Iranian Journal of Fuzzy Systems

Volume 16, Number 5, (2019), pp. 15-27

A research on classification performance of fuzzy classifiers based on fuzzy set theory

Y. L. Yang[1] and X. Y. Bai[2]

[1]*School of Mathematics and Statistics, Xidian University, Xi'an 710126, PR China.*
[2]*School of Mathematics and Statistics, Xidian University, Xi'an 710126, PR China,, and School of Sciences, Northwest A&F University, Yangling 712100, PR China.*

ylyang@mail.xidian.edu.cn, xuyingbai@nwafu.edu.cn

Abstract

Due to the complexities of objects and the vagueness of the human mind, it has attracted considerable attention from researchers studying fuzzy classification algorithms. In this paper, we propose a concept of fuzzy relative entropy to measure the divergence between two fuzzy sets. Applying fuzzy relative entropy, we prove the conclusion that patterns with high fuzziness are close to the classification boundary. Thus, it plays a great role in classification problems that patterns with high fuzziness are classified correctly. Meanwhile, we draw a conclusion that the fuzziness of a pattern and the uncertainty of its class label are equivalent. As is well known, entropy not only measures the uncertainty of random variable, but also represents the amount of information carried by the variable. Hence, a fuzzy classifier with high fuzziness would carry much information about training set. Therefore, in addition to some assessment criteria such as classification accuracy, we could study the classification performance from the perspective of the fuzziness of classifier. In order to try to ensure the objectivity in dealing with unseen patterns, we should make full use of information of the known pattern set and do not make too much subjective assumptions in the process of learning. Consequently, for problems with rather complex decision boundaries especially, under the condition that a certain training accuracy threshold is maintained, we demonstrate that a fuzzy classifier with high fuzziness would have a well generalization performance.

Keywords: Fuzziness, fuzzy classifier, fuzzy relative entropy, flassification boundary, generalization.

1 Introduction

Classification is one of the most well-known tasks in supervised learning and data mining[2, 33, 14]. This is fundamental to many application domains like computer vision, decision making, information retrieval, natural language processing, bioinformatics, pattern recognition, etc. Classification refers to predicting the class label y ($y \in C$) of a pattern x ($x \in \mathcal{X}$) based on its features. Here C is the space of class labels, and $\mathcal{X}$ is the space of patterns. There is a comprehensive introduction to many popular classification methods in literatures[13][17]. When $C = \{y_1, y_2\}$, this is an important kind of classification problems, called binary classification.

We assume that there is some "correct" labeling function, $F : \mathcal{X} \to C$, and for all i, $c_i = F(x_i)$, where $x_i \in \mathcal{X}$, $c_i \in C$. However, the labeling function F is unknown for learners. In fact, this is just what the learners try to find out. A classification algorithm f is obtained by learning training set S which has a distribution identical to the distribution of $\mathcal{X}$. The goal of the algorithm is to minimize the error between F and f. Since we don't know what F is, the true error is not directly available. The training error — the error between F and f over the training set— can be used as an indicator to estimate the performance of f [27]. However, our ultimate goal is to predict unseen pattern over $\mathcal{X}$ by f. The ability of f predicting unseen data is called generalization ability. That is to say, the purpose of learning training set is to acquire a classifier with better generalization performance. Actually, a classification algorithm with lower training error does not imply that it has a better generalization. Sometimes, an algorithm with lower training error may occur overfitting. There are many studies on the generalization abilities of classifiers being expressed from different points of view[29, 7, 1, 15, 11, 23, 8].

Corresponding Author: X. Y. Bai

Received: November 2017; Revised: August 2018; Accepted: February 2019.

Crisp classifiers which the result is or isn't a member of one class ignore the differences among various patterns (the differences exist, though these patterns belong to the same class). Although they have obtained significant achievements, they appear powerless for some types of problems. Especially, the problems are imprecise and indefinite naturally. Since fuzzy sets theory was introduced by Zadeh [37], a new vitality is endowed to the research of that problems. Researchers have found numerous ways to utilize this theory to generalize existing techniques and to develop new algorithms in pattern recognition and decision analysis [16, 3, 19, 31, 5]. In literature [5], Bezdek comes up with fuzzy classification algorithms. The output of an input pattern by a fuzzy classifier is a fuzzy set vector of which each component represents the membership degree of the input pattern belonging to the corresponding class. A great deal of fuzzy classification algorithms have been proposed [20, 24, 22, 6, 36] and applied to a variety of fields[4, 25, 28]. And better results have been acquired.

As is well known, patterns near the classification boundary are easily misclassified. In order to get a classifier with good classification performance, many classification algorithms attempt to learn the borderline of each class as exactly as possible in the process of learning training set [18, 26, 30, 32, 34, 38]. For some types of problems, the classification boundaries are easy to determine, and even could be expressed by formulas explicitly. As a matter of fact, for most classification problems, the boundaries are quite complex or cannot be clearly delineated. According to the conclusion that patterns with high fuzziness are close to the classification boundary, we should consider the fuzziness of patterns rather than decision boundaries directly.

One focal point of this article is the conclusion that the fuzziness of a pattern and the uncertainty of its class label are equivalent. Namely, the higher fuzziness a fuzzy set is, the larger uncertainty it is. As is well known, entropy not only measures the uncertainty of random variable, but also represents the amount of information carried by the variable [9]. Therefore, a fuzzy set with high fuzziness carries much information. It is universally known that patterns near the boundaries contain information about two or more different classes. This theoretically explains why the decision boundary pattern points have the maximum fuzziness. Furthermore, a fuzzy classifier with high fuzziness would carry much information about training set. In order to try to ensure the objectivity in dealing with unseen patterns, we should make full use of information of the known pattern set and do not make too much subjective assumptions in the process of learning. Consequently, for problems with rather complex decision boundaries especially, under the condition that a certain training accuracy threshold is maintained, we demonstrate that a fuzzy classifier with high fuzziness would have a well generalization performance.

In order to measure the divergence between two fuzzy sets, we propose a concept of fuzzy relative entropy. Fuzzy relative entropy $FD(\mu(x)\|\eta(x))$ indicates the invalidity of substituting fuzzy set $\eta(x)$ for $\mu(x)$. Applying fuzzy relative entropy, we prove the conclusion that a pattern with high fuzziness is close to the classification boundary. Through the study of this article, in the process of learning fuzzy classifiers, beside of some assessment criteria such as classification accuracy, it is known that from the perspective of fuzziness of classifiers we should study their classification performance.

The paper is organized as following: in section 2, basic concepts and definitions are given. In section 3, we not only discuss the relationship between the fuzziness of patterns and the decision boundary, but also analyze the influence of fuzziness of fuzzy classifier on classification performance. Numerical experiments of our relevant conclusions are presented in section 4. This further demonstrates our conclusions. Section 5 concludes this article with recommendations of further study.

2 Definition and preliminaries

In this section we introduce some basic concepts required for subsequent development of the theory.

2.1 Fuzzy set and its fuzziness

Fuzzy set is a generalization of the crisp set that is well known to math and engineering students. Let $\mathcal{X} = \{x_1, x_2, \cdots, x_m\}$ be a finite domain of pattern points set. A characteristic function of set A defined on domain $\mathcal{X}$ assumes the following form:

$$A(x_i) = \begin{cases} 1 & if\ x_i \in A, \\ 0 & if\ x_i \notin A. \end{cases}$$

Characteristic functions $A : \mathcal{X} \to \{0, 1\}$ induce a constraint with well-defined boundaries on the elements of the domain $\mathcal{X}$ that can be assigned to a set A. The fundamental idea of fuzzy set is to relax this requirement by admitting intermediate values of class membership [37][10].

Definition 2.1. *A fuzzy set $\mathcal{A}$ is described by a membership function mapping the elements of domain $\mathcal{X}$ to the unit interval* $[0, 1]$, $\mu_{\mathcal{A}} : \mathcal{X} \to [0, 1]$.

Fuzzy set $\mathcal{A}$ can be considered as a set of the form $\{\mu_{\mathcal{A}}(x_1), \mu_{\mathcal{A}}(x_2), \cdots, \mu_{\mathcal{A}}(x_m)\}$, where $\mu_{\mathcal{A}}(x_i)$ denotes the corresponding degree of membership of x_i. The nearer the value of $\mu_{\mathcal{A}}(x)$ to 1, the higher the degree of membership of x in $\mathcal{A}$. If $\forall x_i \in \mathcal{X}$, $\mu_{\mathcal{A}}(x_i) = 1$ or $\mu_{\mathcal{A}}(x_i) = 0$, i.e., x_i does or does not belong to $\mathcal{A}$. Hence, $\mu_{\mathcal{A}}(x_i)$ reduces to the familiar characteristic function of a set A that is a crisp set. The membership functions are therefore synonymous of fuzzy sets. In a nutshell, membership functions generalize characteristic function in some way as fuzzy sets generalize sets.

Fuzzy set theory deals with ambiguity and imprecision of certain sets. For every fuzzy set, a measure of the degree of its "fuzziness" should be introduced. Let $R^+ = [0, +\infty)$, $\mathcal{X}$ is the domain set. $F(\mathcal{X})$ consists of all fuzzy set of $\mathcal{X}$. $\mu_{\mathcal{A}}(x)$ is the membership function of $\mathcal{A} \in F(\mathcal{X})$. The fuzziness of a fuzzy set could be measured by fuzzy entropy. Similar to Shannon's entropy, the fuzzy entropy of a finite fuzzy set $\mathcal{A}$ can be defined as follows [35].

Definition 2.2. *Let $\mathcal{A} = (\mu_1, \mu_2, \cdots, \mu_m)$ be a fuzzy set. The fuzziness of $\mathcal{A}$ can be defined as*

$$e(\mathcal{A}) = -\frac{1}{m}\sum_{i=1}^{n}(\mu_i \log \mu_i + (1-\mu_i)\log(1-\mu_i)) \tag{1}$$

The fuzziness of a fuzzy set defined by formula (1) attains its minimum when every element absolutely belongs to the fuzzy set or absolutely does not, i.e. $\mu_i = 1$ or $\mu_i = 0, \forall i \in \{1, 2, \cdots, m\}$. The fuzziness attains its maximum when the membership degree of each element is equal to 0.5, i.e. $\mu_i = 0.5, i = 1, 2, \cdots, m$.

2.2 Fuzziness of fuzzy classifier

Fuzzy classification algorithm is an important application of fuzzy set theory [5]. Let x be a pattern vector in the domain $\mathcal{X} = \{x_1, x_2, \cdots, x_m\}$, and $C = \{y_1, \cdots, y_c\}$ be a set of class labels. In this paper, we consider the fuzzy classifier of which the output for an input pattern $x \in \mathcal{X}$ is membership degree in each class. Namely, every components of this output vector describes the degree of membership that x belongs to corresponding class. The output vector could be denoted by $\mu(x) = (\mu_1(x), \cdots, \mu_c(x))^T$, satisfying $\sum_{i=1}^{c} \mu_i(x) = 1$. Compared to the crisp classifier, the advantage of fuzzy version is that no arbitrary assignment is made. Moreover, the pattern point's membership values provide a level of assurance to conform the resultant classification.

Given a set of training patterns $S = \{(x_1, \omega_1), (x_2, \omega_2), \cdots, (x_m, \omega_m)\}, \omega_i \in C, i = 1, \cdots, m$, a fuzzy partition of these patterns assigns the degree of membership of each sample in each of the c classes. The partition can be described by a membership matrix $M = (\mu_{ij})_{c\times m}$, where μ_{ij} denotes the membership degree of the j−th pattern x_j belonging to the i−th class. The elements in the membership matrix satisfy the following properties:

$$a) \quad \mu_{ij} \in [0, 1], \qquad b) \quad \sum_{i=1}^{c}\mu_{ij} = 1, \qquad c) \quad 0 \le \sum_{j=1}^{m}\mu_{ij} \le m. \tag{2}$$

Thus, if we have completed the training procedure of a classifier, we could obtain membership degree matrix M upon the m training patterns. For the j−th pattern x_j, the trained classifier will give an output vector represented as a fuzzy set $\mu_j = (\mu_{1j}, \mu_{2j}, \cdots, \mu_{cj})^T$. Based on formula (1), the fuzziness of the classifier on x_j is given by

$$e(\mu_j) = -\frac{1}{c}\sum_{i=1}^{c}(\mu_{ij} \log \mu_{ij} + (1-\mu_{ij})\log(1-\mu_{ij})) \tag{3}$$

Therefore, the fuzziness of the trained classifier can be defined as follows.

Definition 2.3. *Let the membership degree matrix of a fuzzy classifier on m training patterns with c classes be $M = (\mu_{ij})_{c\times m}$. The fuzziness of this classifier is given by*

$$e(M) = -\frac{1}{mc}\sum_{j=1}^{m}\sum_{i=1}^{c}(\mu_{ij} \log \mu_{ij} + (1-\mu_{ij})\log(1-\mu_{ij})) \tag{4}$$

Equation (4) defines the fuzziness of a trained fuzzy classifier. It plays a central role in investigating the performance of fuzzy classifiers. From definition 2.3, we could obtain that the fuzziness of a trained fuzzy classifier is the averaged fuzziness of the classifier's fuzzy outputs on all training patterns. Actually, it is more reasonable that the definition of a classifier's fuzziness should be in the entire sample space containing training patterns and unseen testing patterns. Nevertheless, the fuzziness for unseen patterns is generally unknown. But, for any supervised learning problem, a well-known assumption is that the training patterns have a distribution identical to the distribution of patterns in the entire space. Therefore, we use formula (4) as the definition of a classifier's fuzziness.

In order to further understanding fuzzy classifier, fuzzy K-nearest neighbor (K-NN) algorithm is introduced as an example.

Example 2.4. *The training set is S which has mentioned above. Before utilizing fuzzy classifier, a fuzzy partition of these labeled training patterns should be made, and a membership degree matrix $M = (\mu_{ij})_{c\times m}$ is obtained. For instance, one of the fuzzy partition method is K-NN rule. For any labeled pattern x_j, say x_j in $t-th$ class, we can find the K nearest neighbors of x_j in S. And then membership degree in each class is assigned according to the following equation:*

$$\mu_{ij} = \begin{cases} 0.51 + (n_i/K) * 0.49, & if \quad i = t \\ (n_i/K) * 0.49, & if \quad i \neq t \end{cases} \tag{5}$$

Where n_i is the number of the neighbors found belong to the $i-th$ class. According to equation (5), we could calculate the membership degree matrix $M = (\mu_{ij})_{c\times m}$.

In fuzzy K-NN algorithm, it must be searched K the nearest neighbors of an unseen pattern from the labeled pattern set. On the basis of the membership matrix $M = (\mu_{ij})_{c\times m}$, the class membership degree of a pattern x is calculated by the equation (6).

$$\mu_i(x) = \frac{\sum\limits_{j=1}^{K} \mu_{ij} ||x - x_j||^{-2(m-1)}}{\sum\limits_{j=1}^{K} ||x - x_j||^{-2(m-1)}} \tag{6}$$

Then $\mu(x) = (\mu_1(x), \mu_2(x), \cdots, \mu_c(x))^T$ is a membership degree vector of pattern x belonging to each class. As it shown in (6), the assigned membership degrees of x are influenced by the inverse of the distance from the K nearest neighbors and their class membership degrees. The inverse distance provides to weight a pattern vector's membership degree more if it is closer and less if it is farther from the considered pattern vector. The variable m determines how heavily the distance is weighted, when we compute each neighbor's contribution on the value of membership degree.

2.3 Entropy and relative entropy

In order to explore the relationship between the fuzziness of a pattern and the uncertainty of a random variable. Let us introduce the concepts of entropy and relative entropy [9].

The entropy of a random variable measures the uncertainty of the random variable. Let X be a discrete random variable with $\mathscr{X}$ and probability density function $p(x) = Pr\{X = x\}, x \in \mathscr{X}$.

Definition 2.5. *The entropy $H(X)$ of a discrete random variable X is defined as*

$$H(X) = -\sum_{x\in\mathscr{X}} p(x) \log p(x) \tag{7}$$

We use the convention that $0 \log 0 = 0$, which is justified by continuity since $x \log x \to 0$ as $x \to 0$. It does not change the entropy when terms of zero probability is added.

The relative entropy is a measure of the "distance" between two distributions.

Definition 2.6. *The relative entropy between two probability density functions $p(x)$ and $q(x)$ is defined as*

$$D(p||q) = \sum_{x\in\mathscr{X}} p(x) \log \frac{p(x)}{q(x)} \tag{8}$$

In the definition 2.6, we use the convention that $0 \log \frac{0}{0} = 0$, the convention that $0 \log \frac{0}{q} = 0$ and $p \log \frac{p}{0} = \infty$. Thus, if there is any symbol $x \in \mathscr{X}$ such that $p(x) > 0$ and $q(x) = 0$, then $D(p||q) = \infty$.

From the formula (8), we obtain that relative entropy is not a true distance between two distributions since it is not symmetric and does not satisfy the triangle inequality principal. $D(p||q)$ is a measure of the inefficiency of presuming that the distribution is q when the true distribution is p. Actually, we could consider that relative entropy measures the divergence between two distributions.

3 Discussion of classification performance based on fuzzy relative entropy

3.1 Fuzzy relative entropy

In a fuzzy classification problem, U and V are two different fuzzy classifiers on domain $\mathcal{X}$, C denotes the class labels set. For a pattern $x \in \mathcal{X}$, the outputs by classifier U and classifier V can be described as $\mu(x) = (\mu_1, \mu_2, \cdots, \mu_c)^T$ and $\eta(x) = (\eta_1, \eta_2, \cdots, \eta_c)^T$ respectively. Taking fuzzy set $\mu(x)$ for example, since $\sum_{i=1}^{c} \mu_i = 1, \mu_i \in [0, 1], i = 1, \cdots, c$, we can consider fuzzy set $\mu(x)$ as conditional probability distribution $Pr(Y|X = x)$, i.e., $Pr(Y = y_i|X = x) = \mu_i, i = 1, \cdots, c$. Therefore, the probability that x belongs to the $i-$th class is μ_i in $\mu(x)$ and η_i in $\eta(x)$. Of course, the probability that x doesn't belong to the ith class is $1 - \mu_i$ in $\mu(x)$ and $1 - \eta_i$ in $\eta(x)$. According to the definition of relative entropy, the divergence that x belongs to the $i-$th class between fuzzy set $\mu(x)$ and $\eta(x)$ is

$$D(\mu_i \| \eta_i) = \mu_i \log \frac{\mu_i}{\eta_i} + (1 - \mu_i) \log \frac{1 - \mu_i}{1 - \eta_i} \tag{9}$$

On the basis of formula (9), we can give a concept of fuzzy relative entropy between fuzzy set $\mu(x)$ and $\eta(x)$. It is defined as following.

Definition 3.1. *Fuzzy sets $\mu(x)$ and $\eta(x)$ are as above mentioned. The fuzzy relative entropy between them is defined as*

$$FD(\mu(x) \| \eta(x)) = \sum_{i=1}^{c} (\mu_i \log \frac{\mu_i}{\eta_i} + (1 - \mu_i) \log \frac{1 - \mu_i}{1 - \eta_i}) \tag{10}$$

According to formula (9), fuzzy relative entropy $FD(\mu(x) \| \eta(x)$ is the sum of relative entropy $D(\mu_i \| \eta_i), i = 1, 2, \cdots, c.$ i.e. $FD(\mu(x) \| \eta(x) = \sum\limits_{i=1}^{c} D(\mu_i \| \eta_i)$. Therefore, the divergence between fuzzy sets $\mu(x)$ and $\eta(x)$ can be measured by fuzzy relative entropy. The fuzzy relative entropy $FD(\mu(x) \| \eta(x)$ is a measure of the inefficiency of assuming that the fuzzy set is $\eta(x)$ when the true fuzzy set is $\mu(x)$. We use the convention that $0 \log \frac{0}{0} = 0$, $0 \log \frac{0}{q} = 0$ and $p \log \frac{p}{0} = \infty$. Therefore, if $\exists\, \mu_i \in (0, 1)$ and $\eta_i \in \{0, 1\}$, then $FD(\mu(x) \| \eta(x)) = \infty$. Therefore, suppose $\mu(x)$ is a fuzzy set even with lower fuzziness and $\eta(x)$ is a crisp set, then the divergence between $\mu(x)$ and $\eta(x)$ is infinite. The main property of fuzzy relative entropy is as follows.

Theorem 3.2. *Let $\mu(x)$ and $\eta(x)$ be two fuzzy sets as above mentioned. Then $FD(\mu(x) \| \eta(x)) \geq 0$ with equality if and only if $\mu(x) = \eta(x)$, i.e. $\mu_i = \eta_i, i = 1, \cdots, c$.*

Before proving this theorem, let us introduce some relevant knowledge [9].

Definition 3.3. *A function $f(x)$ is defined over an interval (a, b). If for any $x_1, x_2 \in (a, b)$ and $0 \leq \lambda \leq 1$, we have*

$$\lambda f(x_1) + (1 - \lambda) f(x_2) \leq f(\lambda x_1 + (1 - \lambda) x_2) \tag{11}$$

Then, function $f(x)$ is said to be concave over interval (a, b). A function $f(x)$ is strictly concave if equality holds only if $\lambda = 0$ or $\lambda = 1$.

Lemma 3.4. *(Jensen's Inequality) If $f(x)$ is a real continuous function that is concave, and $x_i \in R, p_i \leq 0, i = 1, 2, \cdots, n$, ,satisfying $\sum\limits_{i=1}^{n} p_i = 1, R$ is real number, then*

$$\sum_{i=1}^{n} p_i f(x_i) \leq f(\sum_{i=1}^{n} p_i x_i) \tag{12}$$

The equality case holds for all linear functions.

Now, we present a proof of theorem 3.2 as follows.

Proof. Since $p \log \frac{p}{0} = \infty$, without loss of generality, let $\eta_i \neq 0$ and $\eta_i \neq 1, i = 1, \cdots, c$

$$\begin{aligned} -FD(\mu(x) \| \eta(x)) &= -\sum_{i=1}^{c} (\mu_i \log \frac{\mu_i}{\eta_i} + (1 - \mu_i) \log \frac{1 - \mu_i}{1 - \eta_i}) = \sum_{i=1}^{c} (\mu_i \log \frac{\eta_i}{\mu_i} + (1 - \mu_i) \log \frac{1 - \eta_i}{1 - \mu_i}) \leq \sum_{i=1}^{c} \log(\mu_i \cdot \frac{\eta_i}{\mu_i} + (1 - \mu_i) \cdot \frac{1 - \eta_i}{1 - \mu_i}) \\ &= \sum_{i=1}^{c} \log(\eta_i + (1 - \eta_i)) = 0 \end{aligned} \tag{13}$$

Where the inequality of (13) follows from lemma 3.4.

Because $f(x) = \log(x)$ is a strictly concave function of x, we have equality (13) if and only if $\frac{\mu_i}{\eta_i} = \frac{1 - \mu_i}{1 - \eta_i} = t$ (t is a constant) and $\mu_i, \eta_i \in (0, 1)$, thus, $t = 1, \mu_i = \eta_i, i = 1, \cdots, c$. i.e., $\mu(x) = \eta(x)$. Hence, $FD(\mu(x) \| \eta(x)) = 0$ if and only if $\mu(x) = \eta(x)$. □

From definition 3.1 and theorem 3.2, we can obtain that fuzzy relative entropy measures the cost of substituting fuzzy set $\mu(x)$ for fuzzy set $\eta(x)$. As we expected, the cost is always nonnegative. Furthermore, the cost is equal to zero, if and only if $\mu(x) = \eta(x)$.

3.2 Classification boundary

The determination of decision boundaries plays a key role in classification tasks [21]. We could predict the class label of a pattern easily, when we have acquired the decision boundary. For instance, a linear classification problem, the decision boundary is described as $\omega^T x + b = 0$ by utilizing SVM classification algorithm. For a pattern x', if $\omega^T x' + b > 0$, it belongs to positive class, while $\omega^T x' + b < 0$, it is attributed to negative class. From reference [32], we obtained that the classification error rate for patterns near to the decision boundary is larger than that far from the boundary. That is to say, compared with patterns far from the classification boundary, the patterns near to the boundary may be misclassified easily. Therefor, we would prefer the algorithm which could classify patterns near to the classification boundary well. It is of great significance to study the classification boundary.

For classification problems with not very complex decision boundary, some algorithms could describe the boundaries explicitly with formulas. As aforementioned linear classification problem, the decision boundary can be described as a formula by using SVM classification algorithm. However, some algorithms could not give a specific expression of classification boundary. For example, fuzzy K-NN algorithm, it only provides the locus of points of classification boundary. Between these two type of classifiers, it is difficult to say which one is better. Nevertheless, there are a lot of classification problems with complex and highly nonlinear boundaries or without an explicitly delineated boundary. In this instance, fuzzy classification algorithms which can not give a specific formula to description the decision boundary may be better. Taking binary classification problem as an example, for an input pattern x, the fuzzy output is $\mu(x) = (\mu_1(x), \mu_2(x))^T$ by a fuzzy classifier. It means that the degree of x belonging to each class is equal when $\mu_1(x) = \mu_2(x) = 0.5$. Now, we could consider pattern x as a classification boundary point. Similarly, for classification problems with c classes, the fuzzy output is $\mu(x) = (\mu_1(x), \mu_2(x), \cdots, \mu_c(x))^T$. Let $\{\mu_1^*(x), \mu_2^*(x), \cdots, \mu_c^*(x)\}$ be a permutation of $\{\mu_1(x), \mu_2(x), \cdots, \mu_c(x)\}$ in decreasing order, the classification boundary is the locus $\{x|\mu_1^*(x) = \mu_2^*(x)\}$.

Entropy is a measure of uncertainty of a random variable. In c classes classification tasks, let Y be a random variable with set $C = \{y_1, \cdots, y_c\}$. In this paper, we focus on binary classification problems unless specified stated. For pattern x, $\mu(x) = (\mu_1, \mu_2)$ represents the fuzzy set of x by a fuzzy classifiers. As section 2.2 shows, $0 \leq \mu_1, \mu_2 \leq 1, \mu_1 + \mu_2 = 1$. Therefor, we could consider μ_i as posteriori probability $Pr(Y = y_i|X = x)$. Then, the fuzziness of x is

$$e(\mu(x)) = -\frac{1}{2}(\mu_1 \log \mu_1 + (1-\mu_1)\log(1-\mu_1) + \mu_2 \log \mu_2 + (1-\mu_2)\log(1-\mu_2)) = -(\mu_1 \log \mu_1 + \mu_2 \log \mu_2) \tag{14}$$

Meanwhile, under the condition $X = x$, the uncertainty of random variable Y could describe as

$$H(Y|X=x) = -(Pr(Y=y_1|X=x)\log Pr(Y=y_1|X=x) + Pr(Y=y_2|X=x)\log Pr(Y=y_2|X=x)) = -(\mu_1 \log \mu_1 + \mu_2 \log \mu_2) \tag{15}$$

From equations (14) and (15), we have the following conclusion.

Theorem 3.5. *In binary classification problems, for pattern x, the two statements, the fuzziness of its fuzzy output vector and the uncertainty of its class label, are equivalent. i.e., for a pattern x, the larger the fuzziness is, the larger the class uncertainty is.*

In two-class classification problems, pattern x may be a decision boundary point when $Pr(Y = y_1|X = x) = Pr(Y = y_2|X = x) = 0.5$. At this time, from the formula (7) of the definition of entropy, $H(Y \mid X = x)$ attains its maximal value. Therefore, the uncertainty of class random variable Y of pattern x attains maximal value when x is a classification boundary point.

Corollary 3.6. *In binary classification problems. Suppose pattern x is a boundary point. Then the uncertainty of class random variable Y of the pattern x attains maximal value, namely, $H(Y|X = x)$ attains maximal value, $Y \in \{y_1, y_2\}$.*

According to theorem 3.5 and corollary 3.6, we obtain the following conclusion.

Corollary 3.7. *The fuzziness of the output vector of an input pattern x by a fuzzy classifier reaches maximal value when x lies on classification boundary.*

In reference [32], Wang et al. propose that the fuzziness of x_1 is higher than that of x_2 when the distance from x_2 to the classification boundary is larger than that from x_1 to the boundary. For a two-class classification problem, Wang provide a proof through Euclidean distance 1- norm between decision boundary point and pattern point.

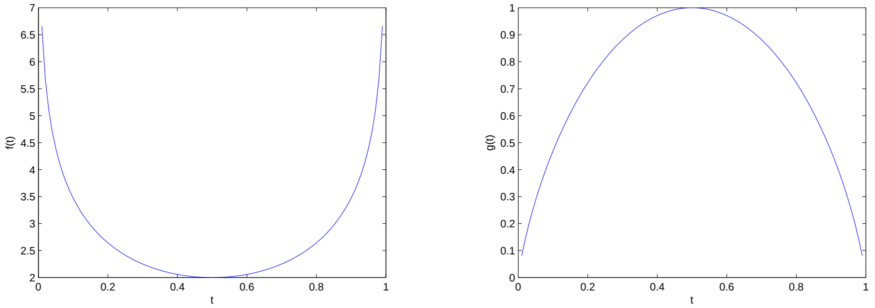
Now, from the angle of fuzzy theory to understand the fuzziness of pattern with fuzzy output. According to corollary 3.7, we obtain that the fuzziness of classification boundary points reaches maximal value. Then there may have the conclusion that the smaller the divergence between the fuzzy set of decision boundary points and the fuzzy set of pattern x is, the higher fuzziness of the pattern x is. As mentioned in section 3.1, the divergence between two fuzzy sets can be measured by fuzzy relative entropy. From above mentioned, the fuzzy vector of decision boundary points is $\mu = (0.5, 0.5)^T$. Thus, we have the following conclusion.

Theorem 3.8. *In binary classification problems, the fuzzy output of decision boundary points is fuzzy set $\mu = (0.5, 0.5)^T$. The fuzzy output of pattern x_1 and pattern x_2 are $\mu(x_1) = (\mu_{11}, \mu_{21})^T$ and $\mu(x_2) = (\mu_{12}, \mu_{22})^T$ respectively. Then $FD(\mu\|\mu(x_1)) \leq FD(\mu\|\mu(x_2))$ if and only if $e(\mu(x_1)) \geq e(\mu(x_2))$.*

Proof.

$$FD(\mu\|\mu(x_1)) = 2(0.5\log\frac{0.5}{\mu_{11}} + 0.5\log\frac{0.5}{\mu_{21}}) = -(\log\mu_{11} + \log(1-\mu_{11}))e(\mu(x_1)) = -(\mu_{11}\log\mu_{11} + (1-\mu_{11})\log(1-\mu_{11}))$$

$$FD(\mu\|\mu(x_2)) = 2(0.5\log\frac{0.5}{\mu_{12}} + 0.5\log\frac{0.5}{\mu_{22}}) = -(\log\mu_{12} + \log(1-\mu_{12}))e(\mu(x_2)) = -(\mu_{12}\log\mu_{12} + (1-\mu_{12})\log(1-\mu_{12}))$$



(a) $f(t) = -(\log t + \log(1-t))$ (b) $g(t) = -(t\log t + (1-t)\log(1-t))$

Figure 1: Figures of function $f(t)$ and $g(t)$.

If $\mu_{11} = 0$ or 1, $FD(\mu\|\mu(x)) = \infty$. Without loss of generality, let $\mu_{11} \in (0,1), \mu_{12} \in (0,1)$. Suppose $f(t) = -(\log t + \log(1-t))$, $g(t) = -(t\log t + (1-t)\log(1-t))$, $t \in (0,1)$. Functions $f(t)$ and $g(t)$ are symmetric, as shown in figure 1. The axis of symmetry of them are equal and can be expressed by $t = 0.5$. $f(t)$ is monotonic decreasing in interval $(0, 0.5)$ and monotonic increasing in interval $(0.5, 1)$. $g(t)$ is monotonic increasing in interval $(0, 0.5)$ and monotonic decreasing in interval $(0.5, 1)$. Therefore, we have

$$f(t_2) \leq f(t_1) \Longleftrightarrow t_1 \leq t_2 \Longleftrightarrow g(t_1) \leq g(t_2)$$

Thus,

$$FD(\mu \parallel \mu(x_1)) \leq FD(\mu \parallel \mu(x_2)) \Longleftrightarrow e(\mu(x_1)) \geq e(\mu(x_2))$$

From theorem 3.2, we get that the equality holds if and only if $\mu(x_1) = \mu(x_2)$. □

In fuzzy classification problems, each component of fuzzy set vector $\mu(x) = (\mu_1, \cdots, \mu_c)^T$ shows the degree of membership that pattern x belongs to the corresponding class. If pattern x_1 is near to pattern x_2, each feature of x_1 and x_2 is similar. Thus, there are similar in membership degrees of x_1 and x_2 belonging to each class. That is to say, the divergence between fuzzy set $\mu(x_1)$ and $\mu(x_2)$ is very little. Then, we have the following corollary.

Corollary 3.9. *Let patterns $x_1, x_2 \in \mathcal{X}$, fuzzy sets $\mu(x_1) = (\mu_{11}, \mu_{21})^T$ and $\mu(x_2) = (\mu_{12}, \mu_{22})^T$ are obtained by a fuzzy classifier. If pattern x_1 is near to x_2, the divergence between fuzzy sets $\mu(x_1)$ and $\mu(x_2)$ is small, namely, fuzzy relative entropy $FD(\mu(x_1)\|\mu(x_2))$ is small.*

Remark 3.10. *If $\exists\ i \in \{1, 2\}$ such that $\mu_{i1} \neq 0$ but $\mu_{i2} = 0$ or $\mu_{i2} = 1$, we have $FD(\mu(x_1)\|\mu(x_2)) = \infty$. For example, $\mu(x_1) = (0.99, 0.01)^T, \mu(x_2) = (1, 0)^T$, from the view point of class membership of classification problem, the divergence between $\mu(x_1)$ and $\mu(x_2)$ is small, but $FD(\mu(x_1)\|\mu(x_2)) = \infty$. The fuzzy relative entropy measures the divergence between two fuzzy sets, but $\mu(x_2) = (1, 0)^T$ is a crisp output. Because we focus on the patterns near to the boundary, without loss of generality, let us suppose $\mu_{i2} \neq 0, i = 1, 2$.*

From corollary 3.9, we obtain that the smaller the distance between pattern x and decision boundary point is, the smaller the divergence between the fuzzy set of pattern x and the fuzzy set of boundary points is. Combining with theorem 3.8, we have the following conclusion.

Corollary 3.11. *If the distance between pattern x_1 and decision boundary is smaller than that between pattern x_2 and boundary, the fuzziness of x_1 is larger than that of x_2.*

Corollary 3.11 is the conclusion in accordance with Wang et al. proposing in literature [32]. This further shows that it is more reasonable and more general to deal with fuzzy classification problems from fuzzy set theory view point rather than from the crisp idea.

3.3 Generalization abilities of fuzzy classifiers

It is general recognized that generalization capability of classifier is the most important index in classification learning process. From statistical learning theory, the generalization capability of a classifier obtained from training set is the ability to generalize well on unseen testing patterns which follows the same distribution as the training patterns. We note that most of the papers concerning generalization focus on the representational capabilities of classifier systems, such as upper error bound and VC dimension etc. [27].

According to theorem 3.8 and corollary 3.11, we have learned that the fuzziness of patterns nearing the decision boundary is high. Actually, as is known to all, the patterns with high fuzziness are difficult to assign class label to them. Hence, this type of patterns is liable to misclassification. In the same way, the patterns far away from decision boundaries own a low fuzziness. As Wang etc. stated in literature [32], in the training set, the patterns far away from the boundaries have little or no effect on the result of classification. That is to say, the training patterns with high fuzziness play a key role in learning classifiers. Therefore, we could start from the angle of fuzziness of patterns to study fuzzy classifier. And correctly classifying patterns with high fuzziness would be considered as one of the main goals.

For classification problems with simple decision boundaries such as linear separable cases, the boundaries should be determined easily in the process of learning training set. Because of problems simply, the uncertainty of class label of training patterns is low averagely. According to theorem 3.5, the fuzziness of training patterns is low averagely. Therefor, for this type of problems, we would like to find the decision boundaries directly rather than considering the fuzziness of patterns.

For problems with rather complex classification boundaries, it is very difficult to determine the boundaries. The uncertainty of class label of training patterns is much high averagely. Therefore, we should lend the fuzziness of patterns to learn classification model, instead of just thinking about decision boundaries.

As literature [9] shown, entropy not only could measure the uncertainty of random variable, but also measure the amount of information carried by the variable. The higher the uncertainty of random variable is, the more information the variable carries. In the light of theorem 3.5, we could say that the higher the fuzziness of fuzzy set is, the more information the fuzzy set carries. That is to say, a high fuzziness of fuzzy classifier means that it carries much information about the training pattern set. In order to try to ensure the objectivity in dealing with unseen patterns, we should make full use of information of the known pattern set and do not make too much subjective assumptions in the process of learning. Consequently, under the condition that a certain training accuracy threshold is maintained, a fuzzy classifier with high fuzziness will be selected. In other words, a higher fuzziness of fuzzy classifier would lead a better generalization performance.

As mentioned in section 3.1, the divergence between two fuzzy sets could be measured by fuzzy relative entropy. Now, the fuzzy relative entropy can be used to measure the divergence between two fuzzy classifiers.

Let us discuss binary classification problems. Suppose training pattern set $S = \{(x_1, \omega_1), (x_2, \omega_2), \cdots, (x_m, \omega_m)\}, \omega_i \in \{y_1, y_2\}, i = 1, 2, \cdots, m$. Two fuzzy classifiers U, V could be obtained by assigning class membership degree to each training pattern. The classifiers U and V can be expressed as follows.

$$U = (\mu_1, \mu_2, \cdots, \mu_m) = \begin{pmatrix} \mu_{11} & \mu_{21} & \cdots & \mu_{m1} \\ \mu_{12} & \mu_{22} & \cdots & \mu_{m2} \end{pmatrix}, \quad V = (\eta_1, \eta_2, \cdots, \eta_m) = \begin{pmatrix} \eta_{11} & \eta_{21} & \cdots & \eta_{m1} \\ \eta_{12} & \eta_{22} & \cdots & \eta_{m2} \end{pmatrix}$$

and

$$\begin{array}{lll} \mu_{i1} + \mu_{i2} = 1, & \eta_{i1} + \eta_{i2} = 1, & i = 1, 2, \cdots, m. \\ 0 \leq \mu_{1j} + \mu_{2j} + \cdots + \mu_{mj} \leq m, & 0 \leq \eta_{1j} + \eta_{2j} + \cdots + \eta_{mj} \leq m, & j = 1, 2. \end{array}$$

According to equation (10), the fuzzy relative entropy between two fuzzy classifiers U and V can be defined as follows

$$FD(U\|V) = 2\sum_{i=1}^{m}(\mu_{i1}\log\frac{\mu_{i1}}{\eta_{i1}} + (1-\mu_{i1})\log\frac{1-\mu_{i1}}{1-\eta_{i1}})$$

Let fuzzy classifier

$$W = (\omega_1, \omega_2, \cdots, \omega_m) = \begin{pmatrix} \omega_{11} & \omega_{21} & \cdots & \omega_{m1} \\ \omega_{12} & \omega_{22} & \cdots & \omega_{m2} \end{pmatrix}$$

The fuzziness of W is the highest of all fuzzy classifiers that have similar training accuracy and are obtained from learning training pattern set S. From the previous analysis, the fuzzy classifier with higher fuzziness carries more information from training pattern set when the two classifiers U and V have similar training accuracy. Furthermore, U and V are learned from training set S. Then, the fuzzy classifier with smaller divergence from W would have better generalization performance. Thus, we have the following conclusion.

Table 1: Date sets used in the experiments with the number of features (m), the number of classes (c) and the number of patterns (n).

No.	Data sets	m	c	n
1	synthetic	2	2	200
2	blood	4	2	748
3	haberman	3	2	306
4	heart	13	2	270
5	column3C	6	3	310
6	credit	14	2	690
7	wdbc	30	2	569

Theorem 3.12. *For a binary classification problem with rather complex and highly nonlinear decision boundaries. Let S be a training pattern set. Fuzzy classifiers U, V and W presented previously have the similar training accuracy threshold. If the fuzziness of fuzzy classifier U is higher than that of classifier V, i.e., $e(U) \geq e(V)$, we have $FD(W\|U) \leq FD(W\|V)$.*

Proof.

$$FD(W\|U) - FD(W\|V) = 2\sum_{i=1}^{m}(\omega_{i1}\log\frac{\omega_{i1}}{\mu_{i1}} + (1-\omega_{i1})\log\frac{1-\omega_{i1}}{1-\mu_{i1}}) - 2\sum_{i=1}^{m}(\omega_{i1}\log\frac{\omega_{i1}}{\eta_{i1}} + (1-\omega_{i1})\log\frac{1-\omega_{i1}}{1-\eta_{i1}})$$

$$= 2\sum_{i=1}^{m}(\omega_{i1}\log\frac{\eta_{i1}}{\mu_{i1}}+(1-\omega_{i1})\log\frac{1-\eta_{i1}}{1-\mu_{i1}}) \leq 2\sum_{i=1}^{m}\log(\omega_{i1}\cdot\frac{\eta_{i1}}{\mu_{i1}}+(1-\omega_{i1})\cdot\frac{1-\eta_{i1}}{1-\mu_{i1}}) \qquad (16)$$

The inequality of formula (16) follows by Lemma 3.4. Since $e(U) \geq e(V)$ and $e(W) \geq e(U)$, then $\eta_{i1} \leq \mu_{i1}$ and $\mu_{i1} \leq \omega_{i1}$.

$$\omega_{i1}\cdot\frac{\eta_{i1}}{\mu_{i1}} + (1-\omega_{i1})\cdot\frac{1-\eta_{i1}}{1-\mu_{i1}} = \frac{\omega_{i1}\eta_{i1}(1-\mu_{i1}+\mu_{i1}(1-\omega_{i1})(1-\eta_{i1})}{\mu_{i1}(1-\mu_{i1})} = \frac{\omega_{i1}\eta_{i1}+\mu_{i1}-\mu_{i1}\omega_{i1}-\mu_{i1}\eta_{i1}}{\mu_{i1}(1-\mu_{i1})}$$

Furthermore,

$$(\omega_{i1}\eta_{i1}+\mu_{i1}-\mu_{i1}\omega_{i1}-\mu_{i1}\eta_{i1}) - (\mu_{i1}(1-\mu_{i1})) = (\omega_{i1}-\mu i1)(\eta_{i1}-\mu_{i1}) \leq 0.$$

Thus, $\omega_{i1}\cdot\frac{\eta_{i1}}{\mu_{i1}} + (1-\omega_{i1})\cdot\frac{1-\eta_{i1}}{1-\mu_{i1}} \leq 1$. We have $\log(\omega_{i1}\cdot\frac{\eta_{i1}}{\mu_{i1}} + (1-\omega_{i1})\cdot\frac{1-\eta_{i1}}{1-\mu_{i1}}) \leq 0$ and $2\sum_{i=1}^{m}\log(\omega_{i1}\cdot\frac{\eta_{i1}}{\mu_{i1}} + (1-\omega_{i1})\cdot\frac{1-\eta_{i1}}{1-\mu_{i1}}) \leq 0$. From above all, we have the following inequality $FD(W\|U) - FD(W\|V) \leq 0, i.e., FD(W\|U) \leq FD(W\|V)$ Equality holds if and only if $\mu_{i1} = \eta_{i1}, i = 1, \cdots, m$. □

For the complex classification problem, theorem 3.12 shows that the fuzzy classifier with higher fuzziness has better generalization performance under the condition of satisfying certain classification accuracy.

4 Experimental demonstration

In this section, we present numerical experiments to validate the related conclusions of this paper. Fuzzy K-NN algorithm is used to realize our purpose by varying the value of K. As shown in example 1, we utilize formula (5) to fuzzify training set, and a membership matrix is obtained. Then, the membership degree of unseen pattern to each class is calculated according to formula (6). In order to clarify the problem more clearly, we first use an synthetic data set containing 200 patterns. Then, six benchmark data sets coming from UCI Repository of machine learning [12] are put to use, see table 1.

From the analysis of section 3.2, we obtained that the patterns near to classification boundary have high fuzziness and easily misclassified. In binary classification, for a pattern x, the fuzzy output is $\mu(x) = (\mu_1(x), \mu_2(x))^T$. Then, we estimate its boundary as $\{x|\mu_1(x) = \mu_2(x) = 0.5\}$. From figure 2, we could see that misclassified patterns and high fuzziness patterns are all near to the boundary. Almost, the misclassified patterns have high fuzziness. Now, according to the size of fuzziness, we divide each data set into two parts – high fuzziness patterns and low fuzziness patterns. As shown in figure 4, for each data set, the classification accuracy of patterns with high fuzziness is smaller than the patterns with low fuzziness. At the same time, figure 5 show that the fuzziness of misclassification pattern is the highest, the fuzziness of correct classified patterns is the lowest. Therefore, it is very important to correctly handle patterns with high fuzziness.

Definition 4.1. *Let set $A = \{a_1, a_2, \ldots, a_n\}$, and $a = max\{a_1, a_2, \ldots, a_n\}$, then the complementary of A is defined as,*

$$A^c = \{a - a_1, a - a_2, \ldots, a - a_n\}$$

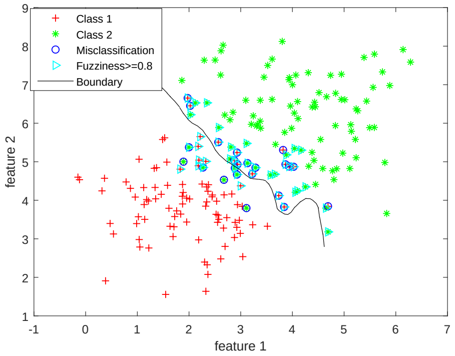


Figure 2: Relationship among high fuzziness patterns, misclassification patterns and boundary in synthetic data set.

The divergence between two fuzzy sets can be measured by fuzzy relative entropy. The smaller the divergence of two fuzzy sets, the smaller value of the fuzzy relative entropy. In fuzzy classification problems, we utilize fuzzy relative entropy to explore the divergence between patterns and classification boundary. From figure 2, patterns with high fuzziness are selected. Here, we present the fuzziness of patterns and the fuzzy relative entropy between these patterns and boundary. In figure 3, in order to remain consistency, we utilize the complementary set of fuzzy relative entropy set to substitute it. As shown in figure 3, it illustrates that the higher fuzziness of a pattern is, the smaller value of the fuzzy relative entropy between the pattern and classification boundary point is.

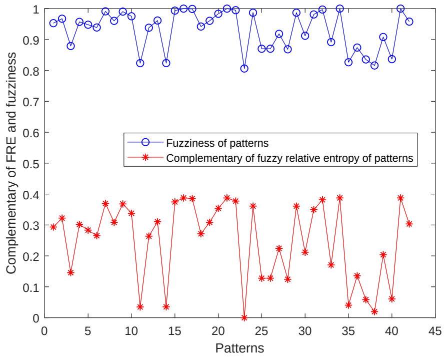


Figure 3: The fuzzy relative entropy and the fuzziness of patterns.

From figure 6, for each data set, the accuracy increases when the number of neighbors increase in general. Meanwhile, the fuzziness of fuzzy classifiers increases as well. For wdbc data set, we observe that the fuzziness of fuzzy classifiers is the lowest, while the classification accuracy is the largest. For heart data set, the fuzziness of fuzzy classifier is the highest, while the accuracy is the smallest. This again illustrates that patterns with high fuzziness are more difficult to classify correctly. In wdbc data set, the fuzziness of classifier does not change much with the variation of K, but in heart data set, the fuzziness of classifier changes greatly. This fully explicates that it is very well to consider the complex classification problems from the view of fuzziness of patterns. Meanwhile, fuzzy classifier with high fuzziness will lead a well generalization performance.

Through experimental analysis, we further understand the relationship between fuzziness and misclassification of patterns. We recognize that the fuzziness of patterns near to boundary is high. Consequently, it is very important to correctly classify patterns with high fuzziness. Meanwhile, the fuzziness of fuzzy classifiers plays a key role in quite complicated classification problems.

5 Conclusions

Because of the complexities of problems and the ambiguity of the data sets, fuzzy classification algorithms have paid more attention by researchers. In this paper, from the view of fuzziness, we discuss the performance of fuzzy classifiers of which the output for an input pattern is a fuzzy set. One of our main conclusion is that we put forward a concept of fuzzy relative entropy to measure the divergence between two fuzzy sets. As mentioned in section 3.2, in the binary classification problems, the decision

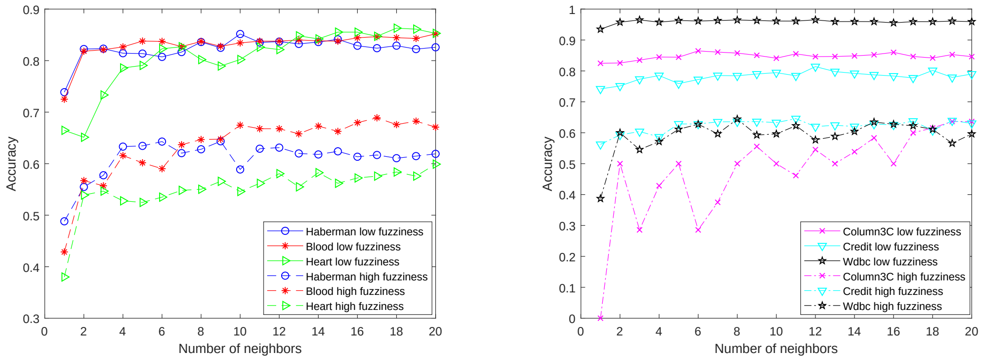


Figure 4: Classification accuracy of patterns with high fuzziness and low fuzziness.

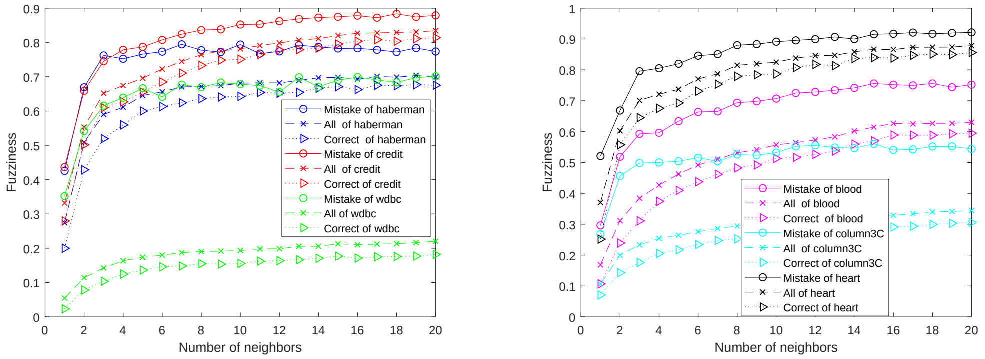


Figure 5: The fuzziness of misclassified patterns, all patterns and correct patterns.

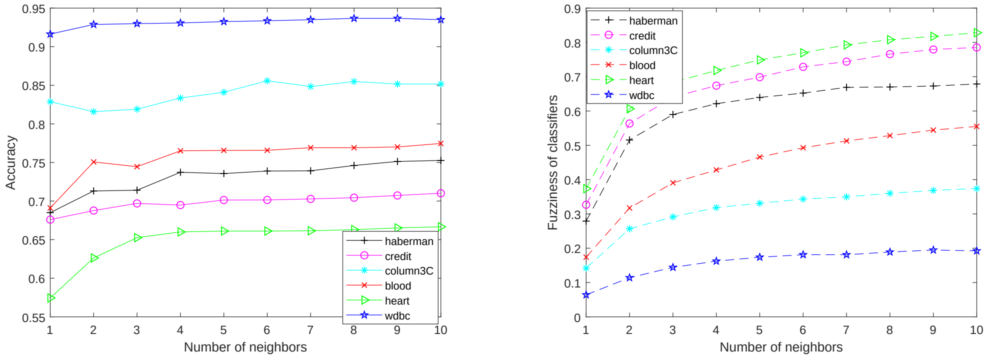


Figure 6: The accuracy and fuzziness of fuzzy $K-NN$ classifiers by varying K from 1 to 10.

boundary is the locus of the point where the output fuzzy set is $\mu(x) = (0.5, 0.5)$. Applying fuzzy relative entropy, we prove that ”the patterns with the higher fuzziness is closer to the classification boundary”. This conclusion is also in line with people's conventional ideas.

Meanwhile, we present that the fuzziness of a pattern's output fuzzy set and the uncertainty of its class label are equivalent. This conclusion further demonstrate that the higher the fuzziness of a pattern is, the closer the pattern to the boundary is. It also provides a theoretical basis for the conclusion that the classifier with higher fuzziness would own better generalization capability under the condition of maintaining a certain training accuracy.

It is not as traditionally thought that the fuzziness should be eliminated or reduced as much as possible in the process of learning . The lower the fuzziness is, the better the learning effect is. Actually, through the study of this article, one may know that sometimes fuzzy classifier with high fuzziness may lead to a better result, especially for problems with rather complex classification boundaries. And This is consistent with the principle of maximum entropy.

Acknowledgement

Authors would like to express their sincere thanks to the editors and the anonymous reviewers for helpful comments and suggestions. This work was supported by the National Natural Science Foundation of China (Grant No. 61573266).

References

[1] A. Agarwal, J. C. Duchi, *The generalization ability of online algorithms for dependent data*, IEEE Transactions on Information Theory, **59**(1) (2013), 573–587.

[2] E. Alpaydin, *Introduction to machine learning*, 3nd edition, MIT Press (2014), 1–5.

[3] G. Atalik, S. Senturk, *A new approach for parameter estimation in fuzzy logistic regression*, Iranian Journal of Fuzzy Systems, **15**(1) (2018), 91–102.

[4] R. Batuwita, V. Palade, *FSVM-CIL: Fuzzy support vector machine for class imbalance learning*, IEEE Transactions on Fuzzy Systems, **18**(3) (2010), 558–571.

[5] J. C. Bezdek, *Pattern recognition with fuzzy objective function algorithms*, New York: Plenum Press (1981), 203–239.

[6] J. J. Buckley, Y. Hayashi, *Fuzzy neural networks: A survey,* Fuzzy Sets and Systems, **66** (1994) 1–13.

[7] O. Chapelle, V. Vapnik, O. Bousquet, S. Mukherjee, *Choosing multiple parameters for support vector machines*, Machine learning, **46**(1) (2002), 131–159.

[8] S. Y. Chong, P. Tino, X. Yao, *Relationship between generalization and diversity in coevolutionary learning*, IEEE Transactions on Computational Intelligence and AI in Games, **1**(3) (2009), 214–232.

[9] T. M. Cover, J. A. Thomas, *Elements of information theory.* Jhon wiley & Sons, INC., Publication, Sec. edition (2006), 13-29.

[10] A. De Luca, S. Termini, *A definition of a nonprobabilistic entropy in the setting of fuzzy sets theory*, Information and Control, **20** (1972), 301-312.

[11] S. Decherchi, S. Ridella, R. Zunino, P. Gastaldo, D. Anguita, *Using unsupervised analysis to constrain generalization bounds for support vector classifiers*, IEEE Transactions on Neural Networks, **21**(3) (2010), 424–438.

[12] D. Dua, Karra Taniskidou, *UCI machine learning repository*, [http://archive.ics.uci.edu/ml], Irvine, CA: University of California, School of Information and Computer Science, 2017.

[13] R. O. Duda, P. E. Hart, D. G. Stork. *Pattern Classification.* Wiley Press, (2001), 394–453.

[14] S. Ezghari, A. Zahi, K. Zenkouar, *A new nearest neighbor classification method based on fuzzy set theory and aggregation operators*, Expert Systems with Applications, **80** (2017), 58–74.

[15] C. C. Gavin, L. C. Nicola, *On over-fitting in model selection and subsequent selection bias in performance evaluation*, Journal of Machine Learning Research, **11** (2010), 2079–2107.

[16] M. Gupta, R. Ragade, P. Yager, *Advances in fuzzy set theory and applications*, North-Holland Publishing Company, **22**(4) (1979), 623–633.

[17] T. Hastie, R. Tibshirani, J. H. Friedman. *The elements of statistical learning: Data mining, inference and prediction*, Springer (2009), 295–481.

[18] L. Hu, K. C. C. Chan, *Fuzzy clustering in a complex network based on content relevance and link structures*, IEEE Transactions on Fuzzy Systems, **24**(2) (2016), 456–470.

[19] A. Kandel, *Fuzzy techniques in pattern recognition.* New York, John Wiley, 1982.

[20] J. M. Keller, M. R. Gray, J. A. Givens, *A fuzzy K-nearest neighbor algorithm*, IEEE Transactions on Systems Man and Cybernetics, **SMC-15**(4) (1985), 580–585.

[21] C. Lee, D. A. Landgrebe, *Decision boundary feature extraction for neural networks*, IEEE Transactions on Neural Networks and Learning, **8**(1) (1997), 75–83.

[22] C. F. Lin, S. D. Wang, *Fuzzy support vector machines*, IEEE Transactions on Neural Networks and Learning, **13**(2) (2002), 464–471.

[23] O. Ludwig, U. Nunes, B. Ribeiro, C. Premebida. *Improving the generalization capacity of cascade classifers*, IEEE Transactions on Systems, Man and Cybernetics, **43**(6) (2013), 2135–2146.

[24] R. K. Nowicki, J. T. Starcaewski, *A new method for classification of imprecise data using fuzzy rough fuzzification*, Information Sciences, **414**(5) (2017), 33–52.

[25] A. H. M. Pimenta, H. de A. Camargo, *Genetic interval type-2 fuzzy classifier generalization: A comparative approach*, Eleventh Brazilian Symposium on Neural Networks, (2010), 194–199.

[26] X. Qiao, L. Zhang, *Flexible high-dimensional classification machines and their asymptotic properties*, Journal of Machine Learning Research, **16**(8) (2015), 1547–1572.

[27] S. S. Shwartz, S. B. David, *Understanding machine learning-from theory to algorithms*, New York, NY, USA: Cambridge Press (2014), 43–54.

[28] S. K. Shukla, M. K. Tiwari, *GA guided cluster based fuzzy decision tree for reactive ion etching modeling: A data mining approach*, IEEE Transactions on Semiconductor Manufacturing, **25**(1) (2012), 45–56.

[29] M. Stone, *Cross-validatory choice and assessment of statistical predictions*, Journal of the Royal Statistical Society, Ser. B, (Statist. Methodol.), **36**(2) (1974), 111–147.

[30] G. Varando, C. Bielza, P. Larranaga, *Decision boundary for discrete bayesian network classifiers*, Journal of Machine Learning Research, **16**(12) (2015), 2725–2749.

[31] P. P. Wang, S. K. Chang, *Theory and applications to policy analysis and information systems*. New York: Plenum Press, 1980.

[32] X. Z. Wang, H. J. Xing, Y. Li, Q. Hua, C. R. Dong, W. Pedrycz, *A study on relationship between generalization abilities and fuzziness of base classifiers in ensemble learning*, IEEE Transactions on Fuzzy Systems, **23**(5) (2015), 1638–1653.

[33] I. H. Witten, E. Frank, *Data mining: practical machine learning tools and techniques*, 2nd edition, Morgan Kaufmann Publication (2005), 1–36.

[34] Z. Yan, C. Xu, *Studies on classification models using decision boundaries*, in Proc. 8th IEEE International Conference of Cognitive Information, (2009), 287-294.

[35] D. S. Yeung, E. Tsang, *Measures of fuzziness under different uses of fuzzy sets*, Advanced Computing Intelligent Communication Computing Information Sciences, **298** (2012), 25–34.

[36] Y. Yuan, M. J. Shaw, *Induction of fuzzy decision trees.* Fuzzy Sets and Systems, **69** (1995), 125–139.

[37] L. A. Zadeh, *Fuzzy sets*, Information Control, **8** (1965) 338–353.

[38] Y. Zhang, R. Maciejewski, *Quantifying the visual impact of classification boundaries in choropleth maps*, IEEE Transactions on Visualization and Computer Graphics, **23**(1) (2017), 371–380.