BioMedical Engineering
OnLine

RESEARCH

CrossMark

Imbalanced biomedical data classification using self-adaptive multilayer ELM combined with dynamic GAN

Liyuan Zhang, Huamin Yang* and Zhengang Jiang

*Correspondence:
yhm@cust.edu.cn
School of Computer Science
and Technology, Medical
Imaging Engineering
Laboratory, Changchun
University of Science
and Technology, No.7089,
Weixing Road, Changchun,
China

Abstract

Background: Imbalanced data classification is an inevitable problem in medical intelligent diagnosis. Most of real-world biomedical datasets are usually along with limited samples and high-dimensional feature. This seriously affects the classification performance of the model and causes erroneous guidance for the diagnosis of diseases. Exploring an effective classification method for imbalanced and limited biomedical dataset is a challenging task.

Methods: In this paper, we propose a novel multilayer extreme learning machine (ELM) classification model combined with dynamic generative adversarial net (GAN) to tackle limited and imbalanced biomedical data. Firstly, principal component analysis is utilized to remove irrelevant and redundant features. Meanwhile, more meaningful pathological features are extracted. After that, dynamic GAN is designed to generate the realistic-looking minority class samples, thereby balancing the class distribution and avoiding overfitting effectively. Finally, a self-adaptive multilayer ELM is proposed to classify the balanced dataset. The analytic expression for the numbers of hidden layer and node is determined by quantitatively establishing the relationship between the change of imbalance ratio and the hyper-parameters of the model. Reducing interactive parameters adjustment makes the classification model more robust.

Results: To evaluate the classification performance of the proposed method, numerical experiments are conducted on four real-world biomedical datasets. The proposed method can generate authentic minority class samples and self-adaptively select the optimal parameters of learning model. By comparing with W-ELM, SMOTE-ELM, and H-ELM methods, the quantitative experimental results demonstrate that our method can achieve better classification performance and higher computational efficiency in terms of ROC, AUC, G-mean, and F-measure metrics.

Conclusions: Our study provides an effective solution for imbalanced biomedical data classification under the condition of limited samples and high-dimensional feature. The proposed method could offer a theoretical basis for computer-aided diagnosis. It has the potential to be applied in biomedical clinical practice.

Keywords: Imbalanced data classification, Limited biomedical samples, High-dimensional feature, Multilayer ELM, Dynamic GAN

Zhang *et al. BioMed Eng OnLine*   (2018) 17:181

Page 2 of 21

Background

In the biomedical domain, machine learning techniques can make computer-aided diagnosis (CAD) [1] more intelligent in diagnoses of breast cancer, liver disorder, and other diseases. While imbalanced class distribution frequently occurs in real-world biomedical datasets, which causes the loss of essential pathological information from abnormal class [2]. Indeed, the misdiagnosis of abnormal class is more severe than that of a normal class in medical disease diagnosis [3]. Additionally, the training set sometimes contains high-dimensional feature and small samples. These factors further result in a lower classification accuracy of abnormal class and incorrect diagnosis result [4]. Therefore, establishing an effective classification model is an urgently necessary task for limited and imbalanced biomedical dataset.

To solve class-imbalanced classification problem, many studies [5–12] have been proposed. These methods mainly focus on three strategies: the algorithm level, the data level, and hybrid method. For the first strategy, the algorithm-based method often needs to amend the model parameters. Among numerous classifiers, ELM is famous owing to its analytical solution and fast learning speed, which is applicable to the engineering applications [13]. Various scholars have proposed some improved ELM models for imbalanced data classification [14–16]. So far, the weighted extreme learning machine (W-ELM) [17] is the most representative learning method for the class-imbalanced classification. The samples belonging to different classes are assigned different weights. This method attaches great importance to the minority class samples and alleviates the bias towards the majority class. A computationally efficient cost-sensitive method [18] has been developed by integrating a cost factor into the fuzzy rule-based classifier. The misclassified cost of majority class is set to one, while the penalty value of minority class equals to the imbalanced ratio. It is well suitable for a larger dataset. To extract hidden pathological features, forming a deep representation may be more meaningful [19]. Hierarchical ELM (H-ELM) [20] as a multilayer neural network has stable hierarchical structure. And it can produce a better feature representation by unsupervised feature learning. In view of the second strategy, the data-based method [21–24] concentrates on generating new samples for minority class (oversampling) or removing samples from majority class (undersampling). Resampling techniques are often employed as a preprocessing process. Different from cost-sensitive method, it is much easier to be implemented. The synthetic minority oversampling technique (SMOTE) [25] is a typical method. It creates synthetic samples to oversample the minority samples rather than mere data duplicating, thus avoiding the overfitting. Also, it is more helpful in recognizing outliers. Despite the goodness, this resampling method is prone to neglect the sample distribution and lead to the information loss.

The last strategy is the widely-employed hybrid method. Apart from the preprocessing methods, a better classification algorithm is beneficial for class-imbalanced classification task. For example, Yu et al. [26] proposed a combination method of asymmetric bagging ensemble classifier and feature subspace (Bagging-FSS). This method adopts random projection to establish the relationship between feature selection and ensemble classifier. The single classifier performance is improved by combining advantages of data pre-processing and ensemble learning methods in practical tasks. Similarly, Krawczyk et al. [27] combined the boosting scheme and evolutionary undersampling (EUS)

Zhang *et al. BioMed Eng OnLine*   (2018) 17:181

Page 3 of 21

technology for imbalanced classification of breast cancer malignancy. The usage of EUS allows selecting the most representative samples for boosting classifier, thereby improving the diversity of base classifiers. In fact, if the training sample is limited, this model will be difficult to guarantee the diversity of base classifiers. Moreover, this ensemble learning method largely depends on the performance of base classifier. In [28], synthetic minority oversampling technique and ELM (SMOTE-ELM) are integrated to provide an efficient solution for the imbalanced data classification. To produce a balanced dataset, the distribution of majority class samples is taken into consideration. Then, the oversampling of minority samples is conducted. SMOTE-ELM method has the lower bound of model reliability and reduces the information loss of majority samples. However, when addressing smaller dataset, particularly less training samples, the aforementioned works face some issues. How to establish the quantitative relationship between feature extraction and model selection should be considered to reduce manual parameters tuning. For this purpose, a specifically designed method to address the imbalanced biomedical data classification has important meanings in medical intelligent diagnosis.

In this paper, a self-adaptive multilayer ELM model with dynamic generative adversarial net (GAN) (for short PGM-ELM) is proposed to solve the class-imbalanced classification problem. The proposed method makes biomedical data classification more efficient and robust in the context of small-data and high-dimensional feature. The main contributions of this paper are summarized as follows: Principal component analysis (PCA) is used to remove irrelevant and redundant features from raw feature set, thereby extracting more effective features; Dynamic GAN is introduced to generate the realistic-looking minority class samples and balance the class distribution, thus alleviating effect of the imbalanced dataset and avoiding overfitting; The analytic expression for numbers of hidden layer and node is determined by establishing the quantitative relationship among the changes of imbalance ratio, the sample distribution, and the hyper-parameters of model. This provides a solution for reducing the parameter sensitivity of multilayer ELM. The effectiveness of the PGM-ELM model is validated and evaluated on four biomedical datasets. The obtained experimental results can help guide us to construct the optimal classification model for practical biomedical applications.

The remaining of the paper is organized as follows. "Related works" section simply introduces the basic principles of hierarchical ELM and classical GAN. Then, the detailed process of the proposed method is described in "Methods" section. Afterwards, the dataset description, evaluation metrics, and experimental results are presented in "Results" section. Comparative analyses of the proposed method with other state-of-the-art methods are given in "Discussions" section. Finally, "Conclusions" section provides the conclusion and future research directions of this paper.

Related works

Some variants of the ELM model have been employed effectively. Here the basic principles of H-ELM and classic GAN are briefly described. They can contribute to the solving of imbalanced biomedical data classification.

Hierarchical extreme learning machine framework

In network structure of H-ELM, the original input is decomposed into multiple hidden layers. The output of the previous layer is regarded as the input of the current one. The learning of hidden layer represents more abstract information. By doing so, the hidden information can be exploited for deeper feature representation.

Assume that we have a training set $\{(x_i, t_i)\}_{i=1}^{N}$, where x_i denotes the input node i, and t_i stands for the output of the ith sample. A single hidden layer feedforward neural network with L hidden nodes is used to fit N training samples. Then, the corresponding output function of ELM can be expressed as [29]
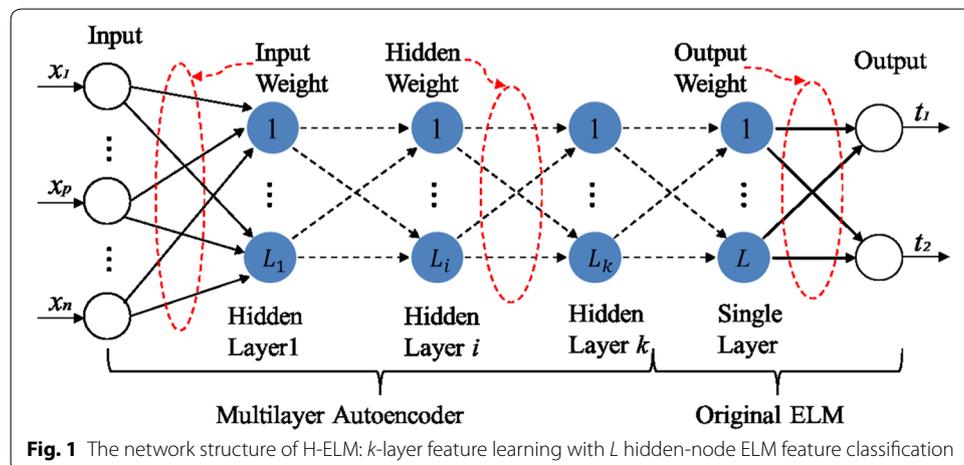
$$f(x) = h(x)\beta = h(x)\left(\frac{\mathbf{I}}{C} + \mathbf{H}\mathbf{H}^T\right)^{-1}\mathbf{H}^T\mathbf{T}, \tag{1}$$

where $\mathbf{H} = [h(x_1), \ldots, h(x_N)]^T$ is the randomized output matrix of hidden layer, and $\mathbf{T} = [t_1, \ldots, t_N]^T$ is the target matrix of the output layer. β denotes the connection weight from a hidden layer node to each output node. C is a regularization coefficient. $\mathbf{I}$ is a unit matrix. The input weight and bias will be assigned randomly. Desired outputs of minority and majority classes are set to 1 and 0. Figure 1 shows the basic network structure of H-ELM, which consists of two separate parts: unsupervised and supervised training.

The first phase is unsupervised feature learning by ELM-based autoencoder (ELM-AE) [30]. ELM-AE based ℓ_1-norm optimization is employed to form a multi-layer feature learning model. By recovering the input data as much as possible, new features can be learned to represent the input data. A fast iterative shrinkage-thresholding algorithm (FISTA) [31] is utilized to obtain weight β of each hidden layer. The optimization model of ELM-AE is given by

$$O_\beta = \arg\min\left\{\|\mathbf{H}\beta - \mathbf{X}\|^2 + \|\beta\|_{\ell_1}\right\}, \tag{2}$$

where $\mathbf{X}$ is the original input data. $\mathbf{H}$ represents the random initialized output.

Fig. 1 The network structure of H-ELM: *k*-layer feature learning with *L* hidden-node ELM feature classification

Next, the second phase is supervised feature classification. The original ELM is performed for final decision making. The output of the H-ELM is calculated by using the last layer output of the ELM-AE as the input of the parallel ELM. Mathematically, the output of each hidden layer can be represented as

$$\mathbf{H}_i = g(\mathbf{H}_{i-1} \cdot \beta), \tag{3}$$

where $\mathbf{H}_i (i \in (1, \dots, K))$ is the output of the ith hidden layer. $g(\cdot)$ denotes the activation function of the hidden layers, and β represents the output weight. Here, the node number L_k of the kth hidden layer equals to the node number L_{k-1} of the $(k-1)$th hidden layer. Different from deep back propagation (BP) network, all hidden neurons in H-ELM as a whole are not required to be iteratively tuned. The parameter of the last hidden layer will be adjusted no longer.
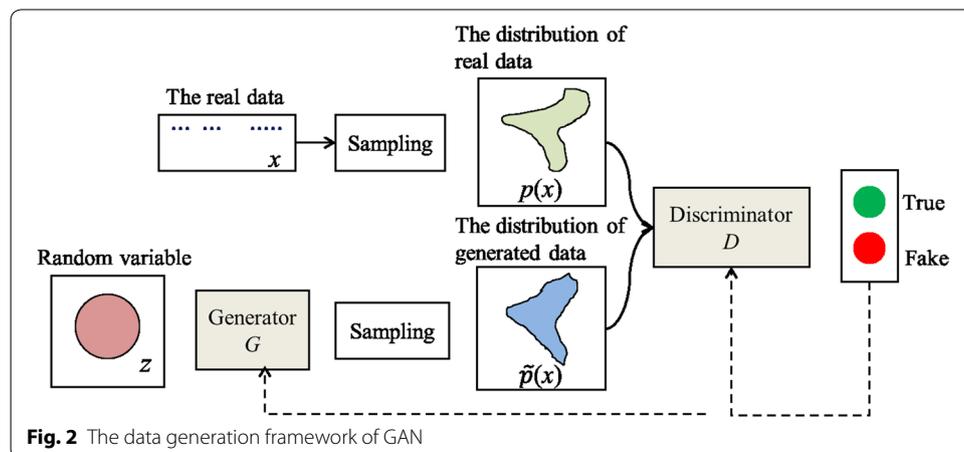
Generative adversarial net

GAN [32] is a combination method of simulation and unsupervised learning, and it largely depends on the adversarial relationship among competitive neural networks. GAN can generate entirely new data like the observed data based on the probability distribution model. Figure 2 presents the whole data generation process. GAN simultaneously trains the generative model G and the discriminative model D by playing a non-cooperative game. G can capture the data distribution to generate samples, while D assists G to classify these samples as true or fake. By discriminator D to optimize, the parameters of G are adjusted to make the probability distribution $\tilde{p}(x)$ and the real data distribution $p(x)$ as close as possible.

This process can be expressed by minimizing an objective function. The overall objective function of GAN model is given as follows

$$\min_G \max_D V(D, G) = E_{x \sim p_{data}(x)} \big[\log D(x) \big] + E_{z \sim p_z(z)} \big[\log (1 - D(G(z))) \big], \tag{4}$$

where $p_{data}(x)$ is the distribution of the training set. $p_z(z)$ is the distribution of noise. E denotes the expectation. If the generator G is fixed, the optimal discriminator D is depicted by the following formula.

Fig. 2 The data generation framework of GAN

Zhang *et al. BioMed Eng OnLine* (2018) 17:181

Page 6 of 21

$$D_G^*(x) = \frac{p_{data}(x)}{p_{data}(x) + p_g(x)},$$

(5)

where $p_g(x)$ expresses the probability distribution of the generator. The training objective for D can be interpreted as maximizing the log-likelihood for estimating the conditional probability $P(Y = y|x)$. The Y makes clear whether the x comes from the real data or the generated data. Therefore, the minimax game in Eq. (4) can be rewritten as
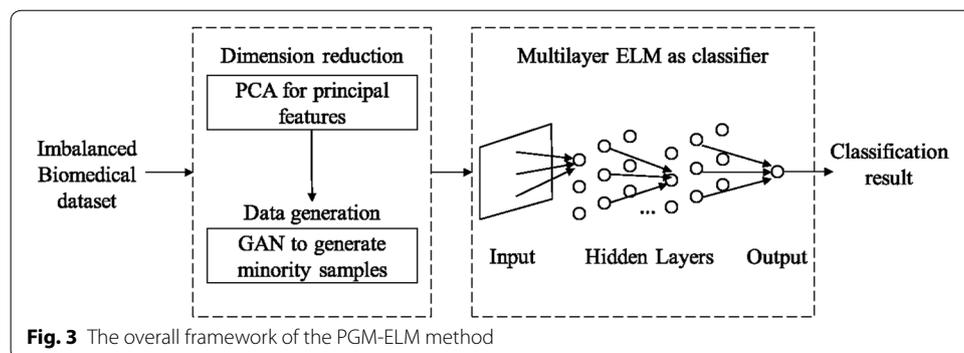
$$\max_D V(G, D) = E_{x \sim p_{data}} \left[\log \frac{p_{data}(x)}{p_{data}(x) + p_g(x)} \right] + E_{x \sim p_g} \left[\log \frac{p_g(x)}{p_{data}(x) + p_g(x)} \right].$$

(6)

G and D will reach a balance after conducting several times training, that is $p_g = p_{data}$. The discriminator is incapable to distinguish the difference between two distributions, such that $D_G^*(x) = 1/2$ [33].

Methods

Throughout this paper, aiming at the limited and imbalanced biomedical data, a hybrid PGM-ELM classification model is proposed. Figure 3 shows the whole process of the proposed method. In Fig. 3, the model first employs PCA to extract the principal features and reduce dimensionality. Afterwards, we use GAN to dynamically generate real minority class samples, thus balancing the class distribution. Lastly, considering the numbers of samples and features, once the quantitative relationship between the imbalance ratio and the hyper-parameters of multilayer ELM is established. A self-adaptive PGM-ELM classification model is constructed for imbalanced classification.

For a given training set with N samples $DS = \left\{ (\mathbf{x}_i, y_i) \right\}_{i=1}^N$, $\mathbf{x}_i$ denotes the feature vector of the ith sample, and y_i is the class label of the ith sample. In our study, the medical diagnosis with or without lesions is identified as a binary classification problem. For convenience, N^+ represents the number of the minority class samples, and N^- represents the number of the majority class samples. $N = N^- + N^+$ is the total number of all samples in training set.

Fig. 3 The overall framework of the PGM-ELM method

Zhang *et al. BioMed Eng OnLine*   (2018) 17:181

Page 7 of 21

Principal features extraction

Most of original biomedical datasets have lots of noise and redundant features. PCA is adopted to remove the irrelevant and redundant information [34]. For the original feature set $X = \left\{x^{(1)}, x^{(2)}, \ldots, x^{(M)}\right\}$, the matrix $\tilde{X}$ is obtained through standardized processing. This transform relation is given by

$$\tilde{x}^{(i)} = \frac{x^{(i)} - \mu^{(i)}}{\delta^{(i)}}, \tag{7}$$

where $\tilde{x}^{(i)}$ is the ith feature of standardized matrix. $x^{(i)}$ is the ith sample in original feature set. $\mu^{(i)}$ and $\delta^{(i)}$ are the mean value and the variance of the original features. The covariance matrix is calculated as follows

$$R = \tilde{X}^T \tilde{X} / (M - 1). \tag{8}$$

The eigenvalue decomposition is applied to solve the eigenvalues and corresponding eigenvectors of the covariance matrix. The eigenvalues are arranged from large to small, and the contribution rate is computed. The formula is described as follows

$$\alpha = \sum_{k=1}^{r} \lambda_k \Big/ \sum_{k=1}^{M-1} \lambda_k, \tag{9}$$

where λ_k denotes the kth eigenvalue. The threshold of cumulative contribution rate of the eigenvalue is selected as 85%. When the proportion of the largest M' eigenvalues is greater than this threshold, M' is viewed as the number of the principal components. By calculating the product of the standard feature matrix and eigenvector, we get the corresponding principal component vector, which is expressed as follows

$$z^{(i)} = \sum_{j=1}^{M'} \tilde{x}^{(j)} \boldsymbol{\eta}_i^T, \tag{10}$$

where $\boldsymbol{\eta}_i$ represents the standard orthogonal eigenvectors corresponding to the ith eigenvalues. $Z = \left\{z^{(1)}, z^{(2)}, \ldots, z^{(M')}\right\}$ is new feature set after analyzing the principal components.

Samples generation

From the perspective of the data, dynamic GAN generates new samples to change the imbalanced ratio. To fully make use of the data distribution, all minority class samples as a whole chunk are input into GAN model. And then, dynamic GAN is executed multiple times to balance class samples. It is worthy note that the execution number of GAN is set to $num = \left\lfloor \frac{N^-}{N^+} \right\rfloor$ according to initial imbalanced ratio, where $\lfloor \cdot \rfloor$ is on behalf of the round down. That is to say, the samples generation procedure using GAN is repeated until the imbalanced ratio is closer to 1. By doing so, the class distribution is balanced as much as possible.

For the minority class samples $\mathbf{X}^+$, the initial condition is noise $\mathbf{Z}$ with the same size as the whole target fragment. The objective function of GAN can be depicted by the following formula.

Zhang *et al. BioMed Eng OnLine* (2018) 17:181

Page 8 of 21

$$\min_{G} \max_{D} V(D, G) = E_{\mathbf{x}_{\mathbf{k}}^{+} \sim p_{data}(num \cdot \mathbf{X}^{+})} \left[\log D \left(\left\lfloor \frac{N^{-}}{N^{+}} \right\rfloor \cdot \mathbf{X}^{+} \right) \right]$$
$$+ E_{\mathbf{z} \sim \mathbf{p_z(z)}}[\log (1 - D(G(\mathbf{Z})))]. \tag{11}$$

The optimal discriminator D equals to $\frac{p_{data}(\mathbf{X}^{+})}{p_{data}(\mathbf{X}^{+}) + p_g(\tilde{\mathbf{X}}^{+})}$. $p_g(\tilde{\mathbf{X}}^{+})$ denotes the distribution of generated data. The discriminator D can be updated by whole target segment.

$$\nabla_{\theta_d} \frac{1}{num \cdot N} \sum_{i=1}^{num \cdot N} [\log D(x_i) + \log(1 - D(G(z_i)))], \tag{12}$$

where, x_i and z_i denote the samples of $\mathbf{X}^{+}$ and $\mathbf{Z}$. θ_d is the parameter of discriminator D. Generator G is updated by

$$\nabla_{\theta_g} \frac{1}{num \cdot N} \sum_{i=1}^{num \cdot N} [\log(1 - D(G(z_i)))], \tag{13}$$

where θ_g is the parameter of generator G. If G recovers data distribution, and D equals to 0.5 in any instance, the new samples $\tilde{\mathbf{X}}^{+}$ will be generated. The sample number of the training set is increased to $N' = \left\lfloor \frac{N^{-}}{N^{+}} \right\rfloor \cdot N^{+} + N^{-}$. $IR = \frac{N^{+}}{N^{-}}$ is initial imbalanced ratio of the training set, while $IR' = \left\lfloor \frac{N^{-}}{N^{+}} \right\rfloor \cdot N^{+}$ represents new imbalanced ratio after samples generation. For clear representation, the change of imbalanced ratio ΔIR can be obtained as follows

$$\Delta IR = IR' - IR = \frac{\left\lfloor \frac{N^{-}}{N^{+}} \right\rfloor \cdot N^{+}}{N^{-}} - \frac{N^{+}}{N^{-}} = \frac{\left(\left\lfloor \frac{N^{-}}{N^{+}} \right\rfloor - 1 \right) \cdot N^{+}}{N^{-}}. \tag{14}$$

Self-adaptive multilayer ELM modeling

In last phase of the PGM-ELM, using the multilayer ELM model is to classify the balanced dataset. The network structure of the classification model is first determined. In fact, multilayer ELM is sensitive to the numbers of hidden layer and node. Sometimes it is difficult for users to specify an appropriate number of nodes without prior knowledge. If the number of nodes is too small, the classifier is unable to learn feature well, causing the under-fitting performance. If the number of nodes is too big, the time complexity of the network structure will be increased. Generally, it is related to the numbers of sample and feature. Therefore, the change of the imbalanced ratio and the number of new features are considered in our multilayer ELM model. Mathematically, the number of hidden nodes is obtained by

$$P = \left\lceil (1 - \Delta IR) \times \frac{N}{M} + \Delta IR \times \frac{N'}{M'} \right\rceil. \tag{15}$$

Simultaneously, the number of hidden layers is determined by

Zhang *et al. BioMed Eng OnLine*   (2018) 17:181

Page 9 of 21

$$Q = \lceil \Delta IR \times M' \rceil, \tag{16}$$

where $\lceil \cdot \rceil$ shows the round up.

It can be found that, on the one hand, the bigger the change of imbalanced ratio is, the greater the number of hidden layers is. On the other hand, the more numbers of the feature and generated samples are, the larger the number of hidden nodes is. This specific relationship can self-adaptively adjust the parameters of model for different datasets. After that, the designed network is learned layer by layer using the M–P generalized inverse. And the functional relationship of each layer is achieved as follows

$$\boldsymbol{\beta} = \mathbf{H}_Q^T \left(\frac{\mathbf{I}}{C} + \mathbf{H}_Q \mathbf{H}_Q^T \right)^{-1} \mathbf{T}_Q, \tag{17}$$

where $\mathbf{H}_Q = \begin{bmatrix} g(a_1 \cdot x_1 + b_1) & \dots & g(a_L \cdot x_1 + b_P) \\ \vdots & \dots & \vdots \\ g(a_1 \cdot x_{N'} + b_1) & \dots & g(a_L \cdot x_{N'} + b_P) \end{bmatrix}_{N' \times P}$ is the output matrix of the

Qth hidden layer. a is the orthogonal random weight vector between input nodes and hidden nodes. b is the orthogonal random threshold of the hidden neurons. The sigmoid function is selected as the activation function $g(\cdot)$. This function expression is

$$g(u) = \frac{1}{(1 + \exp(-au))}. \tag{18}$$

Finally, the output matrix $\boldsymbol{\beta}$ is obtained, and the entire hybrid model is established. Pseudo-code description for the process of hybrid approach is shown as Algorithm 1.

Algorithm 1 The training process of PGM-ELM model

Input: The training dataset T_{train} and the activation function $g(\cdot)$.
Output: The parameter β of the PGM-ELM model.
  1: Extracting the principal component vector z from the original dataset $X = \{x^{(1)}, x^{(2)}, \dots, x^{(M)}\}$.
  2: **if** $N^- - N'^+ > N^+$ **then**
  3:     Generating new minority samples to balance the class distribution using dynamic GAN.
  4: **else**
  5:     Calculating the new imbalanced ratio.
  6: **end if**
  7: Determining the number of hidden node P and the number of hidden layer Q.
  8: Randomly assigning the input weight a_j and the bias weight b_j.
  9: **for** $i = 1 \rightarrow Q$ **do**
 10:     Calculating the output matrix $\mathbf{H}_i = g(\mathbf{H}_{i-1} \cdot \beta)$ of the ith ($i \in [1, Q]$) hidden layer.
 11:     Obtaining the output weight vector β.
 12:     Computing the target matrix $\mathbf{T}_i = \mathbf{H}_i \beta$.
 13: **end for**
 14: Inputing the output $\mathbf{H_Q}$ of the Qth layer into the original ELM model.
 15: Calculate the weight vector $\boldsymbol{\beta}$ of output layer.
 16: **return** The weight vector β

Results

In this section, to validate the effectiveness of the proposed PGM-ELM method, extensive experiments have been performed. We first describe four real-world imbalanced biomedical datasets derived from the UCI machine learning repository [35]. Then we

present the classification results of our method. Also, the obtained results are discussed adequately. Our experimental computer configurations are listed as follows: Intel(R) dual-core, 3.20 GHz, 8 GB RAM with Windows 7 Operating System. All algorithms in this study are programmed with MATLAB R2014a.

Datasets description

For constructing a small training sample set, each dataset are divided into the training and test sets via a random sampling process. The breast cancer diagnostic dataset provides information on the discrimination of benign and malignant. Each instance has one ID number, 30 real value variables and one diagnosis label. The Indian liver dataset describes liver patient or not, which is made up of two patient information, eight real-valued features and a class label. The diabetic retinopathy Debrecen dataset with 19 numerical features contains the sign of diabetic retinopathy or not. The Pima diabetes dataset collects pathologic data from diabetes patients, including eight real-valued features and a class label. Table 1 summarizes the detailed information of the four biomedical datasets.

From Table 1 we can see that these four datasets are imbalanced since the imbalance ratios are much less than 1. Besides, they have different feature dimensionalities and smaller instances. It is noticeable that all datasets should be normalized to facilitate processing. Furthermore, only real-valued features are used as the input of the model in all experiments. Considering the fact that the distinction between normal and abnormal is a typical two-class classification task, so the labels containing majority and minority classes are specified as 0 and 1, respectively.

Performance evaluation metrics

In order to evaluate the classification performance of the proposed model, there are several commonly considered measurement criteria that can be used in imbalanced classification task [36]. First, Table 2 gives the confusion matrix of a two-class problem for explaining the performance measures. TP and TN are the numbers of correctly classified positive and negative samples, respectively. FP and FN are the numbers of the

Table 1 Description of the experimental datasets

Datasets	Attributes	Minority	Majority	Imbalance ratio	Training samples	Test samples
Breast cancer	32	212	357	0.59	100	469
Liver patient	11	26	224	0.12	50	200
Diabetic retinopathy	20	100	300	0.30	100	300
Pima diabetes	9	117	517	0.23	200	434

Table 2 Confusion matrix for a two-class problem

	Positive prediction	Negative prediction
Positive class	True positive (TP)	False negative (FN)
Negative class	False positive (FP)	True negative (TN)

Zhang *et al. BioMed Eng OnLine* (2018) 17:181

Page 11 of 21

misclassified negative and positive samples, respectively. The confusion matrix gives the quantitative classification results on each dataset.

And then, receiver operator characteristic (ROC) is a graphical method to intuitively show the compromise between the true positive rate and false positive rate for the classification models. Area under the ROC curve (AUC) can describe the performance of classifiers in different decision thresholds. The AUC value is larger, the better the performance of classifier is. G-mean is a popular measure to indicate the geometric mean of sensitivity and specificity. F-measure is the harmonic mean of precision and recall. They can be effective to evaluate generalization performance than overall classification accuracy, and their definitions are expressed as follows.

$$G\text{-}mean = \sqrt{TPR \cdot TNR}, \tag{19}$$

$$F\text{-}measure = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{20}$$

where, true positive rate (TPR) represents the proportion of positive samples to be correctly classified as positive class, whose definition is the same as Recall. True negative rate (TNR) indicates the proportion of negative samples to be correctly classified as negative class. Precision denotes the proportion of positive samples to be correctly classified and all positive samples. They are defined in the following.
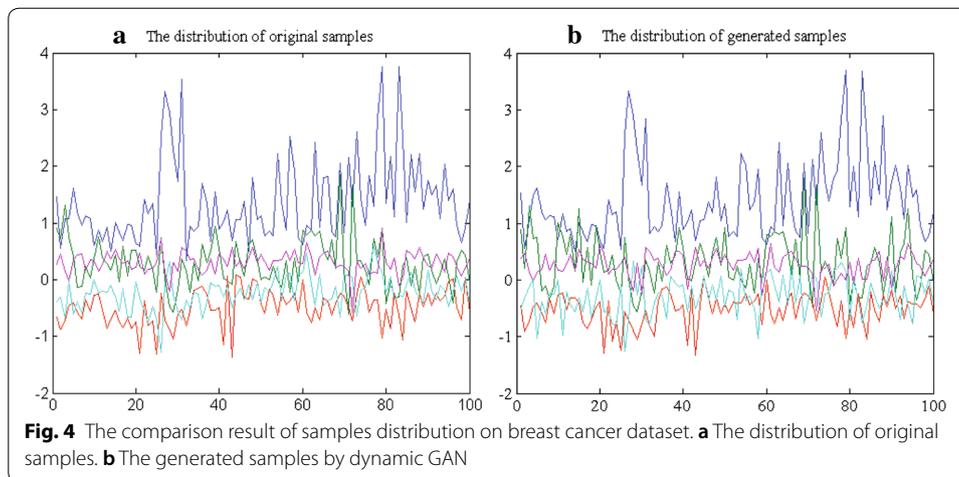
$$TNR = \frac{TN}{FP + TN}. \tag{21}$$

$$TPR = \text{Recall} = \frac{TP}{TP + FN}. \tag{22}$$

$$\text{Precision} = \frac{TP}{TP + FP}. \tag{23}$$

The result analysis of dynamic GAN

First of all, the principal components of original feature set are extracted from a given imbalanced training set by using PCA. Thereafter, new balanced dataset are achieved after generating minority class samples using dynamic GAN. In the network structure of dynamic GAN, several appropriate parameters are selected to generate realistic minority class samples. The number of hidden nodes is set to 100. The learning rate is set to 0.01. Dropout fraction of discriminator D and generator G are set to 0.9 and 0.1, respectively. The activation function of GAN is given as follows: the generator G uses ReLU and Sigmoid, while the discriminator D employs Maxout and Sigmoid. Figure 4 depicts the comparative distributions of the original samples and the generated samples after performing the dynamic GAN.

In Fig. 4, five different colors represent five principal components after performing PCA. There are 100 minority class samples derived from breast cancer dataset. In general, similar dataset should be represented by similar distribution. We can easily observe that, the distribution of the generated samples is consistent with the original sample distribution. This visually proves that the dynamic GAN is capable to capture

Zhang *et al. BioMed Eng OnLine*   (2018) 17:181

Page 12 of 21

Fig. 4 The comparison result of samples distribution on breast cancer dataset. **a** The distribution of original samples. **b** The generated samples by dynamic GAN

Table 3 Comparison result of Kernel MMD on four test sets

The dataset	50 samples	100 samples	200 samples
Breast cancer	0.104	0.096	0.089
Liver patient	0.181	0.118	0.124
Diabetic retinopathy	0.219	0.143	0.078
Pima diabetes	0.072	0.057	0.060

the distribution of actual data to generate convincing samples, thus balancing the class distribution and avoiding the overfitting.

To quantify the quality of generated data, we compute the dissimilarity between the distributions of generated data and original data by means of kernel maximum mean discrepancy (MMD). Kernel MMD [37] is a popular sample-based evaluation metric for quantitatively evaluating GANs model. A lower MMD means that the distribution of generated data is consistent with that of the real data. Table 3 reports the comparison results of Kernel MMD on four datasets. All MMD values are calculated over 50, 100 and 200 samples generated by dynamic GAN. In Table 3, as increasing the number of samples, the smaller the MMD value is, the higher the quality of generated samples is. Based on this quantitative result, we can conclude that the dynamic GAN can capture the training data distribution. GAN can be appropriate for producing samples without the information loss of majority class in class-imbalanced classification.
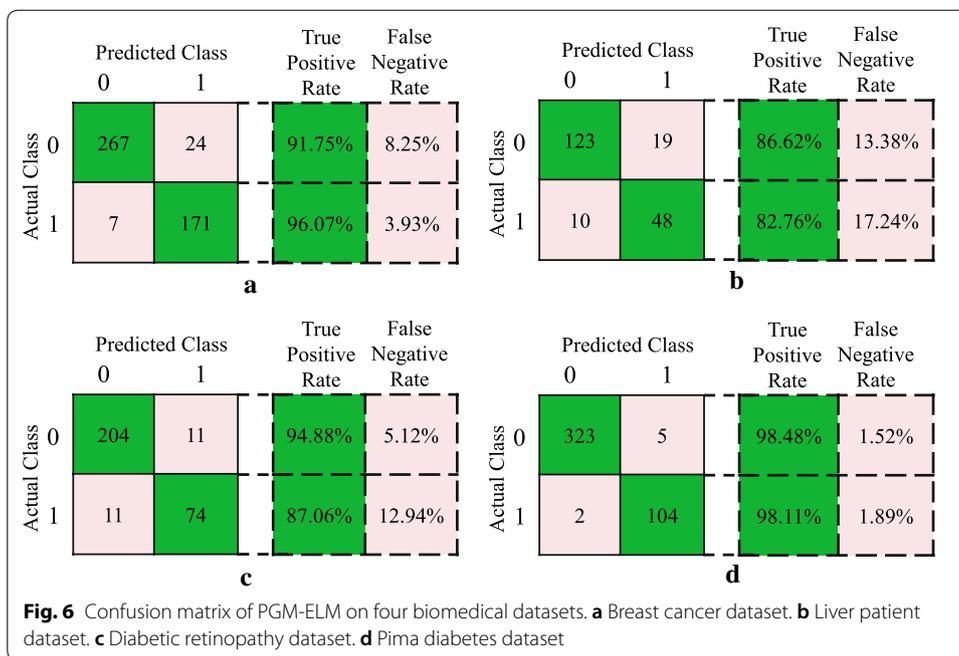
Analysis of the classification results

In order to examine the classification results of PGM-ELM against other constructive algorithms: W-ELM, SMOTE-ELM, and H-ELM. We give the corresponding results of these algorithms on four biomedical datasets. Considering fact that the weight of ELMs model is randomly chosen, four methods are ran 20 independent monte carlo trials. The final result is from the average of the 20 results. For fair comparison, these methods use same sigmoid activation function for learning.

Consequently, Fig. 5 displays the spatial distribution of classification results on four datasets after performing one monte carlo trial. The correctly classified samples and the misclassified samples are visualized. From Fig. 5 can be seen that the correctly classified samples are much more compared to the misclassified ones on each dataset. Obviously, Pima diabetes dataset yields the best classification result of PGM-ELM model. And its misclassified samples number is much less than those of other datasets. This reflects better classification ability of the PGM-ELM for most of biomedical datasets.

Apart from the spatial distribution results, the result of confusion matrix (two-class case: 0 for majority class and 1 for minority class) on four biomedical datasets is presented in Fig. 6. The numbers of correctly classified and misclassified samples are shown. Corresponding true positive rate (TPR) and false negative rate (FNR) are computed. Taking breast cancer dataset as an example, given a classification of the minority class 1, 171/178 will be correct (class 1). Moreover, the number of misclassified minority sample is smaller than the misclassified rate of the majority class. It can be seen that most of predicted samples are classified as actual class on each dataset. Therefore, the proposed PGM-ELM significantly improves the classified rate of minority class samples. This reflects a superior classification capacity for imbalanced biomedical dataset.

Fig. 5 Spatial distribution of sample using PGM-ELM. (Blue circles describe the correctly classified samples, while the red stars mean the misclassified samples.) **a** Breast cancer dataset. **b** Liver patient dataset. **c** Diabetic retinopathy dataset. **d** Pima diabetes dataset

Fig. 6 Confusion matrix of PGM-ELM on four biomedical datasets. **a** Breast cancer dataset. **b** Liver patient dataset. **c** Diabetic retinopathy dataset. **d** Pima diabetes dataset
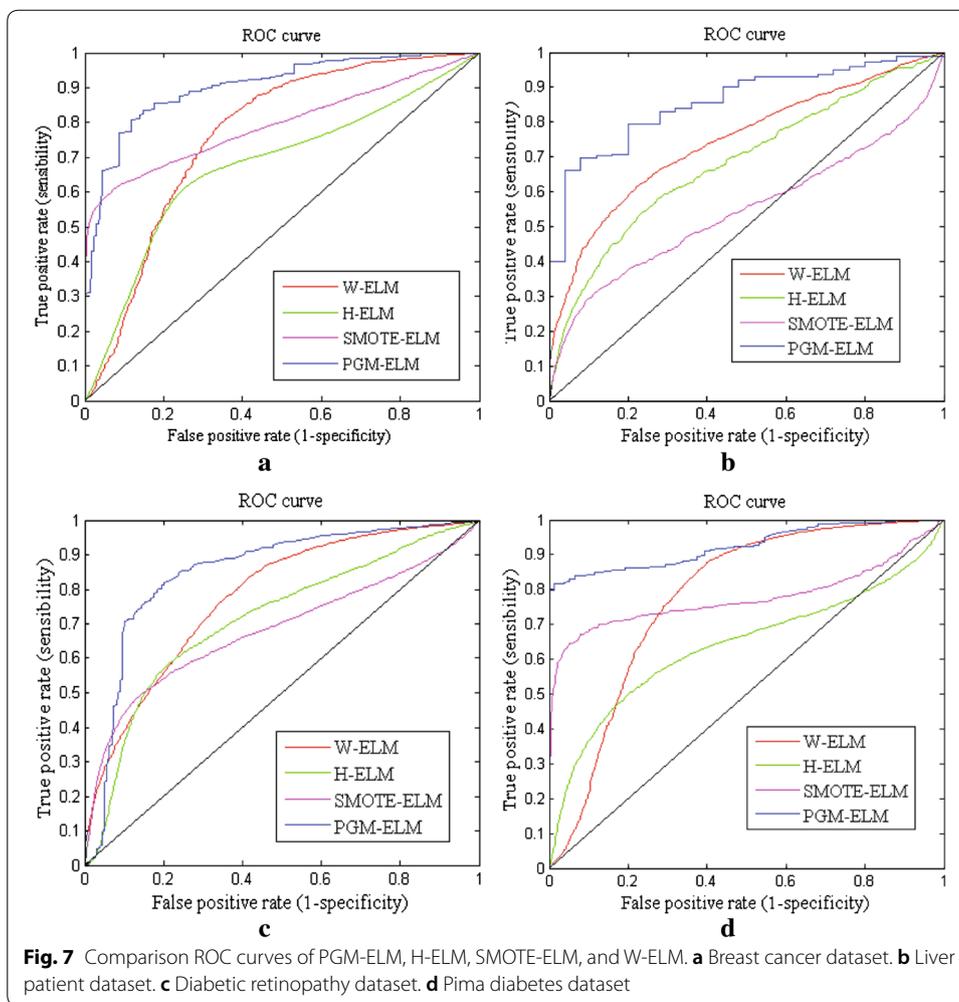
Meanwhile, we assess the classification performance of four models in terms of ROC curve. Figure 7 shows comparison results of the averaged ROC curve on four datasets. From almost most of results of Fig. 7a–d can be seen that, by comparing with other three algorithms, the PGM-ELM method has much higher ROC curve on each dataset. However, H-ELM has a relatively poor performance, especially on small training set, which is showed in Fig. 7a, d. It can explain that H-ELM is sometimes difficult to control the optimal hyper-parameters by manually tuning parameter. In Fig. 7b, the ROC curve of SMOTE-ELM is higher at first and tends to the obvious decline at last. Generally, SMOTE method uses local information to generate synthetic samples. When the training set is smaller and severe imbalanced, it usually ignores the overall class distribution, leading to some information loss. By contrast, although W-ELM reveals a merely superior recognition ability to these two algorithms on breast, liver, and diabetes datasets. But if data dimensionality is greater, W-ELM poorly performs the classification due to some redundant features. The PGM-ELM can present better performance thanks to the realistic-looking samples generation and the information loss reduction by dynamic GAN. More importantly, biomedical hidden features are learned by using layer wise unsupervised learning.

Now onto a discussion about the number of hidden nodes in ELMs model. Limited availability of the training samples necessitates careful selection of the parameters of the hidden layer, thereby achieving well-generalizing model. To this end, we give the analytic expression for numbers of layer and hidden node in PGM-ELM. The accumulated G-means and F-measures of four models as changing the number of hidden nodes are illustrated in Figs. 8 and 9.
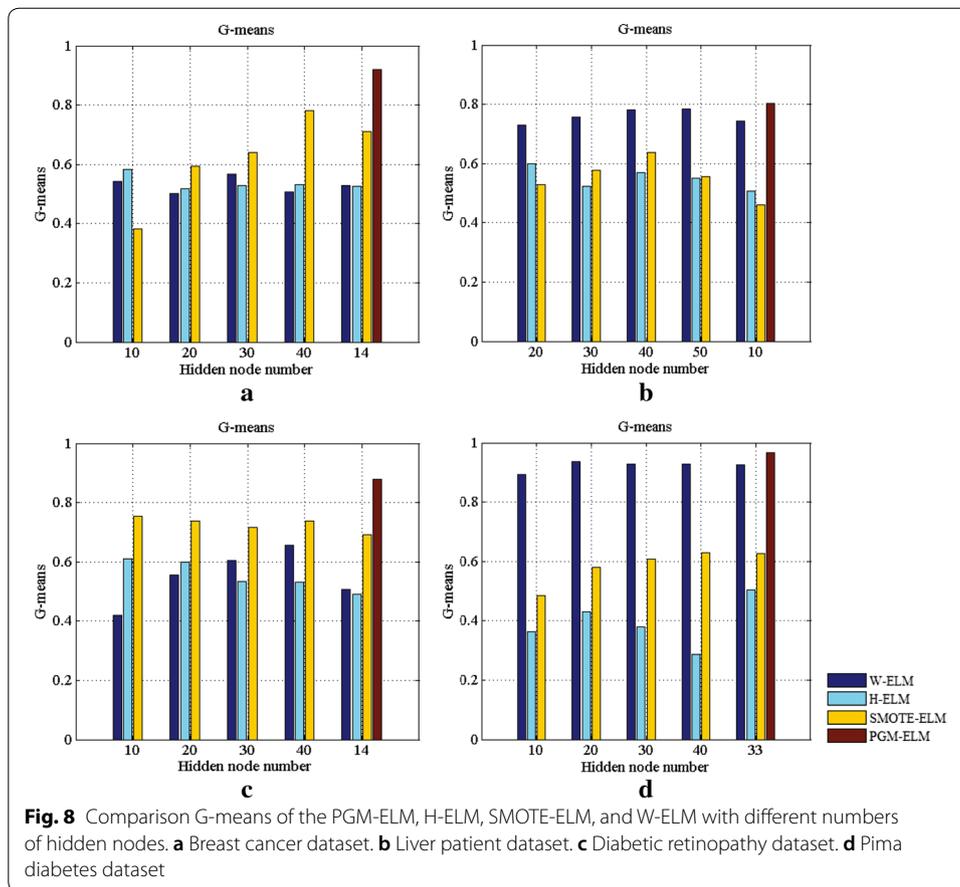
As can be seen from Figs. 8a and 9a, taking breast cancer dataset as an example, the PGM-ELM gets the highest G-mean and F-measure when the number of hidden nodes

Fig. 7 Comparison ROC curves of PGM-ELM, H-ELM, SMOTE-ELM, and W-ELM. **a** Breast cancer dataset. **b** Liver patient dataset. **c** Diabetic retinopathy dataset. **d** Pima diabetes dataset

is 14. It suggests that our method obtains better classification accuracy and robustness. Besides, we can easily observe that, compared with H-ELM, PGM-ELM shows superior performance in case of same number of hidden nodes on most of datasets. This indicates that PGM-ELM is not sensitive to the hyper-parameter of hidden layer by considering the changes of imbalance ratio and sample distribution. This is explained by the fact that the analytical solution for parameters of the hidden layer makes classification results more accurate. For W-ELM and SMOTE-ELM, G-mean and F-measure only slightly change with different hidden nodes. This is perhaps because that simpler single layer network is also less sensitive to the number of hidden nodes. As a consequence, these results demonstrate the adaptability of the proposed PGM-ELM in dealing with small sample and imbalanced data.

Discussions

In this study, we have developed a self-adaptive multilayer ELM model combining with dynamic GAN to classify the limited and imbalanced dataset for the biomedical engineering application. Representative W-ELM, SMOTE-ELM, and H-ELM models are also implemented to solve the biomedical data classification in our work. In this

Zhang *et al. BioMed Eng OnLine*   (2018) 17:181

Page 16 of 21

Fig. 8 Comparison G-means of the PGM-ELM, H-ELM, SMOTE-ELM, and W-ELM with different numbers of hidden nodes. **a** Breast cancer dataset. **b** Liver patient dataset. **c** Diabetic retinopathy dataset. **d** Pima diabetes dataset
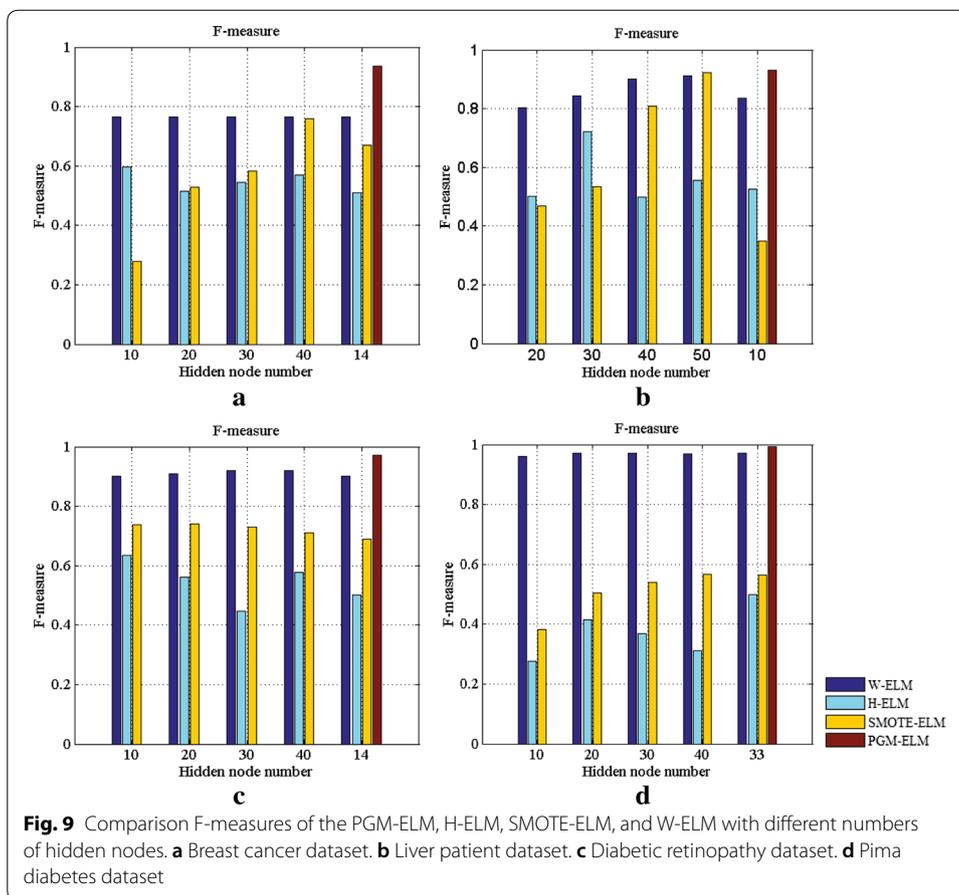
section, we discuss the classification performance, the statistical significance, and the computational time of these four models. At last, the advantages and limitations of the PGM-ELM method are summarized.

Evaluation of the classification performance

To further objectively verify the superiority of the proposed method, extensive evaluations are conducted on four datasets. We compute G-mean, AUC, and F-measure metrics of four methods. Table 4 tabulates the quantitative comparison results of different methods on four biomedical datasets in terms of G-mean, F-measure, and AUC.

From the AUC values in Table 4, we can clearly observe through the comparison and analysis, the proposed PGM-ELM has a much larger value than SMOTE-ELM and H-ELM, while a little higher than W-ELM for most of the test sets. The reason calls for PGM-ELM, the input of the model is changed from the original imbalanced data to a more balanced one by dynamic GAN. From the values of G-mean and F-measure, we also can find that our approach has a significant improvement against the others on four datasets. Especially, for Pima diabetes dataset, the value of F-measure nearly tends to 1. The hyper-parameter analytic expression of hidden layer helps

Fig. 9 Comparison F-measures of the PGM-ELM, H-ELM, SMOTE-ELM, and W-ELM with different numbers of hidden nodes. **a** Breast cancer dataset. **b** Liver patient dataset. **c** Diabetic retinopathy dataset. **d** Pima diabetes dataset

to achieve a better performance by providing more robust features extract from the balanced data. Meanwhile, an important observation is that fewer parameters need to be chosen adaptively in the training process. The whole performance of the algorithm is not only high but also stable.

The statistical significance testing

In the statistical hypothesis testing, the Friedman test and post-hoc Nemenyi test [38] are used to further analyze whether our method is statistically significant than other compared methods. Combining these two hypothesis testing methods are to compare the performances of various classification methods on multiple datasets. After Friedman hypothesis testing, the null hypothesis (i.e. the performances of all four methods are equivalent) is rejected at $\alpha = 0.05$ since the p-values for G-mean, AUC, and F-measure are 0.0256, 0.0129, and 0.0112. This result indicates that our method has a significant difference than the others.

Then, the post-hoc Nemenyi test is adopted to observe the differences among the four models. A critical difference (CD) of 2.345 is computed at $p = 0.05$. For G-mean metric, the average ranks of PGM-ELM, W-ELM, SMOTE-ELM, and H-ELM are 1, 2.75, 2.5, and 3.75, respectively. From these rank differences among PGM-ELM, W-ELM and SMOTE-ELM, they are lower than the CD value. So PGM-ELM has no statistically

Zhang *et al. BioMed Eng OnLine*   (2018) 17:181

Page 18 of 21

Table 4 Performance comparison results of testing on different datasets

Biomedical datasets	Methods	G-mean	AUC	F-measure
Breast cancer	W-ELM	0.5679	0.8093	0.7658
	SMOTE-ELM	0.7816	0.7981	0.7584
	H-ELM	0.5835	0.6584	0.5967
	PGM-ELM	0.9212	0.9013	0.9354
Liver patient	W-ELM	0.7827	0.7439	0.9127
	SMOTE-ELM	0.6379	0.5198	0.9218
	H-ELM	0.5980	0.6919	0.7226
	PGM-ELM	0.8016	0.8581	0.9304
Diabetic retinopathy	W-ELM	0.6555	0.7849	0.9207
	SMOTE-ELM	0.7554	0.7220	0.7404
	H-ELM	0.6112	0.7493	0.6346
	PGM-ELM	0.8778	0.8619	0.9715
Pima diabetes	W-ELM	0.9360	0.9151	0.9724
	SMOTE-ELM	0.6277	0.8792	0.5655
	H-ELM	0.5041	0.8580	0.5000
	PGM-ELM	0.9657	0.9324	0.9922

significant difference in terms of G-mean, despite our method wining on most of the datasets. While PGM-ELM is statistically different from H-ELM. This explains why our method is suitable for the imbalanced data classification problem.

Comparison of the computational time

The classification efficiency of the W-ELM, SMOTE-ELM, H-ELM, and PGM-ELM algorithms are compared, which is presented in Fig. 10. By analyzing the computational times, we can find that the training time of PGM-ELM is slightly higher than that of W-ELM. And it is obviously lower than those of H-ELM and SMOTE-ELM. The reason for this is that a lot of time is costed for the sample generation process using GAN. W-ELM has a computational advantage owing to its fast weighting process. Nevertheless, if the imbalanced ratio is extremely low, the W-ELM usually leads to an excessive learning. It is difficult to control the optimal parameter. Anyway, the computational time of PGM-ELM method on each dataset is below 2s. In a word, the proposed method can quickly and accurately alleviate the class-imbalanced problem. These findings demonstrate that the algorithm presented here has a potential significance for the clinical practice.

Based on the above analysis, we can summarize the advantages and limitations of the proposed method. Our method attempts to tackle the classification of limited and imbalanced biomedical dataset. In the proposed method, dynamic GAN takes the data distribution into account for producing authentic minority class samples. Furthermore, the parameters of hidden layer are adaptively chosen according to the change of the imbalanced ratio. It avoids the drawback of manual parameter adjustment. Under imbalanced scenarios, different types of biomedical data (e.g. protein dataset, gene expression data, and medical images) have similar properties, such as high-dimensional and small samples. For example, image data can be converted to numerical attributes by using some segmentation methods [39, 40]. In this way, the proposed method can effectively address

Fig. 10 Comparison result of the running time
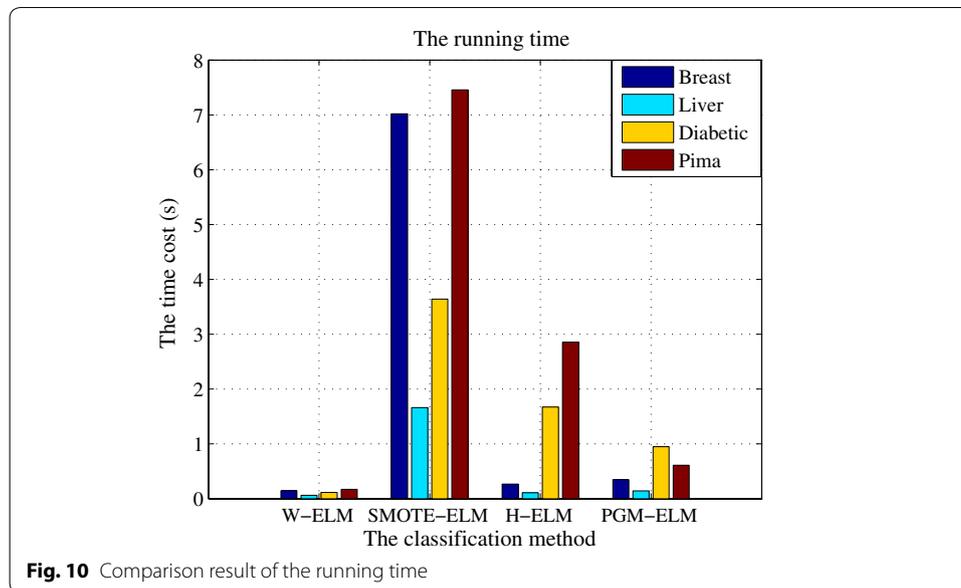
the class-imbalanced classification problem with respect to different biomedical datasets. Despite this goodness, the proposed method has also two potential weakness. One limitation is that the time cost of our method is slightly higher than W-ELM, mainly due to extra cost of the samples generation process. The other is, if a large of missing values occur in biomedical dataset, GAN model will generate some ineffective samples. The proposed model also will suffer from worse classification performance. In future works, these two limitations will be addressed.

Conclusions

In this paper, a self-adaptive multilayer ELM with dynamic GAN has been proposed for the imbalanced biomedical classification. Different from traditional deep network, self-adaptive multilayer ELM gives the analytic expression for numbers of layer and hidden node according to the changes of the imbalanced ratio and sample distribution. This is helpful for avoiding the hyper-parameter sensitivity. Furthermore, principal components of the original features are extracted by PCA, thus removing irrelevant features and obtaining more effective feature set. Then, dynamic GAN generates the real-looking samples to balance the class distribution. It fully considers the sample distribution and reduces overfitting. The proposed method has been evaluated on four real-world biomedical datasets. Qualitative and quantitative results show that the proposed method is quite promising than other representative methods in terms of ROC curve, AUC, G-mean, and F-measure metrics. The generality and capability of the proposed model are further confirmed under the condition of small sample and high-dimensional feature. We will make efforts to provide multi-class classification model for multiclass imbalanced classification problem in our future works.

Zhang *et al. BioMed Eng OnLine*   (2018) 17:181

Page 20 of 21

Publisher's Note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References
1. Kooi T, Litjens G, van Ginneken B, Gubern-Merida A, Sanchez CI, Mann R. Large scale deep learning for computer aided detection of mammographic lesions. Med Image Anal. 2017;35:303–12.
2. Krawczyk B. Learning from imbalanced data: open challenges and future directions. Prog Artif Intell. 2016;5(4):221–32.
3. De Bruijne M. Machine learning approaches in medical image analysis: from detection to diagnosis. Med Image Anal. 2016;33:94–7.
4. Soleymani R, Granger E, Fumera G. Progressive boosting for class imbalance and its application to face re-identification. Expert Syst Appl. 2018;101:271–91.
5. Zhu M, Xia J, Jin XQ. Class weights random forest algorithm for processing class imbalanced medical data. IEEE Access. 2018;6:4641–52.
6. Guo HX, Li YJ, Shang J. Learning from class-imbalanced data: aeview of methods and applications. Expert Syst Appl. 2017;73:220–39.
7. Lee W, Jun CH, Lee JS. Instance categorization by support vector machines to adjust weights in AdaBoost for imbalanced data classification. Inf Sci. 2017;381:92–103.
8. Zhang Y, Xin Y, Li Q. Empirical study of seven data mining algorithms on different characteristics of datasets for biomedical classification applications. Biomed Eng Online. 2017;16(1):125.
9. Wu Z, Lin W, Ji Y. An integrated ensemble learning model for imbalanced fault diagnostics and prognostics. IEEE Access. 2018;6:8394–402.
10. Cao P, Yang J, Li W. Ensemble-based hybrid probabilistic sampling for imbalanced data learning in lung nodule CAD. Comput Med Imaging Graphics. 2014;38(3):137–50.
11. Jiang J, Liu X, Zhang K. Automatic diagnosis of imbalanced ophthalmic images using a cost-sensitive deep convolutional neural network. Biomed Eng Online. 2017;16(1):132.
12. Zhang Y, Liu B, Cai J, Zhang S. Ensemble weighted extreme learning machine for imbalanced data classification based on differential evolution. Neural Comput Appl. 2016;28(1):1–9.
13. Huang GB, Wang DH, Lan Y. Extreme learning machines: a survey. Int J Mach Learn Cybernet. 2011;2(2):107–22.
14. Lin SJ, Chang C, Hsu MF. Multiple extreme learning machines for a two-class imbalance corporate life cycle prediction. Knowl Based Syst. 2013;39:214–23.

Zhang *et al. BioMed Eng OnLine*   (2018) 17:181

Page 21 of 21

15. Kasun CLL, Zhou H, Huang GB, Vong CM. Representational learning with extreme learning machine for Big Data. IEEE Intell Syst. 2013;28(6):31–4.
16. Song G, Dai Q. A novel double deep ELMs ensemble system for time series forecasting. Knowl Based Syst. 2017;134:31–49.
17. Zong W, Huang GB, Chen Y. Weighted extreme learning machine for imbalance learning. Neurocomputing. 2013;101:229–42.
18. Lopez V, del Rio S, Benitez JM, Herrera F. Cost-sensitive linguistic fuzzy rule based classification systems under the MapReduce framework for imbalanced big data. Fuzzy Sets Syst. 2015;258:5–38.
19. Huang C, Li Y, Change Loy C, et al. Learning deep representation forimbalanced classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016. p. 5375–84.
20. Tang J, Deng C, Huang GB. Extreme learning machine for multilayer perceptron. IEEE Trans Neural Netw Learning Syst. 2016;27(4):809–21.
21. Castellanos FJ, Valero-Mas JJ, Calvo-Zaragoza J. Oversampling imbalanced data in the string space. Pattern Recogn Lett. 2018;103:32–8.
22. Zhang L, Zhang D. Evolutionary cost-sensitive extreme learning machine. IEEE Trans Neural Netw Learning Syst. 2017;28(12):3045–60.
23. Oh S, Lee MS, Zhang BT. Ensemble learning with active example selection for imbalanced biomedical data classification. IEEE/ACM Trans Comput Biol Bioinform. 2011;8(2):316–25.
24. Galar M, Fernandez A, Barrenechea E. EUSBoost: enhancing ensembles for highly imbalanced datasets by evolutionary undersampling. Pattern Recogn. 2013;46(12):3460–71.
25. Chawla NV, Bowyer KW, Hall LO. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res. 2002;16:321–57.
26. Yu H, Ni J. An improved ensemble learning method for classifying high-dimensional and imbalanced biomedicine data. IEEE/ACM Trans Comput Biol Bioinform. 2014;11(4):657–66.
27. Krawczyk B, Galar M, Jelen L, Herrera F. Evolutionary undersampling boosting for imbalanced classification of breast cancer malignancy. Appl Soft Comput. 2016;38:714–26.
28. Gong CL, Gu LX. A novel SMOTE-based classification approach to online data imbalance problem. Math Problems Eng. 2016;2016:1–14.
29. Huang G, Huang GB, Song S. Trends in extreme learning machines: a review. Neural Netw. 2015;61:32–48.
30. Tissera MD, McDonnell MD. Deep extreme learning machines: supervised autoencoding architecture for classification. Neurocomputing. 2016;174:42–9.
31. Kamilov US, Mansour H, Wohlberg B. A plug-and-play priors approach for solving nonlinear imaging inverse problems. IEEE Signal Process Lett. 2017;24(12):1872–6.
32. Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets. In: Advances in neural information processing systems. 2014. p. 2672–80.
33. Calimeri F, Marzullo A, Stamile C, et al. Biomedical data augmentation using generative adversarial neural networks. International conference on artificial neural networks. Berlin: Springer; 2017. p. 626–34.
34. Ait-Sahalia Y, Xiu D. Using principal component analysis to estimate a high dimensional factor model with high-frequency data. J Econometr. 2017;201(2):384–99.
35. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml/datasets.
36. Kraiem MS, Moreno MN. Effectiveness of basic and advanced sampling strategies on the classification of imbalanced data. A comparative study using classical and novel metrics. International conference on hybrid artificial intelligence systems. Cham: Springer; 2017. p. 233–45.
37. Xu Q, Huang G, Yuan Y, et al. An empirical study on evaluation metrics of generative adversarial networks. arXiv preprint arXiv:1806.07755. 2018.
38. Mirza B, Lin Z, Toh KA. Weighted online sequential extreme learning machine for class imbalance learning. Neural Process Lett. 2013;38(3):465–86.
39. Wang JK, Cheng YZ, Guo CY. Shape-intensity prior level set: combining probabilistic atlas and probability map constrains for automatic liver segmentation from abdominal CT images. Int J Comput Assisted Radiol Surg. 2016;11(5):817–26.
40. Shi CF, Cheng YZ, Wang JK. Low-rank and sparse decomposition based shape model and probabilistic atlas for automatic pathological organ segmentation. Med Image Anal. 2017;38(30):30–49.