A Study on Relationship Between Generalization Abilities and Fuzziness of Base Classifiers in Ensemble Learning

Xi-Zhao Wang, *Fellow, IEEE*, Hong-Jie Xing, *Member, IEEE*, Yan Li, *Member, IEEE*, Qiang Hua, *Member, IEEE*, Chun-Ru Dong, *Member, IEEE*, and Witold Pedrycz, *Fellow, IEEE*

***Abstract*—We investigate essential relationships between generalization capabilities and fuzziness of fuzzy classifiers (viz., the classifiers whose outputs are vectors of membership grades of a pattern to the individual classes). The study makes a claim and offers sound evidence behind the observation that higher fuzziness of a fuzzy classifier may imply better generalization aspects of the classifier, especially for classification data exhibiting complex boundaries. This observation is not intuitive with a commonly accepted position in "traditional" pattern recognition. The relationship that obeys the conditional maximum entropy principle is experimentally confirmed. Furthermore, the relationship can be explained by the fact that samples located close to classification boundaries are more difficult to be correctly classified than the samples positioned far from the boundaries. This relationship is expected to provide some guidelines as to the improvement of generalization aspects of fuzzy classifiers.**

***Index Terms*—Classification, decision boundary, fuzziness, fuzzy classifier, generalization.**

I. Introduction

CLASSIFICATION refers to a task of assigning objects to one of several predefined class labels and is one of the most pervasive problems in data mining and pattern recognition. The input to the classification scheme is a certain object (pattern) to be labeled, and each object is typically described by a set of attributes. More formally, the classification problem is about determining (estimating) a target function F that maps each object to a class label y. Then, finding this estimate is completed through a process of learning. Learning is usually completed by minimizing some error between F and its estimate f (classifier) on training samples.Wu *et al*. [1] list the top-ten learning algorithms in data mining.

For the evaluation of a learning algorithm, generalization is the most important index because the ultimate goal of learning is to reduce the testing error on unseen samples and produce an accurate prediction. In statistical learning theory, generalization originally refers to the model's ability to well generalize the results obtained from the training set to a set of unseen samples drawn from the distribution same as that of the training set [2]. Following this way, in the literature, one can highlight many studies on the generalization abilities of classifiers being expressed from different points of view.

1) *Generating training/testing sample set:* Focusing on the relation between the training and testing sets, much research investigates a way how to generate training and testing samples so that they directly affect the evaluation output of generalization performance. This type of studies includes resampling methods [3]–[6], leave-one-out cross-validation [7]–[9] approaches to assuming a specific distribution of testing samples and correspondingly developing generalization error formulation [10]–[14], online learning models on samples coming from a dependent source of data [15], etc.
2) *Estimating error bounds:* From references, one can find a number of theoretical studies on the estimation of generalization error bounds, for example, the discussion on the performance bounds to overcome overfitting problems [16], structural risk minimization to link the generalization to the error on training samples and the classifier complexity [17], [18], necessary and sufficient conditions on the number of required training examples [19], the theoretical analysis for classifier ensemble bounds [20], [21], the biased regularization approach to computing the generalization bound [22], and the bounds on the false and truth positive rates based on a VC-style analysis [23].
3) *Relating diversity to generalization:* The relationship between generalization ability and diversity of learning strategies in coevolutionary learning systems is investigated in [25] and [26]. The diversity in the population

Manuscript received February 11, 2014; revised June 9, 2014 and August 19, 2014; accepted October 18, 2014. Date of publication November 20, 2014; date of current version October 2, 2015. This work was supported by the National Natural Science Fund of China under Grant 61170040 and Grant 71371063) and by the Hebei NSF under Grant F2013201110, Grant F2013201060, Grant F2014201100, and Grant ZD2010139.

X.-Z. Wang is with the College of Computer Science and Software, Shenzhen University, Shenzhen 518060, China (e-mail: xizhaowang@ieee.org).

H.-J. Xing and Y. Li are with the College of Mathematics and Information Science, Hebei University, Baoding 071002, China (e-mail: hjxing@hbu.edu.cn; ly@hbu.edu.cn).

Q. Hua is with the College of Mathematics and Information Science, Hebei University, Baoding 071002, China, and also with the Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China (e-mail: huaq@hbu.edu.cn).

C.-R. Dong is with the College of Mathematics and Information Science, Hebei University, Baoding 071002, China, and also with the South China University of Technology, Guangzhou 510640, China (e-mail: dongcr@hbu.edu.cn).

W. Pedrycz is with the Department of Electrical & Computer Engineering, University of Alberta, Edmonton T6R 2V4 AB, Canada, and with Department of Electrical and Computer Engineering, Faculty of Engineering, King Abdulaziz University Jeddah, 21589, Saudi Arabia, and also with Systems Research Institute, Polish Academy of Sciences Warsaw, Poland (e-mail: wpedrycz@ualberta.ca).

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.2014.2371479

is shown to have the potential in improving the generalization performance of coevolutionary learning if the coevolved strategies can be combined, for example, the diversity measure near training samples based on the outputted sensitivity of its member neural networks [24] and the random initializations of network architecture's impact on the generalization ability [27]. This randomness can be understood as a manifestation of diversity of the neural network structure.

Most of the above-mentioned studies on generalization focus on some specific types of classifiers such as neural networks and support vector machines (SVMs). In contrast, this paper offers a study on generalization aspects from a different angle, which does not limit us to a certain types of classifiers but focus on any classifier with fuzzy outputs. We study the issue of measuring and improving the generalization ability by discovering the relationship between the generalization and the uncertainty of the outputs of the classifier. The nature of this uncertainty, i.e., fuzziness, is quantified by fuzzy entropy of the output vectors, and then, it is related to the generalization ability for different types of classifiers with fuzzy vector output (see, e.g., [28]–[32]).

In the literature, except [33] and [34], we have not found other studies on the generalization considered from the perspective of fuzziness of classifier outputs. The works in [33] and [34], however, do not analyze the relationship between the generalization and the fuzziness of the classifier outputs, and their proposed methods are limited only to rule-based systems.

In this paper, for any classifier with fuzzy vector outputs (fuzzy classifier), we attempt to associate fuzziness with the generalization performance of a classifier. Our idea can be explained as follows. Suppose that there are two trained classifiers: Classifier A and Classifier B. If we process training samples by them, two groups of output vectors can be obtained, respectively, from A and B, where each element in the output vector indicates the membership with which the input sample belongs to the corresponding class. Consider a case that the outputted result of Classifier A is the same as that of B for each sample using the maximum membership grade, i.e., the two classifiers cannot be distinguished by their training accuracies, but the fuzziness of classifier A's output vector is different from that of classifier B's for samples. Assume that classifier A has lower fuzziness level than classifier B. The question is: Which classifier will we select? Intuitively, we prefer the classifier having lower fuzziness since this classifier contains less uncertainty and can classify samples more profoundly. In contrast with this opinion, we show that this intuitive view is not always true. For some types of classification problems, the classifier with larger level of fuzziness, i.e., classifier B, may achieve better generalization performance. Therefore, the answer to the previous question can vary. For some types of classification problems, when the classifiers have the same or similar training accuracies, we would like to select one or several with largest fuzziness. Furthermore, this idea can be explained via boundary samples that are considered to have a key impact on the classifier performance [36], [37]. This is essentially consistent with the idea of AdaBoost [38], which assigns heavier weights to the training samples that are hard to train.

Our main contributions in this paper include the following.

1) The establishment of statistical relationship between boundary samples and fuzziness of the samples' outputs. It is demonstrated that boundary samples' outputs given by a classifier have higher fuzziness.
2) The first attempt to investigate the relationship between classifier's fuzziness and the generalization ability. For some certain types of problems, when the classifiers obtain similar training accuracies, higher fuzziness implies higher generalization ability.
3) The finding that samples with higher fuzziness exhibit higher risk of misclassification, which leads to a divide-and-conquer handling strategy of classification.

This paper is organized as follows. Section II describes the classification boundaries and their side effect on classification performance. Section III introduces the fuzziness of a classifier and analyzes the relationship between fuzziness and boundary samples. Section IV discusses the relationship between generalization and fuzziness of a classifier and provides the experimental verifications. Section V gives conclusions and further discussions of this research.

II. Classification Boundary

The theme of this study is to investigate the relationship between generalization of a classifier and fuzziness of the classifier's output. One way to investigate this relation is the analysis on boundary points including their fuzziness and their classification performance. This section will discuss the classifier's generalization from viewpoint of boundary points and their side effect on classification performance.

A. Boundary and Its Estimation Given by a Leaned Classifier

Generally, a hypersurface in n-dimensional space can partition the input space into disjoint subsets called decision regions, and each region has points (samples) belonging to the same class. Decision boundary usually refers to the hypersurface between decision regions with different classes. In many real classification problems, the real decision boundary objectively exists but is usually unknown. One purpose of learning for the classification problem is to find an approximation of the real boundary such that the difference between the real boundary and its estimation is as small as possible. The difference between the real boundary and its estimated boundary is called approximation error, and a training algorithm is required to find the estimated boundary. An ideal algorithm tries to make the error equal to zero, but practically it is impossible. The estimated boundary is usually acquired based on a classifier, which is trained from a set of training samples according to a training algorithm.

Theoretically, the estimated boundary can be determined if the classifier has been trained well. It means that we can obtain the class label of each sample in our considered area if the classifier has been well trained. For some classifiers, the mechanism to obtain the label of a sample is clear. In these cases, the estimated boundary is explicitly expressed by a certain formula. A

simple illustration to indicate these cases is the linear boundary of decision.

Consider a binary classification task with $y = \pm 1$ labels. When the training samples are linearly separable, we can set the parameters of a linear classifier so that all the training samples are classified correctly. Let $\mathbf{w}$ denote a vector orthogonal to the decision boundary, and b denote a scalar offset term; then, we can write the decision boundary as

$$\mathbf{w}^T \mathbf{x} + b = 0. \tag{1}$$

A typical case of (1) is the classifier given by SVMs for linearly separable samples. It is easy to judge whether a sample is near to or far from the boundary. The distance between a sample and the boundary is computed as $\left|\mathbf{w}^T \mathbf{x} + b\right|$. A certain threshold value imposed on the distance can be used to judge whether a sample is near to or far from the boundary.

Some classifiers do not have a clear mechanism to obtain the class label for each sample. In other words, we can use the trained classifier to calculate the label for each sample, but the pertinent formula cannot be provided explicitly. One example of this case is the Bayes decision boundary [38]. Given a sample $\mathbf{x}$, a prior probability $P(y_i)$ of class, and the conditional probability $p(\mathbf{x}|y_i)$, we convert the prior probability to the posteriori probability$P(y_i|\mathbf{x})$ through Bayes' theorem. The Bayes' decision rule reads as

$$\begin{cases} \mathbf{x} \in \text{class } y_1, \text{ if } P(y_1)p(\mathbf{x}|y_1) > P(y_2)p(\mathbf{x}|y_2) \\ \mathbf{x} \in \text{class } y_2, \text{ if } P(y_1)p(\mathbf{x}|y_1) < P(y_2)p(\mathbf{x}|y_2). \end{cases} \tag{2}$$

This decision boundary for a two-class problem can be determined by the point locus $\{\mathbf{x}|P(y_1)p(\mathbf{x}|y_1) - P(y_2)p(\mathbf{x}|y_2) = 0\}$, which is difficult to be explicitly expressed as a formula except for few certain special data distributions.

Another example of this situation is the fuzzy *K*-nearest neighbor (*K*-NN) classifier [31], which outputs a vector of class membership. Each component of the vector is a number in [0,1], representing a membership of the sample belonging to the corresponding class. If the components are equal to either 0 or 1, then it degrades to the traditional *K*-NN. Fuzzy *K*-NN acquires the membership of a sample $\mathbf{x}$ by the formula

$$\mu_i(\mathbf{x}) = \frac{\sum_{j=1}^{K} \mu_{ij} \left\|\mathbf{x} - \mathbf{x}_j\right\|^{-2(m-1)}}{\sum_{j=1}^{K} \left\|\mathbf{x} - \mathbf{x}_j\right\|^{-2(m-1)}} \tag{3}$$

where $(\mu_1(\mathbf{x}), \mu_2(\mathbf{x}), \ldots, \mu_c(\mathbf{x}))^T$ is a membership vector (and the other symbols remain to be specified in next subsection). Its decision boundary is the locus $\{\mathbf{x}|\mu_1^*(\mathbf{x}) = \mu_2^*(\mathbf{x})\}$, where $\{\mu_1^*(\mathbf{x}), \mu_2^*(\mathbf{x}), \ldots, \mu_c^*(\mathbf{x})\}$ is a permutation of $\{\mu_1(\mathbf{x}), \mu_2(\mathbf{x}), \ldots, \mu_c(\mathbf{x})\}$ in a decreasing order. Obviously, it is impossible to explicitly express the classification boundary.

These two examples indicate that it is difficult to judge whether a sample is near to or far from the boundary when the boundary cannot be explicitly expressed as a formula.

Due to the difference between classifier design objectives, the estimated boundary is dependent strongly on the selection of classifier for the same training set. The difference between the estimated boundary and the real boundary is considered as a key index to evaluate the generalization performance of a classifier. From references (e.g., [39]), one can find the study on the classifier design according to the estimated boundary. The good estimated decision boundary could give an insight into the high-performance classifier design, which cannot be supplied by accuracy only. It can be applied to select proper classifiers, to discover possible overfitting, and to calculate the similarity among the models generated by different classifiers (see [39]).

For a well-trained classifier with high performance, it is reasonable to believe that the estimated boundary has sufficiently approximated the real boundary, but since the real boundary is unknown, it is hard to judge which one is better based only on estimated boundaries. Therefore, there is a need to find a new index to measure the generalization. Perhaps, the ability of a classifier correctly classifying boundary samples is a crucial index.

B. Two Types of Methods for Training a Classifier

Usually, there are two types of classifiers: one can explicitly give the analytic formula of the estimated decision boundary, while the other cannot but provide the approximation by locus of some points. SVM and fuzzy *K*-NN are two typical representatives of the two types of classifiers, respectively.

SVMs select a boundary according to the maximization of margin, which is based on the statistical learning theory [40]. SVM supposes an implicit function φ mapping the data from the input space X into a high-dimensional feature space F. The mapping is associated with a kernel function $K(\mathbf{x}_i, \mathbf{x}_j)$, which satisfies $K(\mathbf{x}_i, \mathbf{x}_j) = \langle \varphi(\mathbf{x}_i), \varphi(\mathbf{x}_j) \rangle$, where $\mathbf{x}_i$ and $\mathbf{x}_j$, respectively, denote the *i*th and *j*th training samples, and $\langle \cdot, \cdot \rangle$ denotes the inner product. The decision boundary is explicitly given by

$$f(\mathbf{x}) = \operatorname{sgn}\left(\sum_{i=1}^{N} \alpha_i y_i K(\mathbf{x}, \mathbf{x}_i) + b\right) \tag{4}$$

where α_i and b are unknown parameters that are determined by solving a quadratic programming. From [41], one can find several open SVM tools such as LIBSVM [42]. For a more detailed description of SVM, see [40].

The fuzzy *K*-NN classifier [31] considers fuzzy classification problems and assigns each unseen sample $\mathbf{x}$ with a membership vector (grades), which can be determined by using the neighbors' class memberships and computing the distances between $\mathbf{x}$ and its *K*-NNs. For every training sample, fuzzy *K*-NN assumes that the class information has been given by the memberships of the sample belonging to the predefined classes. Let $(\mu_1(\mathbf{x}), \mu_2(\mathbf{x}), \ldots, \mu_c(\mathbf{x}))^T$ denote the output vector which fuzzy *K*-NN outputs for an unseen sample $\mathbf{x}$, where $\mu_i(\mathbf{x}) \in [0, 1]$ is the membership of $\mathbf{x}$ belonging to the ith class. $\mu_i(\mathbf{x})$ is given by formula (3), where $\mathbf{x}_j \in X$ is a labeled training sample that falls in the set of *K*-NNs of the unseen sample $\mathbf{x}$, $(\mu_{ij})_{j=1,2,\cdots,K} \in [0, 1]$ is the known class membership of $\mathbf{x}_j$ to the *i*th class y_i, $\|\mathbf{x} - \mathbf{x}_j\|$ is the distance between $\mathbf{x}$ and $\mathbf{x}_j$, and m is a parameter to adjust the weights that indicate neighbors' contribution to the membership value. As the parameter m is increasing, the neighbors are more evenly weighted, and their relative distances from the sample being classified have less ef-

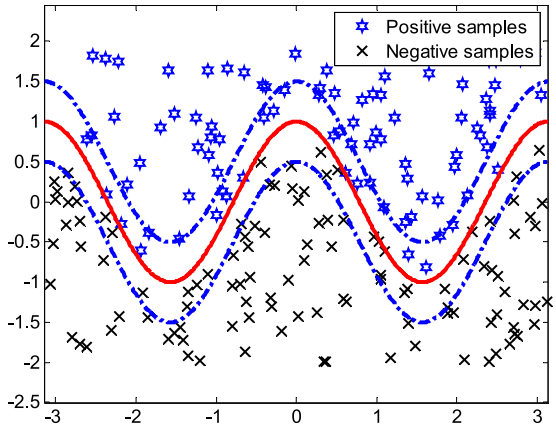

Fig. 1. Simple two class data and its boundary.

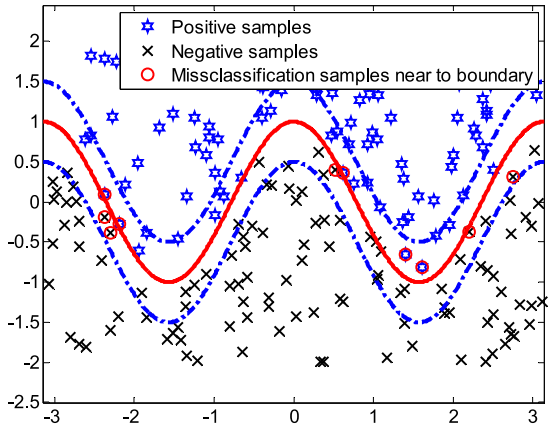

Fig. 2. Relationship between the real boundary and the classification error.

fect. In the experiments in Sections III and IV, without otherwise specified, we set $m = 2$ and train 49 fuzzy *K*-NN classifiers by varying the value of *K* from 2 to 50 with a step size of 1. The impact of different values of *m* on the classifier performance will also be discussed in Section IV.

It is worth noted that if we only know the class label of each training sample, i.e., μ_{ij} is equal to either 0 or 1, then formula (3) will degenerate and can be considered as the weighted summation of samples in the *K*-NNs that belong to the *i*th class, where the weight is the inverse of the distance between the sample and its neighbors.

In this paper, we focus on the second type of classifiers considered above.

C. Side Effect of Boundary and Experimental Verification

It is experimentally observed that for classification problems with continuous attributes in supervised learning, a sample near to boundary usually has the statistical testing error higher than a sample positioned far from boundary. Here, the boundary refers originally to real one, which is replaced by an estimated boundary encountered in real-world problems. We call this phenomenon a side effect of boundary. The classifier we use in this study to estimate the boundary is fuzzy *K*-NN. As mentioned in Section II-A, judging whether a sample is located near to or far from the boundary is more difficult for fuzzy *K*-NN than for other classifiers, which have an explicit expression of the boundary.

The following simple simulation confirms the side effect phenomenon for a known boundary. Consider a two-class problem in the xy plane, and suppose that the real boundary is given by the function $y = \cos(2x)$ via the following rule: A sample (x, y) is considered as positive if $y > \cos(2x)$ and negative if $y < \cos(2x)$. The boundary is shown in Fig. 1. Uniformly, we select 200 samples from the rectangular area $\{(x, y)| -\pi < x < \pi,\ -2 < y < 2\}$ to form a sample set from which we randomly select 70% as the training set. Since the real boundary is known, we artificially split the entire set of all samples as two categories: samples near to boundary $\{(x, y)||y - \cos(2x)| < 0.5\}$ and sample far from boundary $\{(x, y)||y - \cos(2x)| > 0.5\}$. Using fuzzy *K*-NN ($K = 5$) to train a classifier and then apply it to classify the samples near to and far from the boundary, respectively, we have the experimental result that the classification error rate for samples near to the boundary is 20%, while the error rate for samples far from the boundary is zero. Fig. 2 clearly shows the experimental result.

TABLE I
DATASETS USED IN EXPERIMENTS

Databases	N_{sample}	N_{cat}	N_{con}	N_{class}
Banknote	*1372*	*0*	*4*	*2*
Blood	*748*	*0*	*4*	*2*
Breast Cancer	*263*	*9*	*0*	*2*
Cleverland Heart	*297*	*7*	*6*	*2*
Diabetes	*768*	*8*	*0*	*2*
Flare Solar	*144*	*9*	*0*	*2*
German	*1000*	*7*	*13*	*2*
Glass	*214*	*10*	*0*	*6*
Heart	*270*	*7*	*6*	*2*
Housing	*506*	*1*	*12*	*2*
Ionosphere	*351*	*0*	*34*	*2*
New Thyroid	*215*	*0*	*5*	*3*
Parkinsons	*195*	*0*	*22*	*2*
Seeds	*210*	*7*	*3*	*3*
Sonar	*208*	*0*	*60*	*2*
Vowel	*990*	*0*	*10*	*11*
Wall-Following	*5456*	*0*	*2*	*4*
Wdbc	*569*	*0*	*9*	*2*
Wholesale	*440*	*0*	*7*	*2*
Yeast	*1484*	*0*	*8*	*10*

Note: N_{sample}—Number of samples; N_{cat}—Number of categorical features; N_{con}—Number of continuous features; N_{class}—Number of classes.

More numerical experiments are conducted to confirm this side effect phenomenon for the fuzzy *K*-NN classifier on a number of selected datasets that are obtained from UCI Machine Learning Repository [43] and summarized in Table I.

Basically, the experiments have three steps: 1) training the classifiers and estimating boundaries; 2) splitting all samples as two categories, i.e., samples near to or far from boundaries; and 3) computing the classification error rates, respectively, for the two categories. A difficulty for the three steps is how to estimate the boundary and then judge a sample near to or far from the boundary for fuzzy *K*-NN. We have a simple scheme

TABLE II
EXPERIMENTAL RESULTS FOR THE FUZZY *K*-NN CLASSIFIER

Databases	$N_{MSNTB}\,(Err)$	$N_{MSFFB}\,(Err)$	Threshold
Banknote	0 (0.0000)	0 (0.0000)	1.0000
Blood	38 (0.3393)	21 (0.1858)	0.8499
Breast Cancer	12 (0.3077)	8 (0.2000)	0.6340
Cleverland Heart	15 (0.3750)	3 (0.0600)	0.6821
Diabetes	45 (0.3913)	17 (0.1466)	0.6238
Flare Solar	8 (0.3636)	5 (0.2273)	0.3115
German	44 (0.2933)	26 (0.1733)	0.6118
Glass	1 (0.0312)	0 (0.0000)	1.6667
Heart	7 (0.1842)	4 (0.0930)	0.7355
Housing	23 (0.3026)	8 (0.1053)	0.8299
Ionosphere	10 (0.1887)	2 (0.0377)	1.0000
New Thyroid	4 (0.1250)	0 (0.0000)	1.3333
Parkinsons	4 (0.1379)	1 (0.0333)	1.0000
Seeds	6 (0.1935)	0 (0.0000)	1.3333
Sonar	9 (0.2903)	1 (0.0312)	0.8425
Vowel	9 (0.0608)	1 (0.0067)	1.7618
Wall-Following	19 (0.0232)	3 (0.0037)	1.5000
Wdbc	14 (0.1647)	2 (0.0233)	1.0000
Wholesale	8 (0.1212)	2 (0.0303)	1.0000
Yeast	118 (0.5291)	75 (0.3363)	1.6000

Note: N_{MSNTB}—Number of samples near to boundary; Err—Error rate; N_{MSFFB}—Number of samples far from boundary.

to overcome this difficulty for fuzzy *K*-NN without an explicit expression of the estimated boundary. Since the output of fuzzy *K*-NN for a sample is a vector $(\mu_1, \mu_2, \ldots, \mu_n)^T$ in which the component is a number in [0,1] representing the membership of the sample belonging to the corresponding class, we estimate its boundary as $\{\mathbf{x}|\mu_1(\mathbf{x}) = \mu_2(\mathbf{x}) = 0.5\}$ for a two-class problem and define the distance between the boundary and a sample with output $(\mu_1, \mu_2)^T$ as $(|\mu_1 - 0.5| + |\mu_2 - 0.5|)$. This way, a threshold can also be set to judge a sample near to or far from the boundary. Experimental results are listed in Table II, from which one can see that the classification error rate for samples near to boundaries is much higher than that for samples located far from boundaries.

III. FUZZINESS OF CLASSIFIERS

The final aim of this study is to make clear the statistical relationship between fuzziness of a classifier and generalization of the classifier. This section first shows an investigation to the classifier's fuzziness and then discusses the fuzziness's impact on misclassification.

A. Fuzziness of Fuzzy Set

In [44], Zadeh first mentioned the term "fuzziness" in conjunction with the proposed concept of fuzzy set. The term refers to the imprecision existing in ill-defined events, which cannot be described by sharply defined collection of points. He also generalized a probability measure of an event to fuzzy event and suggested using entropy in information theory to interpret the uncertainty associated with a fuzzy event. Luca and Termini [45] considered fuzziness as the indefiniteness connected with the situations described by fuzzy sets and defined a quantitative measure of fuzziness by a nonprobabilistic entropy that did not use any probabilistic concepts. For the first time, they clearly proposed three properties that fuzziness measure should satisfy, and these properties indicate that the degree of fuzziness should attain its maximum and minimum when all the memberships are equal to each other and equal to either 0 or 1, respectively. In [46], Luca and Termini extended their definition of entropy to measure the fuzziness of L-fuzzy sets, where the entropy was no longer a numerical quantity but a column matrix or a vector. In the above-mentioned references, it seems that the term of "fuzziness" is interchangeable with "ambiguity," "uncertainty," "indefiniteness," "imprecision," etc., which may cause confusion. Klir and Folger [47], [48] stated that vagueness or fuzziness is different from ambiguity and gave two cognitive uncertainty measures. In general, vagueness or fuzziness is associated with the difficulty of making sharp or precise distinctions in the world. Ambiguity, on the other hand, is associated with one-to-many relations, i.e., situations with two or more alternatives such that the choice between them is left unspecified.

In this paper, we consider fuzziness as a type of cognitive uncertainty which results from the uncertainty transition from one linguistic term to another, where a linguistic term is a value of linguistic variable. A linguistic variable is a word or a phrase, which could take linguistic values. For example, temperature is a linguistic variable that can take the linguistic terms/values, say hot, cool, middle, or etc. Essentially, a linguistic term is a fuzzy set defined on a certain universe of discourse (space).

A mapping from a space $X \to [0, 1]$ is called a fuzzy set and all fuzzy set on X is denoted by $F(X)$. As stated in the literature [49], the fuzziness of a fuzzy set can be measured by a function $E : F(X) \to [0, +\infty)$ that satisfies the following axioms.

1) $E(\mu) = 0$ if and only if μ is a crisp set.
2) $E(\mu)$ attains its maximum value if and only if $\mu(x) = 0.5\ \forall x \in X$.
3) If $\mu \leq_S \sigma$, then $E(\mu) \geq E(\sigma)$.
4) $E(\mu) = E(\mu')$, where $\mu'(x) = 1 - \mu(x)$ for $\forall x \in X$.
5) $E(\mu \cup \sigma) + E(\mu \cap \sigma) = E(\mu) + E(\sigma)$.

Among the third axiom, the sharpened order $\leq_S$ is defined as [45]

$$\mu \leq_S \sigma \Leftrightarrow \min(0.5, \mu(x)) \geq \min(0.5, \sigma(x)) \\ \& \max(0.5, \mu(x)) \leq \max(0.5, \sigma(x)). \quad (5)$$

.

Definition 3.1: Let $B = \{\mu_1, \mu_2, \ldots, \mu_n\}$ be a fuzzy set. According to [45], the fuzziness of *B* can be defined as

$$E(B) = -\frac{1}{n}\sum_{i=1}^{n}(\mu_i \log \mu_i + (1 - \mu_i)\log(1 - \mu_i)). \quad (6)$$

It is easy to verify that formula (6) indeed satisfies axioms 1–5. The fuzziness of a fuzzy set defined by (6) attains its minimum when every element absolutely belongs to the fuzzy set or absolutely not, i.e., $\mu_i = 1$ or $\mu_i = 0$ for each i $(1 \leq i \leq n)$; the fuzziness attains its maximum when the membership degree of each element is equal to 0.5, i.e., $\mu_i = 0.5$ for every $i = 1, 2, \ldots, n$.

B. Fuzziness of Classifier

Given a set of training samples $\{\mathbf{x}_i\}_{i=1}^N$, a fuzzy partition of these samples assigns the membership degrees of each sample to the c classes. The partition can be described by a membership matrix $\mathbf{U} = (\mu_{ij})_{c\times N}$, where $\mu_{ij} = \mu_i(\mathbf{x}_j)$ denotes the membership of the jth sample $\mathbf{x}_j$ belonging to the ith class. The elements in the membership matrix have to obey the following properties:

$$\sum_{i=1}^{c} \mu_{ij} = 1,\ 0 < \sum_{j=1}^{N} \mu_{ij} < N, \quad \mu_{ij} \in [0,1]. \tag{7}$$

Therefore, once the training procedure of a classifier completes, the membership matrix $\mathbf{U}$ upon the N training samples can be obtained. For the jth sample $\mathbf{x}_j$, the trained classifier will give an output vector represented as a fuzzy set $\mu_j = (\mu_{1j}, \mu_{2j}, \ldots, \mu_{cj})^T$. Based on (6), the fuzziness of the trained classifier on $\mathbf{x}_j$ is given by

$$E(\boldsymbol{\mu}_j) = -\frac{1}{c}\sum_{i=1}^{c} (\mu_{ij}\log\mu_{ij} + (1-\mu_{ij})\log(1-\mu_{ij})). \tag{8}$$

Furthermore, the fuzziness of the trained classifier can be given as follows.

Definition 3.2: Let the membership matrix of a classifier on the N training samples with c classes be $\mathbf{U} = (\mu_{ij})_{c\times N}$. The fuzziness of the trained classifier is given by

$$E(\mathbf{U}) = -\frac{1}{cN}\sum_{i=1}^{c}\sum_{j=1}^{N} (\mu_{ij}\log\mu_{ij} + (1-\mu_{ij})\log(1-\mu_{ij})). \tag{9}$$

Equation (9) defines the fuzziness of a trained classifier that has fuzzy vector output. It plays a central role in investigating the classifier's generalization. From the above definition, one can view that the fuzziness of a trained classifier is actually defined as the averaged fuzziness of the classifier' outputs on all training samples. In other words, it is the training fuzziness of the classifier. The most reasonable definition of a classifier's fuzziness should be the averaged fuzziness over the entire sample space including training samples and unseen testing samples. However, the fuzziness for unseen samples is generally unknown, and for any supervised learning problem, there is a well-acknowledged assumption, that is, the training samples have a distribution identical to the distribution of samples in the entire space. Therefore, we use (9) as the definition of a classifier's fuzziness.

C. Relationship Between Fuzziness and Misclassification

To observe the relationship between misclassified samples and their fuzziness, Ripley's synthetic dataset [50] is utilized in the following experiment. There are 250 2-D samples in the dataset. Moreover, the samples are generated from mixtures of two Gaussian distributions. Fig. 3 visualizes the dataset.

The number of neighbors used in the fuzzy *K*-NN classifier, i.e., the value of *K* in our experiment, ranges from 2 to 50 with a step size of 1. The experimental results are shown in Fig. 4, where we report the averaged fuzziness over 1) the set of correctly classified samples, 2) the set of all samples, and 3) the set of misclassified samples. For the fuzzy *K*-NN classifier, the values of fuzziness reported over misclassified samples are significantly higher than the values reported for the correctly classified samples.

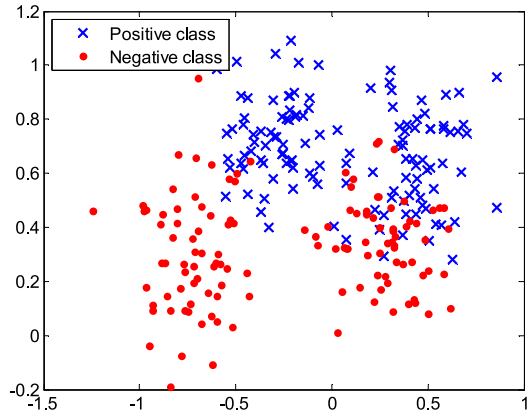

Fig. 3. Ripley's synthetic dataset.

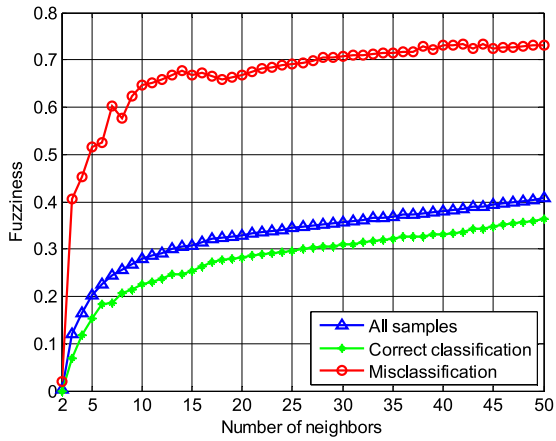

Fig. 4. Fuzziness of fuzzy *K*-NN produced for the Ripley's synthetic dataset.

To further verify the relationship between fuzziness and misclassification, more experiments are conducted on the 20 benchmark datasets taken from UCI Machine Learning Repository [43]. Two illustrations are shown in Fig. 5, where one can still see that the fuzziness on misclassified samples is much larger than that on correctly classified samples, which once again experimentally confirms the mentioned relationship. One worth noting point is that the mentioned relationship is not sensitive to the classifier change. That is, the relationship still holds if the classifier-training algorithm changes from one to another.

D. Relationship Between Fuzziness and Classification Boundary

Furthermore, from the study on the relationship between fuzziness and misclassification, it is found that samples with higher fuzziness are near to the classification boundary, while samples with lower fuzziness are relatively far from the classification boundary. The following experiment gives an illus-

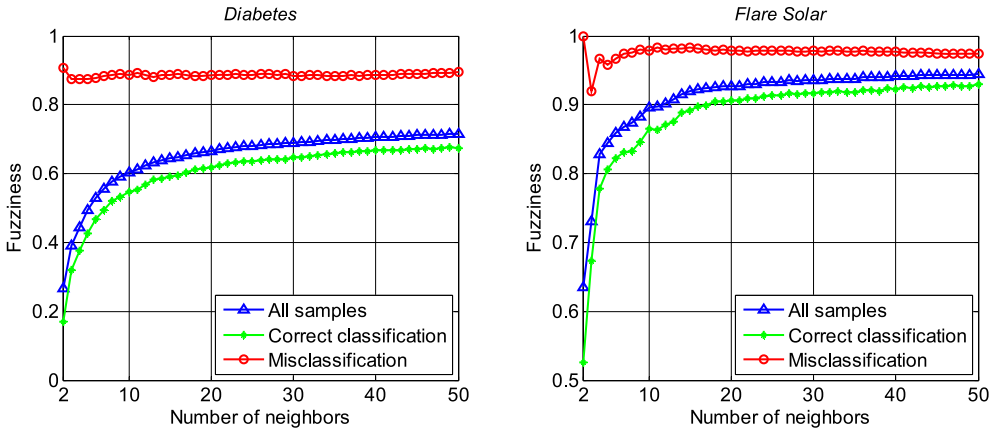

Fig. 5. Relationship between fuzziness and misclassification on the two benchmark datasets.

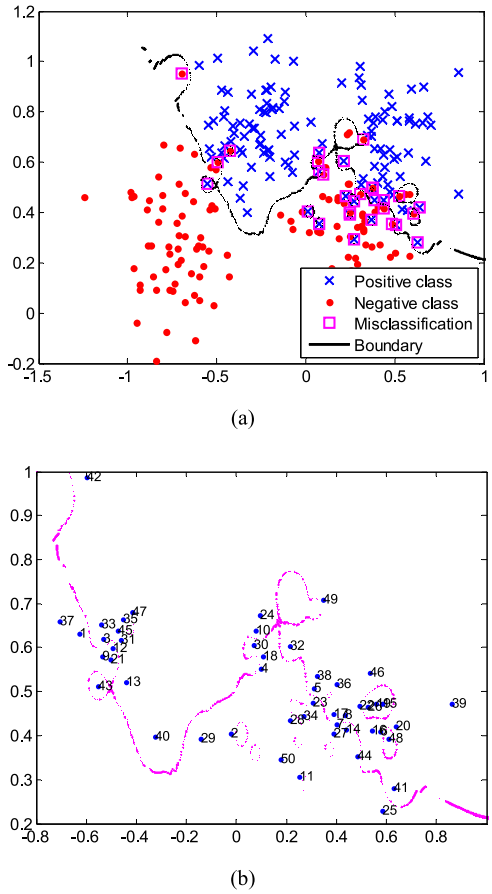

Fig. 6. Classification results of fuzzy *K*-NN with $K = 20$ upon the Ripley's synthetic dataset. (a) Misclassified samples and classification boundary. (b) Fifty samples with highest values of fuzziness.

tration in which the fuzzy *K*-NN is still used as our classifier, the dataset is the Ripley's synthetic data given in Section III-C, and the number of neighbors for fuzzy *K*-NN is fixed as 20. The misclassified samples and the classification boundary are demonstrated in Fig. 6(a). Fig. 6(b) shows the 50 samples with highest value of fuzziness. It can be observed from Fig. 6 that both the misclassified samples and the samples with the larger fuzziness are all near to the classification boundary.

Furthermore, regarding the fuzzy *K*-NN classifier, Proposition 1 relates a sample's fuzziness to the distance between the sample and the classification boundary.

Proposition 1: For a two-class problem, let D_1 be the distance between the sample $\mathbf{x}_1$ and the classification boundary, while D_2 be the distance between the sample $\mathbf{x}_2$ and boundary. Moreover, $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ are the outputs of the classifier on $\mathbf{x}_1$ and $\mathbf{x}_2$, respectively. If $D_1 \leq D_2$, then the fuzziness of $\mathbf{x}_1$ is no less than that of $\mathbf{x}_2$, i.e., $E(\boldsymbol{\mu}) \geq E(\boldsymbol{\sigma})$.

Proof: Let the outputs of the trained classifier on $\mathbf{x}_1$ and $\mathbf{x}_2$ be $\boldsymbol{\mu} = (\mu_1, \mu_2)^T$ and $\boldsymbol{\sigma} = (\sigma_1, \sigma_2)^T$, respectively. According to the distance metric defined in Section II-C, we have $D_1 = |\mu_1 - 0.5| + |\mu_2 - 0.5|$ and $D_2 = |\sigma_1 - 0.5| + |\sigma_2 - 0.5|$. The value of D_1 keeps unchanged if the values of μ_1 and μ_2 are exchanged, while the value of D_2 remains fixed if the values of σ_1 and σ_2 are exchanged. Therefore, without losing generality, we suppose that $\mu_1 \geq \mu_2$ and $\sigma_1 \geq \sigma_2$. It implies that $\mu_1 \geq 0.5$ and $\sigma_1 \geq 0.5$, which result in $D_1 = 2(\mu_1 - 0.5)$ and $D_2 = 2(\sigma_1 - 0.5)$. Since $D_1 \leq D_2$, we have $\mu_1 \leq \sigma_1$. According to formula (5), it can be obtained that $\mu_1 \leq_s \sigma_1$, and further according to the axiom (c) in Section III-A, the inequality $E(\mu_1) \geq E(\sigma_1)$ holds. Since $E(\mu_1) = E(\mu_2)$ and $E(\sigma_1) = E(\sigma_2)$, we finally arrive at the result $E(\boldsymbol{\mu}) = E(\mu_1) \geq E(\boldsymbol{\sigma}) = E(\sigma_1)$.

E. Divide-and-Conquer Strategy

As an experimental observation in Section III-D, we view that the risk of misclassification becomes higher as the fuzziness of training samples gets larger, while the risk is relatively decreasing as the fuzziness of training samples gets statistically smaller. This analysis on misclassification risk inspires us to handle samples with large fuzziness separately from samples with small fuzziness. For most classification problems, samples with more fuzziness are more difficult to be correctly classified in comparison with samples having less fuzziness. Equivalently to say, that boundary points are more difficult to be correctly classified in comparison with inner points. However, the boundary points are often more important than inner points for most

classification problems. Our idea is to use a usual classifier to deal with the samples with less fuzziness while to use a particularly trained classifier to cope with the samples exhibiting higher fuzziness. This is the strategy of divide-and-conquer.

According to the magnitude of fuzziness, all samples are categorized as two groups. One group is of high fuzziness, while the other is of low fuzziness. A number of experiments on both simulated data and on real datasets have been conducted to verify the difference of performance (the correct classification rate) between the two groups. Fig. 7 gives four illustrations, which clearly indicate the significant difference upon the Ripley's synthetic, *Diabetes*, *Flare Solar*, and *German* datasets. The experimental results show that, upon all datasets, the difference is significant for any number of neighbors K ($1 < K < 50$). To save space, we do not report here the difference obtained on some other datasets.

One may argue that the difference tells nothing about the improvement of the classification performance because samples users are really interested in are ones with high fuzziness. In fact, this difference is to make users pay particular attention to samples with high fuzziness and to tell users that the classification for samples with small fuzziness is much possibly correct even they use a simple trained classifier. Due to the limit of paper length, we will report in the next study the handling strategy of high-fuzziness samples separating from the low-fuzziness samples, and the improvement the strategy brings.

F. Impact of the Weighting Exponent m on the Fuzziness of Fuzzy K-Nearest Neighbor Classifier

It is obvious to see from (3) that the output of the fuzzy *K*-NN classifier with respect to a sample is a membership vector. Each component of the membership vector depends on m ($m > 1$), i.e., the parameter of weighting exponent. According to definition 3.2, the fuzziness of a classifier is computed based on the membership vectors, and therefore, the fuzziness of a Fuzzy K-NN classifier changes with value of parameter m. Fuzzy *K*-NN approaches the traditional *K*-NN as m is decreasingly tending to 1. We experimentally examine the impact of m on the fuzziness of fuzzy *K*-NN classifier on the 20 selected databases. All experiments show a consistent trend for the change of the fuzziness value in fuzzy *K*-NN classifiers with a different weighting exponent m. To save space, we only list two illustrations in Fig. 8. It can be observed from Fig. 8 that the fuzziness of fuzzy *K*-NN drastically increases as m increases from 1.05 to 4, and the increase of fuzziness of fuzzy *K*-NN classifier becomes more saturated as $m > 8$.

IV. Relationship Between Generalization and Fuzziness

This section will discuss our main concern, i.e., the relationship between the generalization of a classifier and the fuzziness of the classifier, based on the fuzziness definition of a classifier given in (9) and the properties of boundary points listed in Section II.

A. Definition of Generalization and Its Elaboration

Generally, the task of a learning model is to construct a function $f(\mathbf{x})$ based on a training set $D : (\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)$ in order to approximate an objective function $y = F(\mathbf{x})$ at future observations of $\mathbf{x}$. The use of f to approximate F on future observations is called "generalization." The learned function $f(\mathbf{x})$ is called a classifier for classification problems. The difference between f and F is called generalization error, which is considered as the measurement of generalization ability of the involved learning model.

Theoretically, the generalization error can be investigated from many different angles. One typical method is to estimate an upper bound for the generalization error. The true generalization error reported on the entire input space can be denoted as $R_{\text{true}} = \int_S [f(\mathbf{x}) - F(\mathbf{x})]^2 p(\mathbf{x}) d\mathbf{x}$, where S denotes the entire input space, and $p(\mathbf{x})$ is the probability density function of input $\mathbf{x}$. Since both target outputs and distributions of the unseen samples are unknown, it is impossible to compute R_{true} directly. Many researchers want to find an upper bound to estimate the generalization error. For example, from the angle of structural risk minimization, Vapnik *et al*. [17], [18] gave a bound that depends on the training error and the complexity of the classifier. Here, the complexity of a classifier is described by the size of training set and the VC dimension of a function group including the learned function f. Another example is the localized generalization error model proposed by Ng *et al.* [11], in which the derived error bound is mainly composed of the training error within a neighborhood of training samples and the stochastic sensitivity of classifier outputs.

Experimentally, the generalization error is often verified by observing the prediction accuracy of a classifier on a set of samples, called testing samples, which are not used in the process of training the classifier. This is the testing accuracy, which is regarded as the most crucial index for experimentally measuring the generalization of a classifier.

This paper makes an attempt to study on the generalization of a classifier from a new viewpoint. Different from search for an upper error bound, we try to find a relationship between the generalization of a classifier and the fuzziness of the classifier outputs. The relationship is expected to provide some useful guidelines for improving the generalization ability of a classifier.

B. Classifier Selection

When the membership μ_j of a fuzzy set A is equal to 0.5 for all j, the fuzziness of the fuzzy set attains the maximum. The fuzziness maximization implies that for drawing a fuzzy set as our conclusion, we prefer a fuzzy set with bigger fuzziness to other fuzzy sets. In other words, we consider that an event with much uncertainty (fuzziness) will bring us more information when it occurs [33].

We now consider the output of a trained classifier. Suppose that there are c classes and the output of the classifier for an unseen sample can be represented as $\mu = (\mu_1, \mu_2, \ldots, \mu_c)^T$ in which each component is the degree of the unseen sample belonging to the corresponding class. The final class label C_{i_0} for the unseen sample is determined by $i_0 = \arg\max_{1 \le i \le c} \mu_i$.

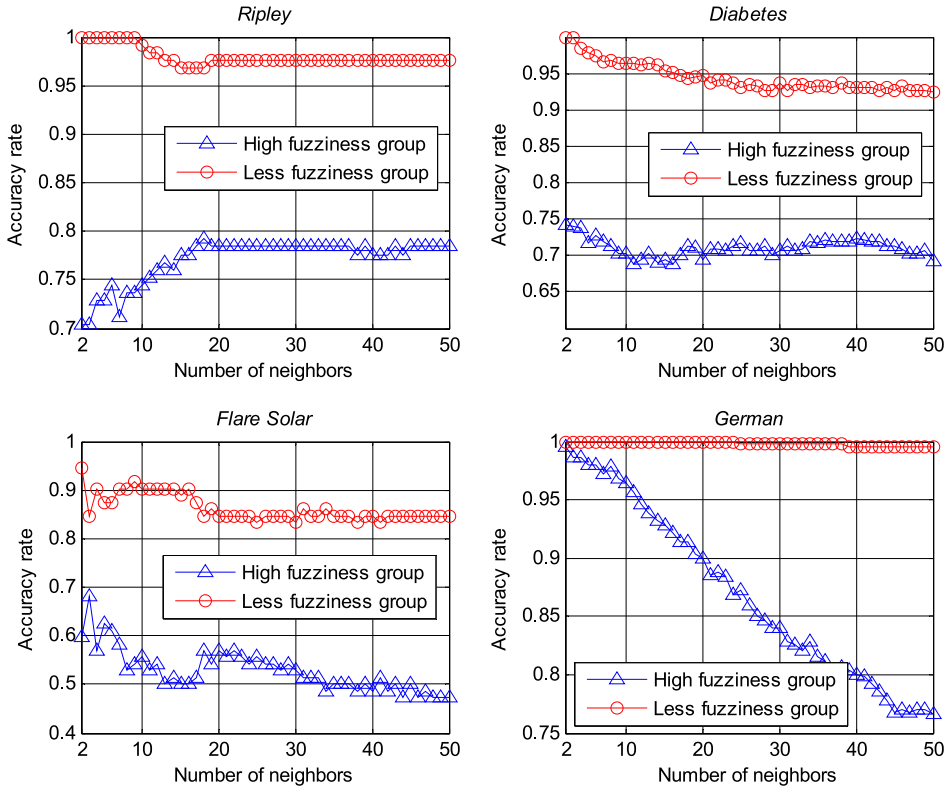

Fig. 7. Difference of testing accuracy rates between the high-fuzziness group and low-fuzziness group.

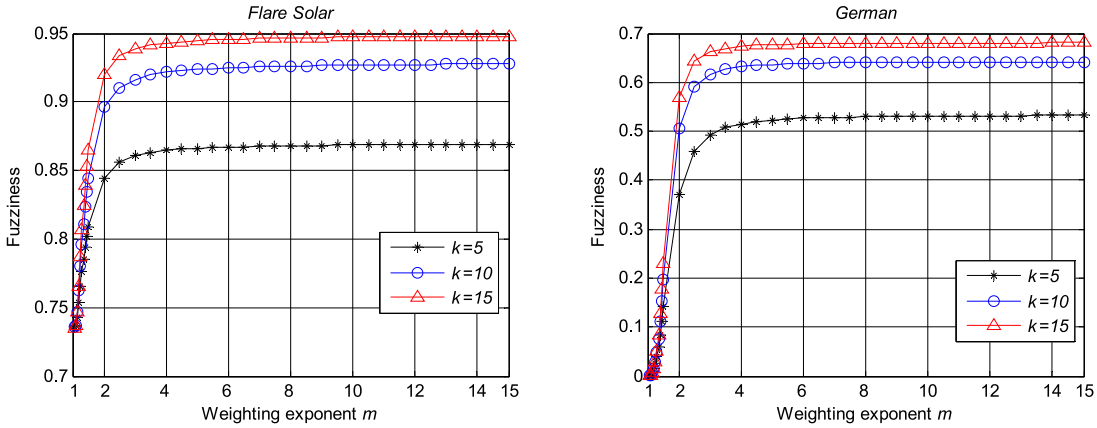

Fig. 8. Impact of the weighting exponent m on the fuzziness of fuzzy K-NN classifier.

Our basic idea is described as Fig. 9, where classifiers A and B denote two trained classifiers, respectively.

Focusing on Fig. 9, we consider the two classifiers having the same training accuracy but generally having the different predictive accuracy. Our problem is: Which one has the better generalization?

It is impossible to provide a general answer since it depends on the specific problem. Nevertheless, from the viewpoint of "traditional" pattern recognition, one definitely prefers classifier A. The reason is at least twofold. The first is that the uncertainty in the training set for classifier A is smaller than for classifier B. People always prefer the one with the lower uncertainty, since it seems making the decision easier. The second is that classifier A has the training accuracy the same as classifier B. In fact, in many approaches to the design of classifiers, the design objective to be optimized is usually the one associated with some constraints using which we tend to minimize the uncertainty of the entries of the output vector while retaining accuracy on the training set. Implicitly, it acknowledges that for two classifiers with the same training accuracy, the classifier with the lower uncertainty has better generalization than the classifier with the higher uncertainty level. However, through a large number of experiments carried out for classification problems with complex and highly nonlinear boundaries or without a clearly delineated

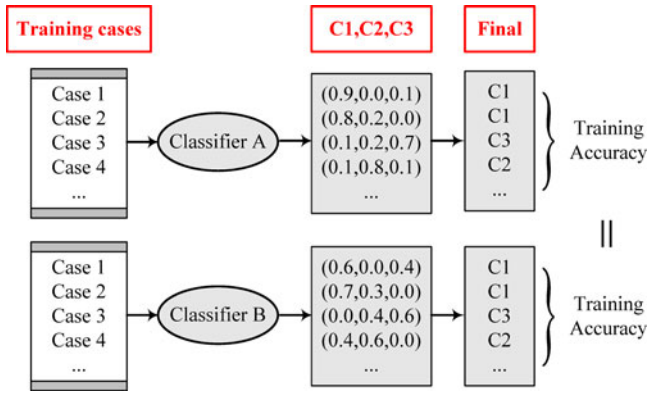

Fig. 9. Underlying idea of the proposed approach—we argue that classifier B has higher generalization capabilities batter than those of classifier A.

boundary, we found that this traditional viewpoint is not always true. In this study, we propose an alternative that for some types of classification problems, classifiers with higher uncertainty for the training set exhibit higher generalization abilities.

C. Explanation Based on Extreme (Max/Min) Fuzziness

We recall the crux of the training algorithm as follows. Basically, the idea of this algorithm is very simple. First, we generate an ensemble of base classifiers, and then, from this ensemble, we find the first m classifiers with the highest fuzziness values while keeping an acceptable individual training accuracy.

Training Phase: Suppose that we have had a training algorithm for generating an ensemble of classifiers by setting up different parameters of this algorithm. Given a training set, we first generate an ensemble of classifiers based on the training algorithm. With respect to any sample $\mathbf{x}_j$, each classifier is required to have an output vector $\boldsymbol{\mu}_j = (\mu_{1j}, \mu_{2j}, \ldots, \mu_{cj})^T$. According to (9), the fuzziness level is computed for each classifier. We sort these classifiers in a decreasing order based on the fuzziness magnitudes of these classifiers and select the first M classifiers with highest fuzziness while keeping an acceptable individual training accuracy. Here, the fuzzy K-NN is used as our classifier generation algorithm.

Reasoning Phase: For any testing sample, match this sample to each of M trained classifiers and get M vectors in which the component represents a possibility of the sample belonging to corresponding class. Take an average for each component of the M vectors and assign the class label with maximum component to the considered sample.

Basically, this algorithm is to raise the fuzziness during the classifier training process under the condition that an acceptable training accuracy is kept. The central idea behind this algorithm is that, regarding the improvement of generalization performance, the classifiers with big fuzziness play an important role more than other classifiers. This idea is identical to our discussion in Section II regarding the relationship between boundary area and high-fuzziness samples. Section II provides us with a result that basically samples near to boundary have the outputted fuzziness higher than samples far from boundary. Usually, we believe that samples near to the boundary are more difficult to be correctly classified than samples far from the boundary. The idea implies such a preference that we would like to select classifier C1 if, for two classifiers C1 and C2, C1 can correctly a set of boundary samples (denoted by A) and C2 can correctly another set of boundary samples (denoted by B), the number of samples in A is as same as in B, and samples in A are nearer to the boundary than samples in B averagely. One may argue that for a classification problem with a simple real boundary, this sufficient consideration of boundary points will give a very complex boundary, which possibly results in the overfitting phenomenon. In fact, for classification problems with simple boundaries such as the linearly separable cases, the fuzziness of a well-trained classifier is usually very small. However, for classification problems with complex boundaries, experimental results do not show the overfitting.

Interestingly, this idea coincides with the following maximum fuzziness principle.

Maximum fuzzy entropy principle [33]: Consider a reasoning process that includes a number of parameters to be determined. With respect to a given fact, the reasoning conclusion will be a parametric fuzzy set, which implies that the reasoning conclusion can be changing with diverse parameters. We prefer the parametric fuzzy set with maximum fuzziness (to other fuzzy sets) as our reasoning conclusion, subject to the given constraints.

We have the following explanations and remarks regarding the maximum fuzziness principle in classifier generalization improvement.

Remark 1: Why does the fuzziness maximization can improve the generalization capability of a classifier? Intuitively, we can offer the following explanation. The explanation is similar to that in [33]. Suppose that there is a classification problem with c classes and A is an object to be classified. If there is not any additional available information for classification, the most reasonable classification result for A should be that the membership degree of A belonging to each of the c classes is identical (i.e., $1/c$). This can be achieved by maximizing the entropy of A, according to the maximum entropy principle in traditional probability theory. If some additional information for classification is available (i.e., there exists a training set in which each example's class is known and an acceptable training accuracy is required to be kept), then in order to get a reasonable and fair classification for A, we should maximize the fuzziness of A subject to some constraints; each constraint represents that a training example can be classified correctly. These constraints mean that the available information for classification has been utilized, but there still exists uncertain factors such that the classification for other objects is uncertain. A reasonable way to handle the remaining uncertain information for classification is to use the maximum uncertainty principle. The reasonable and fair classification for A is expected to result in an increase of generalization capability.

Remark 2: The maximum fuzziness principle is more suitable for classifiers in which the classification uncertainty exists inherently. These cases indicate that the problem may be of crisp classification, but its essence of classes for samples is fuzzy. For example, a crisp classification problem in which any positive

TABLE III
TRAINING ACCURACY (%) OF THREE FUZZY CLASSIFIERS (F-*K*NN, F-ELM, AND F-DCT) WITH DIFFERENT FUZZINESS

Databases	Fuzzy *K*-NN (low fuzziness)	Fuzzy *K*-NN (high fuzziness)	Fuzzy ELM (low fuzziness)	Fuzzy ELM (high fuzziness)	Fuzzy DCT (low fuzziness)	Fuzzy DCT (high fuzziness)
	Acc_{train}	$Acc_{\text{train}}(P_1, P_2)$	Acc_{train}	$Acc_{\text{train}}(P_1, P_2)$	Acc_{train}	$Acc_{\text{train}}(P_1, P_2)$
Banknote	100 ± 0.00	99.77 ± 0.22(1.431E-004,4.883E-004)	100 ± 0.00	100 ± 0.00(1,1)	93.15 ± 1.06	95.86 ± 0.64(6.487E-009,1.316E-004)
Blood	86.63 ± 1.13	79.43 ± 0.76(1.409E-015,8.782E-005)	84.86 ± 0.80	81.89 ± 0.72(1.380E-012,8.696E-005)	76.73 ± 0.33	77.95 ± 0.87(2.855E-011,2.130E-005)
Breast Cancer	88.83 ± 1.60	79.43 ± 1.11(1.357E-012,8.720E-005)	97.73 ± 0.93	$87.98 \pm \pm 1.27$(1.811E-017,8.720E-005)	77.49 ± 1.88	91.86 ± 1.31(1.209E-016,8.696E-005)
Cleverland Heart	99.93 ± 0.18	93.33 ± 1.00(4.514E-016,8.282E-005)	99.83 ± 0.24	92.71 ± 1.26(5.143E-016,8.211E-005)	86.64 ± 3.57	99.35 ± 0.53(6.206E-012,8.720E-005)
Diabetes	90.08 ± 0.56	81.86 ± 0.93(4.922E-14,8.782E-005)	87.12 ± 0.80	82.07 ± 1.19(4.455E-015,8.832E-005)	77.17 ± 1.10	84.17 ± 1.27(1.016E-014,8.807E-005)
Flare Solar	76.65 ± 2.54	66.20 ± 1.11(7.869E-018,8.858E-005)	85.50 ± 2.72	84.95 ± 2.39(0.061,0.056)	76.05 ± 2.31	79.05 ± 2.24(8.393E-006,1.847E-004)
German	99.76 ± 0.18	89.57 ± 0.77(4.279E-025,8.832E-005)	86.13 ± 0.68	80.03 ± 0.78(4.187E-017,8.820E-005)	79.34 ± 4.49	98.81 ± 0.51(4.189E-014,8.858E-005)
Glass	100 ± 0.00	100 ± 0.00(0.3299,1)	100 ± 0.00	99.53 ± 0.50(4.103E-004,9.766E-004)	93.11 ± 3.33	96.76 ± 1.50(2.747E-005,1.803E-004)
Heart	99.60 ± 0.38	91.14 ± 1.43(1.866E-016,8.745E-005)	99.86 ± 0.24	92.46 ± 1.36(1.803E-015,8.609E-005)	77.35 ± 1.26	99.07 ± 1.27(1.544E-025,8.414E-005)
Housing	100 ± 0.00	100 ± 0.00(1,1)	97.54 ± 0.64	92.17 ± 0.89(5.586E-014,8.820E-005)	83.57 ± 1.10	92.71 ± 0.73(1.839E-018,8.795E-005)
Ionosphere	97.29 ± 0.50	85.08 ± 2.52(7.193E-022,8.858E-005)	99.74 ± 0.20	96.12 ± 0.82(5.219E-014,8.560E-005)	99.89 ± 0.20	93.92 ± 0.86(1.312E-014,8.832E-05)
New thyroid	100 ± 0.00	98.00 ± 0.61(8.653E-012,6.757E-005)	100 ± 0.00	99.73 ± 0.34(0.002,0.008)	81.00 ± 2.87	93.80 ± 1.18(2.286E-007,0.002)
Parkinsons	100 ± 0.00	100 ± 0.00(1,1)	100 ± 0.00	97.11 ± 0.83(2.814E-012,7.557E-005)	84.15 ± 1.21	99.19 ± 0.90(2.223E-022,8.271E-005)
Seeds	90.82 ± 1.58	91.09 ± 1.43(0.0029,0.0079)	100 ± 0.00	99.73 ± 0.34(0.002,0.008)	87.11 ± 1.61	95.27 ± 1.31(4.227E-012,8.646E-005)
Sonar	89.38 ± 1.78	81.22 ± 2.57(5.249E-012,8.795E-005)	100 ± 0.00	98.68 ± 0.74(1.904E-007,1.019E-004)	99.12 ± 0.00	98.11 ± 0.79(0.003,0.006)
Vowel	98.40 ± 0.48	95.12 ± 0.84(3.480E-015,8.770E-005)	97.92 ± 0.52	88.66 ± 1.18(2.003E-017,8.832E-005)	95.75 ± 0.71	97.41 ± 0.50(2.364E-006,0.002)
Wall-Following	98.64 ± 0.08	98.25 ± 0.12(1.215E-016,8.832E-005)	95.55 ± 0.24	94.41 ± 0.19(1.884E-016,8.708E-005)	50.09 ± 0.35	51.47 ± 0.55(3.674E-010,8.8575–005)
Wdbc	99.72 ± 0.26	99.77 ± 0.18(0.0217,0.0242)	98.31 ± 0.36	97.13 ± 0.44(7.498E-010,8.696E-005)	91.79 ± 0.73	95.26 ± 0.37(2.852–008,0.002)
Wholesale	100 ± 0.00	100 ± 0.00(1,1)	97.87 ± 0.63	93.89 ± 0.68(9.567E-017,8.487E-005)	82.56 ± 2.09	83.40 ± 2.01(0.026,0.037)
Yeast	50.95 ± 1.29	59.19 ± 0.86(6.750E-018,8.807E-005)	69.03 ± 0.68	63.14 ± 0.69(2.110E-016,8.770E-005)	55.74 ± 1.29	60.27 ± 1.33(1.033E-011,8.845–005)

Note: Acc_{train}—Training accuracy rate; P_1—P-value for paired t-test; P_2 —P-value for Wilcoxon signed rank test.

(negative respectively) sample has several negative (positive respectively) nearest neighbors will definitely not have a boundary, which separates one class of samples from another even with a very low correct rate of classification. This way, every sample associated with a vector (p, q) (representing the possibilities of the sample belonging to each of both classes, respectively) is more reasonable and more accurate than that associated with a crisp class label. The maximum fuzziness principle makes a classifier to output a vector (p, q) with p and q approaching 0.5 rather than with p and q approaching either 0 or 1.

Remark 3: Since A is an object remaining to classify, we do not know its components before matching A to the trained classifier, and further, we cannot directly maximize its fuzziness. Noting that any supervised learning algorithm has a fundamental assumption that the training set is a sampling from a population of examples and the testing set has the distribution identical to the training set, it is reasonable that we replace the fuzziness maximization of A over the entire sample space with that over the training set. Unfortunately, so far, we still have not yet a formal mathematical formulation for this explanation on maximum fuzziness.

Remark 4: The acceptability threshold is referring to the acceptable training accuracy rate which is problem-dependent. Usually, it is defined by users. Experimentally, we find that this threshold is sensitive to the output of our approach. The ensemble of classifiers with high maximum fuzziness is obtained by selecting individual classifiers, which are required to have a training accuracy rate over the threshold. Therefore, the ensemble varies with the change in the threshold value. We experimentally find that our proposed approach has a better performance when the threshold is smaller in comparison with a big threshold for a given learning problem. One explanation may be that for a smaller threshold, the individual classifier with high fuzziness will have more chance to be selected, which will enhance the diversity of the ensemble.

D. Experimental Results

To validate the proposed training and reasoning algorithm presented in Section IV-C, three classifiers, i.e., the fuzzy *K*-NN, the fuzzy extreme learning machine (ELM) [51], [52], and the fuzzy decision tree (DT) [28], are utilized to generate the base classifier ensemble. For a given training set and an integer *N*, we first train *N* fuzzy *K*-NN classifiers by varying the value of *K* from 2 to $N + 1$ with a step of 1, *N* fuzzy ELM classifiers by repeating the random weight *N* times, and *N* fuzzy DTs by varying the leaf level and the parameters of triangular memberships, respectively. Once the training procedure of the *N* fuzzy classifiers has been completed, these classifiers are sorted in a descending order according to their fuzziness values. Then, the first ten base classifiers with highest fuzziness values and the last ten base classifiers with the lowest fuzziness values are selected.

Two mechanisms of validation are selected. The first is the hold-out validation (70–30), namely, for each dataset, 70% samples are randomly chosen for training, while the rest 30% are used for testing. The second is the DOB-SCV validation scheme [53], [54] in which the concept of class-neighbor is used to generate the partition for increasing the randomness and uniformity of samples.

TABLE IV
TESTING ACCURACY (%) OF THREE CLASSIFIERS (F-*K*NN, F-ELM, AND F-DCT) WITH DIFFERENT FUZZINESS

Databases	Fuzzy *K*-NN (low fuzziness)	Fuzzy *K*-NN (high fuzziness)	Fuzzy ELM (low fuzziness)	Fuzzy ELM (high fuzziness)	Fuzzy DCT (low fuzziness)	Fuzzy DCT (high fuzziness)
	Acc_{test}	$Acc_{\text{test}}(P_1, P_2)$	Acc_{test}	$Acc_{\text{test}}(P_1, P_2)$	Acc_{test}	$Acc_{\text{test}}(P_1, P_2)$
Banknote	99.95 ± 0.10	99.66 ± 0.40(0.006,0.023)	100 ± 0.00	100 ± 0.00 (1,1)	93.28 ± 1.08	95.66 ± 1.09(4.941E-007,1.316E-004)
Blood	66.33 ± 2.66	74.27 ± 1.66(2.806E-012,8.745E-005)	74.16 ± 2.12	77.67 ± 1.85(1.137E-005,1.489E-004)	76.27 ± 0.34	77.09 ± 1.29 (0.012,0.013)
Breast Cancer	66.94 ± 3.79	72.50 ± 2.47(8.425E-006,4.899E-004)	51.94 ± 6.58	72.06 ± 4.06(1.286E-010,8.758E-005)	73.88 ± 2.66	75.75 ± 4.68(2.123E-007,8.330E-005)
Cleverland Heart	85.22 ± 2.88	82.72 ± 2.20(6.917E-006,0.001)	83.67 ± 3.43	82.06 ± 2.82(1.840E-010,8.795E-005)	81.83 ± 2.80	79.83 ± 3.33 (0.009,0.017)
Diabetes	72.51 ± 1.61	74.37 ± 2.16(0.003,0.001)	73.27 ± 2.34	76.39 ± 2.64(3.383E-010,8.758E-005)	73.77 ± 2.03	73.40 ± 2.12 (0.376,0.370)
Flare Solar	61.70 ± 3.85	66.02 ± 0.90(8.727E-005,7.204E-004)	37.05 ± 7.19	39.66 ± 6.66(0.108,0.169)	58.98 ± 4.92	63.98 ± 4.29(9.104E-006,1.924E-004)
German	73.50 ± 2.13	74.17 ± 1.28(0.175,8.795E-005)	75.07 ± 1.80	75.83 ± 1.70(0.066,0.048)	72.41 ± 1.34	72.89 ± 1.96(0.396,0.408)
Glass	98.64 ± 1.19	98.18 ± 1.36(0.083,0.001)	67.65 ± 6.82	86.29 ± 4.46(2.732E-009,8.832E-005)	87.27 ± 5.02	88.56 ± 3.59(0.198,0.238)
Heart	80.19 ± 3.91	84.14 ± 3.50(1.493E-005,0.006)	68.77 ± 5.82	79.57 ± 3.68(1.499E-008,8.597E-005)	75.93 ± 3.89	77.47 ± 3.42 (0.046,0.060)
Housing	80.56 ± 3.33	80.98 ± 2.75(0.418,0.370)	87.09 ± 2.02	86.70 ± 2.06(0.368,0.359)	83.99 ± 2.65	86.21 ± 2.27 (0.001,0.001)
Ionosphere	84.29 ± 2.84	84.81 ± 2.51(2.734E-011,8.858E-005)	85.43 ± 3.05	89.95 ± 2.69(1.892E-004,9.426E-004)	84.34 ± 3.34	83.95 ± 3.88(1.562E-005,1.697E-004)
New thyroid	92.54 ± 4.21	87.08 ± 3.40(6.917E-006,1.379E-004)	77.39 ± 5.07	86.00 ± 4.52(4.428E-006,2.180E-004)	78.46 ± 4.53	92.62 ± 1.75 (1.681E-006,0.002)
Parkinsons	85.33 ± 2.79	85.75 ± 3.13(0.514,0.247)	72.42 ± 9.23	89.42 ± 2.93(2.312E-007,8.795E-005)	84.75 ± 2.31	93.17 ± 2.96(5.370E-012,8.536E-005)
Seeds	90.87 ± 2.41	91.90 ± 2.82(0.019,0.0002)	79.37 ± 4.89	76.43 ± 1.70(2.478E-013,8.745E-005)	87.14 ± 3.13	92.78 ± 2.84(7.266E-008,1.282E-004)
Sonar	81.41 ± 4.59	85.44 ± 3.61(0.015,0.007)	72.27 ± 8.27	81.64 ± 4.44(2.797E-006,2.459E-004)	75.28 ± 9.17	82.66 ± 4.67(3.745E-007,8.832E-005)
Vowel	96.82 ± 1.19	93.45 ± 1.67(5.360E-010,8.782E-005)	89.69 ± 1.59	90.36 ± 2.41(1.141E-014,8.832E-005)	86.33 ± 2.13	89.06 ± 1.75 (4.517E-004,0.002)
Wall-Following	98.66 ± 0.25	98.91 ± 0.34(1.914E-007,8.820E-005)	95.35 ± 0.63	94.17 ± 0.72(2.436E-012,8.621E-005)	50.30 ± 0.74	51.69 ± 1.11(1.965E-007,8.807E-005)
Wdbc	87.18 ± 2.11	88.40 ± 2.12(2.033E-004,0.004)	93.11 ± 2.45	95.44 ± 1.09(7.662E-004,4.743E-004)	91.80 ± 1.76	93.90 ± 1.65(0.002,0.006)
Wholesale	90.19 ± 1.70	89.36 ± 2.29(0.013,0.015)	82.03 ± 2.81	79.55 ± 2.01(3.143E-010,8.795E-005)	82.59 ± 2.73	81.71 ± 2.50 (0.789,0.763)
Yeast	54.99 ± 2.14	60.00 ± 2.03(1.151E-009,8.858E-005)	58.42 ± 1.82	60.37 ± 1.53(1.856E-005,4.267E-004)	53.21 ± 2.17	56.35 ± 2.53(2.215E-007,1.398E-004)

Note: Acc_{test}—Testing accuracy rate.

The average training and testing accuracy rates together with their corresponding standard deviations are recorded. Furthermore, the paired T-test is conducted to examine whether the performance improvement achieved by the proposed ensemble of classifiers with high fuzziness over the ensemble of classifiers with low fuzziness is statistically significant. Experimental results do not show significant difference between the two validation mechanisms except for two datasets with imbalanced classes. The experimental results for DOB-SCV validation are listed in Tables III and IV.

We have conducted the Wilcoxon signed rank test, which is a nonparametrical statistical test provided in [59] and [60]. The testing results are placed in Tables III–V (of the current version). The paired t-test in previous version is also placed in the current version for comparison. It is observed there is no significant difference between both statistical tests.

From Table IV, one can observe that the fuzzy classifier ensemble with higher fuzziness achieves better generalization ability in comparison with the ensemble with lower fuzziness. This occurs on 14 datasets for fuzzy *K*-NN, 15 datasets for fuzzy ELM, and 16 datasets for fuzzy DT respectively. Taking the average testing accuracy rate into consideration, the values of standard deviation in Table IV show that the ensemble with higher fuzziness is more stable than the ensemble with lower fuzziness on most datasets. Moreover, the paired t-test shows that the difference between the ensemble with higher fuzziness and the ensemble with lower fuzziness is statistically significant on all the datasets except for two datasets. It is worth noting that the experimental results in Table IV show that the fuzziness of base classifiers is important for constructing an ensemble, rather than showing that the fuzziness of a classifier is very important for its generalization power.

In addition, from Table IV, it is experimentally observed that the proposed algorithms may be more suitable for tackling classification problems with complex boundaries than for those with simple boundaries. The boundaries estimated by the fuzzy *K*-NN classifier are very difficult to exactly express and visualize for more than 4-D data. We still not yet have an effective way to estimate the complexity of boundaries acquired from *K*-NN on *n*-dimensional data when $n > 3$, but 3-D feature subsets selected from the *n*-dimensional original data can provide some visualized impression about the *K*-NN estimated boundary. For example, we consider in Table IV the *Cleverland Heart* data, which do not support our conclusions (i.e., which does not show an improvement of testing accuracy for high-fuzziness *K*-NNs). In comparison with other datasets in Table IV, the *Cleverland Heart* data may have a relatively simpler boundary of *K*-NN, which can be partially verified via a projection of the original data in a 3-D space. Fig. 10 shows the projections in feature sets (3, 8, 10) and (3, 4, 9) of *Cleverland Heart* data. Although the two projection figures cannot reflect the entire characteristics of the *K*-NN estimated boundary, they partially indicate the less complexity of the boundary from different visualized profiles.

Ho and Basu [56] proposed data complexity framework that defines a number of measures to describe the difficulty of a classification problem and its boundary complexity. It is observed that based on the framework, the behavior of a fuzzy-rule-based classification system and its relationship to data complexity was discovered in [57], and also based on this framework, the performance of three classic neural network models and one SVM

TABLE V
TRAINING AND TESTING ACCURACY OF F- *K*NN ($K = 20$) ENSEMBLES WITH DIFFERENT FUZZINESS INDUCED BY VARYING WEIGHTING EXPONENT m (%)

Databases	Fuzzy *K*-NN (Low fuzziness)		Fuzzy *K*-NN (High fuzziness)	
	Acc_{train}	Acc_{test}	$Acc_{\text{train}}(P_1, P_2)$	$Acc_{\text{test}}(P_1, P_2)$
Banknote	100.00 ± 0.00	99.98 ± 0.07	99.64 ± 0.25(2.937E-006,5.102E-005)	99.47 ± 0.53(6.722E-004,0.002)
Blood	92.81 ± 0.73	72.20 ± 2.12	79.04 ± 0.83(1.086E-024,8.858E-005)	75.62 ± 1.60(4.379E-006,0.004)
Breast Cancer	85.11 ± 1.02	63.19 ± 2.61	77.87 ± 1.28(6.304E-019,8.858E-005)	73.37 ± 3.27(2.525E-011,5.167E-004)
Cleverland Heart	99.18 ± 0.45	72.67 ± 0.56	86.21 ± 1.40(5.720E-020,8.845E-005)	81.89 ± 2.28(2.501E-013,5.167E-004)
Diabetes	91.69 ± 0.64	62.34 ± 2.17	79.41 ± 1.14(1.690E-021,8.845E-005)	74.13 ± 2.37(3.631E-013,5.934E-004)
Flare Solar	66.05 ± 1.32	64.66 ± 1.56	67.60 ± 1.10(7.595E-004,8.858E-005)	65.91 ± 0.00(0.002,0.0002)
German	90.48 ± 0.53	53.11 ± 0.78	78.66 ± 0.89(1.178E-025,8.832E-005)	74.70 ± 1.57(6.587E-028,8.770E-005)
Glass	100.00 ± 0.00	98.86 ± 1.09	95.98 ± 0.71(4.149E-016,8.807E-005)	94.62 ± 1.59(3.183E-010,8.845E-005)
Heart	92.84 ± 0.78	65.43 ± 1.93	86.35 ± 1.56(2.455E-017,8.795E-005)	83.77 ± 3.93(3.502E-019,5.501E-004)
Housing	100.00 ± 0.00	79.64 ± 3.42	81.05 ± 1.37(2.202E-023,8.858E-005)	86.57 ± 3.21(8.178E-004,8.832E-005)
Ionosphere	98.90 ± 0.38	83.49 ± 2.42	83.22 ± 2.40(1.659E-017,8.845E-005)	81.70 ± 2.78(0.002,0.003)
New Thyroid	100.00 ± 0.00	85.00 ± 3.82	85.57 ± 1.50(2.179E-020,8.845E-005)	91.15 ± 3.54(5.066E-008,8.720E-005)
Parkinsons	100.00 ± 0.00	84.42 ± 3.64	82.78 ± 1.46(4.488E-022,8.858E-005)	85.00 ± 2.32(0.0016,0.0023)
Seeds	89.90 ± 1.78	90.79 ± 3.03	90.85 ± 1.64(0.0206,0.0001)	91.19 ± 3.19(0.5664,0.0006)
Sonar	90.52 ± 1.89	81.33 ± 3.88	71.94 ± 3.01(9.255E-019,8.858E-005)	69.22 ± 4.36(7.853E-011,8.858E-005)
Vowel	98.13 ± 0.58	93.86 ± 2.31	78.52 ± 1.51(1.056E-021,8.845E-005)	95.07 ± 2.89(2.817E-019,8.820E-005)
Wall-Following	97.55 ± 0.17	97.05 ± 0.90	97.27 ± 0.17(6.900E-012,8.858E-005)	97.07 ± 0.40(1.070E-006, 8.807E-005)
Wdbc	99.80 ± 0.25	85.96 ± 2.34	89.57 ± 0.69(2.675E-024,8.683E-005)	88.66 ± 1.72(1.515E-006,0.002)
Wholesale	100.00 ± 0.00	88.83 ± 1.83	92.69 ± 0.92(7.319E-019,8.820E-005)	89.59 ± 2.27 (0.0761,0.135)
Yeast	52.24 ± 1.19	52.75 ± 1.78	60.59 ± 0.78(4.680E-017, 8.858E-005)	59.69 ± 1.77(5.839E-013, 8.858E-005)

TABLE VI
DIFFERENT FUZZINESS AND TESTING ACCURACY (%) OF THREE DIFFERENT MODELS ON THE *Breast Cancer* DATASET INDUCED BY VARYING THE PARAMETER *K* VALUES

	Single fuzzy *K*-NN		Low fuzziness group		High fuzziness group	
K	Fuzziness	Acc_{test}	Fuzziness	Acc_{test}	Fuzziness	Acc_{test}
2	0.3013	65.23	0.2674	63.62	0.3412	68.38
5	0.5286	68.40	0.4884	67.63	0.5632	72.13
10	0.6402	70.63	0.6045	68.13	0.6756	72.63
15	0.6774	71.45	0.6426	69.75	0.7037	71.75
20	0.7038	73.18	0.6722	71.12	0.7335	74.25
25	0.7140	73.35	0.6827	72.50	0.7418	74.88
30	0.7269	73.18	0.6951	71.63	0.7621	74.00
35	0.7382	73.68	0.7074	72.75	0.7643	73.25
40	0.7397	73.40	0.7104	72.88	0.7628	74.75
45	0.7468	73.27	0.7224	72.12	0.7717	75.62
50	0.7468	73.55	0.7190	71.87	0.7703	74.12

with respect to a series of data complexity measures was investigated in [58]. We select two metrics F1 and F2 in our revision to measure the boundary complexity of two-class classification problems. For multiple-class problems, we tentatively select two of them. Bigger F1 for a dataset indicates that its boundary has more complexity. The experimental results are given in Table VIII from which one can see that the better performance is achieved on datasets with bigger values of F1.

We now focus on the fuzzy *K*-NN classifier. As discussed in Section III-F, the fuzziness of fuzzy *K*-NN classifiers is greatly affected by its weighting exponent *m*. In the following experiment, the impact of different values of *m* on our proposed algorithm is examined. We first generate an ensemble of fuzzy *K*-NN base classifiers by varying the value of *m* in (1, 15). The parameter *m* varies in (1, 2) with a step size of 0.05 and in [2] and [15] with a step size of 1, and the parameter *K* is fixed as 20 during the change in *m*. Therefore, 46 different fuzzy *K*-NNs can be constructed on each dataset. The average training and testing accuracy rates together with their corresponding standard deviations are summarized in Table V. Worth noting is the difference between Table V and Tables III and IV (fuzzy *K*-NN column). The difference is that the fuzzy *K*-NN ensemble in Tables III and IV is generated by varying the value of *K* for fixed weighting exponent *m*, while the fuzzy *K*-NN ensemble in Table V is generated by varying the weighting exponent *m* for fixed *K*.

Table V shows an experimental result similar to those reported in Table IV. From Table V, one still can note that the ensemble of fuzzy *K*-NNs with higher fuzziness produces better generalization performance and results in higher stability in comparison with the ensemble of fuzzy *K*-NNs with lower fuzziness. Moreover, the paired t-test demonstrates that the difference between the ensemble of fuzzy *K*-NNs with higher fuzziness and the ensembles of fuzzy *K*-NNs with higher fuzziness and with lower fuzziness is statistically significant on all the datasets. The experimental result indicates that the proposed methodology is basically independent of the weighting exponent parameter *m* used in the classifier.

The fuzziness is strongly related to the number *K* of nearest neighbors. When we increase the number of *K* for points on the boundaries between classes, we can obtain a better generalization. It implicitly gives the relation between *K* and the classifiers selected with the highest fuzziness. That is, the classifiers with the highest fuzziness are also the classifiers obtained with the highest number of *K*. We experimentally verify this relation on the 20 selected datasets. Basically, most experimental results show a uniform trend, namely, both the fuzziness and the testing accuracy are increasing with the value of *K* for the single *K*-NN, and low-fuzziness and high-fuzziness *K*-NN ensembles. For any given training set, *K* has a maximum value. When *K*

TABLE VII
TESTING ACCURACY (%) AND DIVERSITY OF CLASSIFIERS WITH DIFFERENT FUZZINESS

	Fuzzy *K*-NN ($Acc_{\rm test}/Q_{\rm av}$)		Fuzzy ELM ($Acc_{\rm test}/Q_{\rm av}$)		Fuzzy DCT ($Acc_{\rm test}/Q_{\rm av}$)	
Databases	Low Fuzziness	High Fuzziness	Low Fuzziness	High Fuzziness	Low Fuzziness	High Fuzziness
Banknote	99.95 ± 0.10/1.00	99.66 ± 0.40/1.00	100 ± 0.00/1.00	100 ± 0.00/0.69	93.28 ± 1.08/0.93	95.66 ± 1.09/0.90
Blood	66.33 ± 2.66/0.89	74.27 ± 1.66/0.99	74.16 ± 2.12/0.99	77.67 ± 1.85/0.98	76.27 ± 0.34/0.99	77.09 ± 1.29/0.98
Breast Cancer	66.94 ± 3.79/0.97	72.50 ± 2.47/0.99	51.94 ± 6.58/0.96	72.06 ± 4.06/0.95	73.88 ± 2.66/0.99	75.75 ± 4.68/0.96
Cleverland Heart	85.22 ± 2.88/0.95	82.72 ± 2.20/0.99	83.67 ± 3.43/0.95	82.06 ± 2.82/0.83	81.83 ± 2.80/0.95	79.83 ± 3.33/0.03
Diabetes	72.51 ± 1.61/0.97	74.37 ± 2.16/0.99	73.27 ± 2.34/0.98	76.39 ± 2.64/0.91	73.77 ± 2.03/0.97	73.40 ± 2.12/0.89
Flare Solar	61.70 ± 3.85/0.95	66.02 ± 0.90/0.99	37.05 ± 7.19/0.86	39.66 ± 6.66/0.87	58.98 ± 4.92/0.94	63.98 ± 4.29/0.97
German	73.50 ± 2.13/0.97	74.17 ± 1.28/0.99	75.07 ± 1.80/0.95	75.83 ± 1.70/0.91	72.41 ± 1.34/0.97	72.89 ± 1.96/0.96
Glass	98.64 ± 1.19/1.00	98.18 ± 1.36/1.00	67.65 ± 6.82/0.93	86.29 ± 4.46/0.86	87.27 ± 5.02/0.87	88.56 ± 3.59/0.62
Heart	80.19 ± 3.91/0.97	84.14 ± 3.50/0.99	68.77 ± 5.82/0.95	79.57 ± 3.68/0.78	75.93 ± 3.89/0.72	77.47 ± 3.42/0.35
Housing	80.56 ± 3.33/1.00	80.98 ± 2.75/1.00	87.09 ± 2.02/0.97	86.70 ± 2.06/0.81	83.99 ± 2.65/0.98	86.21 ± 2.27/0.93
Ionosphere	84.29 ± 2.84/0.97	84.81 ± 2.51/0.99	85.43 ± 3.05/0.95	89.95 ± 2.69/0.88	84.34 ± 3.34/0.95	83.95 ± 3.88/0.91
New Thyroid	92.54 ± 4.21/1.00	87.08 ± 3.40/1.00	77.39 ± 5.07/0.99	86.00 ± 4.52/0.99	78.46 ± 4.53/0.93	92.62 ± 1.75/0.96
Parkinsons	85.33 ± 2.79/1.00	85.75 ± 3.13/1.00	72.42 ± 9.23/0.72	89.42 ± 2.93/0.97	84.75 ± 2.31/0.99	93.17 ± 2.96/0.54
Seeds	90.87 ± 2.41/0.99	91.90 ± 2.82/0.99	79.37 ± 4.89/0.99	76.43 ± 1.70/0.82	87.14 ± 3.13/0.98	92.78 ± 2.84/0.95
Sonar	81.41 ± 4.59/0.95	85.44 ± 3.61/0.99	72.27 ± 8.27/0.17	81.64 ± 4.44/0.41	75.28 ± 9.17/0.95	82.66 ± 4.67/0.27
Vowel	96.82 ± 1.19/0.99	93.45 ± 1.67/0.99	89.69 ± 1.59/0.90	90.36 ± 2.41/0.74	86.33 ± 2.13/0.79	89.06 ± 1.75/0.75
Wall-Following	98.66 ± 0.25/0.99	98.91 ± 0.34/0.99	95.35 ± 0.63/0.99	94.17 ± 0.72/0.90	50.30 ± 0.74/0.99	51.69 ± 1.11/0.99
Wdbc	87.18 ± 2.11/0.98	88.40 ± 2.12/1.00	93.11 ± 2.45/0.99	95.44 ± 1.09/0.95	91.80 ± 1.76/0.93	93.90 ± 1.65/0.98
Wholesale	90.19 ± 1.70/1.00	89.36 ± 2.29/1.00	82.03 ± 2.81/0.99	79.55 ± 2.01/0.95	82.59 ± 2.73/0.99	81.71 ± 2.50/0.98
Yeast	54.99 ± 2.14/0.94	60.00 ± 2.03/0.99	58.42 ± 1.82/0.99	60.37 ± 1.53/0.85	53.21 ± 2.17/0.93	56.35 ± 2.53/0.89

exceeds this maximum, the generalization is no longer increase with change of *K*. So far, we do not have an effective method to estimate the maximum value of *K*. As an illustration, Table VI lists different fuzziness and testing accuracy of three different models on the *Breast Cancer* dataset induced by varying the parameter *K* values from 2 to 50.

Finally, we check the impact of classifier ensemble diversity on our proposed model. From references, we can know that in ensemble learning, the diversity of a classifier ensemble has a direct impact on the generalization ability of the classifier ensemble. In ensemble learning, the generalization of an ensemble system is closely related to the diversity of base classifiers, classification confidence of base classifiers, and approaches to generating base classifiers. In [61], Hu *et al.* proposed a new methodology for generating base classifiers based on rough subspaces, which can lead to a powerful and compact classification system. Furthermore, in [62], Li *et al.* explored the impact of classification confidence of base classifiers on voting mechanism in ensemble learning and obtained some interesting results.

It is interesting to observe the difference of diversity between two classifier ensembles with low and high fuzziness. There are many different definitions of diversity for a classifier ensemble. In this study, we select a widely used form [55]. That is, the Q statistic that can be used to compute the diversity between two classifiers based on the prediction correctness rate of both classifiers. The formula can be expressed as

$$Q_{i,k} = \frac{N^{11}N^{00} - N^{01}N^{10}}{N^{11}N^{00} + N^{01}N^{10}} \tag{10}$$

where N^{ab} denotes the number of samples for which the output of classifier C_i is a and simultaneously the output of classifier C_k is b. Moreover, if a given sample is correctly classified by C_i (C_k), the value of a (b) is taken as 1. Otherwise, the value of a (b) is taken as 0. The averaged diversity for an ensemble is evaluated by

$$Q_{av} = \frac{2}{L(L-1)} \sum_{i=1}^{L-1} \sum_{k=i+1}^{L} Q_{i,k} \tag{11}$$

where L is the number of classifiers. Q statistic's value varies from –1 to +1 denoting negative and positive correlation. Experimental results on the 20 selected datasets for three kinds of classifiers are shown in Table VII. The testing accuracy rates of the three different methods are directly taken from Table IV.

It can be observed from Table VII that the diversity of an ensemble of classifiers with high fuzziness is a little less than

TABLE VIII
BOUNDARY COMPLEXITY AND ACCURACY

Databases	F_1	F_2	$Acc_{\rm train}$	$Acc_{\rm test}$
Banknote	0.5894	0.1563	99.77 ± 0.22	99.66 ± 0.40
Blood	0.0049	0.2706	79.43 ± 0.76	74.27 ± 1.66
Breast Cancer	0.3088	0.1875	79.43 ± 1.11	72.50 ± 2.47
Cleverland Heart	0.9515	0.2120	93.33 ± 1.00	82.72 ± 2.20
Diabetes	0.6633	0.2516	81.86 ± 0.93	74.37 ± 2.16
Flare Solar	1.3786	0	66.20 ± 1.11	66.02 ± 0.90
German	0.4186	0.6619	89.57 ± 0.77	74.17 ± 1.28
Glass	42.2465	9.164E-004	100 ± 0.00	98.18 ± 1.36
Heart	1.0059	0.1959	91.14 ± 1.43	84.14 ± 3.50
Housing	44.8778	0.0144	100 ± 0.00	80.98 ± 2.75
Ionosphere	2.0177	0	85.08 ± 2.52	84.81 ± 2.51
New Thyroid	1.3481	6.998E-004	98.00 ± 0.61	87.08 ± 3.40
Parkinsons	3.957E + 008	7.712E-010	100 ± 0.00	85.75 ± 3.13
Seeds	96.8723	0.0012	91.09 ± 1.43	91.90 ± 2.82
Sonar	1.932E + 003	1.045E-006	81.22 ± 2.57	85.44 ± 3.61
Vowel	0.8790	0.0600	95.12 ± 0.84	93.45 ± 1.67
Wall-Following	9.8487	-6.669E-006	98.25 ± 0.12	98.91 ± 0.34
Wdbc	2.656E + 003	0.0015	99.77 ± 0.18	88.40 ± 2.12
Wholesale	0.0158	4.814E-004	100 ± 0.00	89.36 ± 2.29
Yeast	13.0694	0	59.19 ± 0.86	60.00 ± 2.03

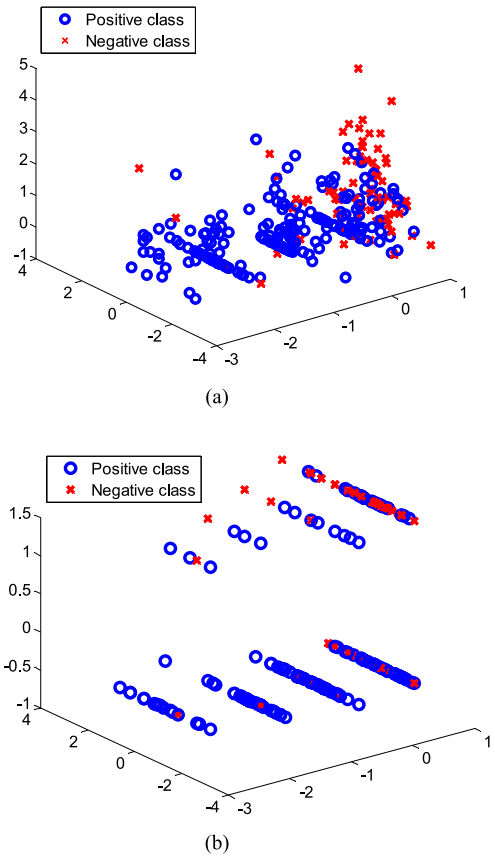

Fig. 10. Visualization of Cleverland Heart on 3-D feature space. (a) Three-dimension F: 3, F: 8, F: 10, (b) Three-dimension F: 3, F: 4, F: 9.

the diversity of an ensemble of classifiers with low fuzziness. From a new angle, it indicates that the diversity has a key impact on the ensemble learning performance, but it does not mean that the more the diversity, the better the performance.

V. Conclusion and Future Works

This paper delivers a study on the relationship between generalization and fuzziness of classifiers with outputs of membership vectors by experimentally observing the high risk of misclassification for samples near to boundaries. We have the following conclusions.

1) For classification problems with complex boundary, big fuzziness samples are more likely misclassified in comparison with small fuzziness samples.
2) The set of samples near to the boundary is identical to the set of samples with high fuzziness, but the one-to-one mapping is difficult to find, which depends on the definition of boundary sample.
3) While a training accuracy is acceptable, we believe that the classifier with higher fuzziness output has a better generalization for complex boundary problems, which is experimentally confirmed in this paper.

Our future works on this topic include the following.

1) For a well-trained classifier that outputs a membership vector, samples with higher fuzziness outputted by the classifier mean a bigger risk of misclassification. One interesting way to promote the correct classification rate is separating the high-fuzziness samples from the low-fuzziness samples and using a particular technique (maybe with more time complexity) to handle the high-fuzziness samples.
2) There are many algorithms to train classifiers with outputs of class memberships. One problem is whether the relationship between the generalization and fuzziness, developed in this paper based on the fuzzy *K*-NN classifier, is sensitive to the selection of classifier. Particularly, is there any difference in the developed relationship between problems with implicit and explicit boundaries?

References

[1] X. D. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. Mclachlan, A. Ng, B. Liu, P. S. Yu, Z. H. Zhou, M. Steinbach, D. J. Hand, and D. Steinberg, "Top 10 algorithms in data mining," *Knowl. Inf. Syst.*, vol. 14, pp. 1–37, 2008.

[2] I. Kushchu, "Genetic programming and evolutionary generalization," *IEEE Trans. Evol. Comput.*, vol. 6, no. 5, pp. 431–442, Oct. 2002.

[3] M. Stone, "Cross-validatory choice and assessment of statistical predictions," *J. Roy. Statist. Soc., Ser. B (Statist. Methodol.)*, vol. 36, no. 2, pp. 111–147, 1974.

[4] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap (Monographs on Statistics and Applied Probability)*. London, U.K.: Chapman & Hall, 1994.

[5] P. A. Lachenbruch and R. M. Mickey, "Estimation of error rates in discriminant analysis," *Technometrics*, vol. 10, pp. 1–11, 1968.

[6] B. Efron, "Bootstrap methods: Another look at the Jackknife," *Ann. Statist.*, vol. 7, no. 1, pp. 1–26, 1979.

[7] A. Luntz and V. Brailovsky, "On estimation of characters obtained in statistical procedure of recognition," (in Russian), *Techicheskaya Kibernetica*, vol. 3, 1969.

[8] V. N. Vapnik, *Statistical Learning Theory (Adaptive and Learning Systems for Signal Processing, Communications and Control Series)*. New York, NY, USA: Wiley, 1998.

[9] O. Chapelle, V. Vapnik, O. Bousquet, and S. Mukherjee, "Choosing multiple parameters for support vector machines," *Mach. Learning*, vol. 46, nos. 1–3, pp. 131–159, Jan. 2002.

[10] M. T. Musavi, K. H. Chan, D. M. Hummels, and K. Kalantri, "On the generalization ability on neural network classifiers," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 16, no. 6, pp. 659–663, Jun. 1994.

[11] W. Y. Ng, P. F. Chan, D. S. Yeung, and C. C. Tsang, "Quantitative study on the generalization error of multiple classifier systems," in *Proc. Int. Conf. Syst., Man, Cybern.*, 2005, pp. 889–894.

[12] D. S. Yeung, W. Y. Ng, D. F. Wang, C. C. Tsang, and X. Z. Wang, "Localized generalization error model and its application to architecture selection for radial basis function neural network," *IEEE Trans. Neural Netw.*, vol. 18, no. 5, pp. 1294–1305, Sep. 2007.

[13] B. B. Sun, W. Y. Ng, D. S. Yeung, and P. K. Chan, "Hyper-parameter selection for sparse LS-SVM via minimization of its localized generalization error," *Int. J. Wavelets, Multiresolution Inf. Process.*, vol. 11, no. 3, 2013.

[14] P. K. Chan, D. S. Yeung, W. Y. Ng, C. M. Lin, and N. K. Liu, "Dynamic fusion method using localized generalization error model," *Inf. Sci.*, vol. 217, pp. 1–20, 2012.

[15] A. Agarwal and J. C. Duchi, "The generalization ability of online algorithms for dependent data," *IEEE Trans. Inf. Theory*, vol. 59, no. 1, pp. 573–587, Jan. 2013.

[16] C. C. Gavin and L. C. Nicola, "On over-fitting in model selection and subsequent selection bias in performance evaluation," *J. Mach. Learning Res.*, vol. 11, pp. 2079–2107, 2010.

[17] V. N. Vapnik, *Estimation of Dependences Based on Empirical Data*. New York, NY, USA: Springer-Verlag, 1982.

[18] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proc. 5th Annu. Workshop Comput. Learn. Theory*, 1992, pp. 144–152.

[19] W. D. Penny and T. J. Stonham, "Generalization in multi-layer networks of sigma-pi units," *IEEE Trans. Neural Netw.*, vol. 6, no. 2, pp. 506–508, Mar. 1995.
[20] N. Littlestone and M. K. Warmuth, "The weighted majority algorithm," *Inf. Comput.*, vol. 108, no. 2, pp. 212–261, Feb. 1994.
[21] R. E. Schapire, Y. Freund, P. Bartlett, and W. S. Lee, "Boosting the margin: A new explanation for the effectiveness of voting methods," *Ann Statist.*, vol. 26, no. 5, pp. 1651–1686, 1998.
[22] S. Decherchi, S. Ridella, R. Zunino, P. Gastaldo, and D. Anguita, "Using unsupervised analysis to constrain generalization bounds for support vector classifiers," *IEEE Trans. Neural Netw.*, vol. 21, no. 3, pp. 424–438, Mar. 2010.
[23] O. Ludwig, U. Nunes, B. Ribeiro, and C. Premebida, "Improving the generalization capacity of cascade classifiers," *IEEE Trans. Cybern.*, vol. 43, no. 6, pp. 2135–2146, Dec. 2013.
[24] J. Yang, X. Q. Zeng, S. M. Zhong, and S. L. Wu, "Effective neural network ensemble approach for improving generalization performance," *IEEE Trans. Neural Netw. Learning Syst.*, vol. 24, no. 6, pp. 878–887, Jun. 2013.
[25] S. Y. Chong, P. Tiňo, and X. Yao, "Relationship between generalization and diversity in coevolutionary learning," *IEEE Trans. Comput. Intell. AI Games*, vol. 1, no. 3, pp. 214–232, Sep. 2009.
[26] S. Y. Chong, P. Tiňo, D. C. Ku, and X. Yao, "Improving generalization performance in co-evolutionary learning," *IEEE Trans. Evol. Comput.*, vol. 16, no. 1, pp. 70–85, Feb. 2012.
[27] D. Sarkar, "Randomness in generalization ability: A source to improve it," *IEEE Trans. Neural Netw.*, vol. 7, no. 3, pp. 676–685, May 1996.
[28] Y. Yuan and M.J. Shaw, "Induction of fuzzy decision trees," *Fuzzy Sets Syst.*, vol. 69, pp. 125–139, 1995.
[29] J. J. Buckley and Y. Hayashi, "Fuzzy neural networks: A survey," *Fuzzy Sets Syst.*, vol. 66, pp. 1–13, 1994.
[30] C. F. Lin and S. D. Wang, "Fuzzy support vector machines," *IEEE Trans. Neural Netw.*, vol. 13, no. 2, pp. 464–471, Mar. 2002.
[31] J. M. Keller, M. R. Gray, and J. A. Givens, "A fuzzy K-nearest neighbor algorithm," *IEEE Trans. Syst., Man Cybern.*, vol. SMC-15, no. 4, pp. 580–585, Jul./Aug. 1985.
[32] D. Dubois and H. Prade, "Rough fuzzy sets and fuzzy rough sets," *Int. J. Gen. Syst.*, vol. 17, no. 2–3, pp. 191–209, 1990.
[33] X. Z. Wang and C. R. Dong, "Improving generalization of fuzzy if-then rules by maximizing fuzzy entropy," *IEEE Trans. Fuzzy Syst.*, vol. 17, no. 3, pp. 556–567, Jun. 2009.
[34] X. Z. Wang, L. C. Dong, and J. H. Yan, "Maximum ambiguity-based sample selection in fuzzy decision tree induction," *IEEE Trans. Knowl. Data Eng.*, vol. 24, no. 8, pp.1491–1505, Aug. 2012.
[35] C. Lee and D. A. Landgrebe, "Decision boundary feature extraction for nonparametric classification," *IEEE Trans. Syst., Man Cybern.*, vol. 23, no. 2, pp. 433–443, Mar./Apr. 1993.
[36] C. Lee and D. A. Landgrebe, "Decision boundary feature extraction for neural networks," *IEEE Trans. Neural Netw.*, vol. 8, no. 1, pp. 75–83, Jan. 1997.
[37] H. Drucker, C. Cortes, L. D. Jackel, Y. LeCun, and V. Vapnik, "Boosting and other ensemble methods," *Neural Comput.*, vol. 6, no. 6, pp. 1289–1301, 1994.
[38] O. D. Richard, P. E. Hart, and D. G. Stork, *Pattern Classification*. New York, NY, USA: Wiley, 2012.
[39] Z. Yan and C. Xu, "Studies on classification models using decision boundaries," in *Proc. 8th IEEE Int. Conf. Cognitive Informat.*, 2009, pp. 287–294.
[40] C. Bishop, *Pattern Recognition and Machine Learning*. Berlin, Germany: Springer-Verlag, 2006.
[41] M. Sewell. (2005). *SVM Software*. [Online]. Available: http://www.svms.org/software.html
[42] C. Chang and C. J. Lin. (2013). *LIBSVM—A Library for Support Vector Machines*. [Online]. Available: http://www.csie.ntu.edu.tw/∼cjlin/
[43] K. Bache and M. Lichma. (2013). *UCI Macine Learning Repository*, School Inf. Comput. Sci., Univ. California, Irvine, CA, USA. [Online]. Available: http://archive.ics.uci.edu/ml
[44] L. A. Zadeh, "Probability measures of fuzzy events," *J. Math. Anal. Appl.*, vol. 23, pp. 421–427, 1968.
[45] A. De Luca and S. Termini, "A definition of a nonprobabilistic entropy in the setting of fuzzy sets theory," *Inf. Control*, vol. 20, pp. 301–312, 1972.
[46] A. De Luca and S. Termini, "Entropy of L-fuzzy sets," *Inf. Control*, vol. 24, pp. 55–73, 1974.
[47] G. Klir, "Where do we stand on measures of uncertainty, ambiguity, fuzziness, and the like?" *Fuzzy Sets Syst.*, vol. 24, no. 2, pp. 141–160, 1987.
[48] G. Klir and T. Folger, *Fuzzy Sets, Uncertainty and Information*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1988.
[49] D. S. Yeung and E. Tsang, "Measures of fuzziness under different uses of fuzzy sets," *Adv. Comput. Intell. Commun. Comput. Inf. Sci.*, vol. 298, pp. 25–34, 2012.
[50] B. D. Ripley, *Pattern Recognition and Neural Networks*. Cambridge, U.K.: Cambridge Univ. Press, 1996.
[51] G. B. Huang, Q. Y. Zhu, and C.-K. Siew, "Extreme learning machine: Theory and applications," *Neurocomputing*, vol. 70, nos. 1–3, pp. 489–501, 2006.
[52] G. B Huang, L. Chen, and C. K. Siew, "Universal approximation using incremental constructive feedforward networks with random hidden nodes," *IEEE Trans. Neural Netw.*, vol. 17, no. 4, pp. 879–892, Jul. 2006
[53] J. G. Moreno-Torres, J. A. Sáez, and F. Herrera, "Study on the impact of partition-induced dataset shift on k-fold cross-validation," *IEEE Trans. Neural Netw. Learning Syst.*, vol. 23, no. 8, pp. 1304–1312, Aug. 2012.
[54] V. López, A. Fernandez, and F. Herrera, "On the importance of the validation technique for classification with imbalanced datasets: Addressing covariate shift when data is skewed," *Inf. Sci.*, vol. 257, pp. 1–13, 2014.
[55] L. I. Kuncheva and C. J. Whitaker, "Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy," *Mach. Learning*, vol. 51, pp. 181–207, 2003.
[56] T. K. Ho and M. Basu, "Complexity measures of supervised classification problems," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 3, pp. 289–300, Mar. 2002.
[57] J. Luengo and F. Herrera, "Domains of competence of fuzzy rule based classification systems with data complexity measures: A case of study using a fuzzy hybrid genetic based machine learning method," *Fuzzy Sets Syst.*, vol. 161, no. 1, pp. 3–19, 2010.
[58] J. Luengo and F. Herrera, "Shared domains of competence of approximate learning models using measures of separability of classes," *Inf. Sci.*, vol. 185, no. 1, pp. 43–65, 2012.
[59] J. Demsar, "Statistical comparisons of classifiers over multiple data sets," *Mach. Learning Res.*, vol. 7, pp. 1–30, 2006.
[60] S. García, A. Fernández, J. Luengo, and F. Herrera, "Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power," *Inf. Sci.*, vol. 180, no. 10, pp. 2044–2064, 2010.
[61] Q. Hu, D. Yu, Z. Xie, and X. Li, "EROS: Ensemble rough subspaces," *Pattern Recognit.*, vol. 40, pp. 3728–3739, 2007.
[62] L. Li, Q. Hu, X. Wu, and D. Yu, "Exploration of classification confidence in ensemble learning," *Pattern Recognit.*, vol. 47, pp. 3120–3131, 2014.

Xi-Zhao Wang (M'03–SM'04–F'12) received the Ph.D. degree in computer science from the Harbin Institute of Technology, Harbin, China, in 1998.

He is currently a Professor with the College of Computer Science and Software Engineering, Shenzhen University, Guangdong, China. From September 1998 to September 2001, he was a Research Fellow with the Department of Computing, Hong Kong Polytechnic University, Hong Kong. He has more than 180 publications, including four books, seven book chapters, and more than 100 journal papers in the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS, IEEE TRANSACTIONS ON FUZZY SYSTEMS, *Fuzzy Sets and Systems*, *Pattern Recognition*, etc. His H-index is 20 (up to June 2014). His current research interests include learning from examples with fuzzy representation, fuzzy measures and integrals, neuro-fuzzy systems and genetic algorithms, feature extraction, multiclassifier fusion, and applications of machine learning.

Dr. Wang has been the PI/Co-PI for 16 research projects supported partially by the National Natural Science Foundation of China and the Research Grant Committee of Hong Kong Government. He was the BoG Member of the IEEE Systems, Man, and Cybernetics Society (IEEE SMCS) in 2005, 2007–2009, and 2012–2014; the Chair of the IEEE SMC Technical Committee on Computational Intelligence, an Associate Editor of IEEE TRANSACTIONS ON CYBERNETICS; an Associate Editor of *Pattern Recognition and Artificial Intelligence*; a Member of the Editorial Board of *Information Sciences*; and an Executive Member of the Chinese Association of Artificial Intelligence. He was the recipient of the IEEE SMCS Outstanding Contribution Award in 2004 and the recipient of IEEE SMCS Best Associate Editor Award in 2006. He is the general Co-Chair of the 2002–2014 International Conferences on Machine Learning and Cybernetics, cosponsored by IEEE SMCS. He is a Distinguished Lecturer of the IEEE SMCS.

Hong-Jie Xing (M'08) received the M.Sc. degree from Hebei University, Baoding, China, in 2003, and the Ph.D. degree in pattern recognition and intelligent system from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2007.

He is currently a Professor with the College of Mathematics and Information Science, Hebei University. His research interests include neural networks, supervised and unsupervised learning, and support vector machines.

Yan Li (M'04) received the Ph.D. degree in computer science from the Department of Computing, The Hong Kong Polytechnic University, Hong Kong, in 2006.

She is currently a Professor with the College of Mathematics and Information Science, Hebei University, Baoding, China. Her research interests include machine learning, game intelligence, fuzzy sets and rough sets.

Dr. Li is a Member of the IEEE Systems, Man, and Cybernetics Society.

Qiang Hua (M'04) received the B.Sc. and M.Sc. degrees from Hebei University, Baoding, China, in 1996 and 1999, respectively. He is currently working toward the Ph.D. degree at Nanjing University of Aeronautics and Astronautics, Nanjing, China.

He is also a Professor with the College of Mathematics and Information Science, Hebei University. His main research interests include pattern recognition and neural networks.

Chun-Ru Dong (M'07) received the B.Sc. degree in computational mathematics and the M.Eng. degree in computer science from Hebei University, Baoding, China, in 2002 and 2005, respectively. He is working toward the Ph.D. degree in computer science at South China University of Technology, Guangzhou, China.

He is currently a Lecturer with the College of Mathematics and Information Science, Hebei University. His research interests include sensitivity analysis for inductive learning, rule-based fuzzy reasoning system, neural networks, evolutionary algorithms and their applications, and adversarial learning.

Witold Pedrycz (M'88–SM'90–F'99) received the M.Sc., Ph.D., and D.Sci. degrees from Silesian University of Technology, Gliwice, Poland, in 1977, 1980, and 1984, respectively.

He is a Professor and Canada Research Chair with the Department of Electrical and Computer Engineering, University of Alberta, Edmonton, AB, Canada. He is also with the Systems Research Institute, Polish Academy of Sciences, Warsaw, Poland. He is actively pursuing research in computational intelligence, human-centric computing, fuzzy modeling, knowledge discovery and data mining, fuzzy control including fuzzy controllers, pattern recognition, knowledge-based neural networks, relational computation, bioinformatics, and software engineering. He has published numerous papers in this area. He is also the author of nine research monographs covering various aspects of computational intelligence and software engineering.

Prof. Pedrycz is the Editor-in-Chief of *Information Sciences* and the Editor-in-Chief of the IEEE TRANSACTIONS ON SYSTEMS MAN AND CYBERNETICS—PART A: SYSTEMS AND HUMANS. He was the President of the International Fuzzy Systems Association and the President of the North American Fuzzy Information Processing Society (NAFIPS). He served as a General Chair of NAFIPS 2004. He is a recipient of the prestigious K.S. Fu Award.