Applied Intelligence (2019) 49:1161–1171
https://doi.org/10.1007/s10489-018-1322-z

Ensemble based fuzzy weighted extreme learning machine for gene expression classification

Yang Wang[1,2] · Anna Wang[1] · Qing Ai[1] · Haijing Sun[1]

Published online: 27 October 2018

Abstract

Multi-class imbalance is one of the challenging problems in many real-world applications, from medical diagnosis to intrusion detection, etc. Existing methods for gene expression classification usually assume relatively balanced class distribution. However, the assumption is invalid for imbalanced data learning. This paper presents an effective method named EN-FWELM for class imbalance learning. First, based on a fast classifier extreme learning machine (ELM), fuzzy membership of sample is proposed in order to eliminate classification error coming from noise and outlier samples, and balance factor is introduced in combination with sample distribution and sample number associated with class to alleviate the bias against performance caused by imbalanced data. Furthermore, ensemble of ELMs is used for making classification performance more stable and accurate. A number of base ELMs are removed based on dissimilarity measure, and the remaining base ELMs are integrated by majority voting. Finally, experimental results on various gene expression classification and real-world classification demonstrate that the proposed EN-FWELM remarkably outperforms other approaches in the literature.

Keywords Gene expression classification · Extreme learning machine · Fuzzy membership · Balance factor · Dissimilarity measure

1 Introduction

Generally, the type of cancer being identified early on can improve the health of people. Because the same cancer resulting from many factors may have different symptoms, traditional diagnostic methods can fail to identify cancer exactly [1]. However, gene expression data based on the microarray technology can achieve more accurate results. Therefore, the relevant research based on gene expression classification attracted more and more attention [2–4].

Extreme learning machine (ELM) was proposed for the single-hidden layer feed-forward networks (SLFNs). It has good generalization performance and fast learning speed by generating randomly input weights and biases of hidden nodes instead of adjusting network parameters iteratively [5, 6]. With the advantages, ELM has been widely applied in various areas [7–12]. Inspired by the ensemble idea [13], the stability and classification performance of single ELM can be improved. For example, Bagging takes different bootstrap samples from training data to construct a parallel ensemble model. AdaBoost runs repeatedly a learning machine on different distribution of training data to construct a serial ensemble model [14]. Cao et al. [15] proposed V-ELM that performs ensemble of ELMs and makes the final decision by majority voting. Li et al. [16] proposed boosting weighted ELM. Weighted ELM is embedded into a modified AdaBoost framework, and the distribution weights can be used as training sample weights. Zhang et al. [17] presented ensemble learning strategy based on differential evolution (DE) that performs ensemble of WELMs with different activation functions and employs DE to optimize the weight of each base classifier. Xu et al. [18] proposed WELM-Ada based on fusion optimization of weighted ELM and AdaBoost. Lu et al. [19] proposed D-D-ELM and DF-D-ELM, which remove some base ELMs

✉ Yang Wang
wangyang0531@163.com

[1] College of Information Science and Engineering, Northeastern University, Shenyang 110819, Liaoning, China

[2] School of Computer and Communication Engineering, Liaoning Shihua University, Fushun 113001, Liaoning, China

based on the dissimilarity and group the remaining ELMs by majority voting on gene expression data.

One is often confronted with multi-class imbalance problem on gene expression data, and this issue brings extreme challenge. On one hand, existing methods for gene expression classification usually ignore the influence of samples distribution on classification, which can incur classification error coming from noise and outlier samples and reduce generalization performance of ELM. Furthermore, existing methods usually assume relatively balanced class distribution and are more concerned with overall accuracy, which can ignore the minority class and tend to be biased against the majority class in dealing with imbalanced data [20, 21]. In other words, they may achieve higher misclassification accuracy of the minority class than that of the majority class.

There are two methods dealing with imbalanced data i.e. resampling technique and algorithmic technique [22]. Resampling technique includes oversampling which duplicates some minority class samples randomly or creates new samples in the neighborhood of minority class samples and undersampling which removes some majority class samples randomly to balance the size of each class [23, 24]. Moreover, resampling technique modifies samples distribution and can lose some useful information. However, algorithmic technique does not change sample distribution and is widely used to cope with imbalanced data [25].

In this study, algorithmic technique is of particular interest, and ensemble based fuzzy weighted extreme learning machine is presented to perform gene expression classification. First, different fuzzy membership is assigned for each sample. Fuzzy membership indicates the importance of sample on classification, and the bigger fuzzy membership is, the greater the influence of sample on classification is. Therefore, noise and outlier samples are assigned low fuzzy membership to improve classification performance. In addition, balance factor is used to alleviate the bias against performance caused by imbalanced data. An extra balance factor, relevant to samples distribution and samples number of each class, is designed for each sample to strengthen the relative impact of the minority class, and G-mean is taken as evaluation measure to monitor classification ability. Furthermore, some base ELMs are removed based on the dissimilarity and the remaining ELMs are integrated by majority voting. Finally, experimental results illustrate that the proposed method named EN-FWELM is effective and robust.

The rest of this paper is organized as follows. Section 2 presents a brief review of relevant preliminary knowledge. In Section 3, the detailed implementations of the proposed method are explained. In Section 4, the experimental design is described and many experiments are completed to demonstrate that the proposed method presents better classification performance than that achieved by some existing methods. Finally, conclusions are summarized in Section 5.

2 Related work

2.1 Extreme learning machine (ELM)

Given a training dataset consisting of N arbitrary samples (x_j, t_j), where $t_j = [t_{j1}, t_{j2}, \cdots, t_{jm}]^T \in R^m$ and $x_j = [x_{j1}, x_{j2}, \cdots, x_{jn}]^T \in R^n$. The jth sample t_j is an $m \times 1$ target vector, and x_j is an $n \times 1$ feature vector. Given hidden nodes $L << N$ and activation function $g(x)$, then the standard mathematical model of SLFNs is as follows:

$$\sum_{i=1}^{L} \beta_i g(a_i \cdot x_j + b_i) = t_j \qquad j = 1, 2, \cdots, N \tag{1}$$

where $\beta_i = [\beta_{i1}, \beta_{i2}, \cdots, \beta_{im}]^T$ is the output weight vector connecting the ith hidden node and output nodes, $a_i = [a_{i1}, a_{i2}, \cdots, a_{in}]^T$ is the input weight vector connecting input nodes and the ith hidden node, $a_i \cdot x_j$ is the inner product of a_i and x_j, and b_i is the bias of the ith hidden node.

SLFNs can approximate the training samples with zero error if the number of hidden nodes L is equal to the number of training samples N. The formula (1) can compactly be rewritten as (2).

$$H\beta = T \tag{2}$$

$$H = \begin{bmatrix} h(x_1) \\ \vdots \\ h(x_N) \end{bmatrix} = \begin{bmatrix} g(a_1 \cdot x_1 + b_1) & \cdots & g(a_L \cdot x_1 + b_L) \\ \vdots & \cdots & \vdots \\ g(a_1 \cdot x_N + b_1) & \cdots & g(a_L \cdot x_N + b_L) \end{bmatrix}_{N \times L}$$

$$\beta = \begin{bmatrix} \beta_1^T \\ \vdots \\ \beta_L^T \end{bmatrix}_{L \times m}, and \quad T = \begin{bmatrix} t_1^T \\ \vdots \\ t_N^T \end{bmatrix}_{N \times m} \tag{3}$$

where H is the hidden layer output matrix, and the jth column of H represents the jth hidden node output vector on all the inputs. T is the output matrix, and β is the output weight matrix.

However, in most cases, it is $L << N$ and there may not exist a β that satisfies (2). The hidden layer biases and input weights need not be tuned at all and can be randomly generated, so the output weights can be determined by finding the Least Square solution $\beta = H^+T$ of $H\beta = T$, where H^+ is the Moore-Penrose generalized inverse of matrix H. In short, ELM algorithm is summarized as follows.

1) Generate randomly input weights a_i and biases b_i, $i = 1, 2, \cdots, L$.
2) Calculate the hidden layer output matrix H.
3) Calculate the output weight $\beta = H^+T$.

2.2 Weighted extreme learning machine (WELM)

According to Bartlett's theory [26], ELM is not only to minimize the training error but also to minimize the norm of the output weights. Meanwhile, an extra weight is designed for each sample to better deal with imbalanced data, so the classification problem can be formulated as (4).

$$
\begin{aligned}
& min\left(\frac{1}{2}||\beta||^2 + CW\frac{1}{2}\sum_{j=1}^{N}\xi_j^2\right) \\
& s.t. \sum_{j=1}^{N}\beta_i g(a_i \cdot x_j + b_i) = t_j - \xi_j
\end{aligned}
\tag{4}
$$

The equivalent dual optimization problem in regard to (4) based on KKT theorem is

$$
\begin{aligned}
L_{ELM} &= \frac{1}{2}||\beta||^2 + CW\frac{1}{2}\sum_{j=1}^{N}\xi_j^2 \\
&\quad - \sum_{j=1}^{N}\alpha_j(\beta_i g(a_i \cdot x_j + b_i) - t_j + \xi_j)
\end{aligned}
\tag{5}
$$

where ξ_j is the training error, C is penalty parameter, and α_j is Lagrange multiplier. W is diagonal matrix relevant to each training sample, namely $W = diag(w_{jj}),\ j = 1, 2, \cdots, N$. For instance, weighted strategy associated with the number of samples in each class can be assigned as follows [20]:

$$
\begin{aligned}
W1 : w_{jj} &= \frac{1}{N(t_j)} \\
W2 : w_{jj} &= \begin{cases} \frac{0.618}{N(t_j)} & if\ N(t_j) > AVG \\ \frac{1}{N(t_j)} & if\ N(t_j) \leq AVG \end{cases}
\end{aligned}
\tag{6}
$$

where $N(t_j)$ is the number of samples in class t_j, and AVG is the average samples number of each class. Then solution of (5) can be formulated as (7).

$$
\beta = \begin{cases} H^T\left(\frac{I}{C} + WHH^T\right)^{-1} WT & when\ N < L \\ \left(\frac{I}{C} + H^T WH\right)^{-1} H^T WT & when\ N >> L \end{cases}
\tag{7}
$$

3 Proposed EN-FWELM model

In this study, construction of the optimal model comprises three main procedures: fuzzy weighted extreme learning machine (FWELM), ensemble learning based on the dissimilarity and performance evaluation.

3.1 FWELM

In this study, balance factor and fuzzy membership are introduced into ELM. WELM is a widely used method dealing with imbalance data. However, it only considers the imbalance of sample numbers in each class. In fact, not only does the imbalance of samples lies in the imbalance of the number of samples but also lies in the imbalance of sample distribution. Therefore, it is crucial that sample number and sample distribution are both considered as balance factors. In this study, sample density, as the important index to measure sample distribution, is used for representing sample distribution [21, 27]. Center of samples in each class can be formulated as (8).

$$
d_k = \frac{1}{N_k}\sum_{j=1}^{N_k} x_j \quad k = 1, \cdots, c
\tag{8}
$$

where c is the number of class, N_k is the number of samples in the kth class, x_j is the jth sample, and d_k is center of samples in the kth class. Accordingly, sample density in each class is defined as (9).

$$
p_k = \frac{\sum_{j=1}^{N_k} ||x_j - d_k||}{N_k} \quad k = 1, \cdots, c
\tag{9}
$$

where p_k is sample density in the kth class. In addition, weighted strategy is also designed for each sample to better deal with imbalanced data, and it can be presented as (10).

$$
W : w_{jj} = \frac{N(c - t_j + 1)}{N}
\tag{10}
$$

where $N(c - t_j + 1)$ is the number of samples in class $c - t_j + 1$. The number of samples in class $1, \cdots, c$ is ranked in the ascending order. Therefore, in this study, balance factor R is diagonal matrix relevant to each training sample and is defined as (11).

$$
R : r_{jj} = w_{jj} \times p_{c-t_j+1}
\tag{11}
$$

Weight W has more enormous influence than weight $W1$ and weight $W2$ on the classification performance. For example, for binary classification problem the reason is as follows.

$$
\begin{aligned}
\Delta W &= \left(\frac{N^-}{N}\right) - \left(\frac{N^+}{N}\right) = \frac{N^- - N^+}{N} \\
&= \frac{N^- - N^+}{N^- + N^+} \\
1)\ \Delta W1 &= \left(\frac{1}{N^+}\right) - \left(\frac{1}{N^-}\right) \\
&= \frac{N^- - N^+}{N^- \times N^+}
\end{aligned}
$$

where N^-, N^+ stand for the number of negative and positive samples, respectively. Moreover, $N^- \times N^+ - (N^- + N^+) = (N^- - 1) \times (N^+ - 1) - 1$. For $N^- > 2$, $N^+ > 2$, so $N^- - 1 > 1$ and $N^+ - 1 > 1$, namely

$(N^- - 1) \times (N^+ - 1) > 1$. Therefore, $\Delta W > \Delta W1$ and weight W has better performance than weight $W1$.

2) $$\Delta W2 = \left(\frac{1}{N^+}\right) - \left(\frac{0.618}{N^-}\right) = \frac{N^- - 0.618 \times N^+}{N^- \times N^+}$$

for $N^- > 2$, $N^+ > 2$, so $(N^- + N^+) - N^- \times N^+ < (N^- - N^+) - (N^- - 0.618 \times N^+)$, namely $\Delta W > \Delta W2$ and weight W has better performance than weight $W2$.

On the other hand, fuzzy membership of sample is proposed in order to eliminate classification error coming from noise and outlier samples [21]. The radius from all samples to center in each class is defined as (12).

$$rd_k = max||x_j - d_k|| \quad j = 1, \cdots, N_k \tag{12}$$

where rd_k is the radius from samples to center in the kth class. Then fuzzy membership is defined as (13).

$$S : s_{jj} = 1 - \frac{||x_j - d_k||}{rd_k + \delta} \quad j = 1, 2, \cdots, N_k \tag{13}$$

where S is diagonal matrix relevant to each training sample, and δ is an arbitrary small positive number. From (13), it can be seen that noise and outlier samples are usually far away from center of the class, and they will be given a minimum fuzzy membership to reduce the influence on classification. Therefore, in this study, the classification problem can be formulated as (14).

$$L_{ELM} = \frac{1}{2}||\beta||^2 + CR\frac{1}{2}\sum_{j=1}^{N} s_{jj}\xi_j^2$$
$$s.t. \quad h(x_j)\beta = t_j - \xi_j \tag{14}$$

Based on KKT theorem, KKT constraint conditions can be formulated as:

$$\frac{\partial L_{ELM}}{\partial \xi_j} = 0 \to \alpha_j = CRs_{jj}\xi_j \quad j = 1, 2, \cdots, N \tag{15}$$

$$\frac{\partial L_{ELM}}{\partial \beta} = 0 \to \beta = \sum_{j=1}^{N} \alpha_j h(x_j)^T = H^T\alpha \tag{16}$$

$$\frac{\partial L_{ELM}}{\partial \alpha_j} = 0 \to h(x_j)\beta - t_j + \xi_j = 0 \tag{17}$$

By substituting (15) and (16) into (17), the output weight of FWELM can be formulated as (18).

$$\beta = H^T\left(\frac{(S)^{-1}}{C} + RHH^T\right)^{-1} RT \tag{18}$$

If the number of training samples is large, the output weight of FWELM can be formulated as (19) by substituting (15) and (17) into (16).

$$\beta = \left(\frac{I}{C} + H^T RSH\right)^{-1} H^T RST \tag{19}$$

3.2 Ensemble learning

In this study, ensemble learning [28] based on the dissimilarity is used for handling class imbalance, and the dissimilarity between the ith ELM and the jth ELM is defined as (20).

$$df_{i,j} = \frac{P^{01} + P^{10} + P^{11}}{P^{01} + P^{10} + P^{11} + P^{00}} \tag{20}$$

where P^{yz} is the number of samples for which samples are separately classified as y by the ith ELM and classified as z by the jth ELM. 0 denotes samples are wrongly classified, while 1 denotes samples are correctly classified. Moreover, the dissimilarity between the ith ELM and other ELMs is defined as (21).

$$D_i = \sum_{j=1}^{K} df_{i,j} \tag{21}$$

where K is the number of classifiers. Inspired by [19], ELMs with larger dissimilarity are selected based on $D = \{D_1, D_2, \cdots, D_K\}$. The one-sided confidence interval of D is calculated to select ELMs whose D_i belong to the confidence interval, and the t distribution with no arguments is constructed to calculate the confidence interval.

The mean value of D is

$$\overline{D} = \frac{1}{K}\sum_{i=1}^{K} D_i \tag{22}$$

The standard deviation of D is

$$SD = \sqrt{\frac{1}{K-1}\sum_{i=1}^{K}(D_i - \overline{D})^2} \tag{23}$$

$$\frac{\overline{D} - \mu}{SD/\sqrt{K}} \sim t(K-1) \tag{24}$$

The one-side confidence interval at 95% confidence level of μ is

$$\left[\overline{D} - \frac{t_{0.05}(K-1)SD}{\sqrt{K}}, \infty\right) \tag{25}$$

This study starts with selecting ELMs whose D_i belong to the one-sided confidence interval, and then integrates them by majority voting.

3.3 Measure metrics

In this study, Accuracy, G-mean and F-score are used for evaluating the performance of proposed EN-FWELM model. These measure metrics are defined below.

$$Accuracy = \frac{\sum_{i=1}^{c} TP_i}{\sum_{i=1}^{c} TP_i + FN_i} \tag{26}$$

$$Precision_i = \frac{TP_i}{TP_i + FP_i} \tag{27}$$

$$Recall_i = \frac{TP_i}{TP_i + FN_i} \tag{28}$$

$$F-measure_i = \frac{2Recall_i Precision_i}{Recall_i + Precision_i} \tag{29}$$

$$F-score = \frac{\sum_{i=1}^{c} F-measure_i}{c} \tag{30}$$

$$G-mean = \left(\prod_{i=1}^{c} Recall_i\right)^{\frac{1}{c}} \tag{31}$$

where *TP, FP, TN, FN* stand for the number of true positive, false positive, true negative and false negative, respectively.

Accuracy is the evaluation measure for correctly classified samples over all samples. F-measure is usually used to assess the performance of imbalanced data classification, and F-score is the average over F-measure. Obviously, G-mean is 0 when the classification accuracy for the *i*th class is 0 [17]. Therefore, G-mean is also used to evaluate classification performance of imbalanced data and makes a more fair comparison.

4 Experiments

4.1 Experiment datasets

To compare proposed EN-FWELM with other learning algorithms, a variety of datasets from GEMS and KEEL repository are used for classification [29, 30]. The detailed information about these datasets is listed in Table 1 and these datasets are ordered according to *IR*. The imbalance degree measured by the imbalance ratio (*IR*) is defined as

$$\begin{aligned} Binary: IR &= \frac{\#majority}{\#minority} \\ Multi-class: IR &= \frac{max(\#i)}{min(\#i)}, i = 1, 2, \cdots, c \end{aligned} \tag{32}$$

Table 1 Description of imbalanced datasets

GEMS/KEEL Datasets	#Atts	#Class	#Train	#Test	*IR*
Leukemia2	11225	3	38	34	1.4
SRBCT	2308	4	43	40	2.64
DLBCL	5469	2	43	34	3.05
Leukemia1	5327	3	38	34	4.22
11_Tumors	12533	11	120	54	4.5
Lung_cancer	12600	5	143	60	23.17
wine	13	3	125	53	1.5
new-thyroid	5	3	150	65	5.0
dermatology	34	6	248	110	5.5
glass2	9	2	150	64	11.59
ecoli-0-1-4-6_vs_5	6	2	196	84	13
shuttle-6_vs_2-3	9	2	160	70	22

The number of these datasets attributes varies from 5 to 12600. The number of these datasets classes varies from 2 to 11, and *IR* varies from 1.4 to 23.17.

4.2 Experimental setting

To evaluate the performance of the proposed approach, it is compared against variants of EN-FWELM and other ensemble learning methods [14–19]. The whole experiment is conducted on MATLAB platform, which runs on windows 8 OS with Intel(R) Core(TM) i5-4460 CPU (3.2 GHz) and 8 GB of RAM. The parameters setting is given as follows. A grid search of penalty parameter C on $\{2^{-18}, 2^{-16}, \cdots, 2^{48}, 2^{50}\}$ and hidden nodes L on $\{10, 20, \cdots, 990, 1000\}$ is used to find the optimal G-mean, and $g(x) = \frac{1}{1+exp(-(a \cdot x+b))}$ is applied as activation function.

The attributes of these datasets are normalized into [0, 1]. Each dataset is randomly divided into a training-testing set. Then each experiment is individually repeated 10 times, and the average of 10 runs is used as the final results.

4.3 Comparison with variants of EN-FWELM

In these experiments, the classification performance of ELM, W1-based weighted learning algorithm (WELM1) and W2-based weighted learning algorithm (WELM2) [20] is evaluated, respectively. To show the effectiveness of EN-FWELM, it is also compared against its variants i.e. WELM and FWELM, which are built to analyze the importance of different parts in EN-FWELM. Meanwhile, the dimension of gene expression data is reduced by information gain (IG) [31] before training the classifier.

Tables 2, 3, 4, 5 and 6 show the detailed results of parameters setting, Accuracy, G-mean, F-score and training time, where the bold indicates the best results. From Tables 4 and 5, we can see that EN-FWELM achieves better G-mean and F-score than other algorithms. In particular, the performance results show that our EN-FWELM can improve significantly G-mean and F-score when datasets are sensitive to class imbalance, such as Lung_cancer, new-thyroid, 11_Tumors, ecoli-0-1-4-6_vs_5 and glass2. On these datasets, G-mean is respectively improved by about 18.43%, 14.18%, 8.46%, 8.39% and 24.32% compared with ELM, then F-score is respectively improved by about 11.53%, 8.24%, 8.23%, 7.83% and 19.35% compared with ELM. The reason is that ELM is based on the assumption that the size of each class is relatively balanced. Therefore, ELM has the bias against the majority class and ignores the minority class. From Tables 4 and 5, it can also be seen that G-mean and F-score of EN-FWELM outperforms WELM1, WELM2 and WELM. G-mean is respectively improved by about 3.01%, 2.67% and 2.53% on average compared

Table 2 Parameters setting

GEMS/KEEL Datasets	ELM	WELM1	WELM2	WELM	FWELM	EN-FWELM
	L	(C,L)	(C,L)	(C,L)	(C,L)	(C,L)
Leukemia2	650	$(2^{14},730)$	$(2^{14},680)$	$(2^{4},920)$	$(2^{34},690)$	$(2^{14},420)$
SRBCT	760	$(2^{30},700)$	$(2^{42},820)$	$(2^{44},770)$	$(2^{10},880)$	$(2^{42},530)$
DLBCL	830	$(2^{-2},730)$	$(2^{0},910)$	$(2^{-4},940)$	$(2^{-16},720)$	$(2^{36},90)$
Leukemia1	930	$(2^{6},900)$	$(2^{0},540)$	$(2^{16},800)$	$(2^{-14},890)$	$(2^{-16},130)$
11_Tumors	600	$(2^{28},850)$	$(2^{6},680)$	$(2^{36},850)$	$(2^{20},540)$	$(2^{-16},590)$
Lung_cancer	720	$(2^{0},740)$	$(2^{0},960)$	$(2^{-2},690)$	$(2^{-10},730)$	$(2^{18},130)$
wine	20	$(2^{12},810)$	$(2^{12},220)$	$(2^{6},450)$	$(2^{4},240)$	$(2^{42},660)$
new-thyroid	30	$(2^{4},50)$	$(2^{14},750)$	$(2^{14},640)$	$(2^{10},320)$	$(2^{30},490)$
dermatology	100	$(2^{10},110)$	$(2^{8},360)$	$(2^{6},100)$	$(2^{0},740)$	$(2^{40},690)$
glass2	430	$(2^{32},10)$	$(2^{38},30)$	$(2^{34},10)$	$(2^{8},110)$	$(2^{24},890)$
ecoli-0-1-4-6_vs_5	60	$(2^{18},240)$	$(2^{16},890)$	$(2^{38},30)$	$(2^{18},50)$	$(2^{-8},610)$
shuttle-6_vs_2-3	30	$(2^{42},40)$	$(2^{44},20)$	$(2^{50},30)$	$(2^{36},10)$	$(2^{24},980)$

Table 3 Performance results (Mean ± SD) in terms of Accuracy(%)

GEMS/KEEL Datasets	ELM	WELM1	WELM2	WELM	FWELM	EN-FWELM
Leukemia2	92.65 ± 5.59	94.41 ± 4.03	94.12 ± 3.67	94.71 ± 3.62	95.00 ± 3.68	**97.65 ± 2.32**
SRBCT	93.75 ± 4.75	94.00 ± 3.57	94.25 ± 3.34	94.50 ± 3.29	95.00 ± 3.54	**98.00 ± 2.58**
DLBCL	93.82 ± 3.78	95.29 ± 4.21	95.29 ± 3.45	95.88 ± 2.48	96.76 ± 2.58	**98.82 ± 1.52**
Leukemia1	91.76 ± 4.76	93.24 ± 4.17	93.82 ± 3.78	94.12 ± 3.10	94.41 ± 3.78	**96.47 ± 2.70**
11_Tumors	89.26 ± 3.68	90.56 ± 3.65	92.41 ± 3.54	92.59 ± 3.15	92.96 ± 3.00	**95.37 ± 2.93**
Lung_cancer	84.33 ± 3.87	85.50 ± 2.94	81.83 ± 3.55	89.33 ± 3.26	91.50 ± 2.28	**93.67 ± 2.19**
wine	97.55 ± 2.19	98.11 ± 2.52	97.92 ± 1.65	98.87 ± 1.32	99.06 ± 1.33	**99.43 ± 0.91**
new-thyroid	90.92 ± 3.28	92.77 ± 3.41	94.92 ± 2.62	93.85 ± 3.16	95.69 ± 1.89	**96.62 ± 1.59**
dermatology	96.82 ± 1.50	97.64 ± 1.30	97.73 ± 1.23	98.09 ± 1.45	98.36 ± 1.03	**99.09 ± 0.96**
glass2	66.09 ± 7.55	75.31 ± 5.97	76.41 ± 5.91	76.72 ± 5.63	80.78 ± 3.30	**85.00 ± 3.06**
ecoli-0-1-4-6_vs_5	96.43 ± 1.68	96.90 ± 1.61	96.90 ± 1.35	96.79 ± 1.87	97.38 ± 1.84	**98.21 ± 1.28**
shuttle-6_vs_2-3	99.00 ± 1.51	99.29 ± 1.01	99.00 ± 1.18	99.57 ± 0.96	99.86 ± 0.45	**99.86 ± 0.45**

Table 4 Performance results (Mean ± SD) in terms of G-mean(%)

GEMS/KEEL Datasets	ELM	WELM1	WELM2	WELM	FWELM	EN-FWELM
Leukemia2	92.99 ± 5.64	94.32 ± 4.58	94.68 ± 3.56	95.15 ± 3.55	95.29 ± 3.52	**97.82 ± 2.18**
SRBCT	94.31 ± 4.62	95.36 ± 3.15	95.42 ± 2.78	95.31 ± 2.57	95.97 ± 2.81	**98.24 ± 2.54**
DLBCL	91.77 ± 5.38	95.87 ± 3.81	95.76 ± 3.50	95.16 ± 3.37	96.18 ± 3.26	**97.67 ± 3.16**
Leukemia1	91.57 ± 6.02	91.76 ± 5.78	91.81 ± 5.56	93.06 ± 5.02	93.33 ± 4.85	**95.49 ± 3.75**
11_Tumors	84.71 ± 5.15	85.87 ± 5.80	88.20 ± 5.67	88.84 ± 4.97	90.48 ± 5.58	**93.17 ± 4.58**
Lung_cancer	65.19 ± 5.97	80.57 ± 3.70	80.43 ± 3.74	79.72 ± 4.26	80.81 ± 3.22	**83.62 ± 2.75**
wine	97.62 ± 2.22	98.57 ± 1.89	98.42 ± 1.26	99.03 ± 1.18	99.27 ± 1.05	**99.56 ± 0.71**
new-thyroid	79.50 ± 7.25	92.32 ± 4.47	93.29 ± 4.83	89.00 ± 5.37	90.77 ± 4.66	**93.68 ± 3.78**
dermatology	96.42 ± 2.11	97.34 ± 1.35	97.44 ± 1.44	98.10 ± 1.38	98.18 ± 1.38	**99.03 ± 1.07**
glass2	59.72 ± 12.13	78.55 ± 11.37	80.72 ± 7.13	81.76 ± 6.33	83.95 ± 5.31	**84.04 ± 4.04**
ecoli-0-1-4-6_vs_5	87.67 ± 6.97	93.30 ± 6.72	91.94 ± 6.36	93.67 ± 6.74	92.28 ± 6.26	**96.06 ± 4.82**
shuttle-6_vs_2-3	94.37 ± 11.56	97.66 ± 4.36	97.52 ± 4.30	98.51 ± 4.21	98.66 ± 4.24	**99.26 ± 2.35**

Table 5 Performance results (Mean ± SD) in terms of F-score(%)

GEMS/KEEL Datasets	ELM	WELM1	WELM2	WELM	FWELM	EN-FWELM
Leukemia2	92.57 ± 5.51	94.07 ± 4.29	93.89 ± 4.08	94.48 ± 3.47	94.84 ± 3.74	**97.72 ± 2.21**
SRBCT	93.82 ± 4.72	94.27 ± 3.46	94.19 ± 3.14	94.40 ± 3.00	95.53 ± 3.17	**97.88 ± 2.79**
DLBCL	91.41 ± 5.64	94.30 ± 4.94	94.05 ± 4.29	94.45 ± 3.26	95.79 ± 3.23	**98.26 ± 2.27**
Leukemia1	90.69 ± 6.75	92.45 ± 5.53	92.82 ± 4.31	93.78 ± 3.88	94.00 ± 4.39	**96.23 ± 2.84**
11_Tumors	85.85 ± 4.17	87.62 ± 4.97	89.49 ± 4.19	90.07 ± 4.09	91.46 ± 4.32	**94.08 ± 3.80**
Lung_cancer	74.10 ± 4.34	77.35 ± 3.90	75.54 ± 3.91	80.75 ± 3.64	83.54 ± 2.99	**85.63 ± 2.65**
wine	97.50 ± 2.11	98.07 ± 2.59	97.98 ± 1.59	98.80 ± 1.38	98.98 ± 1.38	**99.50 ± 0.81**
new-thyroid	86.09 ± 4.94	90.87 ± 4.39	92.67 ± 3.48	91.62 ± 4.21	93.78 ± 2.95	**94.33 ± 2.70**
dermatology	96.40 ± 1.75	97.31 ± 1.33	97.40 ± 1.34	98.00 ± 1.43	98.24 ± 1.19	**99.03 ± 1.03**
glass2	49.78 ± 7.97	59.72 ± 7.04	60.76 ± 6.25	61.38 ± 6.07	63.23 ± 5.82	**69.13 ± 4.25**
ecoli-0-1-4-6_vs_5	87.04 ± 6.31	90.11 ± 5.21	87.40 ± 5.74	89.17 ± 5.74	92.06 ± 5.15	**94.87 ± 4.09**
shuttle-6_vs_2-3	95.82 ± 7.23	96.99 ± 4.47	94.29 ± 7.14	98.17 ± 3.93	99.25 ± 2.38	**99.58 ± 1.34**

with WELM1, WELM2 and WELM. F-score is respectively improved by about 4.43%, 4.65% and 3.43% on average compared with WELM1, WELM2 and WELM. The reason is that in this study FWELM is proposed by modifying WELM, in which balance factor and fuzzy membership are respectively presented. Balance factor is designed for each sample to strengthen the relative impact of the minority class, and fuzzy membership is designed for each sample to improve generalization performance of ELM. On the other hand, multiple FWELMs are trained and FWELMs with high dissimilarity are retained to improve the stability and classification performance of single FWELM. From the results, it can be concluded that it can be better to ensemble some base classifiers instead of all of base classifiers.

To further analyze the results, F-measure of each class on all the datasets is shown in Fig. 1, and the abbreviation of each class name is shown in x-axis to visualize the results more clearly. F-score is the average result over F-measures of each class on a dataset. From Fig. 1, we can see that F-measure of hypo class on new-thyroid acquired by EN-FWELM is worse than that acquired by some other variants. The reason is that F-measures of some classes are improved at the cost of relatively slight decrease in F-measures of other classes. From Fig. 1, it can be concluded that EN-FWELM can improve the classification accuracy of the minority class on the multi-class imbalanced data.

Performance results in terms of Accuracy is shown in Table 3. We can see that EN-FWELM obtains more

Table 6 Comparison (Mean ± SD) of training time(s)

GEMS/KEEL Datasets	ELM	WELM1	WELM2	WELM	FWELM	EN-FWELM
Leukemia2	0.0041 ± 0.00005	0.0030 ± 0.0002	0.0028 ± 0.00001	0.0037 ± 0.00002	0.0028 ± 0.00002	0.0159 ± 0.0017
SRBCT	0.0032 ± 0.0002	0.0022 ± 0.00001	0.0025 ± 0.00006	0.0024 ± 0.0001	0.0029 ± 0.0005	0.0115 ± 0.0018
DLBCL	0.0025 ± 0.0001	0.0023 ± 0.0001	0.0029 ± 0.0002	0.0028 ± 0.00004	0.0023 ± 0.00008	0.0060 ± 0.0006
Leukemia1	0.0034 ± 0.00009	0.0026 ± 0.00004	0.0016 ± 0.00006	0.0024 ± 0.0001	0.0028 ± 0.0005	0.0076 ± 0.0007
11_Tumors	0.0056 ± 0.0002	0.0041 ± 0.0001	0.0034 ± 0.00008	0.0042 ± 0.0007	0.0026 ± 0.00007	0.0127 ± 0.0014
Lung_cancer	0.0294 ± 0.0007	0.0075 ± 0.0007	0.0089 ± 0.0005	0.0070 ± 0.0001	0.0081 ± 0.0004	0.0659 ± 0.0058
wine	0.0005 ± 0.00005	0.0021 ± 0.0002	0.0010 ± 0.00002	0.0014 ± 0.0001	0.0010 ± 0.00002	0.0039 ± 0.0005
new-thyroid	0.0007 ± 0.00002	0.0015 ± 0.00002	0.0022 ± 0.00005	0.0020 ± 0.00005	0.0013 ± 0.00005	0.0032 ± 0.00009
dermatology	0.0033 ± 0.0002	0.0019 ± 0.0008	0.0029 ± 0.0002	0.0015 ± 0.0003	0.0036 ± 0.0005	0.0094 ± 0.0020
glass2	0.0051 ± 0.0005	0.0003 ± 0.00001	0.0004 ± 0.00004	0.0003 ± 0.00003	0.0009 ± 0.00003	0.0025 ± 0.0004
ecoli-0-1-4-6_vs_5	0.0015 ± 0.0002	0.0015 ± 0.00006	0.0025 ± 0.0003	0.0004 ± 0.00004	0.0007 ± 0.00008	0.0050 ± 0.0005
shuttle-6_vs_2-3	0.0007 ± 0.0001	0.0008 ± 0.0002	0.0004 ± 0.0002	0.0007 ± 0.0004	0.0004 ± 0.00002	0.0030 ± 0.00007

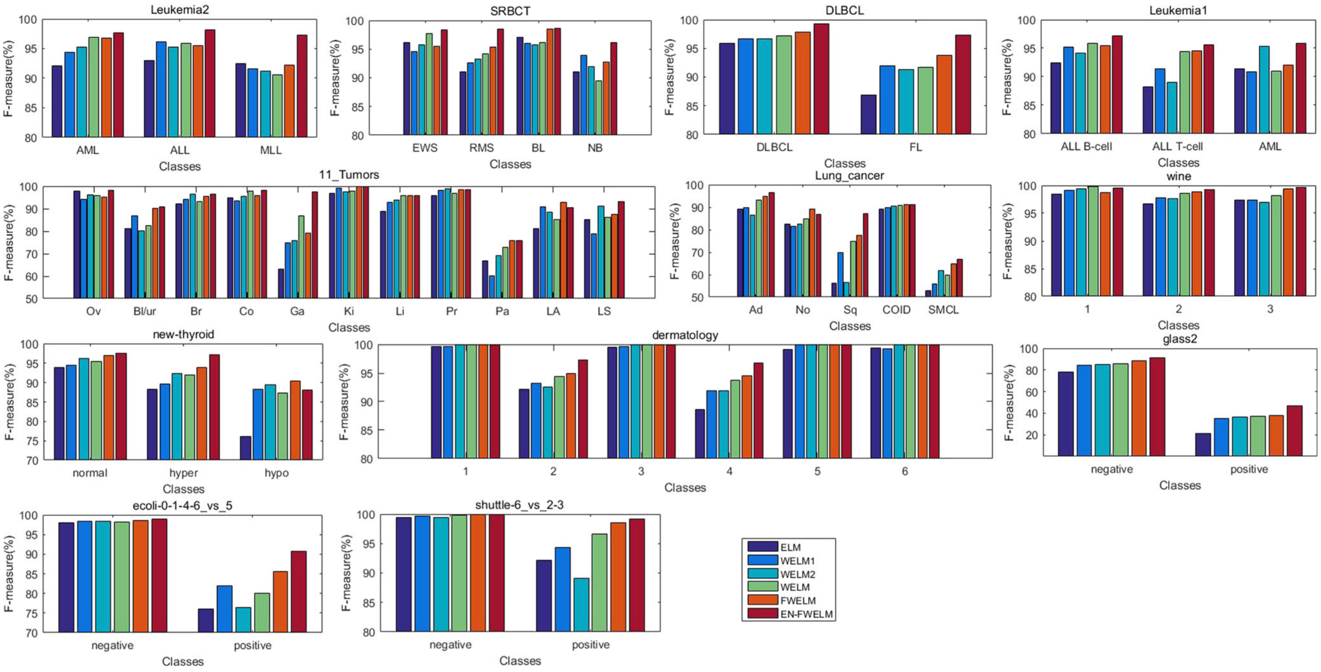

Fig. 1 F-measure of each class on all the datasets by using different methods

than 90% classification accuracy on most of the datasets, and improves significantly the classification accuracy. In particularly, on glass2 dataset, sensitive to class imbalance, Accuracy is improved by about 18.91%, 9.69% and 8.59% compared with ELM, WELM1 and WELM2, respectively. It illustrates that not only can EN-FWELM improve the classification accuracy of the minority class but also can maintain the classification accuracy of the majority class. Moreover, EN-FWELM is applicable to not only imbalanced data, but also relatively balanced data. The standard deviation (SD) acquired by EN-FWELM is also relatively much smaller than that acquired by other variants.

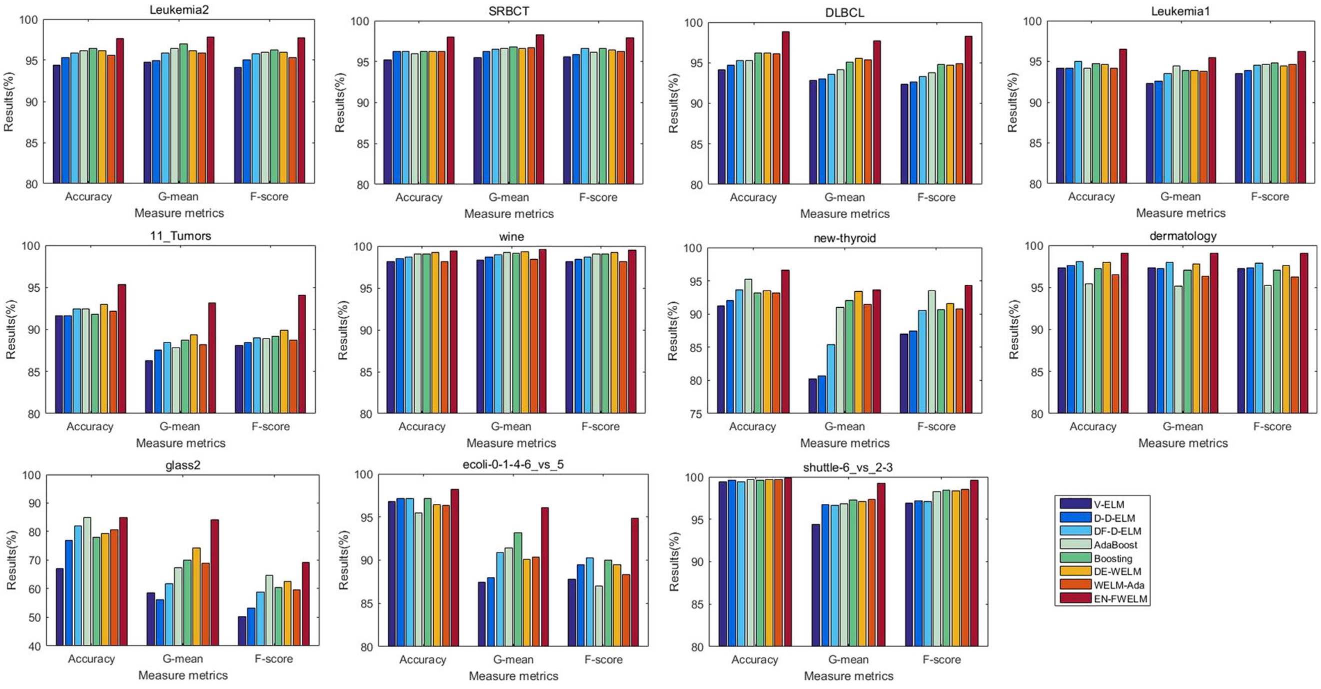

Fig. 2 Comparison of Performance results

From the results, it can be concluded that generalization performance of EN-FWELM outperforms other algorithms.

The computation cost of each method is evaluated by measuring training time. Training time averaged over 10 runs of each method is shown in Table 6. From Table 6, we can see that EN-FWELM learns multiple classifiers and consumes more training time than other variants, which is acceptable because the proposed EN-FWELM is based on ELM algorithm and ELM is fast in learning speed.

4.4 Comparison with other ensemble learning methods

To validate its effectiveness, EN-FWELM is also compared against other ensemble learning methods, including V-ELM [15], D-D-ELM, DF-D-ELM [19], AdaBoost, Boosting [16], DE-WELM [17] and WELM-Ada [18]. Performance results of above methods are shown in Fig. 2.

From Fig. 2, it can be seen that the classification performance of EN-FWELM outperforms other ensemble learning algorithms. Among these ensemble learning algorithms, the classification performance of D-D-ELM and DF-D-ELM based on the dissimilarity measure is better than V-ELM, but the classification results of V-ELM, D-D-ELM and DF-D-ELM are also relatively low. In particular, G-mean is remarkably on the decrease on the datasets sensitive to class imbalance, such as new-thyroid and glass2. The reason is that they are based on ELM algorithm and are more applicable to balanced data. Therefore, these methods ignore the minority class and result in relatively decrease in G-mean. In AdaBoost, multiple classifiers are trained serially. The distribution weights of training samples reflect their relative importance and the samples that are often misclassified will obtain larger distribution weights than the correctly classified samples. In Boosting, the distribution weights of training samples are adjusted according to the performance of the previous classifiers and updated separately for samples coming from different classes. Based on the distribution weights, they perform better than V-ELM, D-D-ELM, and DF-D-ELM. But in some cases, their G-mean is improved at the cost of relatively slight decrease in Accuracy, for instance, Leukemia1 and ecoli-0-1-4-6_vs_5. In DE-WELM, ensemble of WELMs based on $W1$ weighted strategy with different activation functions is constructed and DE is employed to optimize the weight of each WELM. In WELM-Ada, $W2$ weighted strategy is used as initial weight, then the fusion optimization of weighted ELM and AdaBoost is constructed. Similarly, in some cases, their G-mean is also improved at the cost of relatively slight decrease in Accuracy, such as new-thyroid and glass2. However, in all the cases, EN-FWELM achieves the best Accuracy, G-mean, and F-score. The reason is that samples distribution, samples number of each class and the dissimilarity of classifiers are all taken into account to strengthen the classification performance of EN-FWELM. In addition, the results of Lung_cancer are not given, because SMCL class has only 6 samples and there is always no sample in training-testing set.

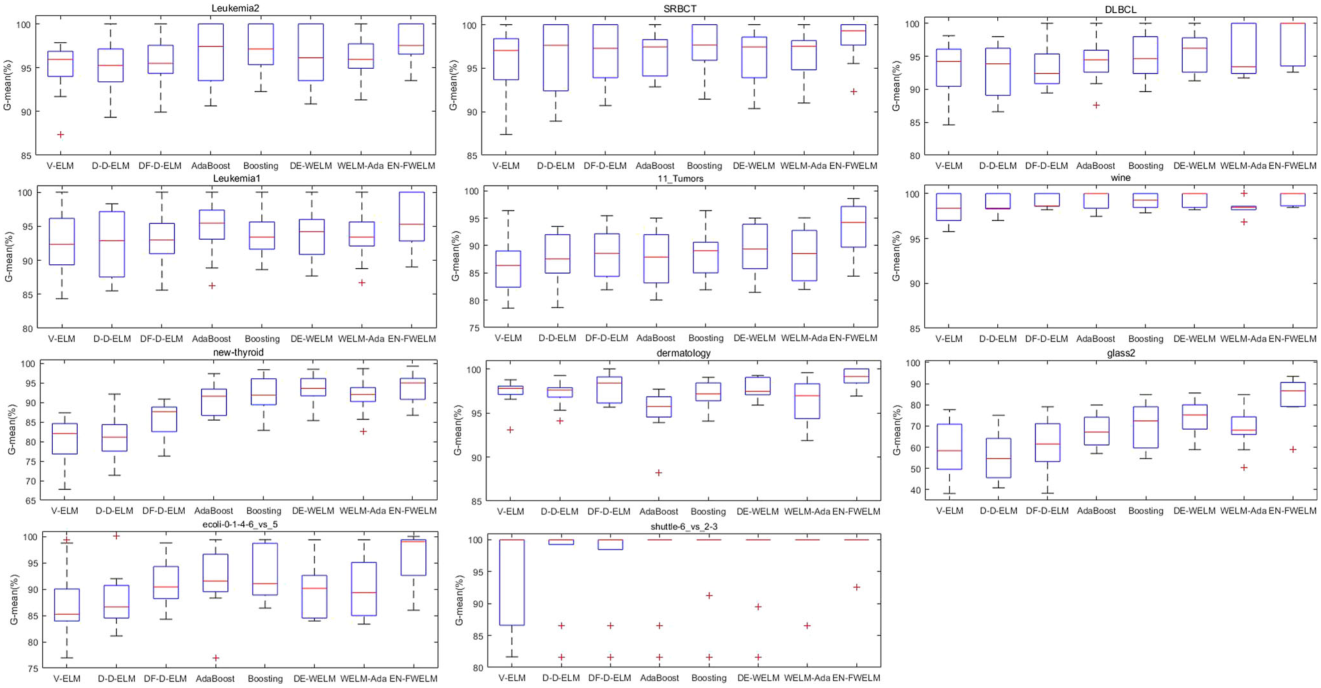

Fig. 3 The box plot representation using different methods

Table 7 Results of the paired t-test

Comparisons	p-value
EN-FWELM vs ELM	0.00230
EN-FWELM vs WELM1	0.00063
EN-FWELM vs WELM2	0.00290
EN-FWELM vs WELM	0.00089
EN-FWELM vs FWELM	0.00040
EN-FWELM vs V-ELM	0.01890
EN-FWELM vs D-D-ELM	0.00200
EN-FWELM vs DF-D-ELM	0.00017
EN-FWELM vs AdaBoost	0.00066
EN-FWELM vs Boosting	0.00290
EN-FWELM vs DE-WELM	0.00130
EN-FWELM vs WELM-Ada	0.00004

Based on the above analysis, the box plot is used to show G-mean results of different algorithms in Fig. 3. From the results, it can be seen that dispersion degree of EN-FWELM is relatively low, which indicates the robustness and stability of the proposed model. Furthermore, statistical testing is a meaningful way to study the difference between EN-FWELM and other algorithms. In this study, the paired t-test at a significance level of 0.05 is used to judge the difference based on the classification accuracy. It illustrates that the difference between two methods is significant if the p-value is less than 0.05. In fact, the p-value results are shown in Table 7. From Table 7, it can be seen that EN-FWELM exists significant difference with other algorithms.

5 Conclusion

In this study, an effective method named EN-FWELM is proposed to handle multi-class imbalance problem. The core components of EN-FWELM are FWELM, ensemble learning based on the dissimilarity and performance evaluation. In FWELM, balance factor is devised in combination with sample distribution and sample number associated with class to alleviate the bias against performance caused by imbalanced data, and fuzzy membership of sample is proposed to eliminate classification error coming from noise and outlier samples. Ensemble of FWELMs is used for making classification results more stable and accurate. Then some base FWELMs are removed based on dissimilarity measure, and the remaining base FWELMs are integrated by majority voting. Experiments are conducted on gene expression classification and real-world classification, and the proposed EN-FWELM is compared against its variants and other ensemble learning methods, respectively. It is proven that EN-FWELM remarkably outperforms other approaches in the literature, and it is applicable to not only imbalanced data, but also relatively balanced data. The future work is that the proposed method is to be evaluated in other medical diagnosis areas.

References

1. Liu JJ, Cai WS, Shao XG (2011) Cancer classification based on microarray gene expression data using a principal component accumulation method. Sci China Chem 54(5):802–811
2. Kar S, Sharma KD, Maitra M (2015) Gene selection from microarray gene expression data for classification of cancer subgroups employing PSO and adaptive K-nearest neighborhood technique. Expert Syst Appl 42(1):612–627
3. Yu HL, Hong SF, Yang XB (2013) Recognition of multiple imbalanced cancer types based on DNA microarray data using ensemble classifiers. BioMed Res Int 2013:1–13
4. Zainuddin Z, Ong P (2011) Reliable multiclass cancer classification of microarray gene expression profiles using an improved wavelet neural network. Expert Syst Appl 38(11):13711–13722
5. Huang GB, Zhu QY, Siew CK (2006) Extreme learning machine: theory and applications. Neurocomputing 70(1):489–501
6. Cao JW, Lin ZP, Huang GB (2012) Self-adaptive evolutionary extreme learning machine. Neural Process Lett 36(3):285–305
7. Huang GB, Ding X, Zhou H (2010) Optimization method based extreme learning machine for classification. Neurocomputing 74(1):155–163
8. Ding SF, Xu XZ, Nie R (2014) Extreme learning machine and its applications. Neural Comput Appl 25:549–556
9. Mohammed AA, Minhas R, Wu QMJ, Sid-Ahmed MA (2012) Human face recognition based on multidimensional pca and extreme learning machine. Pattern Recognit 44(10):2588–2597
10. Kaya Y, Uyar M (2013) A hybrid decision support system based on rough set and extreme learning machine for diagnosis of hepatitis disease. Appl Soft Comput 13(8):3429–3438
11. Li LN et al (2012) A computer aided diagnosis system for thyroid disease using extreme learning machine. J Med Syst 36(5):3327–3337
12. Hu L et al (2015) An efficient machine learning approach for diagnosis of paraquat-poisoned patients. Comput Biol Med 59:116–124
13. Lan Y, Soh YC, Huang GB (2009) Ensemble of online sequential extreme learning machine. Neurocomputing 72(13):3391–3395
14. Shigei N, Miyajima H, Maeda M et al (2009) Bagging and AdaBoost algorithms for vector quantization. Neurocomputing 73(1):106–114
15. Cao JW, Lin ZP, Huang GB, Liu N (2012) Voting based extreme learning machine. Inf Sci 185(1):66–77
16. Li K, Kong X, Lu Z, Liu W, Yin J (2014) Boosting weighted ELM for imbalanced learning. Neurocomputing 128:15–21
17. Zhang Y, Liu B, Cai J, Zhang SH (2016) Ensemble weighted extreme learning machine for imbalanced data classification based on differential evolution. Neural Comput Appl 28(1):1–9
18. Xu Y, Wang QW, Wei ZY (2017) Traffic sign recognition algorithm combining weighted ELM and AdaBoost. JCCS 38(9):2028–2032
19. Lu HJ, An CL, Zheng EH, Lu Y (2014) Dissimilarity based ensemble of extreme learning machine for gene expression data classification. Neurocomputing 128:22–30
20. Zong WW, Huang GB, Chen YQ (2013) Weighted extreme learning machine for imbalance learning. Neurocomputing 101(3):229–242
21. Zhang WB, Ji HB (2013) Fuzzy extreme learning machine for classification. Electron Lett 49(7):448–449

22. He H, Garcia EA (2009) Learning from imbalanced data. IEEE Trans Knowl Data Eng 21(9):1263–1284
23. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 16(1):321–357
24. Liu XY, Wu J, Zhou ZH (2009) Exploratory undersampling for class-imbalance learning. IEEE Trans Syst Man Cybern Part B 39(2):539–550
25. Zhou ZH, Liu XY (2006) Training cost-sensitive neural networks with methods addressing the class imbalance problem. IEEE Trans Knowl Data Eng 18(1):63–77
26. Bartlett PL (1998) The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network. IEEE Trans Inf Theory 44(2):525–536
27. Lin CF, Wang SD (2002) Fuzzy support vector machines. IEEE Trans Neural Netw 13(2):464–471
28. Lin SJ, Chang C, Hsu MF (2013) Multiple extreme learning machines for a two-class imbalance corporate life cycle prediction. Knowl-Based Syst 39(3):214–223
29. GEMS. http://www.gems-system.org/
30. KEEL repository. http://sci2s.ugr.es/keel/imbalanced.php
31. Cover TM, Thomas JA (1991) Elements of information theory. Wiley, New York

Yang Wang received her Master's degree from Liaoning Normal University, China, in 2006. She is currently pursuing the Ph.D. degree at the College of Information Science and Engineering, Northeastern University, China. Her main research interests include pattern recognition, machine learning and optimization theory.

Anna Wang is a Professor and the Ph.D. Supervisor at the College of Information Science and Engineering, Northeastern University, China. Her current research interests include pattern recognition, image processing and fault diagnosis in complex process industry.

Qing Ai received the B.S. and M.S. degrees 2003 and 2007, respectively. Currently he is a ph.D. candidate of pattern recog nition and intelligent system in Northeastern University, China and an associate professor with the University of Science and Technology Liaoning, China. His research interests include pattern recognition, support vector machines, optimization theory and applications, fault diagnosis.

Haijing Sun received the Bachelor's degree from Liaoning Shihua University, Fushun, China, in 2003, and Master's degree from Northeastern University, Shenyang, China, in 2006. She is currently pursuing the Ph.D. degree in Northeastern University, China. Her research interests include image processing and pattern recognition.