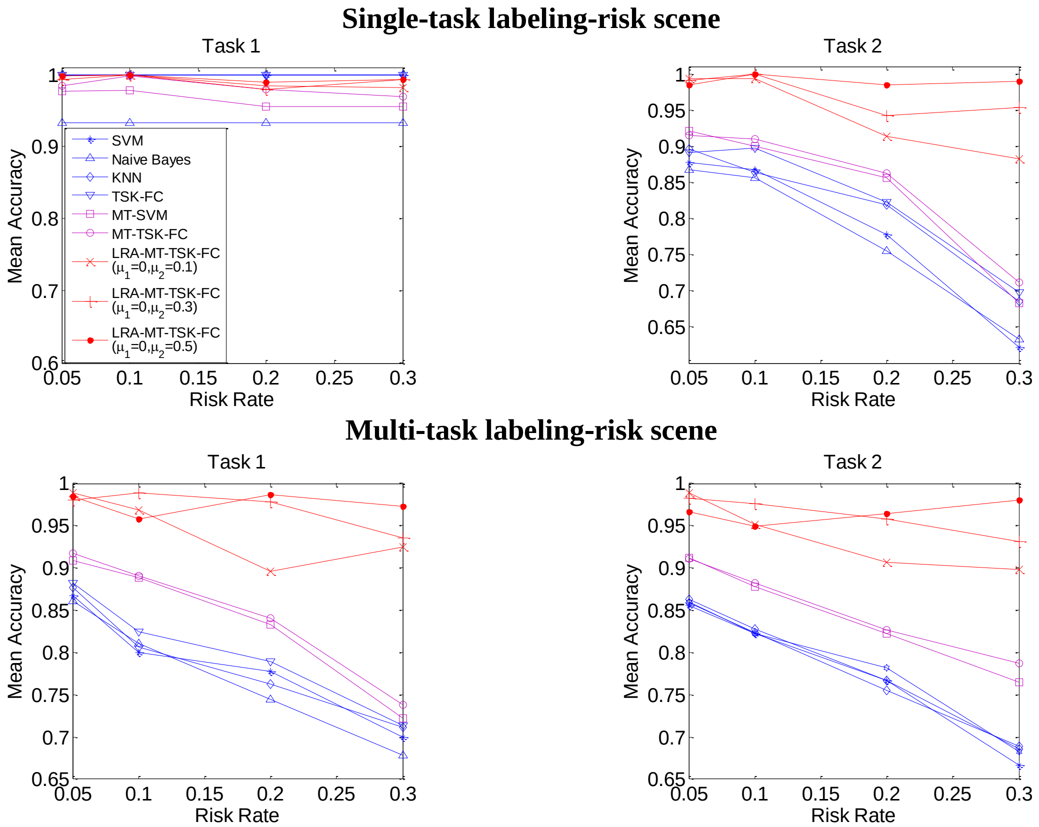


Fig.6 The mean accuracy of TSK-FC, MT-TSK-FC and the benchmarking classifiers under four different labeling-risk situations

5.3. Image datasets

5.3.1 The image dataset

We began with a dataset consisting of 600 greyscale images [28]. The images are pictures of 6 different objects such as coast, forest, inside city, tall building, highway and street. And then, we chosen three sub-datasets into two different binary classification tasks (as shown in Table 5 for example images). The resolution used is 16 × 16 pixels. We created this novel multi-task dataset such that we could have natural images (i.e. not artificially generated or composited) in multiple resolutions, with multiple images of each object. Let us observe the images of task 1 or task 2, there exists a great similarity among each class, especially in task 2. Accordingly, it is easy to labeling error for these images. From this point, we decided to use these datasets to evaluate classification performance and robustness for our classifiers.

Now, let us explain this multi-task dataset, generated from the original 200 16 × 16 image datasets from task 1 to task 2, respectively, for the proposed three classifiers and other benchmarking classifiers. For the adopted data, dimensionality reduction has been applied by using PCA [1] to effectively preprocess the high dimensional data into the final data containing 30 effective features used for multi-task classification.

Table 5 Example images for image classification tasks

Task 1	Class 1: Coast				
	Class 2: Forest				
Task 2	Class 1: Mountain				
	Class 2: Forest				

5.3.2 Comparative Analysis

The experimental results on this multi-task image dataset are reported in Table 6 and Table 7. The findings are similar to those presented in section 5.2 for the experiment performed on the synthetic dataset. As the proposed classifier MT-TSK-FC can effectively exploit not only the *independent information* of each task but also the useful *correlation information* among all tasks, it has demonstrated better accuracies and F1-measures in most cases than single-task classifiers. In addition, the classification accuracy of the proposed LRA-MT-TSK-FC shows stronger robustness than other classifiers under a labeling-risk scene, which demonstrates the effectiveness of the proposed labeling-risk-aware mechanism again.

Table 6. Performance of TSK-FC, MT-TSK-FC, LRA-MT-TSK-FC and the benchmarking classifiers on image dataset under the single-task labeling-risk scene with different labeling-risks

Risk Rate	Classifiers		**Image datasets**					
			Task 1			Task 2		
			Acc	Positive F1	Negative F1	Acc	Positive F1	Negative F1
5%	SVM	*Mean*	0.7826	0.7514	0.8069	0.6425	0.5212	0.7148
		Std.	0	0	0	0	0	0
	Naive Bayes	*Mean*	0.7923	0.7543	0.8201	0.6149	0.6070	0.6225
		Std.	0	0	1.17e-016	0	1.17e-016	1.17e-016
	KNN	*Mean*	0.7681	0.7500	0.7838	0.6404	0.6014	0.6722
		Std.	4.48e-016	6.72e-16	0	0.0199	0.0258	0.0175
	TSK-FC	*Mean*	0.7940	0.7560	0.8244	0.6531	0.6414	0.6606
		Std.	0.0066	0.0052	0.0071	0.0023	0.0033	0.0036
	MT-SVM	*Mean*	0.7633	0.7351	0.7860	0.6873	0.6529	0.7214
		Std.	0	0	0	0	0	0
	MT-TSK-FC	*Mean*	0.7702	0.7347	0.7990	0.7125	0.6863	0.7345
		Std.	0.0150	0.0260	0.0090	0.0090	0.0251	0.0113
	LRA-MT-TSK-FC	*Mean*	0.7828	0.7486	0.8103	0.7525	0.7349	0.7681
	($\mu_1 = 0, \mu_2 = 0.1$)	*Std.*	0.0076	0.0108	0.0087	0.0098	0.0084	0.0170
	LRA-MT-TSK-FC	*Mean*	0.7795	0.7415	0.8098	0.7333	0.7108	0.7529
	($\mu_1 = 0, \mu_2 = 0.3$)	*Std.*	0.0063	0.0104	0.0039	0.0090	0.0116	0.0074
	LRA-MT-TSK-FC	*Mean*	0.7705	0.7338	0.7999	0.7052	0.6976	0.7124
	($\mu_1 = 0, \mu_2 = 0.5$)	*Std.*	0.0049	0.0062	0.0044	0.0076	0.0086	0.0107
10%	SVM	*Mean*	0.7826	0.7514	0.8069	0.6380	0.4118	0.7386
		Std.	0	0	0	0	0	0
	Naive Bayes	*Mean*	0.7923	0.7543	0.8201	0.6466	0.6455	0.6477
		Std.	0	0	1.17e-016	0	1.17e-016	0
	KNN	*Mean*	0.7681	0.7500	0.7838	0.6392	0.6369	0.6411
		Std.	4.48e-016	6.72e-16	0	0.0183	0.0197	0.0203
	TSK-FC	*Mean*	0.7940	0.7560	0.8244	0.6852	0.6628	0.7031
		Std.	0.0066	0.0052	0.0071	0.0249	0.0590	0.0020
	MT-SVM	*Mean*	0.7536	0.7437	0.7628	0.7170	0.7102	0.7235
		Std.	0	0	0	0	0	0
	MT-TSK-FC	*Mean*	0.7595	0.7666	0.7490	0.7164	0.7138	0.7189
		Std.	0.0044	0.0039	0.0050	0.0081	0.0070	0.0100
	LRA-MT-TSK-FC	*Mean*	0.7633	0.7759	0.7492	0.7333	0.7424	0.7235
	($\mu_1 = 0, \mu_2 = 0.1$)	*Std.*	0.0049	0.0062	0.0038	0.0045	0.0043	0.0059
	LRA-MT-TSK-FC	*Mean*	0.7675	0.7803	0.7534	0.7359	0.7509	0.7194
	($\mu_1 = 0, \mu_2 = 0.3$)	*Std.*	0.0088	0.0079	0.0101	0.0059	0.0049	0.0074
	LRA-MT-TSK-FC	*Mean*	0.7624	0.7741	0.7495	0.7223	0.7500	0.6884
	($\mu_1 = 0, \mu_2 = 0.5$)	*Std.*	0.0107	0.0118	0.0092	0.0096	0.0048	0.0170

20%	SVM	*Mean*	0.7826	0.7514	0.8069	0.6135	0.6135	0.6135
		Std.	0	0	0	0	0	0
	Naive Bayes	*Mean*	0.7923	0.7543	0.8201	0.6199	0.6000	0.6379
		Std.	0	0	1.17e-016	1.17e-016	1.17e-016	1.17e-016
	KNN	*Mean*	0.7681	0.7500	0.7838	0.6001	0.5897	0.6097
		Std.	4.48e-016	6.72e-16	0	0.0205	0.0230	0.0206
	TSK-FC	*Mean*	0.7940	0.7560	0.8244	0.6282	0.6031	0.6502
		Std.	0.0066	0.0052	0.0071	0.0090	0.0058	0.0188
	MT-SVM	*Mean*	0.7150	0.6740	0.7468	0.6812	0.6292	0.7203
		Std.	0	0	0	0	0	0
	MT-TSK-FC	*Mean*	0.7370	0.7131	0.7576	0.6779	0.6444	0.7069
		Std.	0.0034	0.0064	0.0018	0.0076	0.0129	0.0049
	LRA-MT-TSK-FC ($\mu_1=0, \mu_2=0.1$)	*Mean*	0.7595	0.7295	0.7845	0.7008	0.6913	0.7091
		Std.	0.0093	0.0107	0.0085	0.0088	0.0083	0.0198
	LRA-MT-TSK-FC ($\mu_1=0, \mu_2=0.3$)	*Mean*	0.7889	0.7610	0.8125	0.7307	0.6811	0.7696
		Std.	0.0026	0.0020	0.0050	0.0082	0.0119	0.0059
	LRA-MT-TSK-FC ($\mu_1=0, \mu_2=0.5$)	*Mean*	0.7895	0.7658	0.8096	0.7390	0.6691	0.7897
		Std.	0.0043	0.0055	0.0027	0.0040	0.0055	0.0019
30%	SVM	*Mean*	0.7826	0.7514	0.8069	0.6244	0.3465	0.7365
		Std.	0	0	0	0	0	0
	Naive Bayes	*Mean*	0.7923	0.7543	0.8201	0.6261	0.6307	0.6214
		Std.	0	0	1.17e-016	1.17e-016	0	0
	KNN	*Mean*	0.7681	0.7500	0.7838	0.5924	0.6048	0.5788
		Std.	4.48e-016	6.72e-16	0	0.0275	0.0290	0.0289
	TSK-FC	*Mean*	0.7940	0.7560	0.8244	0.6520	0.6496	0.6526
		Std.	0.0066	0.0052	0.0071	0.0078	0.0308	0.0220
	MT-SVM	*Mean*	0.7488	0.7347	0.7615	0.6715	0.7094	0.6222
		Std.	0	0	0	0	0	0
	MT-TSK-FC	*Mean*	0.7425	0.7306	0.7500	0.6730	0.6398	0.6707
		Std.	0.0041	0.0087	0.0037	0.0005	0.0052	0.0043
	LRA-MT-TSK-FC ($\mu_1=0, \mu_2=0.1$)	*Mean*	0.7634	0.7496	0.7759	0.7043	0.7088	0.6997
		Std.	0.0055	0.0045	0.0068	0.0061	0.0048	0.0106
	LRA-MT-TSK-FC ($\mu_1=0, \mu_2=0.3$)	*Mean*	0.7837	0.7707	0.7955	0.7299	0.7332	0.7265
		Std.	0.0073	0.0072	0.0112	0.0020	0.0041	0.0015
	LRA-MT-TSK-FC ($\mu_1=0, \mu_2=0.5$)	*Mean*	0.7824	0.7677	0.7957	0.7232	0.7330	0.7134
		Std.	0.0043	0.0023	0.0070	0.0020	0.0020	0.0032

Table 7. Performance of TSK-FC, MT-TSK-FC, LRA-MT-TSK-FC and the benchmarking classifiers on the image dataset under a multiple-task labeling-risk scene with different labeling-risks

Risk Rate	Classifiers		Image datasets					
			Task 1			Task 2		
			Acc	Positive F1	Negative F1	Acc	Positive F1	Negative F1
5%	SVM	*Mean*	0.7971	0.7813	0.8108	0.7138	0.6768	0.7416
		Std.	0	0	0	0	0	0
	Naive Bayes	*Mean*	0.8068	0.7959	0.8165	0.7059	0.6948	0.7162
		Std.	0	0	0	0	1.17e-016	1.17e-016
	KNN	*Mean*	0.7874	0.7755	0.7982	0.6968	0.6599	0.7265
		Std.	3.36e-016	6.73e-016	3.36e-016	1.12e-016	4.48e-016	1.12e-016
	TSK-FC	*Mean*	0.7902	0.7880	0.7924	0.7415	0.7388	0.7441
		Std.	0.0145	0.0126	0.0173	0.0145	0.0168	0.0133
	MT-SVM	*Mean*	0.8184	0.8154	0.8213	0.7363	0.7017	0.7648
		Std.	0	0	0	0	0	0
	MT-TSK-FC	*Mean*	0.8357	0.8225	0.8477	0.7328	0.7120	0.7514
		Std.	0.0105	0.0127	0.0122	0.0160	0.0211	0.0132
	LRA-MT-TSK-FC ($\mu_1=0.1, \mu_2=0.1$)	*Mean*	0.8610	0.8485	0.8720	0.7787	0.7850	0.7722
		Std.	0.0040	0.0038	0.0047	0.0032	0.0024	0.0058
	LRA-MT-TSK-FC ($\mu_1=0.3, \mu_2=0.3$)	*Mean*	0.8610	0.8331	0.8827	0.7661	0.7593	0.7725
		Std.	0.0063	0.0052	0.0083	0.0067	0.0051	0.0093
	LRA-MT-TSK-FC	*Mean*	0.8214	0.8584	0.7677	0.7389	0.7349	0.7428
		Std.	0.0026	0.0016	0.0050	0.0081	0.0110	0.0062

	($\mu_1 = 0.5, \mu_2 = 0.5$)							
10%	SVM	*Mean*	0.7585	0.7253	0.7845	0.6833	0.6111	0.7328
		Std.	0	0	0	0	0	0
	Naive Bayes	*Mean*	0.7523	0.7198	0.7770	0.6471	0.6389	0.6549
		Std.	0	0	1.17e-016	1.17e-016	1.17e-016	1.17e-016
	KNN	*Mean*	0.7343	0.7264	0.7418	0.6968	0.6700	0.7197
		Std.	2.24e-016	1.12e-16	4.48e-16	1.12e-016	0	2.24e-016
	TSK-FC	*Mean*	0.7503	0.7242	0.7733	0.7047	0.6700	0.7337
		Std.	0.0028	0.0037	0.0031	0.0152	0.0261	0.0149
	MT-SVM	*Mean*	0.7670	0.7343	0.7944	0.7212	0.7227	0.7196
		Std.	0	0	0	0	0	0
	MT-TSK-FC	*Mean*	0.7749	0.7526	0.7948	0.7276	0.6779	0.7686
		Std.	0.0073	0.0083	0.0088	0.0109	0.0146	0.0094
	LRA-MT-TSK-FC	*Mean*	0.7979	0.7827	0.8116	0.7852	0.7696	0.7994
	($\mu_1 = 0.1, \mu_2 = 0.1$)	*Std.*	0.0055	0.0041	0.0067	0.0045	0.0048	0.0044
	LRA-MT-TSK-FC	*Mean*	0.8172	0.7913	0.8389	0.8290	0.7618	0.8782
	($\mu_1 = 0.3, \mu_2 = 0.3$)	*Std.*	0.0026	0.0021	0.0030	0.0072	0.0137	0.0054
	LRA-MT-TSK-FC	*Mean*	0.8004	0.7704	0.8244	0.8335	0.7492	0.8911
	($\mu_1 = 0.5, \mu_2 = 0.5$)	*Std.*	0.0026	0.0045	0.0017	0.0096	0.0188	0.0067
20%	SVM	*Mean*	0.7155	0.6729	0.7469	0.6504	0.6032	0.6860
		Std.	0	0	0	0	0	0
	Naive Bayes	*Mean*	0.7357	0.7152	0.7522	0.6833	0.6789	0.6875
		Std.	1.17e-016	2.34e-16	0	1.17e-016	0	0
	KNN	*Mean*	0.7391	0.7128	0.7611	0.6833	0.6635	0.7009
		Std.	3.36e-016	1.12e-016	0	4.48e-016	7.85e-16	1.12e-16
	TSK-FC	*Mean*	0.7602	0.7289	0.7865	0.6811	0.6498	0.7084
		Std.	0.0139	0.0211	0.0094	0.0136	0.0170	0.0112
	MT-SVM	*Mean*	0.7560	0.7332	0.7817	0.7094	0.7082	0.7162
		Std.	0	0	0	0	0	0
	MT-TSK-FC	*Mean*	0.7835	0.6934	0.8431	0.7348	0.7115	0.7556
		Std.	0.0158	0.0265	0.0124	0.0234	0.0205	0.0266
	LRA-MT-TSK-FC	*Mean*	0.8072	0.7853	0.8263	0.7824	0.7648	0.7982
	($\mu_1 = 0.1, \mu_2 = 0.1$)	*Std.*	0.0026	0.0039	0.0024	0.0038	0.0036	0.0044
	LRA-MT-TSK-FC	*Mean*	0.8240	0.7751	0.8604	0.8068	0.8038	0.8096
	($\mu_1 = 0.3, \mu_2 = 0.3$)	*Std.*	0.0040	0.0063	0.0026	0.0067	0.0059	0.0079
	LRA-MT-TSK-FC	*Mean*	0.8337	0.7727	0.8774	0.8141	0.8069	0.8209
	($\mu_1 = 0.5, \mu_2 = 0.5$)	*Std.*	0.0026	0.0043	0.0017	0.0052	0.0055	0.0054
30%	SVM	*Mean*	0.6812	0.6700	0.6916	0.6516	0.6131	0.6831
		Std.	0	0	0	0	0	0
	Naive Bayes	*Mean*	0.6633	0.6487	0.6763	0.6516	0.6351	0.6667
		Std.	0	1.17e-016	0	1.17e-016	1.17e-016	1.17e-016
	KNN	*Mean*	0.6957	0.6897	0.7014	0.5882	0.5381	0.6286
		Std.	1.12e-016	2.24e-016	2.24e-016	5.60e-016	4.48e-016	4.48e-016
	TSK-FC	*Mean*	0.7151	0.6785	0.7459	0.6806	0.6510	0.7058
		Std.	0.0101	0.0114	0.0096	0.0183	0.0045	0.0303
	MT-SVM	*Mean*	0.7329	0.7238	0.7415	0.7135	0.7040	0.7226
		Std.	0	0	0	0	0	0
	MT-TSK-FC	*Mean*	0.7366	0.7073	0.7611	0.7291	0.6983	0.7559
		Std.	0.0063	0.0088	0.0046	0.0038	0.0059	0.0028
	LRA-MT-TSK-FC	*Mean*	0.7734	0.7349	0.8045	0.7346	0.6878	0.7728
	($\mu_1 = 0.1, \mu_2 = 0.1$)	*Std.*	0.0073	0.0062	0.0098	0.0050	0.0066	0.0036
	LRA-MT-TSK-FC	*Mean*	0.7969	0.7421	0.8364	0.7766	0.7322	0.8100
	($\mu_1 = 0.3, \mu_2 = 0.3$)	*Std.*	0.0026	0.0017	0.0037	0.0055	0.0110	0.0025
	LRA-MT-TSK-FC	*Mean*	0.8111	0.7502	0.8539	0.8063	0.7319	0.8556
	($\mu_1 = 0.5, \mu_2 = 0.5$)	*Std.*	0.0079	0.0135	0.0050	0.0053	0.0091	0.0033

5.4. Model analysis

In this subsection, we take the model trained by LRA-MT-TSK-FC as an example to show the characteristics of the proposed fuzzy classifier. In Table 8, a multi-task LRA-MT-TSK-FC model with five rules trained in a certain time on the synthetic dataset is presented.

The constructed model by LRA-MT-TSK-FC contains two TSK fuzzy systems for different tasks as shown in Table 8. The first fuzzy system is trained for task 1. Similar to the first system, the second one is constructed for the task 2. With the fuzzy rule base obtained for two tasks, the model can be linguistically interpreted with expert knowledge.

In Fig. 7, the corresponding membership functions of all fuzzy subsets in the antecedent of the first fuzzy rule are shown for the two tasks, respectively. For each membership function, it corresponds to a fuzzy subset that can be explained by the expert knowledge.

Although the proposed fuzzy classifiers have shown the better interpretability than many existing methods, such as SVM, the interpretation is not the focus in this study. In future, we will consider how to further improve interpretability of the proposed methods.

Table 8. Rule bases obtained with five rules for each task by LRA-MT-TSK-FC on the synthetic dataset under a multi-task labeling-risk scene

Fuzzy rules base			
TSK Fuzzy Rule R^k : IF x_1 is $A_1^k(c_1^k,\delta_1^k)$ $\wedge x_2$ is $A_2^k(c_2^k,\delta_2^k)$ $\wedge\cdots\wedge x_d$ is $A_d^k(c_d^k,\delta_d^k)$, Then $f_k(\mathbf{x}) = p_{k0} + p_{k1}x_1 + \cdots + p_{kd}x_d$.			
Task	**No. of rules**	**Antecedent parameters** (Gaussian membership function parameters)	**Consequent parameters** (linear function parameters)
Task 1	k	$\mathbf{c}^k = (c_1^k,\cdots,c_d^k)^T$, $\boldsymbol{\delta}^k = (\delta_1^k,\cdots,\delta_d^k)^T$	$\mathbf{p}_k = (p_{k0}, p_{k1},\cdots, p_{kd})^T$
	1	$\mathbf{c}^1$ = [-6.9678, 1.2660], $\boldsymbol{\delta}^1$ = [4.7103, 2.9172]	$\mathbf{p}_1$ = [0.2569, -0.0440, -0.0350]
	2	$\mathbf{c}^2$ = [-10.1171, -4.5138], $\boldsymbol{\delta}^2$ = [4.8673, 3.8396]	$\mathbf{p}_2$ = [0.1865, 0.0345, -0.0157]
	3	$\mathbf{c}^3$ = [3.6223, -3.0175], $\boldsymbol{\delta}^3$ = [6.0426, 3.0230]	$\mathbf{p}_3$ = [0.2574, -0.1983, 0.0473]
	4	$\mathbf{c}^4$ = [3.5561, 2.8597], $\boldsymbol{\delta}^4$ = [6.0460, 4.0642]	$\mathbf{p}_4$ = [0.2581, 0.1545, 0.0488]
	5	$\mathbf{c}^5$ = [-1.6661, -4.4722], $\boldsymbol{\delta}^5$ = [5.5391, 3.8129]	$\mathbf{p}_5$ = [-0.0906, -0.3987, 0.0175]
Task 2	1	$\mathbf{c}^1$ = [-3.9621, -10.3456], $\boldsymbol{\delta}^1$ = [3.5171, 5.1898]	$\mathbf{p}_1$ = [0.0970, 0.2278, -0.0258]
	2	$\mathbf{c}^2$ = [1.9843, -4.3404], $\boldsymbol{\delta}^2$ = [4.7475, 4.6044]	$\mathbf{p}_2$ = [0.1120, -0.0320, -0.0808]
	3	$\mathbf{c}^3$ = [-5.8222, -4.0317], $\boldsymbol{\delta}^3$ = [3.6057, 4.0217]	$\mathbf{p}_3$ = [0.2942, 0.1804, -0.0520]
	4	$\mathbf{c}^4$ = [4.6950, 0.4278], $\boldsymbol{\delta}^4$ = [4.6425, 4.4230]	$\mathbf{p}_4$ = [0.2119, 0.1241, 0.0319]
	5	$\mathbf{c}^5$ = [0.4924, 4.5367], $\boldsymbol{\delta}^5$ = [3.3541, 6.7561]	$\mathbf{p}_5$ = [0.2122, 0.3256, -0.0355]

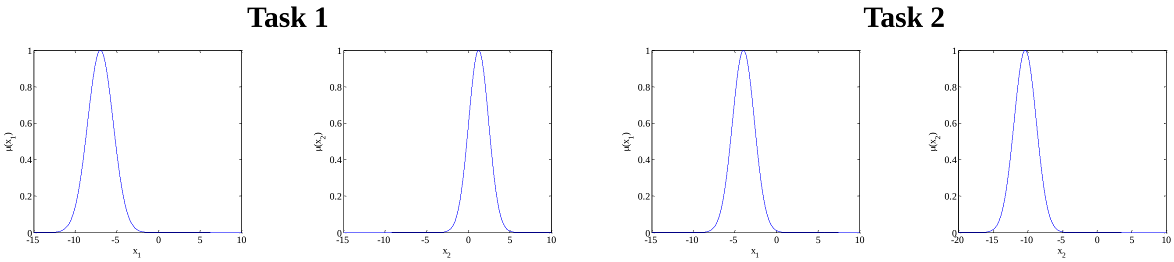


Fig.7 The corresponding membership functions of each fuzzy subset in the antecedent of the 1st fuzzy rule.

6. Conclusions

In this study, a novel single-task fuzzy classifier called TSK-FC is first presented for a single classification task. TSK-FC exhibits some distinctive characteristics inheriting from the conventional fuzzy systems, such as high interpretability. Furthermore, we extend TSK-FC to its multi-task version called MT-TSK-FC by using the multi-task learning mechanism, which can not only take full advantage of independent information for each task, but also effectively mine the correlation information among multiple tasks. However, when labeling-risk scenarios are considered, the performance of both TSK-FC and MT-TSK-FC deteriorate a lot. This situation will become more serious for more learning tasks in multi-task classification problems. To address this problem, we further extend MT-TSK-FC into its enhanced version LRA-MT-TSK-FC by using the proposed labeling-risk-aware mechanism. The labeling-risk-aware mechanism enhances the classification performance and robustness of LRA-MT-TSK-FC under a labeling-risk scene. It is worthy to mention that the training problems of the proposed three classifiers, i.e., TSK-FC, MT-TSK-FC and LRA-MT-TSK-FC are still classical QP problems and they can automatically derive the margin for each task. Extensive experiments on multi-task synthetic and real image classification datasets demonstrate the effectiveness and robustness of the proposed fuzzy classifiers, especially LRA-MT-TSK-FC.

As in LRA-MT-TSK-FC, the labeling-risk means that parameter μ was a critical issue influencing the robustness of LRA-MT-TSK-FC. In this paper, we just fixed three values to test the performance of our classifiers. How to adaptively learn is an interesting work in the future. Nevertheless, seeking the optimal value of labeling-risk means μ in labeling-risk-aware learning is still an open problem worth studying, and further establishing a solid theory regarding with it is absolutely necessary, it naturally becomes an important future work for us.

Acknowledgements

This work was supported in part by the Hong Kong Polytechnic University under Grant G-UA3W, and by the National Natural Science Foundation of China under Grants 61170122, 61272210 and by the Natural Science Foundation of Jiangsu Province under Grant BK2011003, BK2011417, JiangSu 333 expert engineering grant (BRA2011142), the Fundamental Research Funds for the Central Universities (Grant JUSRP111A38) and 2011 and 2012 Postgraduate Student’s Creative Research Fund of Jiangsu Province.

References

[1] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(4) (2010) 433-459.

[2] R. Babuska , *Fuzzy Modeling for Control* . Boston, MA: Kluwer, 1998.

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

[4] C. C. Chuang , S. F. Su and S. S. Chen, "Robust TSK fuzzy modeling for function approximation with outliers", *IEEE Trans. Fuzzy Systems*, 9(6) (2001) 810 -821.

[5] P. C. Chang and C. Y. Fan, “A hybrid system integrating a wavelet and TSK fuzzy rules for stock price forecasting,” *IEEE Trans. Syst., Man Cybern., Part C*, 38(6) (2008) 802–815.

[6] R. Caruana, "Multitask learning," *Machine Learning* , 28(1) (1997) 41–75.

[7] C. Cortes and V. Vapnik, “Support vector networks,” *Machine Learning*, 20(3) (1995) 273–297.

[8] 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*, 19(2) (2011) 210-226.

[9] Z.H. Deng, Y.Z. Jiang, K.S. Choi, F.L. Chung and S.T. Wang, “Knowledge-Leverage based TSK fuzzy system modeling,” *IEEE Trans. Neural Networks and Learning Systems*, 24(8) (2013) 1200-1212.

[10] R. Duda, P.Hart, andD. Stork, Pattern Classification. NewYork:Wiley, 2000.

[11] T. Evgeniou and M. Pontil, "Regularized Multi-Task Learning", *Proc. 10th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining*, (2004) 109 -117.

[12] 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*, 6 (2005) 1889-1918.

[13] O. Guenounou , A. Belmehdi and B. Dahhou, "Multi-objective optimization of TSK fuzzy models,” *Expert Syst. Appl.*, 36(4) (2009) 7416 -7423.

[14] D. Grossman, and P. Domingos, “Learning Bayesian network classifiers by maximizing conditional likelihood,” *Proceedings of the twenty-first international conference on Machine learning*, (2004) 46.

[15] F. Hoffmann and O. Nelles, ‘‘Structure identification of TSK-fuzzy systems using genetic programming,” *in Proceedings of IPMU 2000*, Madrid, Spain, (2000) 438-445.

[16] K. Ito, R. Nakano, “Optimizing support vector regression hyperparameters based on cross-validation,” *Proceedings of the International Joint Conference on Neural Networks*, (2003) 2077-2082.

[17] C. F. Juang and C. D. Hsieh "TS-fuzzy system based support vector regression", *Fuzzy Sets and Systems*, 160(17) (2009) 2486-2504.

[18] 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* , 37(6) (2007) 1077–1087.

[19] J.-S. R. Jang, “ANFIS: Adaptive- network-based fuzzy inference systems,” *IEEE Trans. Syst., Man, Cybern.* , 23(3) (1993) 665–685.

[20] C.F. Juang, "A TSK-type recurrent fuzzy networks for dynamic systems processing by neural network and genetic algorithms," *IEEE Trans. Fuzzy Systems*, 10(2) (2002) 155-170.

[21] Y. Ji and S. Sun, "Multitask multiclass support vector machines: Model and experiments," *Pattern Recognition*, 46(3) (2013) 914-924.

[22] C. Lee, W. Lai and Y. Lin. "A TSK type fuzzy neural network systems for dynamic systems identification," *In Proceedings of the IEEE-CDC*, (2003) 4002-4007.

[23] J. Leski, “TSK-fuzzy modeling based on ε-insensitive learning,” *IEEE Trans. Fuzzy Systems*, 13(2) (2005) 181-193.

[24] G. Li, K. Chang, and S. C.H. Hoi, "Multi-view semi-supervised learning with consensus," *IEEE Transactions on Knowledge and Data Engineering*, 24(11) (2012) 2040-2051.

[25] R. Mikut, O. Burmeister, L. Groll, and M. Reischl, “Takagi-Sugeno-Kang fuzzy classifiers for a special class of time-varying systems,” *IEEE Trans. Fuzzy Systems*,16(4) (2008) 1038-1049.

[26] G. M. Mendez, "Interval type-1 non-singleton type-2 TSK fuzzy logic systems using the hybrid training method RLS-BP," *Advances in Soft Computing: Analysis and Design of Intelligent Systems Using Soft Computing Techniques*, Springer, (2007) 36-44.

[27] G. M. Mendez, M. A. Hernández, "Hybrid learning mechanism for interval A2-C1type-2 non-singleton type-2 Takagi–Sugeno–Kang fuzzy logic systems," *Information Science*, 220(20) (2013) 149–169.

[28] A. Oliva and A. Torralba, “Modeling the shape of the scene: a holistic representation of the spatial envelope,” *IJCV*, 42(3) (2001) 145-175.

[29] S. E. Papadakis and J. B. Theocharis, "A GA-based fuzzy modeling approach for

generating TSK models", *Fuzzy Sets & Systems*, 131 (2002) 121-152.

[30] S. Parameswaran and K.Q. Weinberger, "Large margin multi-task metric learning," *Advances in Neural Information Processing Systems*, (2010) 1867–1875.

[31] Z.K. Qin, Q. Ren, B. Lionel, B. Marek, "Joint friction identification for robots using TSK fuzzy system based on subtractive clustering," *in Annual Meeting of the North American Fuzzy Information Processing Society. Missouri*, (2008) 1–6.

[32] S. Sun, "Multitask learning for EEG-based biometrics," *Proceedings of the 19th International Conference on Pattern Recognition*, (2008) 1–4.

[33] T. Takagi and M. Sugeno, “Fuzzy identification of systems and its application to modeling and control,” *IEEE Trans. Systems Man and Cybernetics*, 15(1) (1985) 116-132.

[34] A. M. Tang, C. Quek, and G. S. Ng, "GA-TSKfnn: Parameters tuning of fuzzy neural network using genetic algorithms," *Expert Systems with Applications*, 29 (2005) 769-781.

[35] W.W. Tan, C.L. Foo and T.W. Chua, “Type-2 fuzzy system for ECG arrhythmic classification,” *In Proceedings of the IEEE International Conference on Fuzzy Systems*, London UK, (2007) 859-864.

[36] J.W. Tao, KF.L. Chung, S.T. Wang, “On minimum distribution discrepancy support vector machine for domain adaptation.” *Pattern Recognition*, 45(11) (2012) 3962-3984.

[37] 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*, 6(4) (1998) 530–537.

[38] X.T. Yuan and S. Yan, "Visual classification with multi-task joint sparse represen-tation," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, (2010) 3493–3500.

[39] S.-M. Zhou and J.Q. Gan, "Extracting Takagi-Sugeno Fuzzy Rules with Interpretable Submodels via Regularization of Linguistic Modifiers," *IEEE Trans. Knowledge and Data Eng.*, 21(8) (2009) 1191-1204.

[40] W. Q. Zhao, K. Li and G. W. Irwin, "A new gradient descent approach for local learning of fuzzy neural models", *IEEE Trans. Fuzzy Systems*, 21(1) (2013) 30-44.

[41] M. A. Sanchez, O. Castillo, J. R. Castro, "Information granule formation via the concept of uncertainty-based information with Interval Type-2 Fuzzy Sets representation and Takagi-Sugeno-Kang consequents optimized with Cuckoo search," *Applied Soft Computing*, 27 (2015) 602-609.

[42] P. Melin, O. Castillo, "A review on type-2 fuzzy logic applications in clustering, classification and pattern recognition," *Applied Soft Computing*, 21 (2014) 568-577.

[43] M. A. Sanchez, O. Castillo, J. R. Castro, P. Melin, "Fuzzy granular gravitational clustering algorithm for multivariate data," *Information Sciences*, 279 (2014) 498-511.

[44] O. Castillo, J. R. Castro, P. Melin, A. R. Díaz, "Application of interval type-2 fuzzy neural networks in non-linear identification and time series prediction," *Soft Computing*, 18(6) (2014) 1213-1224.

[45] Z Deng, K-S Choi, Y Jiang, S Wang, "Generalized Hidden-Mapping Ridge Regression, Knowledge-Leveraged Inductive Transfer Learning for Neural Networks, Fuzzy Systems and Kernel Methods," *IEEE Transactions on Cybernetics*, 44(2) (2014) 2585-2599.

[46] Y Jiang, F L Chung, H Ishibuchi,et al, "Multitask TSK Fuzzy System Modeling by Mining Intertask Common Hidden Structure", *IEEE Transactions on Cybernetics*, 45(3) (2015) 548 - 561.

[47] M. Elkano, M. Galar, J. Sanz, et al. "Enhancing multi-class classification in FARC-HD fuzzy classifier: On the synergy between n-dimensional overlap functions and decomposition strategies," *IEEE trans. on Fuzzy Systems*, 23(5) (2014) 1562 - 1580.

[48] R. Qun, L. Baron, M. Balazinski, "Type-2 Takagi–Sugeno–Kang fuzzy logic modeling using subtractive clustering," *Proceedings of the Annual Conference of the North American Fuzzy Information Processing Society – NAFIPS*, (2006) 120–125.

[49] G. Zheng, J. Wang, W. Zhou, Y. Zhang, "A similarity measure between interval type-2 fuzzy sets," *Proceedings of the 2010 IEEE International Conference on Mechatronics and Automation, ICMA 2010*, (2010) 191–195.

[50] J. Alcalá-Fdez, R. Alcalá, F. Herrera, "A Fuzzy Association Rule-Based Classification Model for High-Dimensional Problems With Genetic Rule Selection and Lateral Tuning," *IEEE Trans. Fuzzy Systems*, 19(5) (2011) 857-872.

[51] M. Fazzolari, R. Alcalá, F. Herrera," A multi-objective evolutionary method for learning granularities based on fuzzy discretization to improve the accuracy-complexity trade-off of fuzzy rule-based classification systems: D-MOFARC algorithm," *Appl. Soft Comput.*, 24 (2014) 470-481.

Appendix A

For Eq.(5.b), the corresponding Lagrangian function is given by

$$\begin{aligned} & L(\mathbf{p}_{\mathrm{g}}, \xi_i, \varepsilon, \boldsymbol{\lambda}, \boldsymbol{\varphi}, \delta) \\ & = \frac{1}{2}(\mathbf{p}_{\mathrm{g}}^{\mathrm{T}}\mathbf{p}_{\mathrm{g}}) + \frac{1}{N\tau}\sum_{i=1}^{N}\xi_i - \frac{1}{\tau}\varepsilon + \sum_{i=1}^{N}\lambda_i\left(\varepsilon - \xi_i - y_i \cdot (\mathbf{p}_{\mathrm{g}}^{\mathrm{T}}\mathbf{x}_{\mathrm{g}i})\right) - \sum_{i=1}^{N}\varphi_i\,\xi_i - \delta\cdot\varepsilon \end{aligned} \tag{A1}$$

From this equation, the optimal values can be computed by setting the derivatives of $L(\bullet)$ w.r.t. $\mathbf{p}_{\mathrm{g}}, \xi_i, \varepsilon, \boldsymbol{\lambda}, \boldsymbol{\varphi}$ and δ to zeros, respectively, i.e.,

$$\frac{\partial L}{\partial \mathbf{p}_{\mathrm{g}}} = \mathbf{p}_{\mathrm{g}} - \sum_{i=1}^{N}\lambda_i\, y_i\mathbf{x}_{\mathrm{g}i} = 0 \tag{A2}$$

$$\frac{\partial L}{\partial \xi_i} = \frac{1}{N\tau} - \lambda_i - \varphi_i = 0 \tag{A3}$$

$$\frac{\partial L}{\partial \varepsilon} = -\frac{1}{\tau} - \delta + \sum_{i=1}^{N}\lambda_i = 0 \tag{A4}$$

From (A2) to (A4), we have

$$\mathbf{p}_{\mathrm{g}} = \sum_{i=1}^{N}\lambda_i\, y_i\mathbf{x}_{\mathrm{g}i} \tag{A5}$$

$$\lambda_i = \frac{1}{N\tau} - \varphi_i \tag{A6}$$

$$\delta = \sum_{i=1}^{N}\lambda_i - \frac{1}{\tau} \tag{A7}$$

Substituting (A5)–(A7) into (A1), the following optimization problem is obtained:

$$L(\boldsymbol{\lambda}) = -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\lambda_i\lambda_j y_i y_j \mathbf{x}_{gi}^{T}\mathbf{x}_{gj} \tag{A8}$$

$$\text{s.t.} \qquad \lambda \in [0, \frac{1}{N\tau}] \qquad \sum_{i=1}^{N}\lambda_i \geq \frac{1}{\tau}$$

where the constraint $\sum_{i=1}^{N}\lambda_i \geq \frac{1}{\tau}$ can be equivalently expressed as $\sum_{i=1}^{N}\lambda_i = \frac{1}{\tau}$.

It is clear that Eq.(A5) and Eq.(A8) are equivalent to Eq.(5.f) and Eq.(5.c), respectively.

Appendix B

For Eq.(7), the corresponding Lagrangian function is given by

$$
\begin{aligned}
&L(\boldsymbol{p}_{g_0},\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_K,\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_K,\varepsilon_1,\ldots,\varepsilon_K,\boldsymbol{\lambda}_1,\ldots,\boldsymbol{\lambda}_K,\alpha_1,\ldots,\alpha_K)\\
&=\frac{1}{2}\boldsymbol{p}_{g_0}^T\boldsymbol{p}_{g_0}+\lambda\frac{1}{K}\sum_{k=1}^{K}\frac{1}{2}\boldsymbol{\theta}_k^T\boldsymbol{\theta}_k+\sum_{k=1}^{K}\frac{1}{N_k\tau_k}\sum_{i=1}^{N_k}\xi_{i,k}-\sum_{k=1}^{K}\frac{1}{\tau_k}\varepsilon_k\\
&+\sum_{k=1}^{K}\sum_{i=1}^{N_k}\lambda_{i,k}\left(\varepsilon_k-\xi_{i,k}-y_{i,k}\cdot((\boldsymbol{p}_{g_0}+\boldsymbol{\theta}_k)^T\boldsymbol{x}_{gi,k})\right)-\sum_{k=1}^{K}\alpha_k\varepsilon_k-\sum_{k=1}^{K}\sum_{i=1}^{N_k}\beta_{i,k}\xi_{i,k}
\end{aligned}
\tag{B1}
$$

From this equation, the optimal values can be computed by setting the derivatives of $L(\bullet)$ w.r.t. $\boldsymbol{p}_{g_0},\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_K,\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_K,\varepsilon_1,\ldots,\varepsilon_K,\boldsymbol{\lambda}_1,\ldots,\boldsymbol{\lambda}_K$ and $\alpha_1,\ldots,\alpha_K$ to zeros, respectively, i.e.,

$$\frac{\partial L}{\partial \boldsymbol{p}_{g_0}}=\boldsymbol{p}_{g_0}-\sum_{k=1}^{K}\sum_{i=1}^{N_k}\lambda_{i,k}y_{i,k}\boldsymbol{x}_{gi,k}=0 \tag{B2}$$

$$\frac{\partial L}{\partial \boldsymbol{\theta}_k}=\frac{\lambda}{K}\boldsymbol{\theta}_k-\sum_{i=1}^{N_k}\lambda_{i,k}y_{i,k}\boldsymbol{x}_{gi,k}=0 \tag{B3}$$

$$\frac{\partial L}{\partial \xi_{i,k}}=\frac{1}{N_k\tau_\mathrm{k}}-\lambda_{i,\mathrm{k}}-\beta_{i,k}=0 \tag{B4}$$

$$\frac{\partial L}{\partial \varepsilon_k}=-\frac{1}{\tau_k}+\sum_{i=1}^{N_k}\lambda_{i,\mathrm{k}}-\alpha_k=0 \tag{B5}$$

From (B2) to (B5), we have

$$\boldsymbol{p}_{g_0}=\sum_{k=1}^{K}\sum_{i=1}^{N_k}\lambda_{i,k}y_{i,k}\boldsymbol{x}_{gi,k} \tag{B6}$$

$$\boldsymbol{\theta}_k=\frac{K}{\lambda}\sum_{i=1}^{N_k}\lambda_{i,k}y_{i,k}\boldsymbol{x}_{gi,k} \tag{B7}$$

$$\lambda_{i,\mathrm{k}}+\beta_{i,k}=\frac{1}{N_k\tau_\mathrm{k}} \tag{B8}$$

$$\sum_{i=1}^{N_k}\lambda_{i,\mathrm{k}}-\alpha_k=\frac{1}{\tau_k} \tag{B9}$$

Substituting (B6)–(B9) into (B1), the following optimization problem is obtained:

$$
\begin{aligned}
&L(\boldsymbol{\lambda}_1,\ldots,\boldsymbol{\lambda}_K)\\
&=\frac{1}{2}(\sum_{k=1}^{K}\sum_{i=1}^{N_k}\lambda_{i,k}y_{i,k}\boldsymbol{x}_{gi,k})^T(\sum_{l=1}^{K}\sum_{j=1}^{N_k}\lambda_{j,l}y_{j,l}\boldsymbol{x}_{gj,l})\\
&+\frac{\lambda}{2K}\sum_{k=1}^{K}(\frac{K}{\lambda}\sum_{i=1}^{N_k}\lambda_{i,k}y_{i,k}\boldsymbol{x}_{gi,k})^T(\frac{K}{\lambda}\sum_{j=1}^{N_k}\lambda_{j,k}y_{j,k}\boldsymbol{x}_{gj,k})\\
&+\sum_{k=1}^{K}\sum_{i=1}^{N_k}\lambda_{i,k}\left(-y_{i,k}(\sum_{l=1}^{K}\sum_{j=1}^{N_k}\lambda_{j,l}y_{j,l}\boldsymbol{x}_{gj,l})^T\boldsymbol{x}_{gi,k}-y_{i,k}(\frac{K}{\lambda}\sum_{j=1}^{N_k}\lambda_{j,k}y_{j,k}\boldsymbol{x}_{gj,k})^T\boldsymbol{x}_{gj,k}\right)\\
&\text{s.t.}\qquad \lambda_{i,k}\in[0,\frac{1}{N_k\tau_k}]\qquad \sum_{i=1}^{N_k}\lambda_{i,k}\geq\frac{1}{\tau_k}\qquad \forall k\quad k=1\ldots K
\end{aligned}
\tag{B10}
$$

After simplifying the above objective function, Eq.(B10) can be equivalently expressed as the following optimization problem:

$$L(\boldsymbol{\lambda}_1,\ldots,\boldsymbol{\lambda}_K)$$

$$= -\frac{1}{2}\sum_{k=1}^{K}\sum_{l=1}^{K}\sum_{i=1}^{N_k}\sum_{j=1}^{N_l}\lambda_{j,l}\lambda_{i,k}y_{j,l}y_{i,k}\boldsymbol{x}_{\mathrm{g}j,l}^T\boldsymbol{x}_{\mathrm{g}i,k} - \frac{K}{2\lambda}\sum_{k=1}^{K}\sum_{i=1}^{N_k}\sum_{j=1}^{N_k}\lambda_{i,k}\lambda_{j,k}y_{i,k}y_{j,k}\boldsymbol{x}_{\mathrm{g}j,k}^T\boldsymbol{x}_{\mathrm{g}i,k} \tag{B11}$$

$$\text{s.t.} \qquad \lambda_{i,k}\in[0,\frac{1}{N_k\tau_k}] \qquad \sum_{i=1}^{N_k}\lambda_{i,k}\ge\frac{1}{\tau_k} \qquad \forall k \quad k=1\ldots K$$

where the constraint $\sum_{i=1}^{N_k}\lambda_{i,k}\ge\frac{1}{\tau_k}$ can be equivalently expressed as $\sum_{i=1}^{N_k}\lambda_{i,k}=\frac{1}{\tau_k}$.

It is clear that Eq.(B6), Eq.(B7) and Eq.(B11) are equivalent to Eq.(9.a), Eq.(9.b) and Eq.(8), respectively.

Appendix C

1): The derivation of Eq.(15.a)

$$E_{\varsigma_k}\left[\tilde{\boldsymbol{K}}_k\right]=E_{\varsigma_k}[\tilde{k}_{ij}] \tag{C1}$$

where $E_{\varsigma_k}[\tilde{k}_{ij}]=E_{\varsigma_k}\left[y_{i,k}(1-2\varsigma_{i,k})y_{j,k}(1-2\varsigma_{j,k})\frac{K}{\lambda}\boldsymbol{x}_{\mathrm{g}j,k}^{\mathrm{T}}\boldsymbol{x}_{\mathrm{g}i,k}\right]$

if $i=j$, we have

$$E_{\varsigma_k}[\tilde{k}_{ij}]=y_{i,k}y_{j,k}\frac{K}{\lambda}\boldsymbol{x}_{\mathrm{g}j,k}^{\mathrm{T}}\boldsymbol{x}_{\mathrm{g}i,k} \tag{C2}$$

otherwise,

$$\begin{aligned}E_{\varsigma_k}[\tilde{k}_{ij}]&=E_{\varsigma_k}\left[y_{i,k}(1-2\varsigma_{i,k})y_{j,k}(1-2\varsigma_{j,k})\frac{K}{\lambda}\boldsymbol{x}_{\mathrm{g}j,k}^{\mathrm{T}}\boldsymbol{x}_{\mathrm{g}i,k}\right]\\&=y_{i,k}y_{j,k}\frac{K}{\lambda}\boldsymbol{x}_{\mathrm{g}j,k}^{\mathrm{T}}\boldsymbol{x}_{\mathrm{g}i,k}E_{\varsigma_k}\left[(1-2\varsigma_{i,k})(1-2\varsigma_{j,k})\right]\\&=y_{i,k}y_{j,k}\frac{K}{\lambda}\boldsymbol{x}_{\mathrm{g}j,k}^{\mathrm{T}}\boldsymbol{x}_{\mathrm{g}i,k}\left[1-2E(\varsigma_{i,k})-2E(\varsigma_{j,k})+4E(\varsigma_{i,k})E(\varsigma_{j,k})\right]\\&=y_{i,k}y_{j,k}\frac{K}{\lambda}\boldsymbol{x}_{\mathrm{g}j,k}^{\mathrm{T}}\boldsymbol{x}_{\mathrm{g}i,k}\left(1-2\mu_k-2\mu_k+4\mu_k^2\right)\\&=y_{i,k}y_{j,k}\frac{K}{\lambda}\boldsymbol{x}_{\mathrm{g}j,k}^{\mathrm{T}}\boldsymbol{x}_{\mathrm{g}i,k}\left(1-4\mu_k(1-\mu_k)\right)\end{aligned} \tag{C3}$$

Accordingly, the Eq.(C1) can be formulated as

$$E_{\varsigma_k}\left[\tilde{\boldsymbol{K}}_k\right]=E_{\varsigma_k}[\tilde{k}_{ij}]_{N_k\times N_k},\ E_{\varsigma_k}[\tilde{k}_{ij}]=\begin{cases}y_{i,k}y_{j,k}\frac{K}{\lambda}\boldsymbol{x}_{\mathrm{g}j,k}^{\mathrm{T}}\boldsymbol{x}_{\mathrm{g}i,k} & i=j\\ y_{i,k}y_{j,k}\frac{K}{\lambda}\boldsymbol{x}_{\mathrm{g}j,k}^{\mathrm{T}}\boldsymbol{x}_{\mathrm{g}i,k}\left(1-4\mu_k(1-\mu_k)\right) & i\ne j\end{cases} \tag{C4}$$

2): The derivation of Eq.(15.b)

$$E_{\varsigma_k}\left[\hat{\boldsymbol{K}}_{k,l}\right]=E_{\varsigma_k}[\tilde{k}_{ij}], \tag{C5}$$

where $E_{\varsigma_k}[\tilde{k}_{ij}] = E_{\varsigma_k}\left[y_{i,l}(1-2\varsigma_{i,l})y_{j,k}(1-2\varsigma_{j,k})\boldsymbol{x}_{\mathrm{g}j,l}^{\mathrm{T}}\boldsymbol{x}_{\mathrm{g}i,k}\right]$

if $k = l$, we have

$$E_{\varsigma_k}[\tilde{k}_{ij}] = E_{\varsigma_k}\left[y_{i,k}(1-2\varsigma_{i,k})y_{j,k}(1-2\varsigma_{j,k})\boldsymbol{x}_{\mathrm{g}j,k}^{\mathrm{T}}\boldsymbol{x}_{\mathrm{g}i,k}\right] \tag{C6}$$

Similar to the derivation of Eq.(C1), we have

$$E_{\varsigma_k}[\tilde{k}_{ij}] = \begin{cases} y_{i,k}y_{j,k}\,\boldsymbol{x}_{\mathrm{g}j,k}^{\mathrm{T}}\boldsymbol{x}_{\mathrm{g}i,k} & i = j \\ y_{i,k}y_{j,k}\,\boldsymbol{x}_{\mathrm{g}j,k}^{\mathrm{T}}\boldsymbol{x}_{\mathrm{g}i,k}\left(1-4\mu_k(1-\mu_k)\right) & i \neq j \end{cases} \tag{C7}$$

if $k \neq l$, we have

$$\begin{aligned} E_{\varsigma}[\tilde{k}_{ij}] &= E_{\varsigma_k}\left[y_{i,l}(1-2\varsigma_{i,l})y_{j,k}(1-2\varsigma_{j,k})\boldsymbol{x}_{\mathrm{g}j,l}^{\mathrm{T}}\boldsymbol{x}_{\mathrm{g}i,k}\right] \\ &= y_{i,l}y_{j,k}\,\boldsymbol{x}_{\mathrm{g}j,l}^{\mathrm{T}}\boldsymbol{x}_{\mathrm{g}i,k}E_{\varsigma_k}\left[(1-2\varsigma_{i,l})(1-2\varsigma_{j,k})\right] \\ &= y_{i,l}y_{j,k}\,\boldsymbol{x}_{\mathrm{g}j,l}^{\mathrm{T}}\boldsymbol{x}_{\mathrm{g}i,k}\left[1-2E(\varsigma_{i,l})-2E(\varsigma_{j,k})+4E(\varsigma_{i,l})E(\varsigma_{j,k})\right] \\ &= y_{i,l}y_{j,k}\,\boldsymbol{x}_{\mathrm{g}j,l}^{\mathrm{T}}\boldsymbol{x}_{\mathrm{g}i,k}\left(1-2\mu_l-2\mu_k+4\mu_l\mu_k\right) \\ &= y_{i,l}y_{j,k}\,\boldsymbol{x}_{\mathrm{g}j,l}^{\mathrm{T}}\boldsymbol{x}_{\mathrm{g}i,k}\left(1-2(\mu_l+\mu_k)+4\mu_l\mu_k\right) \end{aligned} \tag{C8}$$

Accordingly, the Eq.(C5) can be formulated as

$$E_{\varsigma_k}\left[\hat{\boldsymbol{K}}_{k,l}\right] = E_{\varsigma_k}[\tilde{k}_{ij}]_{N_l\times N_k},\ E_{\varsigma_k}[\tilde{k}_{ij}] = \begin{cases} y_{i,k}y_{j,l}\,\boldsymbol{x}_{\mathrm{g}j,k}^{\mathrm{T}}\boldsymbol{x}_{\mathrm{g}i,l} & k = l, i = j \\ y_{i,k}y_{j,k}\,\boldsymbol{x}_{\mathrm{g}j,k}^{\mathrm{T}}\boldsymbol{x}_{\mathrm{g}i,k}\left(1-4\mu_k(1-\mu_k)\right) & k = l, i \neq j \\ y_{i,l}y_{j,k}\,\boldsymbol{x}_{\mathrm{g}j,l}^{\mathrm{T}}\boldsymbol{x}_{\mathrm{g}i,k}\left(1-2(\mu_l+\mu_k)+4\mu_l\mu_k\right) & k \neq l, \forall i, j \end{cases} \tag{C9}$$

Eqs. (C4) and (C9) are just Eqs. (15.a) and (15.b) in the text.