# Minimax Probability TSK Fuzzy System Classifier: A More Transparent and Highly Interpretable Classification Model

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

***Abstract*—When an intelligent model is used for medical diagnosis, it is desirable to have a high level of interpretability and transparent model reliability for users. Compared with most of the existing intelligence models, fuzzy systems have shown a distinctive advantage in their interpretabilities. However, how to determine the model reliability of a fuzzy system trained for a recognition task is still an unsolved problem at present. In this study, a minimax probability Takagi–Sugeno–Kang (TSK) fuzzy system classifier called MP-TSK-FSC is proposed to train a fuzzy system classifier and determine the model reliability simultaneously. For the proposed MP-TSK-FSC, a lower bound of correct classification can be presented to the users to characterize the reliability of the trained fuzzy classifier. Thus, the obtained classifier has the distinctive characteristics of both a high level of interpretability and transparent model reliability inherited from the fuzzy system and minimax probability learning strategy, respectively. Our experiments on synthetic datasets and several real-world datasets for medical diagnosis have confirmed the distinctive characteristics of the proposed method.**

***Index Terms*—Classification, medical diagnosis, minimax probability decision, Takagi–Sugeno–Kang (TSK) fuzzy system.**

## I. Introduction

MANY intelligent models, such as neural networks and fuzzy systems, have been applied to pattern recognition tasks in various fields [1]–[3], [48]–[51], such as medical diagnosis. When these intelligent models are adopted, it is desirable to have a high level of interpretability and transparent model reliability for users [4], [38], [39]. Thus, a model with these characteristics can be viewed as the specialists in the related fields, such as a specialist in oncology. Compared with most of the existing intelligence models, fuzzy systems have demonstrated a distinctive advantage in interpretability [5]–[7], [52]–[54] and have been adopted for many practical modeling tasks, especially for medical diagnosis [40]–[44]. Many studies have effectively addressed the interpretation of fuzzy systems [52]–[54]. However, how to determine the model reliability of a fuzzy system trained for a certain pattern recognition task is still unsolved at present. This issue is addressed in this study.

Recently, the minimax probability decision technique has attracted the attention of many researchers and has been adopted for the development of several pattern recognition methods by considering model reliability. In [8] and [9], it was utilized to design a minimax probability machine (MPM) for novelty detection and classification. The distinctive characteristic of minimax probability decision-based methods is that the lower bound of a correct decision can be obtained for the trained model as its reliability. At present, this technique has been extended to cater for different scenarios in classification and regression problems [10]–[14]. In particular, reliability was studied for evolving fuzzy systems in [55] and [56] in a data-stream context by using "conflict" and "ignorance" concepts. However, the related studies are still limited, and more novel mechanisms are needed to evaluate the reliability of fuzzy models.

In this study, in order to make fuzzy systems more transparent as an advanced expert system in practical applications, such as medical diagnosis, the minimax probability decision technique is introduced to train fuzzy systems for classification tasks. Accordingly, a minimax probability Takagi–Sugeno–Kang fuzzy system (TSK-FS) classifier, i.e., MP-TSK-FSC, is proposed to train a classifier and determine its model reliability simultaneously. For the proposed MP-TSK-FSC, the lower bound of correct classification can be presented to the users as the model reliability. Thus, the MP-TSK-FSC possesses the distinctive characteristics of both a high level of interpretability and transparent model reliability. The proposed method is finally evaluated on synthetic datasets and several medical datasets for medical diagnosis, and its effectiveness has been confirmed accordingly.

The rest of this paper is organized as follows. Concepts related to TSK-FS and the MPM are reviewed in Section II. In Section III, the MP-TSK-FSC is proposed based on the minimax probability decision technique. The experimental results on synthetic datasets and several medical datasets for medical diagnosis are reported in Section IV. Conclusions and the potential of the proposed method are given in the final section.

Manuscript received December 12, 2013; revised March 5, 2014; accepted February 4, 2014. Date of publication June 3, 2014; date of current version July 31, 2015. This work was supported in part by the National Natural Science Foundation of China under Grant 61170122 and Grant 61272210, the Ministry of education program for New Century Excellent Talents under Grant NCET-120882, the Fundamental Research Funds for the Central Universities under Grant JUSRP51321B, the Natural Science Foundation of Jiangsu Province under Grant BK2011003, Australian Research Council Discovery Grants (DP130102691), and Linkage Grants (LP120100566).

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

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

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

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TFUZZ.2014.2328014

## II. Takagi–Sugeno–Kang Fuzzy Systems and Minimax Probability Decision

In this section, the related techniques for the proposed MP-TSK-FSC are reviewed. First, the concepts and principles behind the classical TSK-FS are reviewed briefly, and then, the minimax probability decision technique is introduced.

### A. Concepts and Principles Behind the TSK-FS

Of the three classical fuzzy system models, i.e., the TSK-FS model [15], Mamdani–Larsen fuzzy system (ML-FS) model [16], and generalized fuzzy model [17], the TSK model is the most popular due to its effectiveness. For example, for a modeling task, the TSK-FS model usually requires far fewer rules to obtain an equivalent performance than that needed for an ML-FS. In this study, the TSK-FS model is our focus. For this type of fuzzy model, the most commonly used fuzzy inference rules are defined as follows:

$$\begin{aligned}&\text{TSK Fuzzy Rule } R^k\text{:}\\&\text{IF } x_1 \text{ is } A_1^k \wedge x_2 \text{ is } A_2^k \wedge \cdots \wedge x_d \text{ is } A_d^k\\&\text{Then } f_k(\mathbf{x}) = p_{k0} + p_{k1}x_1 + \cdots + p_{kd}x_d,\ k = 1, \ldots, K\end{aligned} \tag{1}$$

In (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 $\mathbf{x} = [x_1, x_2, \ldots, x_d]^T$ and maps the fuzzy subsets in the input space $A^k \subset R^d$ to a varying singleton denoted by $f_k(\mathbf{x})$. When the commonly used *multiplicative* conjunction, *multiplicative* implication, and *additive* disjunction are employed, respectively, as the conjunction operator, the implication operator, and the disjunction operator, the output of the TSK fuzzy model can be formulated as

$$f_{\mathrm{TSK-FS}}(\mathbf{x}) = \sum_{k=1}^{K} \frac{\mu_k(\mathbf{x})}{\sum_{k'=1}^{K} \mu_{k'}(\mathbf{x})} \cdot f_k(\mathbf{x}) = \sum_{k=1}^{K} \tilde{\mu}_k(\mathbf{x}) \cdot f_k(\mathbf{x}) \tag{2}$$

where $\mu_k(\mathbf{x})$ and $\tilde{\mu}_k(\mathbf{x})$ denote the fuzzy membership and the normalized fuzzy membership associated with the fuzzy subset $A_i^k$, respectively. These two memberships can be calculated by

$$\mu_k(\mathbf{x}) = \prod_{i=1}^{d} \mu_{A_i^k}(x_i) \tag{3a}$$

$$\tilde{\mu}_k(\mathbf{x}) = \frac{\mu_k(\mathbf{x})}{\sum_{k'=1}^{K} \mu_{k'}(\mathbf{x})}. \tag{3b}$$

### B. Minimax Probability Decision Technique

The minimax probability decision technique was first utilized to design an MPM for novelty detection and classification and has been further extended to cater for different scenarios in classification and to regression problems [8]–[14]. The objective of minimax probability principle-based methods is to obtain the maximal lower bound of a correct decision for the trained model in related modeling tasks. Here, we briefly review the principle of MPM [9] for classification since this method is closely related to the proposed MP-TSK-FSC in this study.

A given dataset contains two classes which are sampled from two random variables $\mathbf{x} \sim (\mathbf{u}_+, \Sigma_+)$ and $\mathbf{x} \sim (\mathbf{u}_-, \Sigma_-)$, where $\mathbf{u}_+, \Sigma_+$ and $\mathbf{u}_-, \Sigma_-$ denote the means and covariance matrices of two classes, respectively. MPM defines the following optimization objective to obtain a classification hyperplane $\mathbf{w}^T\mathbf{x} - b = 0$:

$$\begin{aligned}&\max_{\alpha, \mathbf{w}, b} \alpha\\&\text{s.t.} \inf_{\mathbf{x} \sim (\mathbf{u}_+, \Sigma_+)} pr(\mathbf{w}^T\mathbf{x} \geq b) \geq \alpha\\&\qquad \inf_{\mathbf{x} \sim (\mathbf{u}_-, \Sigma_-)} pr(\mathbf{w}^T\mathbf{x} \leq b) \geq \alpha\end{aligned} \tag{4a}$$

where $\inf_{\mathbf{x} \sim (\mathbf{u}_+, \Sigma_+)} pr(\mathbf{w}^T\mathbf{x} \geq b)$ denotes the infimum of probability for the condition: $\mathbf{w}^T\mathbf{x} \geq b$, $\mathbf{x} \sim (\mathbf{u}_+, \Sigma_+)$. The optimization objective in (4a) implies that for two-class data samples from random variables $\mathbf{x} \sim (\mathbf{u}_+, \Sigma_+)$ and $\mathbf{x} \sim (\mathbf{u}_-, \Sigma_-)$, there exists an optimal hyperplane $(\mathbf{w}^*)^T\mathbf{x} - b^* = 0$, which makes the lower bound of correct classification of a future datum point maximal and the upper bound of misclassifying it minimal.

By introducing the kernel trick, the objective of the kernelized version for MPM is proposed as follows:

$$\begin{aligned}&\max_{\alpha, \mathbf{w}, b} \alpha\\&\text{s.t.} \inf_{\mathbf{x} \sim (\mathbf{u}_+, \Sigma_+)} pr(\mathbf{w}^T\varphi(\mathbf{x}) \geq b) \geq \alpha\\&\qquad \inf_{\mathbf{x} \sim (\mathbf{u}_-, \Sigma_-)} pr(\mathbf{w}^T\varphi(\mathbf{x}) \leq b) \geq \alpha\end{aligned} \tag{4b}$$

where $\varphi(\mathbf{x})$ is the mapping function, which maps the data $\mathbf{x}$ in the original space to $\varphi(\mathbf{x})$ in the kernel feature space.

## III. Minimax Probability Takagi–Sugeno–Kang Fuzzy System Classifier

### A. Proposed Takagi–Sugeno–Kang Fuzzy System Classifier Model

Fuzzy systems, as a classical regression model, can be used for classification tasks [18], [45]–[47], [55] with different strategies. A very effective way to do this is to decompose the generalized multiclass classification task into many binary classification tasks [55]. Then, fuzzy systems are used to train the classification model by using the specified learning mechanism for binary classification. A commonly used way to develop a binary classification model by using TSK-FS is to use the following decision function:

$$y = \mathrm{sign}\left(f_{\mathrm{TSK-FS}}(\mathbf{x})\right) = \begin{cases} 1, & \text{if } f_{\mathrm{TSK-FS}}(\mathbf{x}) > 0 \\ -1, & \text{otherwise.} \end{cases} \tag{5}$$

Based on the above decision function, some fuzzy system classifiers for binary classification have been developed [19]–[21]. In this study, a TSK-FS-based classification model, which is called TSK-FSC, is proposed in a similar way.

The proposed classification model, as shown in Fig. 1, consists of two parts, i.e., the classical TSK-FS and a decision threshold. Based on this model, the final decision function for

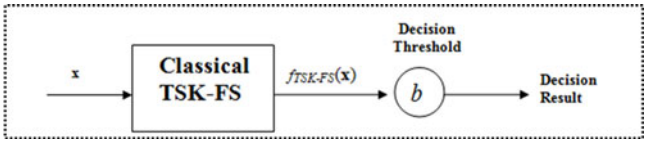



Fig. 1. Proposed TSK-FSC Model.

binary classification can be expressed as follows:

$$y = \text{sign}\left(f_{\text{TSK}-\text{FS}}(\mathbf{x}) - b\right) = \begin{cases} 1, & \text{if} f_{\text{TSK}-\text{FS}}(\mathbf{x}) > b \\ -1, & \text{otherwise.} \end{cases} \tag{6}$$

Please note that the proposed fuzzy classification model has a distinct characteristic, that is, the additional decision threshold has been incorporated into the model, which is different from the traditional fuzzy-system-based classifiers. In addition, the regression coefficients in the proposed model are also not the same as that in the classical TSK-FS to some extent.

### *B. Minimax Probability Objective Criterion for Takagi–Sugeno–Kang Fuzzy System Classifier Training*

Based on the minimax probability decision theory, the following objective is proposed to train the proposed TSK-FSC model:

$$\begin{aligned} &\max_{\boldsymbol{\Theta},\alpha} \alpha \\ &\text{s.t.} \inf_{\mathbf{x}\sim(\mathbf{u}_+,\boldsymbol{\Sigma}_+)} pr(f_{\text{TSK}-\text{FSC}}(\mathbf{x}) - b \geq 0) \geq \alpha \\ &\quad \inf_{\mathbf{x}\sim(\mathbf{u}_-,\boldsymbol{\Sigma}_-)} pr(f_{\text{TSK}-\text{FSC}}(\mathbf{x}) - b \leq 0) \geq \alpha \end{aligned} \tag{7}$$

where $\boldsymbol{\Theta}$ is the parameter set of the proposed TSK-FSC, including the parameters of TSK-FS and the decision threshold $b$. With the above optimization criterion, we expect that the TSK-FSC model can be trained and the corresponding lower bound of correct classification can be obtained as the model reliability. However, it is a difficult task to solve the objective function in (7) directly. We will overcome this issue by replacing (7) with the transformed objective function in order for it to be solved more easily.

Generally, for a TSK-FS, the antecedents and consequents can be determined independently. For the antecedents, a popular way is to construct them by using a certain partitioning technique, such as the self-evolution learning method [19] to partition the input spaces for a modeling task.

In particular, clustering methods have become one kind of popular technique to partition the input space based on input data of a training dataset, which results in corresponding fuzzy sets in the input space [57], [58]. In this study, the classical fuzzy c-means (FCM) clustering algorithm has been adopted due to the following distinctive characteristics: 1) FCM is very popular due to its simplicity and effectiveness in extensive applications; 2) FCM is a fuzzy-set-based clustering algorithm and the obtained clustering partition is a fuzzy partition, which makes it very natural to obtain the fuzzy partitions and then construct the fuzzy sets in the antecedents for fuzzy systems; and 3) it is much easier to control the number of fuzzy rules, i.e., the number of clusters obtained by FCM manually. However, for some other partitioning techniques, such as the grid partition method, the number of rules will increase sharply with the increasing dimensional number of input spaces.

If FCM is adopted, the procedure to construct the antecedents can be described as follows. Given a binary dataset $D_{tr} = \{(\mathbf{x}_i, y_i)\}$, $i = 1, \ldots, N$, by using the clustering algorithms, the data can be partitioned into $K$ clusters with the partition matrix as $\mathbf{U} = [u_{jk}]_{N\times K}$, $k = 1, \ldots, K$, $j = 1, \ldots, N$, where $u_{jk} \in [0, 1]$ denotes the membership of the $j$th input data $\mathbf{x}_j = (x_{j1}, \ldots, x_{jd})^T$, belonging to the $k$th cluster obtained by the FCM algorithm. Then, for the commonly used Gaussian membership function, i.e.,

$$\mu_{A_i^k}(x_i) = \exp\left(\frac{-(x_i - c_i^k)^2}{2\delta_i^k}\right) \tag{8a}$$

the parameters $c_i^k, \delta_i^k$ can be estimated by clustering results. For example, $c_i^k, \delta_i^k$ can be estimated as follows [22]–[24]:

$$c_i^k = \frac{\sum_{j=1}^N u_{jk} x_{ji}}{\sum_{j=1}^N u_{jk}} \tag{8b}$$

$$\delta_i^k = h \cdot \frac{\sum_{j=1}^N u_{jk}(x_{ji} - c_i^k)^2}{\sum_{j=1}^N u_{jk}} \tag{8c}$$

where $h$ is a scale constant and can be set manually or determined with some learning strategy, such as the cross-validation (CV) strategy.

When the antecedents of the TSK fuzzy model in (1) are determined, for a input vector $\mathbf{x}$, let

$$\mathbf{x}_e = (1, \mathbf{x}^T)^T \tag{9a}$$

$$\tilde{\mathbf{x}}_k = \tilde{\mu}_k(\mathbf{x})\mathbf{x}_e, \quad k = 1, \ldots, K \tag{9b}$$

with $\tilde{\mu}_k(\mathbf{x})$ computed by (8a)–(8c), (3a), and (3b)

$$\mathbf{x}_g = (\tilde{\mathbf{x}}_1^T, \tilde{\mathbf{x}}_2^T, \ldots, \tilde{\mathbf{x}}_K^T)^T \tag{9c}$$

$$\mathbf{p}_k = (p_{k0}, p_{k1}, \ldots, p_{kd})^T, \quad k = 1, \ldots, K \tag{9d}$$

$$\mathbf{p}_g = (\mathbf{p}_1^T, \mathbf{p}_2^T, \ldots, \mathbf{p}_K^T)^T. \tag{9e}$$

Then, the output of TSK-FS in (2) can be formulated as the following linear regression problem [23], [24]:

$$f_{\text{TSK}-\text{FS}}(\mathbf{x}) = \mathbf{p}_g^T \mathbf{x}_g. \tag{9f}$$

Thus, the training of a TSK fuzzy model can be transformed into the parameter learning of the corresponding linear regression model [22]–[24]. According to (7), the following objective can be adopted for parameter learning of the proposed TSK-FSC by using the minimax probability decision technique:

$$\begin{aligned} &\max_{\mathbf{p}_g,b,\alpha} \alpha \\ &\text{s.t.} \inf_{\mathbf{x}_g\sim(\tilde{\mathbf{u}}_+,\tilde{\boldsymbol{\Sigma}}_+)} pr\left(\mathbf{p}_g^T\mathbf{x}_g - b \geq 0\right) \geq \alpha \\ &\quad \inf_{\mathbf{x}_g\sim(\tilde{\mathbf{u}}_-,\tilde{\boldsymbol{\Sigma}}_-)} pr\left(\mathbf{p}_g^T\mathbf{x}_g - b \leq 0\right) \geq \alpha \end{aligned} \tag{10}$$

where $\mathbf{x}_g \sim (\tilde{\mathbf{u}}_+, \tilde{\mathbf{\Sigma}}_+)$ denotes the data mapped from $\mathbf{x} \sim (\mathbf{u}_+, \mathbf{\Sigma}_+)$ by the fuzzy inference rules as shown in (9a)–(9c). In the practical application, $\tilde{\mathbf{u}}_+$ and $\tilde{\mathbf{\Sigma}}_+$ can be estimated using the available dataset $\{\mathbf{x}_{gi}\}$ constructed by (9a)–(9c). Here, the obtained lower bound of correct classification can be taken as the model reliability for the trained fuzzy classifier. This means that if the future testing data are sampled from the density distribution with the same means and covariance matrices as that of the training data, the test accuracy for each class is always higher than the obtained lower bound, i.e., $\alpha$, in theory.

### C. Solution of Minimax Probability Takagi–Sugeno–Kang Fuzzy System Classifier

For the parameter solution of (10), we first give the following Theorem 1.

*Theorem 1:* The parameter learning of the consequents in the proposed MP-TSK-FSC in (10) can be taken as a special case of the classical MPM in [9], where the training data $\mathbf{x}$ are mapped as $\mathbf{x}_g$ in a new feature space, which is constructed by the fuzzy inference rules with the strategy in (9a)–(9c).

*Proof:* By comparing (10) with (4a) and (4b), we find that they have the same forms. Thus, (4) can be taken as the special case of the MPM in [9]. The distinctive characteristic of (10) is that the training data are the mapping data in a feature space constructed by using (9a)–(9c) with the fuzzy inference mechanism.

Based on Theorem 1, the conclusions obtained about MPM in [9] can be used for the solution of (10). Thus, we can give the following lemmas for (10) accordingly.

*Lemma 2:* With $\tilde{\mathbf{u}}_-, \tilde{\mathbf{\Sigma}}_-$ positive definite, $\mathbf{p}_g \neq \mathbf{0}$, $b$ given, such that $\mathbf{p}_g^T \tilde{\mathbf{u}}_- - b \leq 0$ and $\alpha \in [0, 1)$, the condition $\inf_{\mathbf{x}_g \sim (\tilde{\mathbf{u}}_-, \tilde{\mathbf{\Sigma}}_-)} pr(\mathbf{p}_g^T \mathbf{x}_g - b \leq 0) \geq \alpha$ holds if and only if $b - \mathbf{p}_g^T \tilde{\mathbf{u}}_- \geq \kappa(\alpha) \sqrt{\mathbf{p}_g^T \tilde{\mathbf{\Sigma}}_- \mathbf{p}_g}$, where $\kappa(\alpha) = \sqrt{\frac{\alpha}{1-\alpha}}$.

*Lemma 3:* With $\tilde{\mathbf{u}}_+, \tilde{\mathbf{\Sigma}}_+$ positive definite, $\mathbf{p}_g \neq \mathbf{0}$, $b$ given, such that $\mathbf{p}_g^T \tilde{\mathbf{u}}_+ - b \geq 0$ and $\alpha \in [0, 1)$, the condition $\inf_{\mathbf{x}_g \sim (\tilde{\mathbf{u}}_+, \tilde{\mathbf{\Sigma}}_+)} pr(\mathbf{p}_g^T \mathbf{x}_g - b \geq 0) \geq \alpha$ holds if and only if $\mathbf{p}_g^T \tilde{\mathbf{u}}_+ - b \geq \kappa(\alpha) \sqrt{\mathbf{p}_g^T \tilde{\mathbf{\Sigma}}_+ \mathbf{p}_g}$, where $\kappa(\alpha) = \sqrt{\frac{\alpha}{1-\alpha}}$.

Based on Lemmas 2 and 3, (10) can be transformed as the following optimization problem:

$$\begin{aligned} &\max_{\mathbf{p}_g, b, \alpha} \alpha \\ &\text{s.t.} \mathbf{p}_g^T \tilde{\mathbf{u}}_+ - b \geq \kappa(\alpha) \sqrt{\mathbf{p}_g^T \tilde{\mathbf{\Sigma}}_+ \mathbf{p}_g} \\ &b - \mathbf{p}_g^T \tilde{\mathbf{u}}_- \geq \kappa(\alpha) \sqrt{\mathbf{p}_g^T \tilde{\mathbf{\Sigma}}_- \mathbf{p}_g}. \end{aligned} \tag{11}$$

Based on (11), Theorem 4 below can be presented for solving the solution variables.

*Theorem 4:* If $\tilde{\mathbf{u}}_+ = \tilde{\mathbf{u}}_-$, then the minimax probability decision problem in (11) is not a meaningful solution: The optimal worst-case misclassification probability that we obtain is $1 - \alpha^* = 1$. Otherwise, an optimal hyperplane $\mathrm{H}(\mathbf{p}_g^*, b^*)$ exists and can be determined by solving the convex optimization problem:

$$\begin{aligned} &\kappa(\alpha)^* = \min_{\mathbf{p}_g} \sqrt{\mathbf{p}_g^T \tilde{\mathbf{\Sigma}}_+ \mathbf{p}_g} + \sqrt{\mathbf{p}_g^T \tilde{\mathbf{\Sigma}}_- \mathbf{p}_g} \\ &\text{s.t.} \mathbf{p}_g^T (\tilde{\mathbf{u}}_+ - \tilde{\mathbf{u}}_-) = 1 \end{aligned} \tag{12a}$$

and setting $b^*$ to the value

$$b^* = (\mathbf{p}_g^*)^T \tilde{\mathbf{u}}_+ - \kappa(\alpha)^* \sqrt{(\mathbf{p}_g^*)^T \tilde{\mathbf{\Sigma}}_+ \mathbf{p}_g^*} \tag{12b}$$

where $\mathbf{p}_g^*$ is the optimal solution of $\mathbf{p}_g$. The optimal worst-case misclassification probability is obtained via

$$\begin{aligned} 1 - \alpha^* &= \frac{1}{1 + (\kappa(\alpha)^*)^2} \\ &= \frac{\left( \sqrt{(\mathbf{p}_g^*)^T \tilde{\mathbf{\Sigma}}_+ \mathbf{p}_g^*} + \sqrt{(\mathbf{p}_g^*)^T \tilde{\mathbf{\Sigma}}_- \mathbf{p}_g^*} \right)^2}{1 + \left( \sqrt{(\mathbf{p}_g^*)^T \tilde{\mathbf{\Sigma}}_+ \mathbf{p}_g^*} + \sqrt{(\mathbf{p}_g^*)^T \tilde{\mathbf{\Sigma}}_- \mathbf{p}_g^*} \right)^2}. \end{aligned} \tag{12c}$$

If either $\tilde{\mathbf{\Sigma}}_+$ or $\tilde{\mathbf{\Sigma}}_-$ is positive definite, the optimal hyperplane $\mathrm{H}(\mathbf{p}_g^*, b^*)$ is unique.

Since (10) can be taken as a special case of (4) as shown in Theorem 1, Lemma 2 and Theorem 3 can be derived by using the same procedure for MPM [9]. To ease the understanding of Lemmas 2 and 3 and Theorem 4, the proof is presented in the Appendix.

Equation (12a) is the second-order core program (SOCP) optimization problem [25] which can be effectively solved by tools such as SeDuMi [26] and a specified algorithm was also proposed in [9] for this problem.

### D. Algorithm

Based on the analysis above, the corresponding learning algorithm of the proposed MP-TSK-FSC is presented below.

**Algorithm of MP-TSK-FSC**

*Step 1:* Use the FCM clustering technique or other partition techniques to determine the antecedents of TSK-FS with (8b) and (8c).

*Step 2:* Construct the dataset in the new feature space mapped by the fuzzy inference rules, $\tilde{D} = \{(\mathbf{x}_{gi}, y_i)\}$, $i = 1, \ldots, N$, where $\mathbf{x}_{gi}$ are obtained by (9a)–(9c).

*Step 3:* Use (12a)–(12c) to solve the consequents parameters $\mathbf{p}_g^*$ of TSK-FS, the decision threshold $b^*$ of the fuzzy classifier and the lower bound of correct classification $\alpha^*$ as the model reliability.

### E. Discussion

The time complexity of the proposed MP-TSK-FSC algorithm is discussed briefly here. The time complexity of Step 1 depends on the adopted clustering technique or partition techniques. For example, if the FCM clustering algorithm is used, the corresponding time complexity is O($N * K * T$), where *N*, *K*, and *T* denote the number of training data, the number of fuzzy rules, and the number of iterations of the FCM algorithm, respectively. The time complexity of Step 2 is O($N * K$) for

constructing the new dataset in the new feature space. The time complexity of Step 3 depends on the adopted SOCP solution algorithm. For example, the SOCP optimization problem in (12a) can be effectively solved by tools such as SeDuMi [26] and a specified algorithm in [9]. In our experimental studies, the SOCP solution in [9] is used and the time complexity is O(*T1*) with T1 as the number of iterations.

The model complexities of the proposed minimax probability-based fuzzy classifier MPM-TSK-FSC and the existing MPM classifier are very different. For an MPM-TSK-FSC with *M* fuzzy rules, the Gaussian membership function-based model trained by the training data with *d*-dimensional inputs will contain *2Md* parameters in the antecedents and $M(d+1)$ parameters in the consequent and a parameter as the decision threshold. Thus, the number of parameters involved in an MPM-TSK-FSC with *M* fuzzy rules are $2Md + M(d+1) + 1$. For a linear MPM classifier, the final model trained with the same training data contains $d+1$ parameters. The model of linear MPM is much simpler, but it only realizes a linear classifier in the original space. For kernelized MPM, the number of parameters involved in the final model is $L + L(d+s) + 1$, which depends on the number of support vectors obtained in the training procedure, i.e., $L$ ($L \leq N$), and the number of parameters in the kernel function, i.e., $s$. For example, if a Gaussian kernel function is adopted, the number of parameters involved in the final model of kernelized MPM is $L + L(d+1) + 1$.

Some additional remarks are given below.

*Remark 1:* It is noted that the purpose of the proposed method is to train a TSK fuzzy classifier with both a high level of interpretability and transparent model reliability, but not to enhance the classification accuracy. Thus, the classification accuracy may be only comparative to the existing methods. However, the model obtained by the proposed method is more transparent to users, which makes the classifier much friendlier and easily acceptable in practical applications, such as medical diagnosis.

*Remark 2:* The proposed algorithm is designed for binary classification. Of course, the multiclassification can be transformed into a combination of many binary classification tasks [55]. Once the above decomposition strategy is adopted for multiclass classification tasks, an average lower bound of correct classification of the obtained models for binary classification tasks can be presented and taken as the model reliability, approximately. In fact, it is more desirable that a lower bound of correct classification can be obtained directly as the model reliability for multiclassification. This is not a trivial task and deserves to be studied in depth in the future.

*Remark 3:* It is interesting that both the minimax probability criterion in the proposed fuzzy classifier and the criterion in the classical SVM aim to maximize the margins. A discussion on the difference and the relationship between them is given as follows. First, two different margin maximization criteria are designed from different views in order to train a classifier with superior generalization abilities. Thus, there is an obvious difference in the physical meanings of both margin maximization criteria. While one is to find the maximal probability lower bound of correct classification for the model, the other is to obtain the maximal geometric margin between the classification hyperplane and the two nearest samples belonging to different classes. Second, it is obvious that the probability margin is more easily understood by users than the geometric margin to observe the model reliability, which means that the former has better interpretability.

*Remark 4:* Although the proposed minimax probability fuzzy classifier is designed based on TSK-FS model in this study, the minimax probability strategy can be extended to some other types of fuzzy system models, such as the ML-fuzzy model [33] and the type-2 fuzzy model [34], [35]. Of course, it is not a trivial thing to address this study, in which many new issues need to be studied in depth. We will address related studies in future.

*Remark 5:* While the proposed fuzzy classifier has a more transparent model than most existing fuzzy classifiers, the interpretability of the proposed one is also enhanced to some extent from the following viewpoints: It is natural that if a fuzzy model is very transparent, the interpretation of the associated fuzzy rules can be understood more confidently. Moreover, for the proposed fuzzy classifier, an additional decision threshold has been introduced as an auxiliary item to enable a final decision to be made. Thus, it can be interpreted as additional information by the experts in related fields.

## IV. Experimental Studies

The proposed MP-TSK-FSC has been evaluated on synthetic datasets and several benchmarking medical datasets and compared with related methods. The experimental studies are organized as follows. In Section IV-A, the experiment settings are described, and the experiment on synthetic datasets is reported in Section IV-B. In Section IV-C, the classification model obtained by the MP-TSK-FSC algorithm is analyzed by using an application to medical diagnosis. Comparative studies on several related methods are reported in Section IV-D.

### A. Experiment Settings

*1) Methods for Comparison:* The proposed MP-TSK-FSC is compared with several related methods, including two mimimax probability-based methods [MPM (linear) and MPM (kernel)], four TSK-FS-based methods (SOTFN-SV, $\varepsilon$-TSK-FS (IQP), $\varepsilon$-TSK-FS (LSSLI), and L2-TSK-FS), and three classical classification methods (KNN, SVC, and Naïve Bayes classifier). The descriptions of these methods are listed in Table I.

As shown in Table I, MP-TSK-FSC, MPM, SOTFN-SV, KNN, SVC, and the Naïve Bayes classifier were developed directly for classification, and the others were originally developed for regression. Different strategies can be adopted for the regression methods to implement classification tasks. In our experiments, we adopted the following simple strategy: class labels are directly used as the outputs of regression datasets for model training. When a future sample is tested, the output of the regression model is compared with different class labels, and the nearest label is taken as the class label of the testing sample.

For kernel technique-based methods, i.e., MPM (kernel) and SVC, the radius basis function (RBF) is adopted as the kernel function due to its effectiveness. For all the fuzzy

TABLE I
METHODS ADOPTED FOR PERFORMANCE COMPARISON

| Method | Description |
|---|---|
| MP-TSK-FSC | The proposed TSK-FS classifier by using minimax probability decision to train the TSK-FS for classification task. |
| MPM(linear) [9] | Linear minimax probability machine by using minimax probability decision. |
| MPM(kernel) [9] | Kernel minimax probability machine by using minimax probability decision. |
| SOTFN-SV [19] | Support vector learning based TSK-type fuzzy neural network, i.e., TSK-fuzzy systems, for classification. |
| $\varepsilon$-TSK-FS(IQP) [22] | $\varepsilon$-insensitive criterion based TSK-FS training method with IQP optimization technique. |
| $\varepsilon$-TSK-FS (LSSLI) [22] | $\varepsilon$-insensitive criterion-based TSK-FS training method with LSSLI optimization technique. |
| L2-TSK-FS [23] | L2 norm penalty and $\varepsilon$-insensitive criterion based TSK-FS training method. |
| KNN [27] | *K*-near neighbor classifier. |
| SVC [28] | Support vector classification. |
| Naïve Bayes classifier [29] | Naïve Bayes classifier. |

TABLE II
HYPERPARAMETERS IN DIFFERENT METHODS AND THE SEARCH GRIDS USED FOR CV

| Method | Description of the hyperparameters and the search grid grid used for cross-validation |
|---|---|
| MP-TSK-FSC | Scale parameter of width in Gaussian membership function: $h \in \{10^{-5}, \ldots, 10^{0}, \ldots, 10^{5}\}$, the number of fuzzy rules: $K \in \{4, 9, 16, 25, 36, 49, 64, 81, 100, 121\}$. |
| MPM(linear) | No hyperparameters |
| MPM(kernel) | RBF kernel width parameter: $\sigma \in \{10^{-12}, \ldots, 10^{0}, \ldots, 10^{12}\}$ |
| SOTFN-SV | Self-organization learning threshold parameter: $\sigma_{th} = \{0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8\}$, regularization parameter for support learning: $C \in \{10^{-12}, \ldots, 10^{0}, \ldots, 10^{12}\}$. |
| $\varepsilon$-TSKFS(IQP) | Scale parameter of width in Gaussian membership function: $h \in \{10^{-5}, \ldots, 10^{0}, \ldots, 10^{5}\}$, the number of fuzzy rules: |
| $\varepsilon$-TSK-FS(LSSLI) | $K \in \{4, 9, 16, 25, 36, 49, 64, 81, 100, 121\}$ |
| L2-TSK-FS | regularization parameter: $\tau \in \{10^{-12}, \ldots, 10^{0}, \ldots, 10^{12}\}$. |
| KNN | The number of near neighbors: $K \in \{1, \ldots, 12\}$ |
| SVC (RBF) | Regularization parameter: $C \in \{10^{-12}, \ldots, 10^{0}, \ldots, 10^{12}\}$; RBF kernel width parameter: $\sigma \in \{10^{-12}, \ldots, 10^{0}, \ldots, 10^{12}\}$. |
| Naïve Bayes classifier | No hyperparameters |

systems-related methods, the commonly used Gaussian function is adopted as the fuzzy membership function in the antecedents.

*2) Datasets:* Four synthetic datasets and three benchmarking medical datasets are adopted for performance evaluation. The three medical datasets are the *epileptic electroencephalograph (EEG)* dataset, *heart disease* dataset, and *breast cancer* dataset. The details of these three medical datasets are described below.

*Epileptic EEG*: The epileptic EEG data used are publicly available on the Web from the University of Bonn, Germany [30]. The complete data archive contains five groups of data (denoted by groups A to E), each containing 100 single-channel EEG segments of 23.6-s duration. The sampling rate of all datasets was 173.6 Hz. Groups A and B consist of segments acquired from surface EEG recordings performed on five healthy volunteer subjects, and groups C, D, and E are data which are obtained from volunteer subjects with epilepsy. In our experiment, groups A and B are used for the healthy class, and groups C–E are used for the patient class. For the *epileptic EEG* data, feature extraction has been conducted by using short-time Fourier transform, and then the data with the five features associated with the energy of different frequency bands are obtained [31].

*Breast cancer*: The breast cancer dataset was obtained from the UCI machine learning repository [32]. It contains 458 instances of the benign class and 241 instances of the malignant class. Each instance is described by nine attributes.

*Heart disease*: The heart disease dataset was also obtained from the UCI machine learning repository [32], which includes 120 instances with heart disease and 150 instances without heart disease. Each instance is described by 13 attributes.

In our experiments, each attribute of the data inputs was normalized into the range $[-1, 1]$ for all datasets.

*3) Parameter Settings:* For all the algorithms, unless specified, the fivefold CV strategy is used to determine the optimal setting within the given grids for the related hyperparameters. The corresponding hyperparameters in different methods and the search grids for CV are listed in Table II.

*4) Evaluation Index:* For the classification task, the following index, i.e., classification accuracy, is used to evaluate the classification performance:

$$J_{clas} = \frac{\text{Number of test samples with correct classification}}{\text{Number of test samples}}. \tag{13}$$

For performance comparison, the means and standard deviations of classification accuracies of different methods under the optimal parameters determined by the CV strategy in the given search grids are reported and compared.

*5) Experimental Environment:* All the algorithms were implemented with the MATLAB codes on a computer with 2-GB RAM and 1.66-GHz CPU.

### B. Synthetic Datasets

In this section, four synthetic datasets, denoted as SD1, SD2, SD3 and SD4, with predetermined class structures are used to evaluate the performance of the proposed minimax probability fuzzy classifier. The parameters used to generate the data are listed in Table III and the generated datasets are shown in Fig. 2. Each dataset contains 600 samples belonging to two different classes, where positive and negative classes are denoted as blue "+" and red "$*$," respectively. The four synthetic datasets have the same means but different covariance matrices for each class. For these synthetic datasets, the covariance matrices of the two classes are adjusted, such that different degrees of correlations could be introduced among the features. From SD1 to SD4, the overlap between the two classes is becoming increasingly severe, which implies that it is more difficult to train a classifier with high generalization abilities.

The performance of the MPM-TSK-FSC with nine rules is reported in Table IV. In particular, the class-wise classification accuracies are also reported by using the classification accuracies of each class, which is used to observe the relationship between the practical classification accuracy of each class and the lower bound of correct classification of the trained model. From the experimental results, we can see that the MPM-TSK-FSC

TABLE III
PARAMETERS USED TO GENERATE THE SYNTHETIC DATASETS

| | SD1 | | SD2 | | SD3 | | SD4 | |
|---|---|---|---|---|---|---|---|---|
| | PC* | NC* | PC | NC | PC | NC | PC | NC |
| Means | [2 20] | [8 20] | [2 20] | [8 20] | [2 20] | [8 20] | [2 20] | [0 20] |
| Covariance | $\begin{bmatrix} 1 & 0 \\ 0 & 5 \end{bmatrix}$ | $\begin{bmatrix} 1 & 0 \\ 0 & 5 \end{bmatrix}$ | $\begin{bmatrix} 1 & 0 \\ 2 & 5 \end{bmatrix}$ | $\begin{bmatrix} 1 & 0 \\ -2 & 5 \end{bmatrix}$ | $\begin{bmatrix} 1 & 0 \\ 4 & 5 \end{bmatrix}$ | $\begin{bmatrix} 1 & 0 \\ -4 & 5 \end{bmatrix}$ | $\begin{bmatrix} 1 & 0 \\ 6 & 5 \end{bmatrix}$ | $\begin{bmatrix} 1 & 0 \\ -6 & 5 \end{bmatrix}$ |
| Size | 600 | | 600 | | 600 | | 600 | |

*PC and NC denote positive class and negative class, respectively.

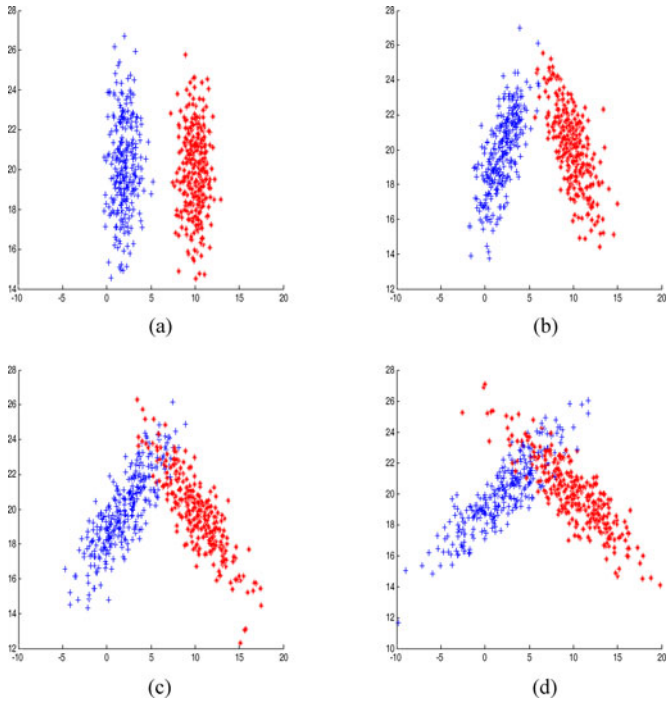



Fig. 2. Four synthetic datasets: (a) SD1; (b) SD2; (c) SD3 and (d) SD4.

can give not only promising classification results but the lower bound of correct classification as well, i.e., $\alpha$, in Table IV. The lower bound of correct classification of the obtained model is particularly significant to users, as it enables users to clearly know the reliability of the adopted model for predicting results.

In addition to the transparent model reliability, the proposed fuzzy classifier can still inherit good interpretability from the fuzzy system. In particular, an additional threshold decision, i.e., $b$, can be provided to further enhance the interpretability of the obtained classification model. More detailed analyses of the proposed fuzzy classifier model will be given by using an application to medical diagnosis in Section IV-C.

### *C. Model Analysis of the Minimax Probability Takagi–Sugeno–Kang Fuzzy System Classifier for Medical Diagnosis*

In this section, the trained MP-TSK-FSC model is analyzed to show its characteristics by using a real world application to medical diagnosis. In Table V, the MP-TSK-FSC with nine rules trained in a certain time on the *epileptic EEG* dataset is presented.

The constructed MP-TSK-FSC contains three parts, as shown in Table V. We explain these parts as follows.

1) The first part is the fuzzy rules base as shown in Part A of Table V, which is used for fuzzy inference and presents a final real value as the TSK-FS output. With the fuzzy rules base, the fuzzy inference rules can be linguistically interpretable with expert knowledge.
2) The second part presents a decision threshold, which is introduced for the classification task in MP-TSK-FSC. The decision threshold and the consequents of the TSK-FS are learned based on the minimax probability decision principle. With the real output of the trained TSK-FS and the decision threshold, the final decision can be given for the classification task.
3) The third part provides the reliability of the trained classification model, where the reliability is characterized by the lower bound of correct classification for the trained fuzzy classifier.

In Fig. 3, the corresponding membership functions of all fuzzy subsets in the antecedent of the second fuzzy rule are shown. Each membership function corresponds to a fuzzy subset, which can be explained by the medical expert in medical terms and with medical knowledge. Fig. 3 shows that all fuzzy sets seem to have a very small width here. The explanation is as follows. As shown in (8c), the width of the fuzzy membership function is obtained by $\delta_i^k = h \cdot \Delta_i^k$ and $\Delta_i^k = \left( \sum_{j=1}^{N} u_{jk}(x_{ji} - c_i^k)^2 \Big/ \sum_{j=1}^{N} u_{jk} \right)$. While $\Delta_i^k$ can be computed with the clustering results of FCM, $h$ is a hyperparameter and needs to be adjusted with a certain strategy. In our experiments, the optimal $h$ has been determined by using a CV strategy within the given search grid of this parameter, as shown in Table II. Since the determined value for $h$ by the CV strategy in our experiment on the Epileptic EEG dataset is small, it makes the corresponding width $\delta_i^k$ much smaller for the fuzzy sets in the antecedents of the fuzzy rules accordingly.

It is also noted that since each specialist may have his own understanding for a given fuzzy membership function, the explanation of the derived fuzzy rules from different specialists will vary. Thus, only a potential explanation for the derived fuzzy rules can be given. For example, the fuzzy subsets in the second fuzzy rule can be expressed with the following linguistic description from the viewpoint of a certain medical expert

TABLE IV
PERFORMANCE ON FOUR SYNTHETIC DATASETS OBTAINED BY MP-TSK-FSC WITH NINE FUZZY RULES

| | | SD1 | | | SD2 | | | SD3 | | | SD4 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | All$^+$ | PC$^+$ | NC$^+$ | All | PC | NC | All | PC | NC | All | PC | NC |
| $J_{clas}$ | Mean | 1 | 1 | 1 | 0.9950 | 0.9934 | 0.9966 | 0.9581 | 0.9674 | 0.9483 | 0.9233 | 0.9298 | 0.9177 |
| | std | 0 | 0 | 0 | 0.0075 | 0.0089 | 0.0076 | 0.0247 | 0.0285 | 0.0346 | 0.0210 | 0.0165 | 0.0442 |
| $\alpha*$ | *Mean* | | 0.9662 | | | 0.9111 | | | 0.8684 | | | 0.8494 | |
| | *std* | | 0.0187 | | | 0.0078 | | | 0.0113 | | | 0.0085 | |

$^*\alpha$ denotes the lower bound of correct classification of the trained classification model. $^+$ "All," "PC," and "NC" denote classification accuracies of all test data, test data of positive class, and test data of negative class, respectively.

TABLE V
MP-TSK-FSC WITH NINE RULES TRAINED IN A CERTAIN TIME ON THE EPILEPTIC EEG DATASET

**Part A: Fuzzy rules base**

TSK Fuzzy Rule *Rk*:
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 $f_k(\mathbf{x}) = p_{k0} + p_{k1}x_1 + \cdots + p_{kd}x_d$.

| **No. of rules** | **Antecedent parameters** (Gaussian membership function parameters) | **Consequent parameters** (linear function parameters) |
|---|---|---|
| $k$ | $\mathbf{c}^k = (c_1^k, \ldots, c_d^k)^T, \boldsymbol{\delta}^k = (\delta_1^k, \ldots, \delta_d^k)^T$ | $\mathbf{p}_k = (p_{k0}, p_{k1}, \ldots, p_{kd})^T$ |
| 1 | $\mathbf{c}^1$ = [0.6274, 0.6622, 0.7091, 0.5682, 0.5147, −0.4565]<br>$\boldsymbol{\delta}^1$ = [1.99e-05, 1.44e-05, 1.23e-05, 2.22e-05, 2.606e-05 , 1.10e-05] | $\mathbf{p}_1$ = [0.4213, −1.0270, 0.5407, −0.0451 1.4879, −1.4751, 0.2835] |
| 2 | $\mathbf{c}^2$ = [−0.0586, 0.7156, 0.6102, 0.3266, 0.3973, −0.2281]<br>$\boldsymbol{\delta}^2$ = [2.11e-05, 1.87e-05, 2.11e-05, 1.85e-05, 1.48e-05, 6.54e-06] | $\mathbf{p}_2$ = [6.6533, –1.4751, −0.0907, 0.4213, −1.0270, 0.1655, −0.0451] |
| 3 | $\mathbf{c}^3$ = [0.5304, −0.1475, 0.4119, 0.5431, 0.9242, −0.9219]<br>$\boldsymbol{\delta}^3$ = [3.57e-05, 3.70e-05, 3.80e-05, 3.45e-05, 3.65e-05, 3.90e-05 ] | $\mathbf{p}_3$ = [–1.0270, 0.1655, 0.3608, 1.4879, −1.4751, 0.2835, 0.4213] |
| 4 | $\mathbf{c}^4$ = [0.4461, 0.0854, 0.2350, 0.1030, 0.0985, −0.8656]<br>$\boldsymbol{\delta}^4$ = [2.43e-05, 1.92e-05, 1.45e-05, 1.38e-05, 1.55e-05, 1.09e-05] | $\mathbf{p}_4$ = [−1.4751, 0.2835, 0.4213, −1.0270, 0.1655, −0.0451, 1.4879 ] |
| 5 | $\mathbf{c}^5$ = [−0.6666, −0.5807, −0.6221, −0.7431, −0.8156, −0.3132]<br>$\boldsymbol{\delta}^5$ = [1.62e-05, 1.201e-05, 1.44e-05, 1.40e-05, 1.11e-05, 1.88e-05] | $\mathbf{p}_5$ = [0.1655, −0.0451, 1.4879, −1.4751, 0.2835, 0.4213, −1.0270] |
| 6 | $\mathbf{c}^6$ = [−0.2397, −0.5931, −0.3827, −0.4711, −0.5062, −0.9506]<br>$\boldsymbol{\delta}^6$ = [2.74e-05, 1.78e-05, 2.47e-05, 2.23e-05, 2.50e-05, 7.49e-06] | $\mathbf{p}_6$ = [0.2835, 0.6366, −1.0270, 0.1655, −0.0451, 1.4879, −1.4751] |
| 7 | $\mathbf{c}^7$ = [−0.2844, 0.3527, 0.2058, −0.0151, −0.2410, −0.2229]<br>$\boldsymbol{\delta}^7$ = [1.62e-05, 1.602e-05, 1.31e-05,9.73e-06, 1.30e-05, 3.71e-06] | $\mathbf{p}_7$ = [−0.0451, 1.4281, −1.4751, 0.2835, 0.4213, −1.0270, 0.1655 ] |
| 8 | $\mathbf{c}^8$ = [−0.5508, −0.5821, 0.0715, −0.2276, −0.3817, 0.3099]<br>$\boldsymbol{\delta}^8$ = [1.60e-05, 1.39e-05, 1.94e-05, 1.08e-05, 1.40e-05, 1.30e-05] | $\mathbf{p}_8$ = [0.4213, −0.8599, 0.1655, −0.0451, 1.4879, −1.4751, 0.2835 ] |
| 9 | $\mathbf{c}^9$ = [0.5521, 0.6120, 0.9045, 0.8486, 0.8170, 0.7474]<br>$\boldsymbol{\delta}^9$ = [ 1.61e-05, 1.01e-05, 2.83e-06, 1.23e-05, 1.07e-05, 2.00e-05] | $\mathbf{p}_9$ = [1.4879, −1.2547, 0.2835, 0.4213, −1.0270, 0.1655, −0.0451 ] |

**Part B: Decision threshold for classification**

Decision threshold of MP-TSK-FSC: $b$ = –0.2693

**Part C: Reliability of the classification model**

Lower bound of correct classification: $\alpha$ =0.8816

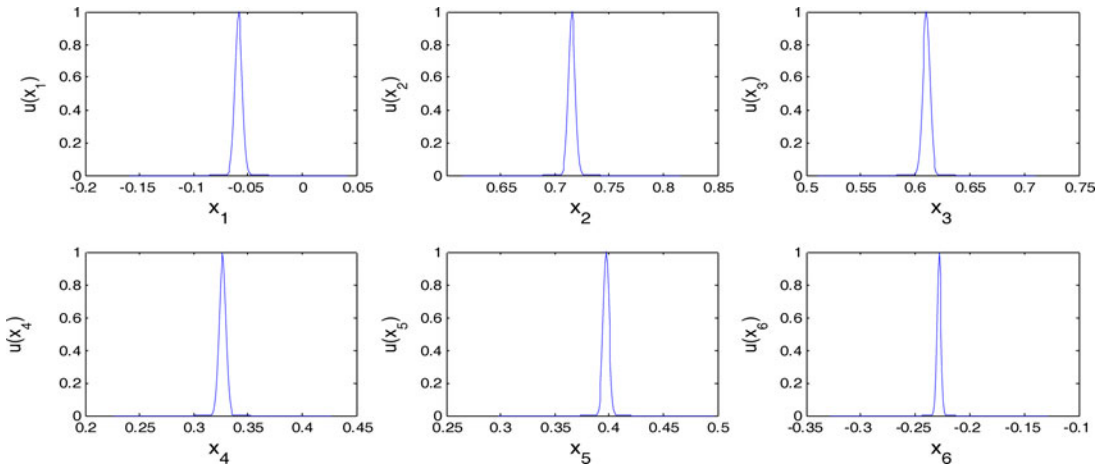



Fig. 3. Corresponding membership functions of each fuzzy subset in the antecedent of the second fuzzy rule.

for EEG signal recognition. (Note that the energy of the EEG signal in the different frequency bands below has been scaled into interval $[-1, 1]$.)

*The second fuzzy rule:*

If the energy of the EEG signal in the frequency band 1 is slightly small,

and if the energy of the EEG signal in the frequency band 2 is rather large,

and if the energy of the EEG signal in the frequency band 3 is much larger,

and if the energy of the EEG signal in the frequency band 4 is large,

and if the energy of the EEG signal in the frequency band 5 is a little larger,

and if the energy of the EEG signal in the frequency band 6 is small,

then this rule gives the decision with the following formula:

$$f_2(\mathbf{x}) = 6.6533 - 1.4751x_1 - 0.0907x_2 + 0.4213x_3 - 1.0270x_4 + 0.1655x_5 - 0.0451x_6.$$

In the above fuzzy rule, each fuzzy subset is linguistically described by the amount of energy in the corresponding frequency band and all the fuzzy subsets are joined with "and." In the consequent of this rule, the simple linear function is directly used as the evaluation formula. Furthermore, from Table V, two conclusions about the trained fuzzy classifier are given below.

1) Decision threshold of MP-TSK-FSC is –0.2693.
2) Lower bound of correct classification, i.e., the model reliability presented for MP-TSK-FSC, is 88.16%.

In particular, the above conclusions also enhance the interpretability of the obtained fuzzy classifier to some extent. While the lower bound of correct classification ensures that users are more confident about the decision results, the decision threshold provides additional information for the specialists to further analyze the diagnosis results. From the analysis above, we can observe that MP-TSK-FSC is a highly interpretable expert system with transparent model reliability, which is very suitable for many practical applications, especially for medical diagnosis.

### D. Comparison With Related Methods

In this section, the classification performance of the proposed method is compared with several related methods, as described in Section IV-A1. For all classification methods, the classification accuracies are reported, and the obtained lower bounds of correct classification for the trained models are also provided for the three minimax probability decision-based methods. In addition, the class-wise classification accuracies of the proposed MPM-TSK-FSC method are provided by using the classification accuracies of each class, in order to observe the relationship between the practical classification accuracies of different classes and the lower bound of correct classification of the trained model. Although the purpose of the proposed method is to enhance the transparency of the classifier and not to improve the classification accuracy, the classification accuracy is also compared with the related method in order to evaluate its generalization abilities. In Tables VI–VIII, the means and standard deviations of the classification accuracies of different methods are presented, which are obtained on three medical datasets under the optimal parameter setting determined by the CV strategy. From these results, we reveal the following observations.

1) The proposed MP-TSK-FSC shows high completive generalization abilities compared with existing state-of-the-art methods.
2) Of all the fuzzy system-based methods, i.e., MP-TSK-FSC, SOTFN-SV, $\varepsilon$-TSK-FS(IQP), $\varepsilon$-TSK-FS(LSSLI) and L2- TSK-FS, although they all have a high level of interpretability, only the proposed MP-TSK-FSC can present the reliability of the trained model to users.
3) Of the three minimax probability-based methods, i.e., MP-TSK-FSC, MPM (linear) and MPM (RBF), the generalization abilities of MP-TSK-FSC are better than that of MPM (linear) and are equivalent to that of MPM (RBF). However, compared with MPM (RBF), MP-TSK-FSC has the following obvious advantage: while MP-TSK-FSC is more transparent to the users and has a high level of interpretability, MPM (RBF) is more like a black box since it corresponds to a hyperplane in an unknown kernel feature space to the users.
4) When compared with the classical classification methods, such as SVC and KNN, the proposed MP-TSK-FSC demonstrates more advantages, including a) highly competitive or better generalization abilities, b) a high level of interpretability, and c) transparent model reliability.

Furthermore, the model complexities of the adopted methods in our experimental studies are compared. To save space, only the models trained on the breast cancer dataset are compared here. The number of parameters in different models obtained in the case that the best generalization abilities have been obtained by the CV strategy is compared in Table IX. While the number of model parameters for MP-TSK-FSC, MPM (linear), and MPM (RBF) are analyzed in Section III-E, the number of model parameters for the other methods are described briefly as follows: 1) For $\varepsilon$-TSK-FS(IQP), $\varepsilon$-TSK-FS(LSSLI), and L2-TSK-FS, the number of model parameters for the obtained fuzzy system with $M$ fuzzy rules trained by the data with $d$-dimensional inputs are $2Md + M(d + 1)$. For SOTFN-SV, an addition threshold is introduced, and thus finally the number of model parameters is $2Md + M(d + 1) + 1$.2) For SVC(RBF), the number of parameters involved in the final model is $L + L(d + s) + 1$, which depends on the number of support vectors involved, i.e., $L$, and the number of parameters in the RBF kernel function, i.e., $s = 1$ here. 3) For KNN, the number of parameters involved in the final model is $Nd + Kd$, where $N$ and $K$ are the number of training data and near neighbors, respectively. 4) For the Naïve Bayes classifier, if the density distribution for each dimension is Gaussian, the number of parameters involved in the final model is $C(ds + 1)$, where $C$ is the number of classes and $s$ is the number of parameters in the Gaussian distribution function. From Table IX, we can see that when the CV strategy is used to determine the hyperparameters for different classifiers, the model complexities of these classifiers obtained based on the breast cancer dataset are very different. The model complexity of the proposed fuzzy classifier is in the middle of the adopted ten methods.

TABLE VI
PERFORMANCE COMPARISON OF SEVERAL METHODS ON THE EPILEPTIC EEG DATASET

| | | MP-TSK-FSC$^+$ | | | MPM (linear) | MPM (RBF) | SOTFN-SV | $\varepsilon$-TSK-FS (IQP) |
|---|---|---|---|---|---|---|---|---|
| | | All | PC | NC | | | | |
| $J_{\text{clas}}$ | Mean | 0.9660 | 0.9600 | 0.9700 | 0.9480 | 0.9600 | 0.9660 | 0.9620 |
| | std | 0.0350 | 0.0652 | 0.0492 | 0.0277 | 0.0346 | 0.0296 | 0.0370 |
| $\alpha*$ | Mean | | 0.8700* | | 0.7724* | 0.8024* | | |
| | std | | 0.0216$^+$ | | 0.0105$^+$ | 0.0200$^+$ | | |
| | | | $\varepsilon$-TSK-FS (LSSLI) | | L2-TSK-FS | SVC (RBF) | KNN | Naïve Bayes |
| $J_{\text{clas}}$ | Mean | | 0.9680 | | 0.9200 | 0.9560 | 0.9580 | 0.9480 |
| | std | | 0.0286 | | 0.0430 | 0.0364 | 0.0311 | 0.0432 |

*$\alpha$ denotes the lower bound of correct classification of the trained classification model. $^+$ "All," "PC," and "NC" denote classification accuracies of all test data, test data of positive class, and test data of negative class, respectively.

TABLE VII
PERFORMANCE COMPARISON OF SEVERAL METHODS ON THE BREAST CANCER DATASET

| | | MP-TSK-FSC$^+$ | | | MPM (linear) | MPM (RBF) | SOTFN-SV | $\varepsilon$-TSK-FS (IQP) |
|---|---|---|---|---|---|---|---|---|
| | | All | PC | NC | | | | |
| $J_{clas}$ | Mean | 0.9715 | 0.9675 | 0.9792 | 0.9685 | 0.9715 | 0.9728 | 0.9629 |
| | std | 0.0166 | 0.0263 | 0.0208 | 0.0128 | 0.01665 | 0.0116 | 0.0144 |
| $\alpha^*$ | Mean | | 0.8412 | | 0.8362 | 0.8576 | | |
| | std | | 0.0063 | | 0.0046 | 0.0063 | | |
| | | | $\varepsilon$-TSK-FS (LSSLI) | | L2-TSK-FS | SVC, (RBF) | KNN | Naïve Bayes |
| $J_{clas}$ | Mean | | 0.9658 | | 0.9200 | 0.9560 | 0.9580 | 0.9480 |
| | std | | 0.0220 | | 0.0430 | 0.0364 | 0.0311 | 0.0432 |

*$\alpha$ denotes the lower bound of correct classification of the trained classification model. $^+$ "All," "PC," and "NC" denote classification accuracies of all test data, test data of positive class, and test data of negative class, respectively.

TABLE VIII
PERFORMANCE COMPARISON OF SEVERAL METHODS ON THE HEART DISEASE DATASET

| | | MP-TSK-FSC$^+$ | | | MPM (linear) | MPM (RBF) | SOTFN-SV | $\varepsilon$-TSK-FS (IQP) |
|---|---|---|---|---|---|---|---|---|
| | | All | PC | NC | | | | |
| $J_{clas}$ | *Mean* | 0.8481 | 0.8533 | 0.8333 | 0.8296 | 0.8222 | 0.8037 | 0.7852 |
| | *std* | 0.0576 | 0.0803 | 0.1102 | 0.0479 | 0.0465 | 0.0712 | 0.0483 |
| $\alpha^*$ | *Mean* | | 0.6018 | | 0.5515 | 0.5561 | | |
| | *std* | | 0.0326 | | 0.0180 | 0.0182 | | |
| | | | $\varepsilon$-TSK-FS (LSSLI) | | L2- TSK-FS | SVC (RBF) | KNN | Naïve Bayes |
| $J_{clas}$ | *Mean* | | 0.8259 | | 0.8481 | 0.8222 | 0.8148 | 0.8222 |
| | *std* | | 0.0663 | | 0.0576 | 0.0426 | 0.0571 | 0.0384 |

*$\alpha$ denotes the lower bound of correct classification of the trained classification model. $^+$ "All," "PC," and "NC" denote classification accuracies of all test data, test data of positive class, and test data of negative class, respectively.

## V. CONCLUSION

In this study, a minimax probability TSK-FS classifier was proposed to train a fuzzy system-based classifier and to provide the model reliability of the trained classifier simultaneously. For the proposed MP-TSK-FSC, a lower bound of correct classification can be presented to the users. Thus, the final TSK-FS classifier has the distinctive characteristics of both a high level of interpretability and transparent model reliability.

Although the proposed minimax probability classifier has shown promising performance, there are still many aspects that deserve further investigation. For example, other minimax probability decision-based fuzzy system models, such as the ML-fuzzy model [33] and the type-2 fuzzy model [34], [35], can be studied for classification tasks. In addition, minimax probability-based fuzzy systems can also be investigated for other modeling tasks, such as outlier detection and regression. These issues will be addressed in our future study.

TABLE IX
MODEL COMPLEXITIES OF THE CLASSIFIERS OBTAINED BASED ON THE BREAST CANCER DATASET WITH DIFFERENT METHODS

| | MP-TSK-FSC | SOTFN-SV | $\varepsilon$-TSK-FS (IQP) | $\varepsilon$-TSK-FS (LSSLI) | L2-TSK-FS |
|---|---|---|---|---|---|
| Number of rules* | 16 | 214 | 9 | 4 | 36 |
| Number of parameters | $16 \times (2 \times 9) + 16 \times (9 + 1) + 1 = 449$ | $214 \times (2 \times 9) + 214 \times (9 + 1) + 1 = 5993$ | $9 \times (2 \times 9) + 9 \times (9 + 1) = 252$ | $4 \times (2 \times 9) + 4 \times (9 + 1) = 84$ | $36 \times (2 \times 9) + 36 \times (9 + 1) = 1008$ |

| MPM (linear) | | MPM (RBF) | | SVC (RBF) | |
|---|---|---|---|---|---|
| Number of parameters | $9 + 1 = 10$ | Number of support vectors | 266 | Number of support vectors | 81 |
| | | Number of parameters | $266 + 266 \times (9 + 1) + 1 = 2927$ | Number of parameters | $81 + 81 \times (9 + 1) + 1 = 8921$ |

| KNN | | Naïve Bayes classifier | |
|---|---|---|---|
| Number of near neighbors# | 5 | Number of parameters | $2\,(9 \times 2 + 1)=38$ |
| Number of parameters | $(216^{\#} + 5) \times 9 = 1989$ | | |

* The number of fuzzy rules and number of near neighbors are all determined by the CV strategy to obtain the optimal generalization abilities. # The number of near neighbors is determined by using the CV strategy.

## APPENDIX A
## PROOF OF LEMMAS 2 AND 3

The second condition in (10), i.e.,

$$\inf_{\mathbf{x}_g \sim (\tilde{\mathbf{u}}_-, \tilde{\mathbf{\Sigma}}_-)} pr(\mathbf{p}_g^T \mathbf{x}_g - b \le 0) \ge \alpha \tag{A1}$$

can be equivalently written as

$$\sup_{\mathbf{x}_g \sim (\tilde{\mathbf{u}}_-, \tilde{\mathbf{\Sigma}}_-)} pr(\mathbf{p}_g^T \mathbf{x}_g - b \ge 0) \le 1 - \alpha. \tag{A2}$$

According to Marshall and Olkin's result [36], as discussed in [37], i.e.,

$$\begin{aligned} \sup_{\mathbf{x} \sim (\mathbf{u}_\mathbf{x}, \mathbf{\Sigma}_\mathbf{x})} pr\{\mathbf{x} \in S\} &= \frac{1}{1 + d^2},\ d^2 \\ &= \inf_{\mathbf{x} \in S} (\mathbf{x} - \mathbf{u}_\mathbf{x})^T \mathbf{\Sigma}_\mathbf{x} (\mathbf{x} - \mathbf{u}_\mathbf{x}) \end{aligned} \tag{A3}$$

with $S$ as a convex set, (A2) can be further expressed as

$$\begin{aligned} &\sup_{\mathbf{x}_g \sim (\tilde{\mathbf{u}}_-, \tilde{\mathbf{\Sigma}}_-)} pr\{\mathbf{x}_g \in S\} = \tfrac{1}{1+d^2} \le 1 - \alpha, \\ &d^2 = \inf_{\mathbf{x}_g \in S} (\mathbf{x}_g - \tilde{\mathbf{u}}_-)^T \tilde{\mathbf{\Sigma}}_- (\mathbf{x}_g - \tilde{\mathbf{u}}_-) \end{aligned} \tag{A4}$$

where $S = \{\mathbf{p}_g^T \mathbf{x}_g - b \ge 0\}$. From (A4), we have

$$d^2 \ge \frac{\alpha}{1 - \alpha}. \tag{A5}$$

Consider $d^2$ in (A4). If $\mathbf{p}_g^T \tilde{\mathbf{u}}_- - b \ge 0$, $d^2 = 0$; otherwise, $d^2 = \frac{(b - \mathbf{p}_g^T \tilde{\mathbf{u}}_-)^2}{\mathbf{p}_g^T \tilde{\mathbf{\Sigma}}_- \mathbf{p}_g}$. Hence

$$\begin{aligned} d^2 &= \inf_{\mathbf{x}_g \in S} (\mathbf{x}_g - \tilde{\mathbf{u}}_-)^T \tilde{\mathbf{\Sigma}}_- (\mathbf{x}_g - \tilde{\mathbf{u}}_-) \\ &= \max \left( \frac{(b - \mathbf{p}_g^T \tilde{\mathbf{u}}_-)^2}{\mathbf{p}_g^T \tilde{\mathbf{\Sigma}}_- \mathbf{p}_g}, 0 \right). \end{aligned} \tag{A6}$$

For $\mathbf{p}_g^T \tilde{\mathbf{u}}_- - b \ge 0$, we will get $d^2 = 0$ and $\alpha = 0$, which is unmeaning. Thus, here only $\mathbf{p}_g^T \tilde{\mathbf{u}}_- - b \le 0$ is considered. If $\mathbf{p}_g^T \tilde{\mathbf{u}}_- - b \le 0$,

$$d^2 = \frac{(b - \mathbf{p}_g^T \tilde{\mathbf{u}}_-)^2}{\mathbf{p}_g^T \tilde{\mathbf{\Sigma}}_- \mathbf{p}_g}. \tag{A7}$$

By substituting (A7) into (A5), we have

$$d^2 = \frac{(b - \mathbf{p}_g^T \tilde{\mathbf{u}}_-)^2}{\mathbf{p}_g^T \tilde{\mathbf{\Sigma}}_- \mathbf{p}_g} \ge \frac{\alpha}{1 - \alpha}. \tag{A8}$$

Let $\kappa(\alpha) = \sqrt{\frac{\alpha}{1-\alpha}}$; (A8) becomes

$$d^2 = \frac{(b - \mathbf{p}_g^T \tilde{\mathbf{u}}_-)^2}{\mathbf{p}_g^T \tilde{\mathbf{\Sigma}}_- \mathbf{p}_g} \ge (\kappa(\alpha))^2 \tag{A9}$$

i.e., if $\mathbf{p}_g^T \tilde{\mathbf{u}}_- - b \le 0$

$$b - \mathbf{p}_g^T \tilde{\mathbf{u}}_- \ge \kappa(\alpha) \sqrt{\mathbf{p}_g^T \tilde{\mathbf{\Sigma}}_- \mathbf{p}_g}. \tag{A10}$$

In a similar way, we can prove that the first condition in (10), i.e.,

$$\inf_{\mathbf{x}_g \sim (\tilde{\mathbf{u}}_+, \tilde{\mathbf{\Sigma}}_+)} \Pr(\mathbf{p}_g^T \mathbf{x}_g - b \ge 0) \ge \alpha \tag{A11}$$

is equivalent to

$$\mathbf{p}_g^T \tilde{\mathbf{u}}_+ - b \ge \kappa(\alpha) \sqrt{\mathbf{p}_g^T \tilde{\mathbf{\Sigma}}_+ \mathbf{p}_g} \tag{A12}$$

when $\mathbf{p}_g^T \tilde{\mathbf{u}}_+ - b \ge 0$ is considered. Thus, Lemmas 2 and 3 are proved.

## APPENDIX B
## PROOF OF THEOREM 4

Recall from (11) that

$$\max_{\mathbf{p}_g, b, \alpha} \alpha$$

$$s.t.\ \mathbf{p}_g^T\tilde{\mathbf{u}}_+ - b \geq \kappa(\alpha)\sqrt{\mathbf{p}_g^T\tilde{\mathbf{\Sigma}}_+\mathbf{p}_g}$$
$$b - \mathbf{p}_g^T\tilde{\mathbf{u}}_- \geq \kappa(\alpha)\sqrt{\mathbf{p}_g^T\tilde{\mathbf{\Sigma}}_-\mathbf{p}_g}. \quad \text{(B1)}$$

Since $\kappa(\alpha) = \sqrt{\frac{\alpha}{1-\alpha}}$ is monotonically increasing with $\alpha$, (B1) can be written as

$$\max_{\mathbf{p}_g, b, \kappa(\alpha)} \kappa(\alpha)$$
$$s.t. \mathbf{p}_g^T\tilde{\mathbf{u}}_+ - b \geq \kappa(\alpha)\sqrt{\mathbf{p}_g^T\tilde{\mathbf{\Sigma}}_+\mathbf{p}_g}$$
$$b - \mathbf{p}_g^T\tilde{\mathbf{u}}_- \geq \kappa(\alpha)\sqrt{\mathbf{p}_g^T\tilde{\mathbf{\Sigma}}_-\mathbf{p}_g}. \quad \text{(B2)}$$

From (B2), we can see that when $\kappa(\alpha)$ approaches to the maximum $\kappa(\alpha)^*$, $\mathbf{p}_g^T\tilde{\mathbf{u}}_+ - \kappa(\alpha)\sqrt{\mathbf{p}_g^T\tilde{\mathbf{\Sigma}}_+\mathbf{p}_g} = \kappa(\alpha)\sqrt{\mathbf{p}_g^T\tilde{\mathbf{\Sigma}}_-\mathbf{p}_g} + \mathbf{p}_g^T\tilde{\mathbf{u}}_-$ must hold, and the optimal bias $b^*$ can be obtained by the following equation:

$$b^* = \mathbf{p}_g^T\tilde{\mathbf{u}}_+ - \kappa(\alpha)\sqrt{\mathbf{p}_g^T\tilde{\mathbf{\Sigma}}_+\mathbf{p}_g}$$
$$= \kappa(\alpha)\sqrt{\mathbf{p}_g^T\tilde{\mathbf{\Sigma}}_-\mathbf{p}_g} + \mathbf{p}_g^T\tilde{\mathbf{u}}_- \quad \text{(B3)}$$

Since the optimal bias can be obtained by the maximum $\kappa(\alpha)^*$ and maximizing $\kappa(\alpha)^*$ can be independent of $b$, (B2) may be equivalent to the following optimization problem:

$$\max_{\mathbf{p}_g, \kappa(\alpha)} \kappa(\alpha)$$
$$s.t.\ \mathbf{p}_g^T\tilde{\mathbf{u}}_+ - \kappa(\alpha)\sqrt{\mathbf{p}_g^T\tilde{\mathbf{\Sigma}}_+\mathbf{p}_g} \geq \kappa(\alpha)\sqrt{\mathbf{p}_g^T\tilde{\mathbf{\Sigma}}_-\mathbf{p}_g} + \mathbf{p}_g^T\tilde{\mathbf{u}}_- \quad \text{(B4)}$$

i.e.,

$$\max_{\mathbf{p}_g, \kappa(\alpha)} \kappa(\alpha)$$
$$\text{s.t. } \frac{\mathbf{p}_g^T(\tilde{\mathbf{u}}_+ - \tilde{\mathbf{u}}_-)}{\left(\sqrt{\mathbf{p}_g^T\tilde{\mathbf{\Sigma}}_-\mathbf{p}_g} + \sqrt{\mathbf{p}_g^T\tilde{\mathbf{\Sigma}}_+\mathbf{p}_g}\right)} \geq \kappa(\alpha). \quad \text{(B5)}$$

Thus, (B5) can be transformed as the following equivalent problem:

$$\kappa(\alpha)^* = \min_{\mathbf{p}_g} \frac{\mathbf{p}_g^T(\tilde{\mathbf{u}}_+ - \tilde{\mathbf{u}}_-)}{\left(\sqrt{\mathbf{p}_g^T\tilde{\mathbf{\Sigma}}_-\mathbf{p}_g} + \sqrt{\mathbf{p}_g^T\tilde{\mathbf{\Sigma}}_+\mathbf{p}_g}\right)}. \quad \text{(B6)}$$

If $\tilde{\mathbf{u}}_+ = \tilde{\mathbf{u}}_-$, then $\mathbf{p}_g = \mathbf{0}$ implies $\kappa(\alpha)^* = 0$, which in turn yields $\alpha^* = 0$. In this case, the minimax probability decision problem (11) does not have a meaningful solution, and the optimal worst-case misclassification probability is $1 - \alpha^* = 1$. Let us proceed with the assumption $\tilde{\mathbf{u}}_+ \neq \tilde{\mathbf{u}}_-$. We observe that condition (B6) is positively homogeneous in $\mathbf{p}_g$. If $\mathbf{p}_g$ satisfies (B6), $s \cdot \mathbf{p}_g$ with $s \geq 0$ does as well. Furthermore, (B6) implies $\mathbf{p}_g^T(\tilde{\mathbf{u}}_+ - \tilde{\mathbf{u}}_-) \geq 0$. Since $\tilde{\mathbf{u}}_+ \neq \tilde{\mathbf{u}}_-$, we can set $\mathbf{p}_g^T(\tilde{\mathbf{u}}_+ - \tilde{\mathbf{u}}_-) = 1$ without loss of generality. This implies $\mathbf{p}_g \neq \mathbf{0}$, and in turn, $\sqrt{\mathbf{p}_g^T\tilde{\mathbf{\Sigma}}_-\mathbf{p}_g} + \sqrt{\mathbf{p}_g^T\tilde{\mathbf{\Sigma}}_+\mathbf{p}_g} \neq 0$. Thus, we can write the optimization problem as

$$\kappa(\alpha)^* = \min_{\mathbf{p}_g} \frac{1}{\left(\sqrt{\mathbf{p}_g^T\tilde{\mathbf{\Sigma}}_-\mathbf{p}_g} + \sqrt{\mathbf{p}_g^T\tilde{\mathbf{\Sigma}}_+\mathbf{p}_g}\right)}$$
$$\text{s.t. } \mathbf{p}_g^T(\tilde{\mathbf{u}}_+ - \tilde{\mathbf{u}}_-) = 1 \quad \text{(B7)}$$

i.e.,

$$\kappa(\alpha)^* = \max_{\mathbf{p}_g} \sqrt{\mathbf{p}_g^T\tilde{\mathbf{\Sigma}}_-\mathbf{p}_g} + \sqrt{\mathbf{p}_g^T\tilde{\mathbf{\Sigma}}_+\mathbf{p}_g}$$
$$\text{s.t. } \mathbf{p}_g^T(\tilde{\mathbf{u}}_+ - \tilde{\mathbf{u}}_-) = 1. \quad \text{(B8)}$$

Let $\mathbf{p}_g^*$ as the optimal solution of $\mathbf{p}_g$, with $k(\alpha)^*$, $\alpha^*$ can be computed by

$$\alpha^* = \frac{(\kappa(\alpha)^*)^2}{1 + (\kappa(\alpha)^*)^2}$$
$$= 1 - \frac{\left(\sqrt{(\mathbf{p}_g^*)^T\tilde{\mathbf{\Sigma}}_+\mathbf{p}_g^*} + \sqrt{(\mathbf{p}_g^*)^T\tilde{\mathbf{\Sigma}}_-\mathbf{p}_g^*}\right)^2}{1 + \left(\sqrt{(\mathbf{p}_g^*)^T\tilde{\mathbf{\Sigma}}_+\mathbf{p}_g^*} + \sqrt{(\mathbf{p}_g^*)^T\tilde{\mathbf{\Sigma}}_-\mathbf{p}_g^*}\right)^2} \quad \text{(B9)}$$

and

$$1 - \alpha^* = \frac{\left(\sqrt{(\mathbf{p}_g^*)^T\tilde{\mathbf{\Sigma}}_+\mathbf{p}_g^*} + \sqrt{(\mathbf{p}_g^*)^T\tilde{\mathbf{\Sigma}}_-\mathbf{p}_g^*}\right)^2}{1 + \left(\sqrt{(\mathbf{p}_g^*)^T\tilde{\mathbf{\Sigma}}_+\mathbf{p}_g^*} + \sqrt{(\mathbf{p}_g^*)^T\tilde{\mathbf{\Sigma}}_-\mathbf{p}_g^*}\right)^2}. \quad \text{(B10)}$$

Thus, Theorem 4 is proved.

## REFERENCES


[1] A. K. Jain, R. P. W. Duin, and J. Mao, "Statistical pattern recognition: A review," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 1, pp. 4–37, Jan. 2000.

[2] C. M. Bishop, *Neural Networks for Pattern Recognition*. London, U.K.: Oxford Univ. Press, 1995.

[3] S. K. Pal and M. Sushmita, *Neuro-Fuzzy Pattern Recognition: Methods in Soft Computing*. Hoboken, NJ, USA: Wiley, 1999.

[4] V. Khatibi and G. A. Montazer, "Intuitionistic fuzzy set vs. fuzzy set application in medical pattern recognition," *Artif. Intell. Med.*, vol. 47, no. 1, pp. 43–52, 2009.

[5] J. M. Mendel, *Uncertain Rule-Based Fuzzy Logic Systems: Introduction and new Directions*. Upper Saddle River, NJ, USA: Prentice-Hall, 2001.

[6] J. S. R. Jang, C. T. Sun, and E. Mizutani, *Neuro-Fuzzy and Soft-Computing*. Upper Saddle River, NJ, USA: Prentice-Hall, 1997.

[7] L. X. Wang, *Adaptive Fuzzy Systems and Control: Design and Stability Analysis*. Upper Saddle River, NJ, USA: Prentice-Hall, 1994.

[8] G. R. G. Lanckriet, L. E. Ghaoui, and M. I. Jordan, "Robust novelty detection with single-class MPM," in *Advances in Neural Information Processing Systems*, vol. 15, Cambridge, MA, USA: MIT Press, 2002.

[9] G. R. G. Lanckriet, L. E. Ghaoui, C. Bhattacharyya, and M. I. Jordan, "A robust minimax approach to classification," *J. Mach. Learning Res.*, vol. 3, pp. 555–582, 2003.

[10] K. Z. Huang, H. Q. Yang, I. King, M. R. Lyu, and L. Chan, "Minimum error minimax probability machine," *J. Mach. Learning Res.*, vol. 5, pp. 1253–1286, 2004.

[11] K. Z. Huang, H. Q. Yang, I. King, and M. R. Lyu, "Imbalanced learning with biased minimax probability machine," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 36, no. 4, pp. 913–923, Aug. 2006.

[12] T. R. Strohmann, A. Belitski, G. Z. Grudic, and D. M. DeCoste, "Sparse greedy minimax probability machine classification," *Proc. Neural Inf. Process.*, vol. 16, 105 pp., 2003.

[13] Z. H Deng, F. L. Chung, and S. T. Wang, "A novel minimax probability based fuzzy hyper-ellipsoid machine," presented at the Int. Joint Conf. Neural Netw., Orlando, FL, USA, 2007.

[14] T. Strohmann and G. Z. Grudic, "A formulation for minimax probability machine regression," *Neural Inf. Process. Syst.: Nat. Synth.*, pp. 769–776, 2002.

[15] T. Takagi and M. Sugeno, "Fuzzy identification of systems and its application to modeling and control," *IEEE Trans. Syst., Man Cybern.*, vol. SMC-15, no. 1, pp. 116–132, Jan./Feb. 1985.

[16] E. H. Mamdani, "Application of fuzzy logic to approximate reasoning using linguistic synthesis," *IEEE Trans. Comput.*, vol. C-26, no. 12, pp. 1182–1191, Dec. 1977.

[17] M. F. Azeem, M. Hanmandlu, and N. Ahmad, "Generalization of adaptive neural-fuzzy inference systems," *IEEE Trans. Neural Netw.*, vol. 11, no. 6, pp. 1332–1346, Nov. 2000.
[18] H. Ishibuchi and T. Nakashima, "Effect of rule weights in fuzzy rule-based classification systems," *IEEE Trans. Fuzzy Syst.*, vol. 9, no. 4, pp. 506–515, Aug. 2001.
[19] C. F. Juang, S. H. Chiu, and S. W. Chang, "A self-organizing TS-type fuzzy network with support vector learning and its application to classification problems," *IEEE Trans. Fuzzy Syst.*, vol. 15, no. 5, pp. 998–1008, Oct. 2007.
[20] C. F. Juang and S. J. Shiu, "Using self-organizing fuzzy network with support vector learning for face detection in color images," *Neurocomputing*, vol. 71, no. 16, pp. 3409–3420, 2008.
[21] G. D. Wu and P. H. Huang, "A maximizing-discriminability-based self-organizing fuzzy network for classification problems," *IEEE Trans. Fuzzy Syst.*, vol. 18, no. 2, pp. 362–373, Apr. 2010.
[22] J. Leski, "TSK-fuzzy modeling based on $\varepsilon$-insensitive learning," *IEEE Trans. Fuzzy Syst.*, vol. 13, no. 2, pp. 181–193, Apr. 2005.
[23] Z. H. Deng, K. S. Choi, F. L. Chung, and S. T. Wang, "Scalable TSK fuzzy modeling for very large datasets using minimal-enclosing-ball approximation," *IEEE Trans. Fuzzy Syst.*, vol. 19, no. 2, pp. 210–226, Apr. 2011.
[24] 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 Netw. Learning Syst.*, vol. 24, no. 8, pp. 1200–1212, Aug. 2013.
[25] F. Alizadeh and D. Goldfarb, "Second-order cone programming," *Math. Program.*, vol. 95, no. 1, pp. 3–51, 2004.
[26] J. Sturm, "Using SeDuMi 1.02, A MATLAB toolbox for optimization over symmetric cones," *Spec. Issue Interior Point Methods, Optim. Methods Softw.*, vols. 11/12, pp. 625–653, 1999.
[27] R. Duda, P. Hart, and D. Stork, *Pattern Classification*. New York, NY, USA: Wiley, 2000.
[28] C. Cortes and V. Vapnik, "Support vector networks," *Mach. Learning*, vol. 20, no. 3, pp. 273–297, 1995.
[29] D. Grossman and P. Domingos, "Learning Bayesian network classifiers by maximizing conditional likelihood," in *Proc. ACM 21st Int. Conf. Mach. Learning*, 2004, pp. 361–368.
[30] R. G. Andrzejak, K. Lehnertz, F. Mormann, C. Rieke, P. David, and C. Elger, "Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: Dependence on recording region and brain state," *Phys. Rev. E*, vol. 64, no. 6, 061907, 2001.
[31] K. S. Choi, Y. Zeng, and J. Qin, "Using sequential floating forward selection algorithm to detect epileptic seizure in EEG signals," in *Proc. 11th Int. Conf. Signal Process.*, Beijing, China, Oct. 21–25, 2012, pp. 1637–1640.
[32] K. Bache and M. Lichman. (2013). "UCI machine learning repository," Sch. Inform. Comput. Sci., Univ. Calif., Irvine, CA, USA. [Online]. Available: http://archive.ics.uci.edu/ml
[33] Z. H. Deng, Y. Z. Jiang, F. L. Chung, H. Ishibuchi, and S. T. Wang, "Knowledge-leverage based fuzzy system and its modeling," *IEEE Trans. Fuzzy Syst.*, vol. 21, no. 4, pp. 597–609, Aug. 2013.
[34] Q. Liang and J. M. Mendel, "Interval type-2 fuzzy logic systems: Theory and design," *IEEE Trans. Fuzzy Syst.*, vol. 8, no. 5, pp: 535–550, Oct. 2000.
[35] J. M. Mendel, "Type-2 fuzzy sets and systems: An overview," *IEEE Comput. Intell. Mag.*, vol. 2, pp. 20–29, 2007.
[36] A. W. Marshall and I. Olkin, "Multivariate chebyshev inequalities," *Ann. Math. Statist.*, vol. 31, pp. 1001–1014, 1960.
[37] S. Boyd and L. Vandenberghe, "Convex optimization. Course notes for EE364," Stanford Univ., 2005.
[38] K. Igor, I. Bratko, and M. Kukar, "Application of machine learning to medical diagnosis," *Mach. Learning Data Mining: Methods Appl.*, vol. 389, 408, 1997.
[39] K. Igor, "Machine learning for medical diagnosis: History, state of the art and perspective," *Artif. Intell. Med.*, vol. 23, no. 1, pp. 89–109, 2001.
[40] M. I. Chacon-Murguia, O. Arias-Enriquez, and R. Sandoval-Rodriguez, "A fuzzy scheme for gait cycle phase detection oriented to medical diagnosis," in *Pattern Recognition* (Lecture Notes in Computer Science). Berlin, Germany: Springer, 2013, vol. 7914, pp. 20–29.
[41] J. C. Obi and A. A. Imianvan, "Fuzzy neural approach for colon cancer prediction," *Sci. Africana*, vol. 11, no. 1, pp. 65–76, 2012.
[42] I. Morsi, A. El Gawad, and Y. Zakria, "Fuzzy logic in heart rate and blood pressure measuring system," in *Proc. Sens. Appl. Symp.*, 2013, pp. 113–117.
[43] K. P. Adlassnig, "Fuzzy set theory in medical diagnosis," *IEEE Trans. Syst., Man Cybern.*, vol. SMC-16, no. 2, pp. 260–265, Mar. 1986.
[44] I. Gadaras and L. Mikhailov, "An interpretable fuzzy rule-based classification methodology for medical diagnosis," *Artif. Intell. Med.*, vol. 47, no. 1, pp. 25–41, 2009.
[45] H. Ishibuchi and Y. Takashi, "Rule weight specification in fuzzy rule-based classification systems," *IEEE Trans. Fuzzy Syst.*, vol. 13, no. 4, pp. 428–435, Aug. 2005.
[46] K. Nozaki, H. Ishibuchi, and H. Tanaka, "Adaptive fuzzy rule-based classification systems," *IEEE Trans. Fuzzy Syst.*, vol. 4, no. 3, pp. 238–250, Aug. 1996.
[47] H. Ishibuchi and T. Nakaskima, "Improving the performance of fuzzy classifier systems for pattern classification problems with continuous attributes," *IEEE Trans. Ind. Electron.*, vol. 46, no. 6, pp. 1057–1068, Dec. 1999.
[48] S. Suresh and K. Subramanian, "A sequential learning algorithm for meta-cognitive neuro-fuzzy inference system for classification problems," *Appl. Soft Comput.*, vol. 12, no. 11, pp. 3603–3614, 2012.
[49] H.-J. Rong, N. Sundararajan, G.-B. Huang, and P. Saratchandran, "Sequential adaptive fuzzy inference system (SAFIS) for nonlinear system identification and prediction," *Fuzzy Sets Syst.*, vol. 157, pp. 1260–1275, 2006.
[50] P. P. Angelov and D. P. Filev, "An approach to online identification of Takagi-Sugeno fuzzy models," *IEEE Trans. Syst., Man Cybern., B, Cybern.*, vol. 34, no. 1, pp. 484–498, Feb. 2004.
[51] J. de Jesus Rubio, "SOFMLS: Online self-organizing fuzzy modified least-squares network," *IEEE Trans. Fuzzy Syst.*, vol. 17, no. 6, pp. 1296–1309, Dec. 2009.
[52] J. Casillas, O. Cordon, F. Herrera, and L. Magdalena, *Interpretability Issues in Fuzzy Modeling*. Berlin, Germany: Springer-Verlag, 2003.
[53] E. Lughofer, "On-line assurance of interpretability criteria in evolving fuzzy systems achievements, new concepts and open issues," *Inf. Sci.*, vol. 251, pp. 22–46, 2013.
[54] M. J. Gacto, R. Alcala, and F. Herrera, "Interpretability of linguistic fuzzy rule-based systems: An overview of interpretability measures," *Inf. Sci.*, vol. 181, no. 20, pp. 4340–4360, 2011.
[55] E. Lughofer and O. Buchtala, "Reliable all-pairs evolving fuzzy classifiers," *IEEE Trans. Fuzzy Syst.*, vol. 21, no. 4, pp. 625–641, Aug. 2013.
[56] E. Lughofer, "Single-pass active learning with conflict and ignorance," *Evolving Syst.*, vol. 3, no. 4, pp. 251–271, 2012.
[57] R. Babuska, *Fuzzy Modeling for Control*. Norwell, MA, USA: Kluwer, 1998.
[58] W. Pedrycz and F. Gomide, *Fuzzy Systems Engineering: Toward Human-Centric Computing*. Hoboken, NJ, USA: Wiley, 2007.

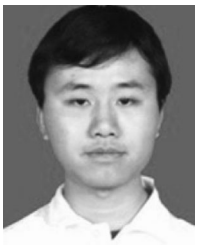
**Zhaohong Deng** (M'12–SM'14) received the B.S. degree in physics from Fuyang Normal College, Fuyang, China, in 2002 and the Ph.D. degree in light industry information technology and engineering from Jiangnan University, Wuxi, China, in 2008.

He is currently an Associate Professor with the School of Digital Media, Jiangnan University, and a Visiting Associate Researcher with the University of California, Davis, CA, USA. His current research interests include computational intelligence and pattern recognition. He is the author or coauthor of more than 50 research papers in international/national journals.

**Longbing Cao** (SM'06) received the Ph.D. degree in intelligent sciences and another in computing science.

He is currently a Professor of information technology with the University of Technology Sydney (UTS), Sydney, Australia, where he is the Founding Director of the Advanced Analytics Institute. He is also the Research Leader of the Data Mining Program with the Australian Capital Markets Cooperative Research Centre and the Chair of IEEE Task Force on Behavior and Social Informatics and of the IEEE Task Force on Educational Data Mining. He has served as an Associate Editor and Guest Editor on many journals. He has published two monographs, four edited books, 15 proceedings, 11 book chapters, and around 170 journal/conference publications, including the International Joint Conference on Artificial Intelligence, the ACM SIGKDD Conference on Knowledge Discovery and Data Mining, the International Conference on Data Engineering, the IEEE International Conference on Data Mining series, the Annual International Conference on Autonomous Agents and Multiagent Systems, the International World Wide Web Conference, and several IEEE TRANSACTIONS in the above areas.

Dr. Cao is a Senior Member of the IEEE Systems, Man, and Cybernetics and Computer Societies.



**Yizhang Jiang** (M'12) is currently pursuing the Ph.D. degree with the School of Digital Media, Jiangnan University, Wuxi, China.

He has been a Research Assistant with the Computing Department, Hong Kong Polytechnic University, Hong Kong, for almost one year. He has published several papers in international journals, including the IEEE TRANSACTIONS ON FUZZY SYSTEMS and the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS. His research interests include pattern recognition, intelligent computation, and their applications.

**Shitong Wang** received the M.S. degree in computer science from the Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 1987.

He has visited London University and Bristol University in U.K.; Hiroshima International University in Japan; Hong Kong University of Science and Technology; and Hong Kong Polytechnic University as a Research Scientist, for over six years. He is currently a Full Professor with the School of Digital Media, Jiangnan University, Wuxi, China. He has published about 80 papers in international/national journals and has authored seven books. His research interests include artificial intelligence, neuro-fuzzy systems, pattern recognition, and image processing.