Information Sciences 535 (2020) 16–27

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

An analysis on the relationship between uncertainty and misclassification rate of classifiers

Xinlei Zhou [a], Xizhao Wang [b,*], Cong Hu [c], Ran Wang [d]

[a] Big Data Institute, College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China
[b] The Guangdong Key Laboratory of Intelligent Information Processing, Shenzhen University, 518060, China
[c] Department of Materials Science and Engineering, College of Materials, Xiamen University, Xiamen, Fujian 361005, China
[d] College of Mathematics and Statistics, Shenzhen University, Shenzhen 518060, China

ARTICLE INFO

Article history:
Received 28 November 2019
Received in revised form 8 April 2020
Accepted 11 May 2020
Available online 21 May 2020

Keywords:
Uncertianty
Supervised learning
Classification problem
Misclassification rate
Statistical distribution

ABSTRACT

This paper provides new insight into the analysis on the relationship between uncertainty and misclassification of a classifier. We formulate the relationship explicitly by taking entropy as a measurement of uncertainty and by analyzing the misclassification rate based on the membership degree difference. Focusing on binary classification problems, this study theoretically and experimentally validates that the misclassification rate will definitely be upgrading with the increase of uncertainty if two conditions are satisfied: (1) the distributions of two classes based on membership degree difference are unimodal, and (2) these two distributions attain peaks when the membership degree difference is less and larger than zero, respectively. This work aims to provide some practical guidelines for improving classifier performance through clearly expressing and understanding the relationship between uncertainty and misclassification of a classifier.

1. Introduction

Classification algorithms have been widely used in pattern recognition, machine learning and many other popular areas of computer science. A classifier trained by a group of labeled objects aims to map each unlabeled object to the correct category. Specifically, the training process of a classification model is to get an approximate function (the classifier) by minimizing the error between the true and the estimated labels of training data. A well-trained classifier performs well in testing data. Misclassification rate is an important index for the evaluation of classification algorithms since the ultimate goal of classification is to reduce the misclassification rate of testing data and produce accurate predictions. To categorize testing objects in a low misclassification rate, it is essential to figure out factors affecting the misclassification rate.

Usually, the misclassification rate mainly depends on the impacts of data and model. (1) Some characteristics of the data have a great impact on the training of the classifier, such as the amount of training data, the noise ratio, the distribution of samples and the balanced ratio of categories; (2) Model selection and parameters of the training algorithm are also sensitive to the prediction results. There are many kinds of classification algorithms, which differ the misclassification rate for the same data set. According to the output value of classifiers, classification algorithms can be categorized into two classes, crisp-output and uncertain-output algorithms. The output of crisp-output algorithms is 0 or 1, such as decision tree (DT)

* Corresponding author.
E-mail address: xizhaowang@ieee.org (X. Wang).

https://doi.org/10.1016/j.ins.2020.05.059
0020-0255/

[22,32] and support vector machine (SVM) [10,29], while it is real values within the interval $[0, 1]$ in uncertain-output algorithms, such as feed-forward neural networks [9], Fuzzy SVM [14,23], Fuzzy DT [12,35], and fuzzy classification systems [18,21]. In this paper, we will intensively investigate the uncertain-output algorithms, focusing on the classifier with fuzzy output.

There are several types of methods to study the misclassification or generalization abilities of classifiers. Part of the researchers put their efforts on generating training and testing samples so that directly affect the evaluation of generalization performance, such as resampling methods [28,4,45], leave-one-out cross-validation [13,17,31], and generalization error formulation [5,19,20], etc. A number of theoretical focus on the estimation of error bounds. This type of studies includes the discussion on the performance bounds to overcome overfitting problems [3], the theoretical analysis for classifier ensemble bounds [15,26], the biased regularization approach to computing the generalization bound [6], and the bounds on the false and truth positive rates based on a VC-style analysis [16]. There are also a large amount of research [1,25,40] relating diversity to generalization. However, most of these studies focus on some specific types of classifiers, rather than get a more general conclusion regarding classification performance.

In the literatures [34,35], Wang et al. consider the classification performance from the perspective of the fuzziness of classifier outputs, but do not analyze the relationship between them, and their methods are limited only to rule-based systems. In 2015, Wang et al. [33] analyses a series of classifiers with fuzzy output, rather than focus on a specific type of classifier. They prove that the uncertainty of the classifier's output has a close relationship with the classification performance, but it is difficult to express explicitly for general cases. Some literatures [24,39] show the analysis on uncertainty is also beneficial to the improvement of performance for clustering algorithms. In 2017, R. W. et al. [36] makes the first attempt to investigate the relationship by incorporating the complexity of classification. However, the conclusion is more rely on empirical results rather than explained exactly with theorems.

Formulating the relationship between uncertainty and misclassification of a classifier would provide a strong support to the improvement and enhancement of machine learning algorithm performance. Moreover, the importance of finding this relationship can be verified from the view of increasing confidence degree and decreasing recognition error in patter classification [7]. Unfortunately, from the existing references, very few investigations to this relationship are found. In this paper, we formulate an explicit and clear expression between the uncertainty and misclassification rate of a classifier, considering from a comprehensive perspective. The formulated relationship can guild improving the performance of classifiers and providing a theoretical background for designing new algorithms. For example, giving guidelines for dynamic classifier selection, or generating new meta-features in the meta-learning system.

As usual, the entropy is used to represent the uncertainty of classifiers. The misclassification rate is analyzed based on value of M-discriminant function which are obtained from the classifier's output. By analyzing the monotonic relationship between the entropy and the misclassification rate, it is concluded that under certain conditions, the probability of misclassification upgrades inevitably with the increase of uncertainty.

There are two examples to illustrate the importance of analyzing the relationship. One example is that, when we suppose trained 2 classifiers based on the same dataset, the 2 classifiers have the same training accuracy but have difference uncertainty. Which classifier do you prefer? The problem can be easily solved if we make clear the relationship between the misclassification rate and the uncertainty of classifiers. Another example is that suppose there is a binary classification task with a testing sample x. A well-trained fuzzy classifier categorizes x as class A. Generally, whether the result is reliable is determined by the misclassification rate of the sample. However, it is difficult to calculate the exact misclassification rate of a specific sample. But the uncertainty of sample classification can be calculated by some means. If the relationship between uncertainty and misclassification is known, we can infer the misclassification rate through the uncertainty to decide whether to accept or reject the classification result.

It is fundamentally crucial to find a relationship between uncertainty and misclassification of a classifier for building a high-performance learning system. In this paper, (1) we confirmed the relationship between uncertainty and misclassification of a classifier from a new viewpoint of probability, (2) we expressed and formulated this relationship explicitly for the first time; (3) we experimentally proved that this relationship is not sensitive to different types of classifiers. This is the primary contribution of this work. In the remainder of this paper, Section 2 describes several fundamental conceptions of this work; Section 3 formulates and discusses the relationship between misclassification rate and uncertainty in detail; Section 4 provides experimental verifications on real data sets, and Section 5 concludes this paper.

2. Definition of related concepts

Generally, there are two processes in supervised learning, modeling(or training) and prediction. As a kind of typical supervised learning method, the key point of classification algorithms is to figure out a discrete-valued function that maps each given object to a class label. Given a training set X that contains N arbitrarily distinct samples with c categories, i.e., $X = \{(x_i, y_i)\}_{i=1}^{N} \subset R^n \times \{0,1\}^c$, where $x_i = [x_{i1}, x_{i2}, \ldots, x_{in}]$ is the ith training samples, $y_i = [y_{i1}, y_{i2}, \ldots, y_{ic}]$ is the label vector of x_i, n is the number of features, and c is the number of classes. Testing set T is formatted as $T = \{(t_i, y_i)\}_{i=1}^{N'} \subset R^n \times \{0,1\}^c$. The training and testing process of a classifier that takes the RWNN(Random Weight Neuronal Network) [2] as an example is shown as Algorithm 1.

Algorithm 1 Train RWNN classifier and compute testing accuracy

Input:

Training set $X = \{(x_i, y_i)\}_{i=1}^{N} \subset R^n \times \{0,1\}^c$; Testing set $T = \{(t_i, y_i)\}_{i=1}^{N'} \subset R^n \times \{0,1\}^c$; Activation function $f(x)$; Number of hidden nodes $\tilde{N}$

Output:

Testing accuracy of the trained classifier.

1. Randomly assign input weight and bias.
2. Calculate the output weight by inputting training set X into the network.
3. Obtain the membership degrees, M_{ij} where $i = 1, 2, \ldots, N, j = 1, 2, \ldots, c$, of testing sample t_i by putting into the well-trained model.
4. Sign each testing data with the label that has the biggest value in membership degrees.
5. Calculate testing accuracy by comparing the true labels with predicted labels of testing set T.

There are many different ways to represent classifiers. One of the most useful is in terms of a set of discriminant functions $g_j(x_i), i = 1, \ldots, n,\ j = 1, \ldots, c$ [7]. The classifier is said to assign a sample x_i to class ω_j if

$$g_j(x_i) > g_k(x_i) \qquad for\ all\ j\neg = k. \tag{1}$$

The membership matrix $M = [m_{ij}]$ can be obtained as the output of classifier, where m_{ij} is the membership degree of the ith sample belonging to the jth category.

For binary classification problem, which contains class ω_1 and ω_2. Given a sample x_i, the membership degrees of x_i belonging to classes ω_1 and ω_2 are m_1 and m_2, respectively. Instead of using two discriminant functions g_1 and g_2, it is more common to define a single discriminant function

$$g(x_i) = g_1(x_i) - g_2(x_i). \tag{2}$$

In this paper, $g_1(x_i)$ and $g_2(x_i)$ is defined as

$$\begin{aligned} g_1(x_i) &= \frac{m_{i1}}{m_{i1}+m_{i2}} \\ g_2(x_i) &= \frac{m_{i2}}{m_{i1}+m_{i2}} \end{aligned} \tag{3}$$

which are based on membership degrees m. In the rest of this paper, we call the discriminant function $g(x_i)$ as M-discriminant function. The decision rule is

$$Decide\ \omega_1\ if\ g(x_i) > 0;\ otherwise\ decide\ \omega_2. \tag{4}$$

Sign the value of M-discriminant function $g(x_i)$ with v. Given a sample of a two categories set which contains classes ω_1 and ω_2. Accroding to Eq. (4), the discriminant rule of binary classification problems can be written as

$$\begin{cases} x \in \omega_1, & v > 0 \\ x \in \omega_2, & v < 0 \\ x \in \text{rand}(\omega_1, \omega_2), & v = 0 \end{cases}. \tag{5}$$

Primary concepts discussed in this paper are uncertainty and misclassification rate. Explanations of these two concepts are given below in detail.

2.1. Uncertainty of classifier output

In this paper, information entropy [27] is used to depict the uncertainty of classifier's output. Information entropy, solving the problem of quantification and measurement of information, was introduced in 1984 by Claude Elwood Shannon, the father of information theory. Information entropy is defined on a probability distribution, evaluating the impurity of classes in a set.

Suppose there are n events in a probability system $S = (p_1, p_2 \ldots, p_n)$, where $\sum_{i=1}^{n} p_i = 1$. p_i is the probability of the ith event, x_i. Information entropy is defined as

$$H(S) = -\sum_{i=1}^{n} p_i log_2 p_i. \tag{6}$$

Specifically, as mentioned in [30], there are some properties of information entropy, summarized as follows

1. *If the probability of an event is 1, then the uncertain degree of the whole system is 0, and the information entropy is determinist*

$$H = H(1,0,0,\ldots,0) = H(0,\ldots,0,1,0,\ldots,0) = H(0,0,\ldots,1) = 0. \tag{7}$$

2. *The calculation of information entropy is independent of the order that events occur in the probability system.* Suppose the probability distribution of n events system is $(p_1, p_2, \ldots, p_n)$ and the order of events change, the new probability distribution is $(p\prime_1, p\prime_2, \ldots, p\prime_n)$. The following relationship can be established

$$H(p_1, p_2, \ldots, p_n) = H(p\prime_1, p\prime_2, \ldots, p\prime_n). \tag{8}$$

3. *The expression of information entropy is unimodality and reaches a maximum when* $p_1 = p_2 = \ldots = p_n = \frac{1}{n}$,

$$max(H) = -\sum_{i=1}^{n} \frac{1}{n} log_2 \frac{1}{n} = -log_2 \frac{1}{n}. \tag{9}$$

Given a sample x with output $(p, 1-p)$ where p and $1-p$ represent the probability of x belonging A and B respectively. The uncertainty(U) of classifier's output can be formulated as

$$U = -plog_2 p - (1-p)log_2(1-p). \tag{10}$$

By solving

$$\frac{dU}{dp} = log_2 \frac{(1-p)}{p} = 0,$$

we get that U attains its maximum at $p = \frac{1}{2}$. The second derivative of U is calculated as

$$\frac{d^2U}{d^2p} = -\frac{1}{p(1-p)ln2}.$$

Obviously, $\frac{d^2U}{d^2p} < 0$ when $0 \leqslant p \leqslant 1$. Thus, we can conclude

$$\begin{cases} U \text{ monotonically increasing}, & 0 \leqslant p < \frac{1}{2} \\ U \text{ monotonically decreasing}, & \frac{1}{2} < p \leqslant 1 \\ max(U), & p = \frac{1}{2} \end{cases}. \tag{11}$$

In this case, the expression of M-discriminant function in Eq. (2) can be rewritten as

$$v = g(x) = p - (1-p) = 2p - 1, \tag{12}$$

we can conclude

$$\begin{cases} v > 0, & p > \frac{1}{2} \\ v < 0, & p < \frac{1}{2} \end{cases}. \tag{13}$$

Combine with Eq. (11), the relationship between uncertainty and the value of M-discriminant function (v) can be explained as

$$\begin{cases} \text{Uncertainty monotonically decreasing}, & v > 0 \\ \text{Uncertainty monotonically increasing}, & v < 0 \end{cases}, \tag{14}$$

which means, the value of M-discriminant function can intuitively quantify the magnitude of the uncertainty of the classifier's output. In addition, the smaller the absolute value of M-discriminant function is, the greater uncertainty is.

2.2. Misclassification rate

Generally speaking, the purpose of training a model is to reduce misclassification rate of unseen samples. Let S be a finite space of samples, and X be a subset of S. Suppose that $F(x)$ is a function defined on S, an estimator function $f(x)$ defined on S can be given by a training algorithm based on values of $F(x)$ in X. The function $f(x)$ has the range of value as same as $F(x)$ has.

As for the classification problem in machine learning field, a classifier $f(x)$ is well-trained on a training set X. The expression $f(x)$ is expected to be infinitely close to the function $F(x)$ on the whole space S including the training set X. The misclassification rate $P_e(f)$ of classifier $f(x)$ on testing set $T(T = \{x | x \in S - X\})$ is the most important index of classifier performance evaluation. Thus, according to the generalization definition in [36], the misclassification rate can be calculated by

$$P_e(f) = \frac{|\{x : x \in T, F(x) \neq f(x)\}|}{|T|}, \tag{15}$$

where $|\ |$ donates the number of elements in a set, $F(x) \neq f(x)$ represents the inconsistency between testing results of the classifier and the original labels.

To analyze the changing tendency of misclassification rate, we define it based on M-discriminant function which is formulated as Eq. (2), and assign value of M-discriminant function with v. There is more detail about the definition below.

Given a binary classification set $X = \{(x_i, y_i)\}_{i=1}^{N} \subset R^n \times \{0, 1\}^2$, where x_i is the ith training samples, $y_i = [y_{i1}, y_{i2}]$ is the label vector of x_i. According to the value of y_i, X can be divided into two subsets, class $A = \{x|x \in X, y = [1, 0]\}$ and class $B = \{x|x \in X, y = [0, 1]\}$. Suppose $f(x)$ is the well-trained classifier, from which membership degrees $[m_1, m_2]$ of sample x are acquired. As defined in Eq. (2), value of M-discriminant fucntion is $v = g(x)$.

Suppose classes A and B of set X are normally distributed, respectively. As shown in Fig. 1, $peak_1$ and $peak_2$ are values of the abscissa where two distributions reach their peaks, respectively. For a new sample x, the value of M-discriminant fucntion is $v(peak_1 < v < peak_2)$, and the probability of sample x belonging to the two classes is denoted as $P_A(v)$ and $P_B(v)$ (values of density function of distribution A and B at v). According to the discriminant rule of classification problems clarified in Eq. (5), a sample would be categorized as class A if it locates in the range of $v > 0$, and be taken as class B when $v < 0$. Once the true label of sample x is class B but x locates in $v > 0, x$ would be misclassified as class A. Similarly, misclassification happens when the original label of the sample is class A but falls in $v < 0$. Therefore, the misclassification rate of sample x with the value of M-discriminant function v can be defined as

$$P_e(v) = \begin{cases} \frac{P_B(v)}{P_A(v)+P_B(v)}, & v > 0 \\ \frac{P_A(v)}{P_A(v)+P_B(v)}, & v < 0 \end{cases}. \tag{16}$$

We have three remarks about the definition of misclassification rate: (1) from the view of pattern recognition, $P_A(v)$ and $P_B(v)$ are probabilities of the sample x belonging to classes A and B. The sample x with the value of M-discriminant fucntion (v) and a true label class B, will be misclassified as class A when $v > 0$. It means, under the condition of $v > 0$, sample x will be misclassified with probability $P_B(v)$ as long as the true label of x is class B. Similarly, when $v < 0$, the misclassifying probability is $P_A(v)$ while sample x is belonging to class A. In Eq. (16), $P_e(v)$ is the form of uniformization of $P_A(v)$ and $P_B(v)$; (2) it is unnecessary to assume that the distributions of two classes based on value of M-discriminant function are normal. The only requirement is that the density functions of two categories are unimodal, respectively. This point is very important since theoretically, we may not know the distribution of data but practically we can check the impinal distribution for given data sets; (3) The study on extension from one dimensional to multidimensional case, which involves a complex formulation for multiple dimensional distributions and its marginal distributions, will be conducted in future work. We won't discuss it further in this article.

3. Relationship between uncertainty and misclassification rate

In this section, we give a detailed analysis on the relationship between uncertainty and misclassification rate based on M-discriminant fucntion.

As mentioned in Section 2.1, we use information entropy to represent the uncertainty in this work. For binary classification problem, assuming the probability of the sample x belonging to class A and class B is p and $1 - p$. The relationship between uncertainty and the value of M-discriminant fucntion (v) can be explained as Eq. (14), which means the value of M-discriminant function can intuitively quantify the magnitude of the uncertainty of the classifier's output. In addition, the smaller the absolute value of M-discriminant function, the greater uncertainty.

To clarify the relationship between both two concepts, the variable ΔP_e is introduced:

$$\Delta P_e = P_e(v') - P_e(v), \tag{17}$$

where $v' = v + \Delta v, \Delta v$ is a positive real number, and $P_e(v')$ is defined as:

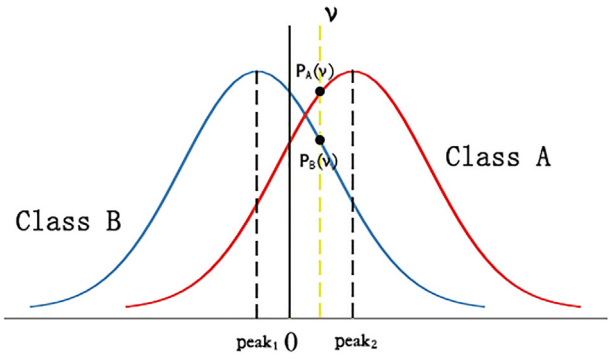

Fig. 1. Distribution of samples based on M-discriminant function.

$$P_e(v') = \begin{cases} \frac{P_B(v+\Delta v)}{P_A(v+\Delta v)+P_B(v+\Delta v)}, & v > 0 \\ \frac{P_A(v+\Delta v)}{P_A(v+\Delta v)+P_B(v+\Delta v)}, & v < 0 \end{cases}, \tag{18}$$

Thus, ΔP_e can be culculated as

$$\begin{aligned} \Delta P_e &= P_e(v') - P_e(v) \\ &= \frac{P_B(v+\Delta v)}{P_A(v+\Delta v)+P_B(v+\Delta v)} - \frac{P_B(v)}{P_A(v)+P_B(v)} \\ &= \frac{P_A(v)P_B(v+\Delta v)-P_B(v)P_A(v+\Delta v)}{(P_A(v+\Delta v)+P_B(v+\Delta v))(P_A(v)+P_B(v))} \\ &= \frac{P_A(v)P_B(v+\Delta v)-P_A(v)P_B(v)+P_A(v)P_B(v)-P_B(v)P_A(v+\Delta v)}{(P_A(v+\Delta v)+P_B(v+\Delta v))(P_A(v)+P_B(v))} \end{aligned},$$

where denominator is always greater than 0. Thus, on the right side of the equation, only the numerator needs to consider. Dividing the numerator into two formulations under the condition of $v > 0$, it is easy to view that the numerator is less than 0, since

$$\begin{cases} P_A(v)P_B(v+\Delta v) - P_A(v)P_B(v) < 0 \\ P_A(v)P_B(v) - P_B(v)P_A(v+\Delta v) < 0 \end{cases}. \tag{19}$$

Thus, $\Delta P_e < 0, \ if \ v > 0$. Similarly, under the condition of $v < 0$, the value of ΔP_e is always over 0. Therefore, we get a conclusion as follows

$$\begin{cases} \Delta P_e < 0, & v > 0 \\ \Delta P_e > 0, & v < 0 \end{cases}, \tag{20}$$

which means, the misclassification rate can be evaluated by M-discriminant function, and the monotonic relationship between them is illustrated as follows

$$\begin{cases} Miscalssification\ rate\ monotonically\ increasing, & v > 0 \\ Misclassification\ rate\ monotonically\ decreasing, & v < 0 \end{cases}. \tag{21}$$

As for two-category data sets, we can view the relationship between uncertainty and misclassification rate by combining Eq. (14) with Eq. (21). Both uncertainty and misclassification rate increase monotonically with the shrinking of M-discriminant fucntion. In other words, the misclassification rate would increase with the growing up of uncertainty under this two condition:

1. The distributions of categories A and B based on value of M-discriminant function are unimodal. In other words, two distributions realize maximums when the shared independent variable v reaches $peak_1$ and $peak_2$, respectively.
2. The inequality holds well, $peak_1 * peak_2 \leqslant 0$, where the $peak_1$ and $peak_2$ are two values the variable v takes, as explained in Section 2.1.

4. Empirical study

This section presents the experiments conducted on 16 data sets from UCI machine learning repository. Table 1 lists details about each data set. In this study, each multiclass data set is transfered into binary one by randomly selecting 50%

Table 1
Selected Data Sets for Experiments.

No.	Data Set	# Samples	# Attributes	# classes	Class/Distribution
1	Australian	690	14	2	383/307
2	Autism	702	21	2	513/189
3	Breast	569	9	2	357/212
4	ClaveVector	10800	20	2	9015/1785
5	Credit	653	24	2	296/357
6	German	1000	24	2	300/700
7	Ionosphere	351	34	2	225/126
8	MAGIC	19020	11	2	6688/12332
9	Mushroom	8124	22	2	4208/3916
10	Pima	768	8	2	500/268
11	Sonar	208	60	2	97/111
12	Spambase	4601	57	2	2788/1813
13	OptDigits	5620	64	10	2822/2798*
14	Pen	10992	16	10	5629/5363*
15	Satellite	6435	36	6	4199/2236*
16	Yeast	1484	8	10	2788/1813*

as positive and the rest 50% as negative. Samples with missing value are deleted in data preprocessing phase, and the rest samples are standardized before passed down as input to the classifier. As mentioned, RWNN is adopted as training algorithm in our experiments. Articles [11,37,42,44,43] give more details about RWNN and mention that the most significant advantage of RWNN is high processing speed since it assigns wight by the random mechanism.

4.1. Expermental design

To reflect the performance of the model more objectively, we repeat 10 experimental trials and calculate the average of accuracy, misclassification rate, and uncertainty based on 10 results. The number of hidden nodes in RWNN is set as 20, and the sigmoid activation function is utilized. We randomly take 30 percent of each data set as the testing set, and the remaining 70 percent data are used for training. Experiments are implemented in Python 3.0 and executed on a computer with the Mac operation system, an i7-8750H CPU, and 32 GB of RAM.

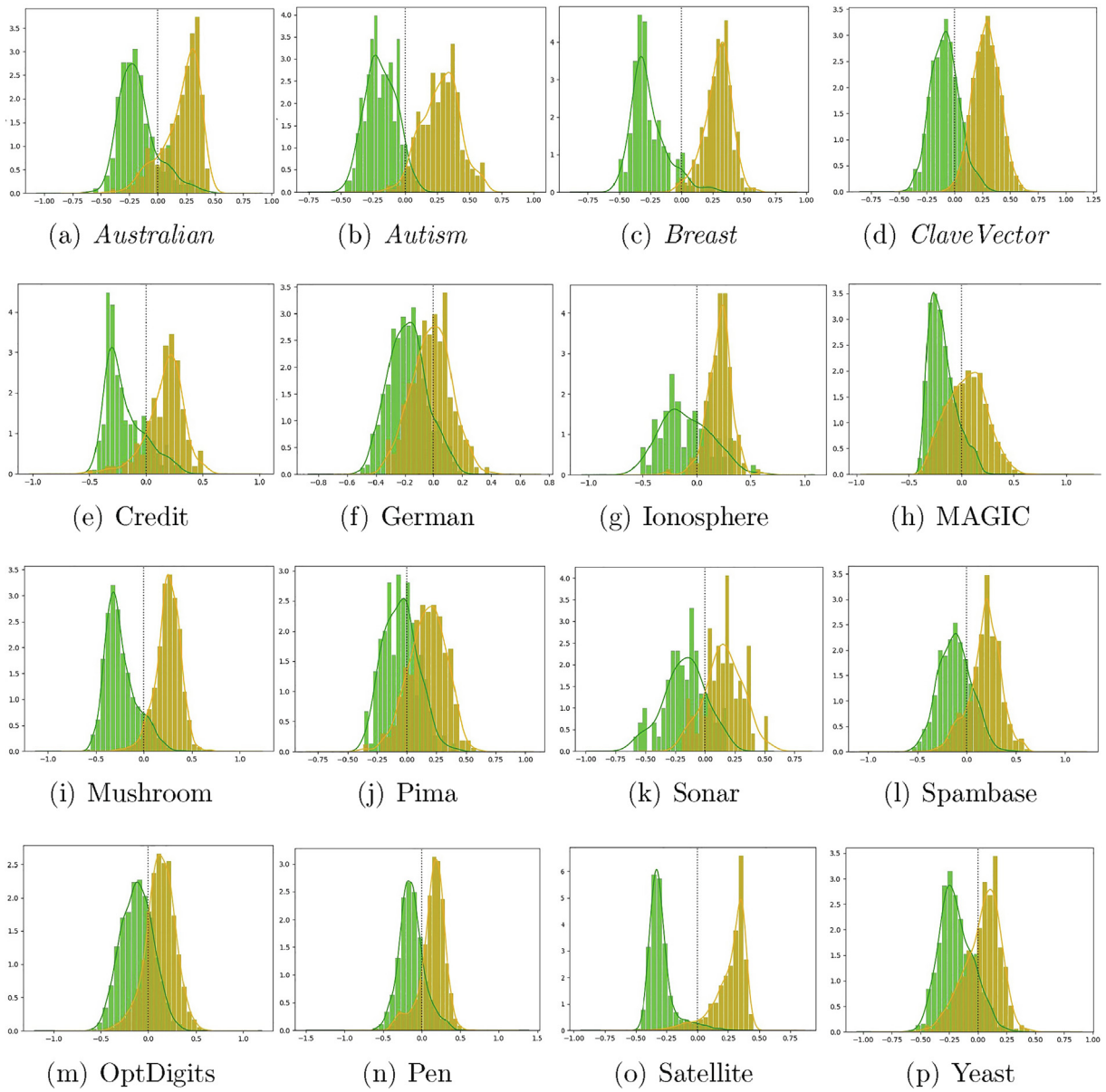

Fig. 2. Distribution of real data sets based on M-discriminant function (taking RWNN as the classifier).

4.2. Expermental analysis

As described in Section 3, we formulate the relationship between uncertainty and misclassification rate under certain conditions. To verify the correctness of our conclusions by experiments, we need to first check whether the selected data sets meet with above conditions. As shown in Fig. 2, the distribution of samples in class A and class B based on M-discriminant function are colored with green and orange, respectively. Each data set consists of two unimodal distributions which attain peaks when the value of M-discriminant function is less and larger than 0, respectively.

For each testing set, 5 uncertainty levels are generated by equally dividing the interval between the maximum and minimum entropy results. According to the division, the average misclassification rate for each uncertainty level can be calculated.

Fig. 3 details the relationship between uncertainty and misclassification rate under certain conditions. It is noteworthy, in all data sets, the misclassification rate is increasing with the climbing of uncertainty, which is well matching the conclusion acquired in Section 3.

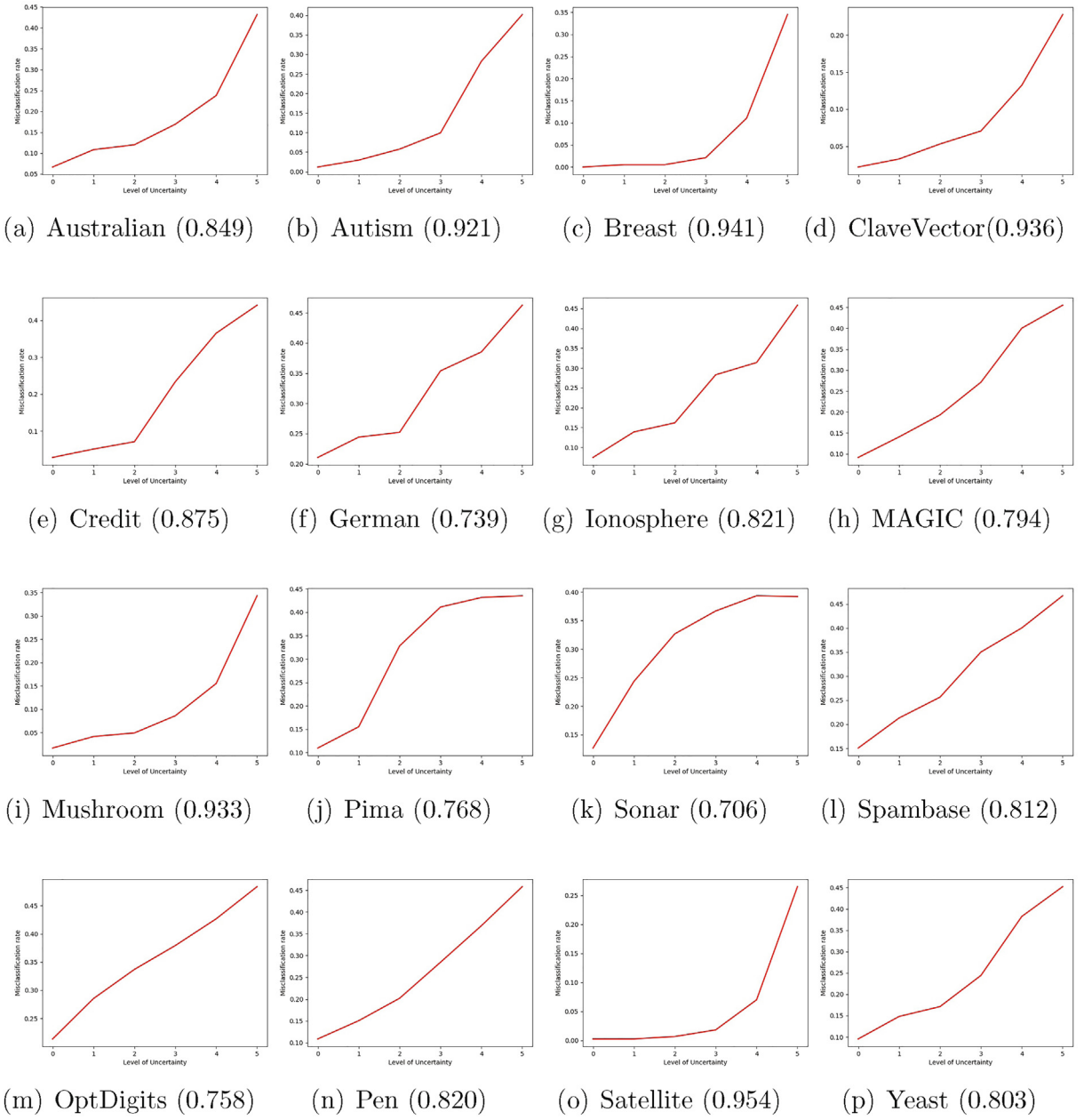

Fig. 3. Relationship between misclassification rate and uncertainty of RWNN classifier on real data sets. The numbers in brackets are testing accuracy of each data set, which is the mean of repeating ten experimental trials.

Table 2
Testing accuracy of RWNN and SVM.

	1	2	3	4	5	6
RWNN	0.849 ★	0.921 ★	0.941 ★	0.936 ★	0.875 ★	0.739 •
SVM	0.865	0.962	0.965	0.999	0.878	0.740
	7	8	9	10	11	12
RWNN	0.821 •	0.794 •	0.933 ★	0.767 •	0.706 •	0.812 •
SVM	0.943	0.796	1.000	0.766	0.706	0.820
	13	14	15	16		
RWNN	0.758 •	0.820 •	0.954 ★	0.803 •		
SVM	0.986	0.997	0.980	0.830		

Note: The "★" and "•" represent the data set with high and low accuracy respectively, according to the result of RWNN.

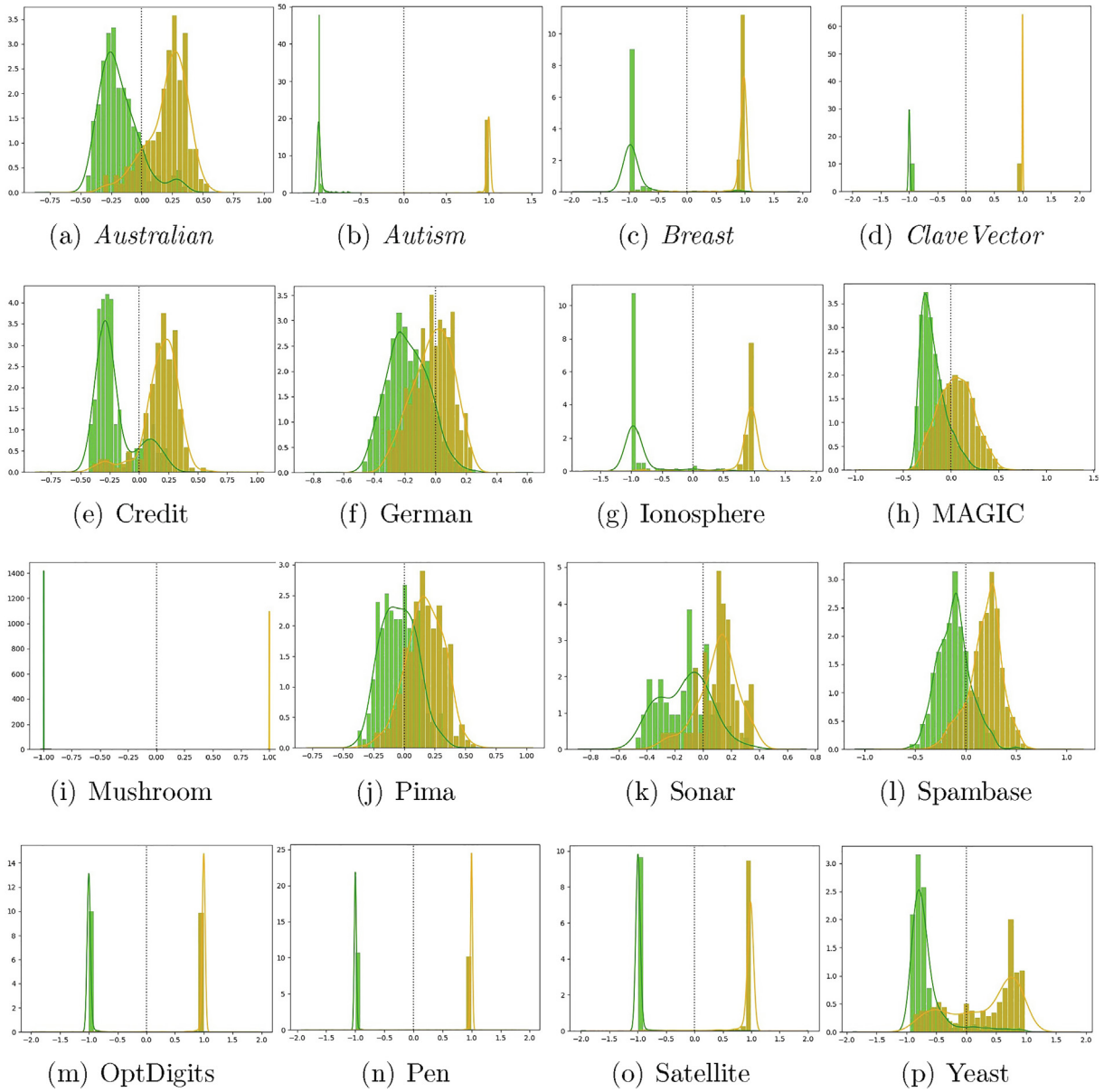

Fig. 4. Distribution of real data sets based on M-discriminant function (taking SVM as the classifier).

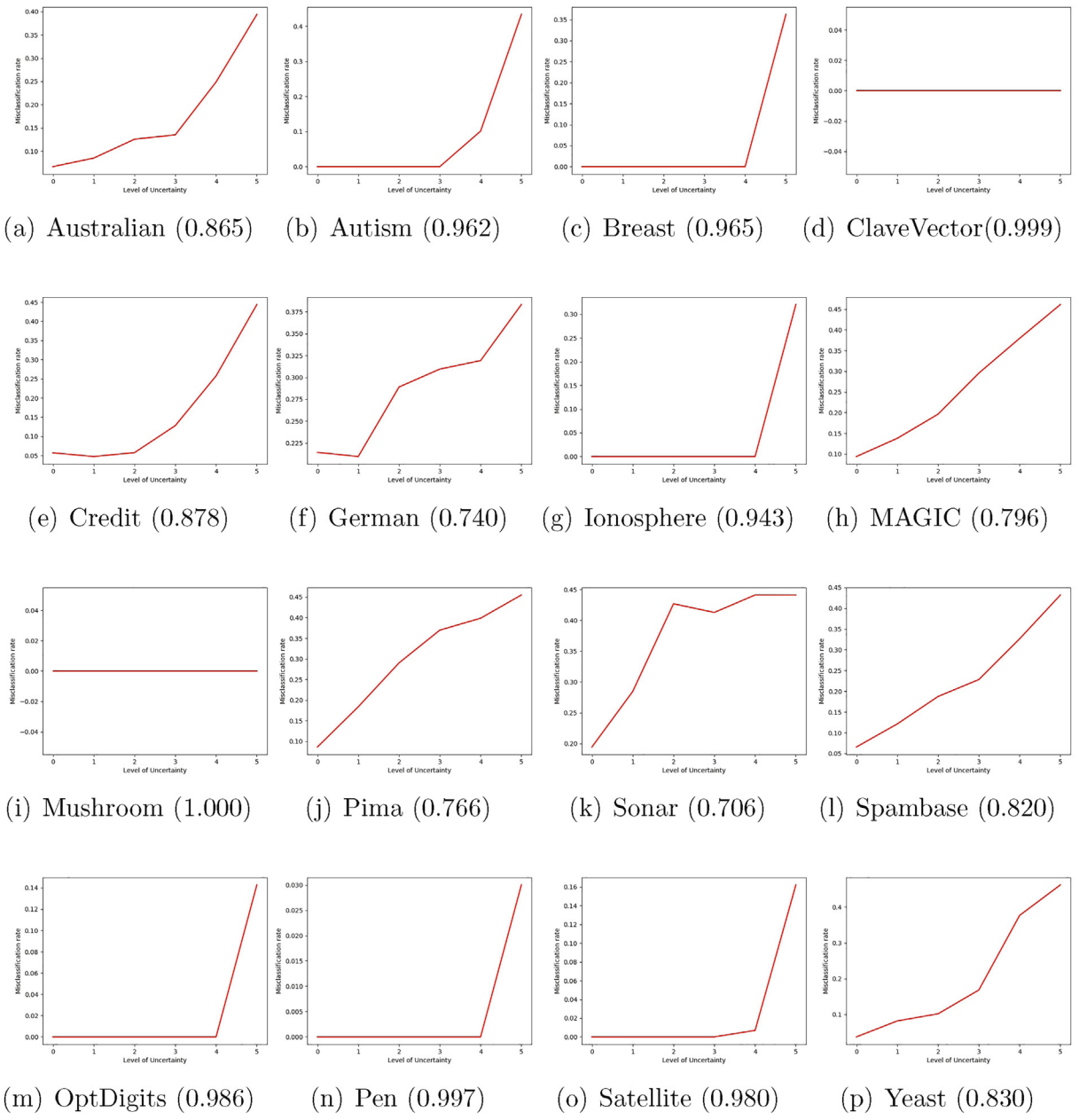

Fig. 5. Relationship between misclassification rate and uncertainty of SVM classifier on real data sets. The numbers in brackets are testing accuracy of each data set, which is the mean of repeating ten experimental trials.

Suppose that k is the changing rate of misclassification with respect to uncertainty. The 12 subgraphs in Fig. 3 can be divided into two groups, according to the changing rule of k. The first group G_1, containing subgraphs (a), (b), (c), (d), (e), (i), and (o), in which k increases slowly in low-level of uncertainty while increases rapidly in high levels. It is noticed that samples G_1 with a low-level of uncertainty will be misclassified in a small chance. Samples with a high level of uncertainty will be misclassified with bigger probability. It means the classifier is performing well in the aspect of accuracy for all data sets in this group.

The second group G_2 is comprising the rest of the data sets. In each subgraph of G_2, the growth of k is large from the very beginning. It means there is a certain probability of misclassification for samples with a lower uncertainty level. In subgraphs (j) and (k), the maximum misclassification rate appears at a lower uncertainty level. Thus, the accuracy of each data set in G_2 would be lower than the sets in G_1.

Actually, the testing accuracy of each data set is summarized in Table 2, in which high and low accuracy data sets are denoted with "★" and "•", respectively. It is observed that all the data sets signed with "★" belong to G_1, while the others belong to G_2, marked with "•". This observation confirms the reliability of our experiments.

4.3. Analysis with SVM classifiers

To check whether or not our scheme is sensitive to classifier selection, this section selects the SVM to replace the RWNN used in previous sections. We demonstrate the results of Support Vector Machine(SVM) classifiers by experimenting in the same computational environment.

As mentioned in [8,38,41], due to the solid mathematical background, high generalization capability and ability to find globally optimal solutions, SVM has been successfully applied to many real-world classification problems. We choose the RBF kernel and set the penalty term C as 1. The relationship between the misclassification rate and uncertainty of each data set in Table 1 is demonstrated in Fig. 5. It is noted that the testing accuracy of data sets *ClaveVector* and *Mushroom* are 0.999 and 1 respectively, which means there are nearly no samples that would be misclassified. Thus, subgraphs (d) and (i) contain a horizontal straight line with a misclassification rate equal to 0, respectively. According to Table 2, the horizontal straight lines in low uncertainty level of subgraphs (b), (c), (g), (m), (n), and (o) result from the high testing accuracy. The distributions of data sets are depicted in Fig. 4, it is noteworthy that distributions of class A or B in data sets *Australian*, *Credit*, *Sonar*, and *Yeast* are multimodal. As shown in Fig. 5, the misclassification rate increases with the growing up of uncertainty for the four data sets, although there is a slight drop during the ascent.

It can be viewed that if data sets meet with the requirements mentioned in Section 4.2, our conclusion is basically correct, i.e., the misclassification rate increases with the growing up of uncertainty, and the scheme is with low sensitivity to classifier selection. It is interesting to find that, for some datasets which cannot meet the mentioned requirements, partial conclusions still look correct.

5. Conclusions and future work

In this paper, the relationship between uncertainty and misclassification rate of a classifier is illustrated explicitly and precisely. It is theoretically and experimentally validated that the misclassification rate increases definitely with the growing up of uncertainty in some cases, which requires that the distributions of different categories based on value of M-discriminant function are unimodal. Without loss of generality, two typical classifiers, RWNN and SVM, are considered in our experiments. The results illustrate that our scheme is insensitive to the selection of different types of classifiers. The limitation of this study is that only binary classification problems are analyzed at present.

Following this work, one can further discuss the descriptions about generalizing the analysis to a multicategory situation, drawing more general conclusions. This work provides a theoretical support to the research on dynamic classifier selection. The conclusion of the relationship is also potentially applicable in meta-learning, i.e., several new meta-features can be generated by further quantifying the rate of change in uncertainty and misclassification rate.

CRediT authorship contribution statement

Xinlei Zhou: Methodology, Software, Investigation, Writing - original draft. **Xizhao Wang:** Conceptualization, Validation, Writing - review & editing. **Cong Hu:** Formal analysis, Validation.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (Grants 61976141, 61732011, and 61811530324) and in part by Basic Research Project of Knowledge Innovation Program in ShenZhen (JCYJ20180305125850156).

References

[1] Yijun Bian, Huanhuan Chen. When does diversity help generalization in classification ensembles? arXiv preprint arXiv:1910.13631, 2019..
[2] Weipeng Cao, Xizhao Wang, Zhong Ming, Jinzhu Gao, A review on neural networks with random weights, Neurocomputing 275 (2018) 278–287.
[3] Gavin C. Cawley, Nicola L.C. Talbot, On over-fitting in model selection and subsequent selection bias in performance evaluation, J. Mach. Learn. Res. 11 (Jul) (2010) 2079–2107.
[4] Rolando de la Cruz, Claudio Fuentes, Cristian Meza, Vicente Núñez-Antón, Error-rate estimation in discriminant analysis of non-linear longitudinal data: a comparison of resampling methods, Stat. Methods Med. Res. 27 (4) (2018) 1153–1167.

[5] Hector De la Torre Gutierrez, D.T. Pham, Estimation and generation of training patterns for control chart pattern recognition, Comput. Ind. Eng. 95 (2016) 72–82.
[6] Sergio Decherchi, Sandro Ridella, Rodolfo Zunino, Paolo Gastaldo, Davide Anguita, Using unsupervised analysis to constrain generalization bounds for support vector classifiers, IEEE Trans. Neural Netw. 21 (3) (2010) 424–438.
[7] Richard O. Duda, Peter E. Hart, David G. Stork, Pattern Classification, John Wiley & Sons, 2012.
[8] Xiaoqing Gu, Fu-lai Chung, Shitong Wang, Extreme vector machine for fast training on large data, Int. J. Mach. Learn. Cybern. (2019) 1–21.
[9] Martin T Hagan, Mohammad B Menhaj, Training feedforward networks with the marquardt algorithm. IEEE transactions on, Neural Netw. 5 (6) (1994) 989–993.
[10] Hu. Lisha, Lu. Shuxia, Xizhao Wang, A new and informative active learning approach for support vector machine, Inf. Sci. 244 (2013) 142–160.
[11] Guang-Bin Huang, Qin-Yu Zhu, Chee-Kheong Siew, Extreme learning machine: theory and applications, Neurocomputing 70 (1–3) (2006) 489–501.
[12] Cezary Z. Janikow, Fuzzy decision trees: issues and methods, IEEE Trans. Syst., Man, Cybern., Part B (Cybern.) 28 (1) (1998) 1–14.
[13] Yiannis Kokkinos, G. Margaritis, Managing the computational cost of model selection and cross-validation in extreme learning machines via cholesky, svd, qr and eigen decompositions, Neurocomputing 295 (2018) 29–45.
[14] Chun-Fu Lin, Sheng-De Wang, Fuzzy support vector machines, IEEE Trans. Neural Netw. 13 (2) (2002) 464–471.
[15] Nick Littlestone, Manfred K. Warmuth, The weighted majority algorithm, Inform. Comput. 108 (1994) 212–261.
[16] Oswaldo Ludwig, Urbano Nunes, Bernardete Ribeiro, Cristiano Premebida, Improving the generalization capacity of cascade classifiers, IEEE Trans. Cybern. 43 (6) (2013) 2135–2146.
[17] Aleksandr Luntz. On estimation of characters obtained in statistical procedure of recognition. Technicheskaya Kibernetica, 3, 1969..
[18] Patricia Melin, Frumen Olivas, Oscar Castillo, Fevrier Valdez, Jose Soria, Mario Valdez, Optimal design of fuzzy classification systems using pso with dynamic parameter adaptation through fuzzy logic, Expert Syst. Appl. 40 (8) (2013) 3196–3206.
[19] Mohamad T Musavi, Khue Hiang Chan, Donald M Hummels, K. Kalantri, On the generalization ability of neural network classifiers, IEEE Trans. Pattern Anal. Mach. Intell. 16 (6) (1994) 659–663.
[20] Wing W.Y. Ng, Aki P.F. Chan, Daniel S. Yeung, Eric C.C. Tsang, Quantitative study on the generalization error of multiple classifier systems, vol. 1, IEEE, 2005, pp. 889–894.
[21] Frumen Olivas, Fevrier Valdez, Oscar Castillo, Fuzzy classification system design using pso with dynamic parameter adaptation through fuzzy logic, Fuzzy Logic Augmentation of Nature-Inspired Optimization Metaheuristics, Springer, 2015, pp. 29–47.
[22] J. Ross Quinlan, Induction of decision trees, Mach. Learn. 1 (1) (1986) 81–106.
[23] Salim Rezvani, Xizhao Wang, Farhad Pourpanah, Intuitionistic fuzzy twin support vector machines, IEEE Trans. Fuzzy Syst. 27 (11) (2019) 2140–2151.
[24] Elid Rubio, Oscar Castillo, Fevrier Valdez, Patricia Melin, Claudia I Gonzalez, Gabriela Martinez, An extension of the fuzzy possibilistic clustering algorithm using type-2 fuzzy logic techniques, Adv. Fuzzy Syst. 2017 (2017).
[25] Dilip Sarkar, Randomness in generalization ability: a source to improve it, IEEE Trans. Neural Netw. 7 (3) (1996) 676–685.
[26] Robert E. Schapire, Yoav Freund, Peter Bartlett, Wee Sun Lee, et al, Boosting the margin: a new explanation for the effectiveness of voting methods, Ann. Stat. 26 (5) (1998) 1651–1686.
[27] Claude Elwood Shannon, A mathematical theory of communication, Bell Syst. Tech. J. 27 (3) (1948) 379–423.
[28] Mervyn Stone, Cross-validatory choice and assessment of statistical predictions, J. Roy. Stat. Soc.: Ser. B (Methodol.) 36 (2) (1974) 111–133.
[29] Johan A.K. Suykens, Joos Vandewalle, Least squares support vector machine classifiers, Neural Process. Lett. 9 (3) (1999) 293–300.
[30] Z.Q. Tian, Yue Zhou, The certification of the fundamental properties of comentropy, J. Inner Mongolia Normal Univ. (Natural Sci. Ed.) 31 (4) (2002) 347–350.
[31] Vladimir N. Vapnik, Adaptive and learning systems for signal processing communications, and control, Stat. Learn. Theory (1998).
[32] Ran Wang, Sam Kwong, Xi-Zhao Wang, Qingshan Jiang, Segment based decision tree induction with continuous valued attributes, IEEE Trans. Cybern. 45 (7) (2014) 1262–1275.
[33] 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 Trans. Fuzzy Syst. 23 (5) (2015) 1638–1654.
[34] Xi-Zhao Wang, Chun-Ru Dong, Improving generalization of fuzzy if–then rules by maximizing fuzzy entropy, IEEE Trans. Fuzzy Syst. 17 (3) (2008) 556–567.
[35] Xi-Zhao Wang, Ling-Cai Dong, Jian-Hui Yan, Maximum ambiguity-based sample selection in fuzzy decision tree induction, IEEE Trans. Knowl. Data Eng. 24 (8) (2011) 1491–1505.
[36] Xi-Zhao Wang, Ran Wang, Xu. Chen, Discovering the relationship between generalization and uncertainty by incorporating complexity of classification, IEEE Trans. Cybern. 48 (2) (2017) 703–715.
[37] Xi-Zhao Wang, Tianlun Zhang, Ran Wang, Noniterative deep learning: Incorporating restricted boltzmann machine into multilayer random weight neural networks, IEEE Trans. Syst., Man, Cybern.: Syst. (2017).
[38] Weichen Wu, Yitian Xu, Accelerating improved twin support vector machine with safe screening rule, Int. J. Mach. Learn. Cybern. (2019) 1–14.
[39] Dasen Yan, Xinlei Zhou, Xizhao Wang, Ran Wang, An off-center technique: learning a feature transformation to improve the performance of clustering and classification, Inf. Sci. 503 (2019) 635–651.
[40] Jing Yang, Xiaoqin Zeng, Shuiming Zhong, Wu. Shengli, Effective neural network ensemble approach for improving generalization performance, IEEE Trans. Neural Netw. Learn. Syst. 24 (6) (2013) 878–887.
[41] Yunfei Ye, A nonlinear kernel support matrix machine for matrix learning, Int. J. Mach. Learn. Cybern. 10 (10) (2019) 2725–2738.
[42] Li Zhao, Jie Zhu, Learning from correlation with extreme learning machine, Int. J. Mach. Learn. Cybern. (2019) 1–11.
[43] Wendong Zheng, Huaping Liu, Bowen Wang, Fuchun Sun, Cross-modal learning for material perception using deep extreme learning machine, Int. J. Mach. Learn. Cybern. (2019) 1–11.
[44] Xinlei Zhou, Dasen Yan, Model tree pruning, Int. J. Mach. Learn. Cybern. (2019) 1–14.
[45] Xiaoyan Zhu, Yueyang He, Long Cheng, Xiaolin Jia, Lei Zhu, Software change-proneness prediction through combination of bagging and resampling methods, J. Softw.: Evol. Process 30 (12) (2018) e2111.