Neurocomputing 138 (2014) 248–259

Contents lists available at ScienceDirect

Neurocomputing

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

PDFOS: PDF estimation based over-sampling for imbalanced two-class problems ☆

Ming Gao [a], Xia Hong [a], Sheng Chen [b,c,*], Chris J. Harris [b], Emad Khalaf [c]

[a] School of Systems Engineering, University of Reading, Reading RG6 6AY, UK
[b] Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ, UK
[c] Electrical and Computer Engineering Department, Faculty of Engineering, King Abdulaziz University, Jeddah 21589, Saudi Arabia

ARTICLE INFO

Article history:
Received 14 October 2013
Received in revised form
8 January 2014
Accepted 1 February 2014
Communicated by K. Li
Available online 16 February 2014

Keywords:
Imbalanced classification
Probability density function based over-sampling
Radial basis function classifier
Orthogonal forward selection
Particle swarm optimisation

ABSTRACT

This contribution proposes a novel probability density function (PDF) estimation based over-sampling (PDFOS) approach for two-class imbalanced classification problems. The classical Parzen-window kernel function is adopted to estimate the PDF of the positive class. Then according to the estimated PDF, synthetic instances are generated as the additional training data. The essential concept is to re-balance the class distribution of the original imbalanced data set under the principle that synthetic data sample follows the same statistical properties. Based on the over-sampled training data, the radial basis function (RBF) classifier is constructed by applying the orthogonal forward selection procedure, in which the classifier's structure and the parameters of RBF kernels are determined using a particle swarm optimisation algorithm based on the criterion of minimising the leave-one-out misclassification rate. The effectiveness of the proposed PDFOS approach is demonstrated by the empirical study on several imbalanced data sets.

1. Introduction

In a typical two-class imbalanced classification problem, the instances in one class outnumber the instances of the other class. The majority class is usually referred to as the negative class, while the minority one as the positive class. Machine learning based on imbalanced data, whereby the imbalance in class distribution renders the positive class instances to be submerged in the negative class, is of great interest. The problem typically arises in life threatening or safety critical applications, such as mammography for breast cancer detection [1], mobile phone fraud detection [2], and detection of oil spills in satellite radar images [3]. In addition, many engineering applications, including information retrieval and filtering [4], direct marketing [5], risk management [6], and so on, are inherently imbalanced. In these applications, the primary objectives are often to target and explore the rare cases/classes which are less probable yet highly risky/costly. The imbalance between two classes is problematic for many standard classification algorithms [7–11]. The performances of these algorithms deteriorate as class imbalance degree increases, or as the data samples of positive class become sparser [9]. For example, the kernel-based methods, which are regarded as robust classifiers [12], construct a decision hyperplane separating two classes. Without special countermeasure, the resultant hyperplane will tend to be placed in favour of the classification performance for the negative class, but the classification performance for the target class becomes unsatisfactory. There exist a large amount of works to deal with the imbalanced learning, and the reader is referred to the excellent survey paper [12] for more information. Typical techniques of tackling the imbalanced problem can be categorised into two categories: resampling methods, also known as external methods, and imbalanced learning algorithms, often referred to as internal methods.

Imbalanced learning algorithms are obtained by modifying some existing learning algorithms internally so that they can deal with imbalanced problems effectively, without 'artificially' altering or re-balancing the original imbalanced data set. For example, the kernel classifier construction or model selection procedure can be modified, in order to cope with the imbalanced distribution during the

☆This work was supported by UK EPSRC.

* Corresponding author at: Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ, UK.
E-mail addresses: ming.gao@pgr.reading.ac.uk (M. Gao), x.hong@reading.ac.uk (X. Hong), sqc@ecs.soton.ac.uk (S. Chen), cjh@ecs.soton.ac.uk (C.J. Harris), ekhalaf@kau.edu.sa (E. Khalaf).

http://dx.doi.org/10.1016/j.neucom.2014.02.006
0925-2312

classifier construction process [11,13]. A well-known radial basis function (RBF) modelling approach is the two-stage procedure [14], in which the RBF centres are first determined using the κ-means clustering [15] and the RBF weights are then obtained using the least squares estimate (LSE). To cope with imbalanced data sets, a natural extension of [14] is to modify the latter stage as the weighted LSE (WLSE), where the same weighted cost function of [13] is used. This κ-means+WLSE algorithm provides a viable technique for this category of imbalanced learning.

The resampling methods are external as they operate on original imbalanced data set, aiming to provide a re-balanced input to train a conventional classifier. One scheme is to assign different weights to the samples of the data set in accordance with their misclassification costs [16,17]. There have been a large number of studies focusing on this simple yet effective methodology to combine with the conventional classifiers for the rebalanced data set. Clearly the ultimate classification performance will be dependent on the adopted resampling strategy as well as the choice of classifier. In terms of classifier development, recently, the particle swarm optimisation (PSO) algorithm [18] has been applied to minimise the leave-one-out (LOO) misclassification rate in the orthogonal forward selection (OFS) construction of tunable RBF classifier [19,20]. PSO [18] is an efficient population-based stochastic optimisation technique inspired by social behaviour of bird flocks or fish schools, and it has been successfully applied to wide-ranging optimisation applications [21–28]. Owing to the efficiency of PSO, the tunable RBF modelling approach advocated in [19,20] offers significant advantages over many existing kernel or RBF classifier construction algorithms, in terms of better generalisation performance and smaller classifier size as well as lower complexity in learning process. With regarding to the choice of resampling strategy, we note that various resampling methods can be divided into the two basic categories, according to whether they re-balance the class distribution by under-sampling or over-sampling.

Random under-sampling is the non-heuristic method aiming to re-balance class distribution by randomly eliminating instances in the negative class [29]. Despite its simplicity, random under-sampling is considered to be one of the most effective re-sampling methods [30]. A major drawback of this technique is that it may discard data potentially important for building the classifier. Thus, many studies focus on heuristic selection techniques [31–40] to eliminate negative class instances. The method presented in [35] selectively under-samples the negative class, while keeping all the samples of the positive class. Specifically, the negative class instances are divided into the four categories: class-label noise instances A that overlap the positive class decision region; borderline instances B that are unreliable and can easily cause misclassification; redundant instances C that do not harm classification accuracy but increase classification costs; and safe instances D that are worthy of being kept for classification process. The categories A and B are detected by the use of Tomek links concept [41], as the instances complying with Tomek links are either borderline or noisy samples. Also a SHRINK [3] system attributes the overlapping regions of both the negative and positive classes as the positive class, and searches for the best positive-class region. Alternatively, Wilson's edited nearest neighbour (ENN) rule [42] is introduced to eliminate noisy instances in the negative class [43]. The ENN rule removes any instance whose class label differs from the class label of at least two of its three nearest neighbours, and a neighbourhood cleaning rule (NCL) [44] modifies the ENN by removing any negative-class instance whose class label differs from that of its 3-nearest neighbours. In order to find a consistent subset, the categories C and D are identified by involving Hart's condensed nearest neighbor (CNN) rule [45].

Under-sampling tends to be an ideal option when the imbalance degree is not very severe. However, as pointed out in [46], the use of over-sampling is necessary when the imbalance degree is high. Random over-sampling is the non-heuristic method aiming to re-balance class distribution by randomly replicating instances in the positive class. Studies [9,29] highlight that this method is simple yet very competitive to more complex over-sampling methods. However, over-fitting is a recognised serious problem for random over-sampling, because the exact copies of the instances in the positive class are made. In the study of imbalanced data sets in marketing analysis, over-sampling the positive instances with replacement is applied to match the number of negative instances [5]. The study [47] proposed a synthetic minority over-sampling technique (SMOTE), which aims to enhance the significance of some specific regions in the feature space by over-sampling the positive class. Instead of mere data oriented duplicating, SMOTE generates synthetic instances in the feature space formed by random samples along the line linking the instance and its k-nearest neighbours (k-NN). Although SMOTE is well acknowledged by the academic community, it still has some drawbacks, including over generalisation and large variance [48]. Thus, SMOTEBoost [49], borderline-SMOTE [50] and adaptive synthetic sampling (ADASYN) [51] were proposed to alleviate its limitations. Despite the empirical evidences that the foregoing methods have been effective in improving the classification performance for the target class, the reason behind the success of the oversampling approaches, such as SMOTE, is not fully understood. In fact, there exist little theoretical studies to justify most of the oversampling methods. This raises the fundamental questions as how to measure the quality of synthetic instances and why these can be used as training samples.

Against this background, we propose a novel oversampling approach based on the kernel density estimation from positive-class data samples. The estimation of the probability density function (PDF) from observed data samples is a fundamental problem in many machine learning and pattern recognition applications [52–54]. The Parzen window (PW) estimate is a simple yet remarkably accurate nonparametric density estimation technique [53–55]. According to the estimated PDF, synthetic instances are generated as the additional training data. The RBF classifier proposed in [20] is then applied to the rebalanced data set, to complete the classification process. In the generic density estimation application, the PW estimator has a well-known drawback, owing to the fact that it employs the full data sample set in defining the density estimate for a subsequent observation and, therefore its computational cost for testing directly scales with the sample size. Note that we apply the PW estimator for estimating the distribution of the minority class, which by nature consists of a small number of data samples. Therefore the potential disadvantage of the PW estimate does not exist in our application. In fact, if the sample size of the positive class is large, there will be no need to oversample it by introducing artificial samples, and the imbalance of the two classes can be better dealt with by removing some samples from the majority class, in other words, by undersampling the negative class.

The significance of our PDFOS+PSO-OFS method is twofold. Firstly, in comparison to the existing oversampling techniques, our PDFOS based oversampling approach has much stronger theoretical justification. This is because an ideal or “optimal” oversampling technique should generate synthetic data according to the same probability distribution which produces the observed positive-class data samples. By using the estimated PDF of the minority class to generate synthetic samples, the generated synthetic data follow the same statistical properties as the observed positive-class data samples. Therefore, the proposed PDFOS technique generates synthetic instances with better quality

than the existing oversampling methods. Secondly, the PSO-OFS based RBF classifier, with its structure and parameters determined using a PSO algorithm based on minimising the LOO misclassification rate in the efficient OFS procedure, has been shown to outperform many existing classifier construction algorithms [20].

To evaluate the proposed PDFOS+PSO-OFS method, an extensive experimental study is carried out, in which three benchmarks are used for the comparison purpose. The first benchmark uses the same PSO-OFS based RBF classifier applied to the SMOTE oversampling data set [56], denoted by the SMOTE+PSO-OFS, which offers a very competitive performance to many existing methods for combating two-class imbalanced classification problems, as demonstrated in [56]. The second benchmark is the algorithm advocated in [13], denoted by the LOO-AUC+OFS, which is a state-of-the-art representative of the internal approach for dealing with imbalanced problems. The third benchmark, the κ-means+WLSE algorithm, as discussed previously, is also a typical imbalanced learning approach. The experimental results obtained demonstrate that the proposed PDFOS+PSO-OF method is competitive to these existing state-of-the-arts methods for two-class imbalanced problems.

The rest of the paper is organised as follows. Section 2 presents the proposed PDF estimation based over-sampling (PDFOS) algorithm. Section 3 describes our chosen classifier, the PSO aided tunable RBF model for two-class classification constructed by minimising the LOO misclassification rate based on the OFS procedure. The effectiveness of our approach is demonstrated by numerical examples in Section 4, and our conclusions are given in Section 5.

2. PDF estimation based over-sampling (PDFOS)

Consider the two-class data set given as

$$D_N=\{\mathbf{x}_k,y_k\}_{k=1}^{N}=D_{N_+}\bigcup D_{N_-}=\{\mathbf{x}_i,y_i=+1\}_{i=1}^{N_+}\bigcup\{\mathbf{x}_l,y_l=-1\}_{l=1}^{N_-} \tag{1}$$

where $y_k\in\{\pm 1\}$ denotes the class label for the feature vector $\mathbf{x}_k\in\mathbb{R}^m$, $N=N_++N_-$ is the total number of instances, while there are N_+ positive-class instances and N_- negative-class instances, respectively. The underlying classification problem is imbalanced, and this manifests as $N_+\ll N_-$. The sample $\mathbf{x}_k$ complies with an unknown PDF, with the assumption that instances are generated independently and identically from the unknown underlying probability distribution.

2.1. Kernel density estimation for positive class

Denote the unknown PDF that generates the positive-class sample set D_{N_+} by $p(\mathbf{x})$. A general kernel-based density estimator $\hat{p}(\mathbf{x})$ for $p(\mathbf{x})$ based on $D_{N_+}=\{\mathbf{x}_i,y_i=+1\}_{i=1}^{N_+}$ is defined by

$$\hat{p}(\mathbf{x})=\frac{1}{N_+}\sum_{i=1}^{N_+}\Phi_\sigma(\mathbf{x}-\mathbf{x}_i) \tag{2}$$

where σ is the window width or smoothing parameter, and $\Phi_\sigma(\mathbf{x}-\mathbf{x}_i)$ is the scaled kernel function which calculates the distance from $\mathbf{x}$ to the training instance $\mathbf{x}_i$, scaled by σ. The normal kernel scaled by a single σ is often chosen as kernel function [57–59]

$$\Phi_\sigma(\mathbf{x}-\mathbf{x}_i)=\frac{\sigma^{-m}}{(2\pi)^{m/2}}e^{-(1/2\sigma^2)(\mathbf{x}-\mathbf{x}_i)^{\mathrm{T}}(\mathbf{x}-\mathbf{x}_i)} \tag{3}$$

Using a single smoothing parameter σ in the above kernel implies that all the dimensions of the feature space are uncorrelated and they have the same spread. To obtain a better estimate of the density distribution for the positive class, the following kernel-based PDF estimate involving the covariance matrix $\mathbf{S}$ of the positive class is adopted in this paper

$$\hat{p}(\mathbf{x})=\frac{(\det\mathbf{S})^{-1/2}}{N_+}\sum_{i=1}^{N_+}\Phi_\sigma(\mathbf{S}^{-1/2}(\mathbf{x}-\mathbf{x}_i)) \tag{4}$$

where

$$\Phi_\sigma(\mathbf{S}^{-1/2}(\mathbf{x}-\mathbf{x}_i))=\frac{\sigma^{-m}}{(2\pi)^{m/2}}e^{-(1/2\sigma^2)(\mathbf{x}-\mathbf{x}_i)^{\mathrm{T}}\mathbf{S}^{-1}(\mathbf{x}-\mathbf{x}_i)} \tag{5}$$

in which $\mathbf{S}$ is the unbiased estimate of the positive class covariance given by

$$\mathbf{S}=\frac{1}{N_+-1}\sum_{i=1}^{N_+}(\mathbf{x}_i-\overline{\mathbf{x}})(\mathbf{x}_i-\overline{\mathbf{x}})^{\mathrm{T}} \tag{6}$$

with $\overline{\mathbf{x}}=(1/N_+)\sum_{i=1}^{N_+}\mathbf{x}_i$ being the mean vector of the positive class. The idea to include $\mathbf{S}$ in (5) is to cope with the situations where the coordinates of the feature space are correlated and the spreads of the coordinates are different. In such situations, if an equal spread parameter σ is applied to all the coordinates as in (3), the estimated PDF could not adequately represent the true distribution of the data set [60,61].

The value of σ in the density estimator $\hat{p}(\mathbf{x})$ of (4) needs to be determined. The most tractable global measure of the discrepancy of $\hat{p}(\mathbf{x})$ from the true density $p(\mathbf{x})$ is the mean integrated square error (MISE) calculated using D_{N_+}, based on which σ can be found by minimising the least-squares cross-validation score function $M(\sigma)$ [52], defined by

$$M(\sigma)=N_+^{-2}\sum_i\sum_j\Phi_\sigma^*(\mathbf{S}^{-1/2}(\mathbf{x}_j-\mathbf{x}_i))+2N_+^{-1}\Phi_\sigma(\mathbf{0}) \tag{7}$$

where

$$\Phi_\sigma^*(\mathbf{S}^{-1/2}(\mathbf{x}_j-\mathbf{x}_i))\approx\Phi_\sigma^{(2)}(\mathbf{S}^{-1/2}(\mathbf{x}_j-\mathbf{x}_i))-2\Phi_\sigma(\mathbf{S}^{-1/2}(\mathbf{x}_j-\mathbf{x}_i)) \tag{8}$$

in which $\Phi_\sigma^{(2)}(\mathbf{S}^{-1/2}(\mathbf{x}_j-\mathbf{x}_i))$ is given by

$$\Phi_\sigma^{(2)}(\mathbf{S}^{-1/2}(\mathbf{x}_j-\mathbf{x}_i))=\frac{(\sqrt{2}\sigma)^{-m}}{(2\pi)^{m/2}}e^{-(1/4\sigma^2)(\mathbf{x}_j-\mathbf{x}_i)^{\mathrm{T}}\mathbf{S}^{-1}(\mathbf{x}_j-\mathbf{x}_i)} \tag{9}$$

The optimal σ can be found by a grid search. The computational cost of this process is $O(N_+^2)$, scaled by the number of the grid points set in the grid search, which is low as N_+ is by nature a small number.

2.2. Over-sampling based on kernel density estimator

Over-sampling on the positive class is performed by drawing data samples according to the PDF estimate $\hat{p}(\mathbf{x})$ in (4), estimated based on the given training data set D_{N_+}. Each synthetic sample can be generated by using the following two steps.

1. Based on the discrete uniform distribution, randomly draw a data sample, $\mathbf{x}_o$, from the positive-class data set.
2. Generate a synthetic data sample, $\mathbf{x}_n$, using the Gaussian distribution with $\mathbf{x}_o$ as the center or mean vector and $\sigma^2\mathbf{S}$ as the covariance matrix.

In Step 2, the synthetic sample $\mathbf{x}_n$ can be generated according to

$$\mathbf{x}_n=\mathbf{x}_o+\sigma\mathbf{R}\cdot\mathbf{randn}() \tag{10}$$

where $\mathbf{R}$ is the upper triangular matrix with strictly positive diagonal entries that is the Cholesky decomposition of $\mathbf{S}$, namely, $\mathbf{R}^{\mathrm{T}}\mathbf{R}=\mathbf{S}$, and $\mathbf{randn}()$ is the m-dimensional pseudorandom vector drawn from the zero-mean normal distribution with the m-dimensional identity matrix $\mathbf{I}_m$ as its covariance matrix. In order

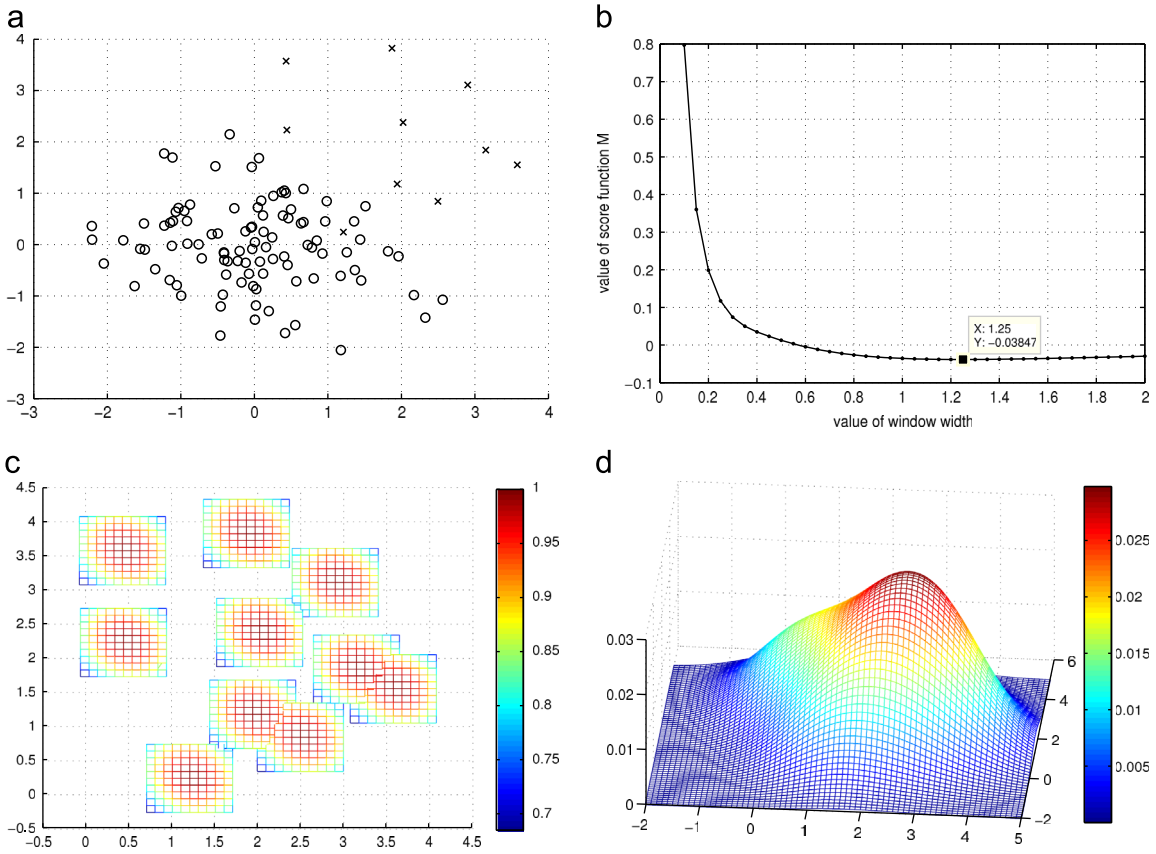

Fig. 1. Illustration of PDF estimation for the synthetic imbalanced data set: (a) the imbalanced synthetic data set with x denoting positive-class instance and ◦ denoting negative-class instance, (b) grid search of σ with step 0.05, (c) the PDF kernel of each instance, and (d) the estimated density distribution of the positive class.

to generate the required amount of synthetic samples specified by the oversampling rate r, which is defined as the ratio of the number of generated instances to that of original positive-class instances, the above procedure is repeated $r \cdot N_+$ times.

A synthetic 2-dimensional imbalanced data set is generated to illustrate the PDF estimation and the over-sampling process. The negative class has 100 instances, with the mean vector $[0\ 0]^{\mathrm{T}}$ and the covariance matrix $\mathbf{I}_2$, while the positive class has 10 instances, with the mean vector $[2\ 2]^{\mathrm{T}}$ and the covariance matrix $\mathbf{I}_2$, as shown in Fig. 1(a). In Fig. 1(b), the minimum value of $M(\sigma)$ is found at $\sigma = 1.25$ by the grid search with step 0.05. In Fig. 1(c), the kernel function placed at each positive-class instance is constructed according to σ and $\mathbf{S}$, in which $\sigma^2\mathbf{S}$ controls the rotation of the kernel and its spread in each dimension. Note that, in this example, however, $\mathbf{S} \approx \mathbf{I}_2$. Fig. 1(d) presents the density estimate for the positive class, which is the mixture of all the density kernels in Fig. 1(c) with an equal weighting for each component.

The over-sampled data distributions for the synthetic imbalanced data set of Fig. 1(a), obtained by the proposed PDFOS and SMOTE methods at the over-sampling rate $r = 1000\%$, are depicted in Fig. 2(a) and (b), respectively, where the solid line $x+y-2=0$ in both Fig. 2(a) and (b) is the ideal decision boundary for this synthetic data set. In the original imbalanced data set, there are only 10 positive-class instances in comparison with 100 negative-class instances. As only a very limited number of instances represent the positive class in the feature space, especially in the crucial region where the decision boundary lies, the trained decision boundary based on the original imbalanced data set will be biased towards the positive class. Both the PDFOS and SMOTE methods increase the positive-class instances, particularly in the decision region, after the positive class has been over-sampled 10 times of its original size. However, it can be seen from Fig. 2(b) that the over-sampled positive-class data set is confined in the region defined by the original positive-class instances, due to the fact that the SMOTE generates the synthetic instances in the line linking the original instance to its k-NN neighbours [56]. As a result, increasing the oversampling rate r only leads to a higher density in this region only. By contrast, the over-sampled positive class generated by the proposed PDFOS expands along the direction of the ideal decision boundary, as can be clearly seen from Fig. 2(a).

3. Tunable RBF modelling for classification

After the positive class has been oversampled with a required oversampling rate r, a tunable RBF classifier can then be constructed based on the expanded or rebalanced training data set using the algorithm proposed in [19,20]. For the completeness, this PSO-OFS algorithm for constructing the tunable RBF classifier is briefly described. For notational simplicity, the oversampled two-class training data set is still denoted as $D_N = \{\mathbf{x}_k, y_k\}_{k=1}^{N}$, where the number of the total instances, N, is understood to have been increased appropriately. Specifically, the RBF classifier to be constructed based on the rebalanced training data set D_N takes the form

$$\hat{y}_k^{(M)} = \sum_{i=1}^{M} w_i g_i(\mathbf{x}_k) = \mathbf{g}_M^{\mathrm{T}}(k)\mathbf{w}_M$$

$$\tilde{y}_k^{(M)} = \mathrm{sgn}(\hat{y}_k^{(M)}) \tag{11}$$

where M is the number of RBF kernels, $\hat{y}_k^{(M)}$ is the output of the M-term classifier with the M kernels, $g_i(\bullet)$, for $1 \le i \le M$, $\mathbf{w}_M = [w_1\ w_2\ \cdots\ w_M]^{\mathrm{T}}$ is the weight vector and $\mathbf{g}_M^{\mathrm{T}}(k) = [g_1(\mathbf{x}_k)\ g_2(\mathbf{x}_k)$

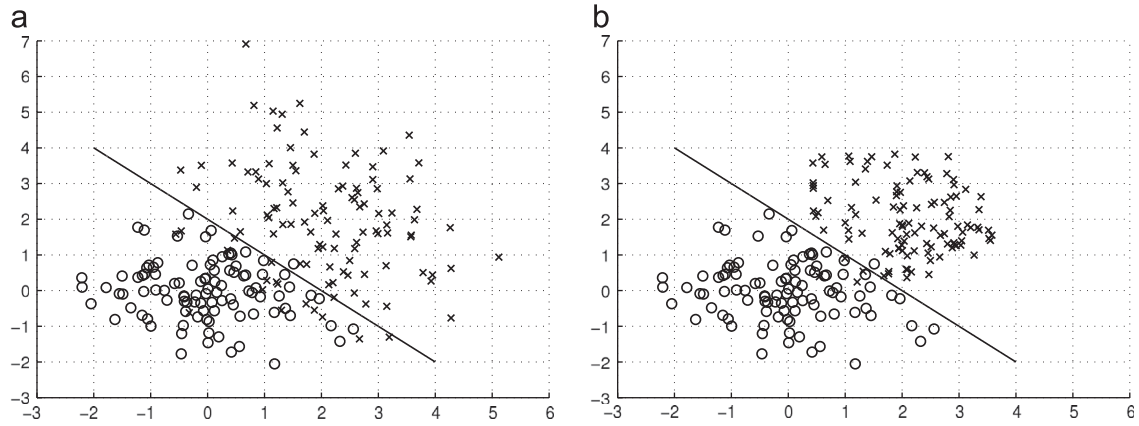

Fig. 2. Comparison between the over-sampled data distributions of the synthetic imbalanced data set by the PDFOS and SMOTE at the over-sampling rate $r = 1000\%$: (a) the PDFOS, and (b) the SMOTE.

$\cdots g_M(\mathbf{x}_k)]$, while $\tilde{y}_k^{(M)}$ denotes the estimated class label for $\mathbf{x}_k$ with

$$\mathrm{sgn}(y) = \begin{cases} -1, & y \le 0 \\ 1, & y > 0 \end{cases} \tag{12}$$

In this study, the Gaussian kernel function

$$g_i(\mathbf{x}) = e^{-(\mathbf{x}-\boldsymbol{\mu}_i)^{\mathrm{T}}\boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)} \tag{13}$$

is adopted, where $\boldsymbol{\mu}_i \in \mathbb{R}^m$ is the center vector of the ith RBF kernel and the ith kernel's covariance matrix takes a diagonal form of $\boldsymbol{\Sigma}_i = \mathrm{diag}\{\sigma_{i,1}^2, \sigma_{i,2}^2, \cdots, \sigma_{i,m}^2\}$. Hence, the position of each kernel, $\boldsymbol{\mu}_i$, and coverage of each kernel, $\boldsymbol{\Sigma}_i$, are both considered as the parameters to be determined in kernel modelling.

From (11), the RBF classifier over D_N can be written in the matrix form as

$$\mathbf{y} = \mathbf{G}_M\mathbf{w}_M + \boldsymbol{\varepsilon}^{(M)} \tag{14}$$

where $\boldsymbol{\varepsilon}^{(M)} = [\varepsilon_1^{(M)}\ \varepsilon_2^{(M)} \cdots \varepsilon_N^{(M)}]^{\mathrm{T}}$ is the error vector with the M-term modelling error $\varepsilon_k^{(M)} = y_k - \hat{y}_k^{(M)}$, $\mathbf{y} = [y_1\ y_2 \ldots y_N]^{\mathrm{T}}$ is the desired class label vector, and the kernel matrix $\mathbf{G}_M = [\mathbf{g}_1\ \mathbf{g}_2 \cdots \mathbf{g}_M]$ with $\mathbf{g}_l = [g_l(\mathbf{x}_1)\ g_l(\mathbf{x}_2) \cdots g_l(\mathbf{x}_N)]^{\mathrm{T}}$ for $1 \le l \le M$. Note that $\mathbf{g}_l$ is the lth column of $\mathbf{G}_M$ while $\mathbf{g}_M^{\mathrm{T}}(k)$ is the kth row of $\mathbf{G}_M$. Consider the orthogonal decomposition $\mathbf{G}_M = \mathbf{P}_M\mathbf{A}_M$, where

$$\mathbf{A}_M = \begin{bmatrix} 1 & a_{1,2} & \cdots & a_{1,M} \\ 0 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & a_{M-1,M} \\ 0 & \cdots & 0 & 1 \end{bmatrix} \tag{15}$$

$$\mathbf{P}_M = [\mathbf{p}_1\ \mathbf{p}_2 \cdots \mathbf{p}_M] \tag{16}$$

and the columns in (16) satisfy $\mathbf{p}_i^{\mathrm{T}}\mathbf{p}_j = 0$ for $i \ne j$. The RBF classifier (14) can alternatively be represented as

$$\mathbf{y} = \mathbf{P}_M\boldsymbol{\theta}_M + \boldsymbol{\varepsilon}^{(M)} \tag{17}$$

where $\boldsymbol{\theta}_M = [\theta_1\ \theta_2 \cdots \theta_M]^{\mathrm{T}}$ satisfies $\boldsymbol{\theta}_M = \mathbf{A}_\mathbf{M}\mathbf{w}_M$. The space spanned by the original model bases $\mathbf{g}_i$, $1 \le i \le M$, is identical to that spanned by $\mathbf{p}_i$, $1 \le i \le M$.

The OFS procedure constructs the RBF kernels one by one by minimising the LOO misclassification rate [19,20]. Specifically, at the nth stage of model construction, the nth RBF kernel, namely, $\mathbf{p}_n$ and θ_n, is determined. Define the LOO model output of the n-term RBF model constructed from the LOO data set $D_N \backslash (\mathbf{x}_k, y_k)$, calculated at $\mathbf{x}_k$, as $\hat{y}_k^{(n,-k)}$. Further define the associated LOO decision variable as

$$s_k^{(n,-k)} = \mathrm{sgn}(y_k)\hat{y}_k^{(n,-k)} = y_k\hat{y}_k^{(n,-k)} \tag{18}$$

Then the LOO misclassification rate is defined by [62]

$$J_{\mathrm{LOO}}^{(n)} = \frac{1}{N}\sum_{k=1}^{N} \mathcal{I}_d(s_k^{(n,-k)}) \tag{19}$$

in which the indicator function $\mathcal{I}_d(s)$ is given by

$$\mathcal{I}_d(s) = \begin{cases} 1, & s \le 0 \\ 0, & s > 0 \end{cases} \tag{20}$$

By making use of Sherman–Morrison–Woodbury theorem [63] as well as the orthogonal property, the LOO decision variable can be efficiently calculated according to [19,20,62]

$$s_k^{(n,-k)} = \frac{\psi_k^{(n)}}{\eta_k^{(n)}} \tag{21}$$

in which $\psi_k^{(n)}$ and $\eta_k^{(n)}$ can be computed recursively by:

$$\psi_k^{(n)} = \psi_k^{(n-1)} + y_k\theta_n p_n(k) - \frac{p_n^2(k)}{\mathbf{p}_n^{\mathrm{T}}\mathbf{p}_n + \lambda} \tag{22}$$

$$\eta_k^{(n)} = \eta_k^{(n-1)} - \frac{p_n^2(k)}{\mathbf{p}_n^{\mathrm{T}}\mathbf{p}_n + \lambda} \tag{23}$$

where $p_n(k)$ is the kth element of $\mathbf{p}_n$ and $\lambda \ge 0$ is a small regularisation parameter.

To determine the nth RBF kernel, its center vector $\boldsymbol{\mu}_n$ and diagonal covariance matrix $\boldsymbol{\Sigma}_n$ can be found by minimising $J_{\mathrm{LOO}}^{(n)}$. The problem of determining the nth RBF kernel's parameters at the nth stage of the OFS procedure is therefore to solve the following optimisation problem

$$\{\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n\}_{\mathrm{opt}} = \arg\min_{\boldsymbol{\mu},\boldsymbol{\Sigma}} J_{\mathrm{LOO}}^{(n)}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{24}$$

The PSO algorithm used to solve this optimisation problem is summarised in Appendix. The construction of the RBF classifier automatically terminates at the size of M when $J_{\mathrm{LOO}}^{(M+1)} \ge J_{\mathrm{LOO}}^{(M)}$ [19,20,62]. The computational complexity of constructing an M-term RBF model tuned by the PSO algorithm can be shown to be $(M+1) \times L \times S \times O(N)$ [20], where L is the number of iterations imposed by the PSO algorithm while S is known as the population size. Note that $M \ll N$ is very small, and L and S are not large integers. Therefore, the PSO-OFS algorithm has a lower complexity than most existing state-of-the-arts kernel classifier construction algorithms.

4. Experimental results

The effectiveness of the PDFOS+PSO-OFS method was examined on the six data sets summarised in Table 1 in the order

Table 1
Summary of the properties of the data sets.

Data set	Attributes $m+1$	Positive N_+	Negative N_-	ID	n-fold CV	σ
Pima Diabetes	8	268	500	1.87	10	0.47 ± 0.03
Haberman's survival	3	81	225	2.78	3	0.52 ± 0.03
Glass(6)	9	29	185	6.38	3	0.42 ± 0.06
ADI	9	90	700	7.78	8	0.56 ± 0.07
Satimage(4)	36	626	5809	9.28	10	0.90 ± 0.00
Yeast(5)	8	44	1440	32.73	3	0.10 ± 0.00

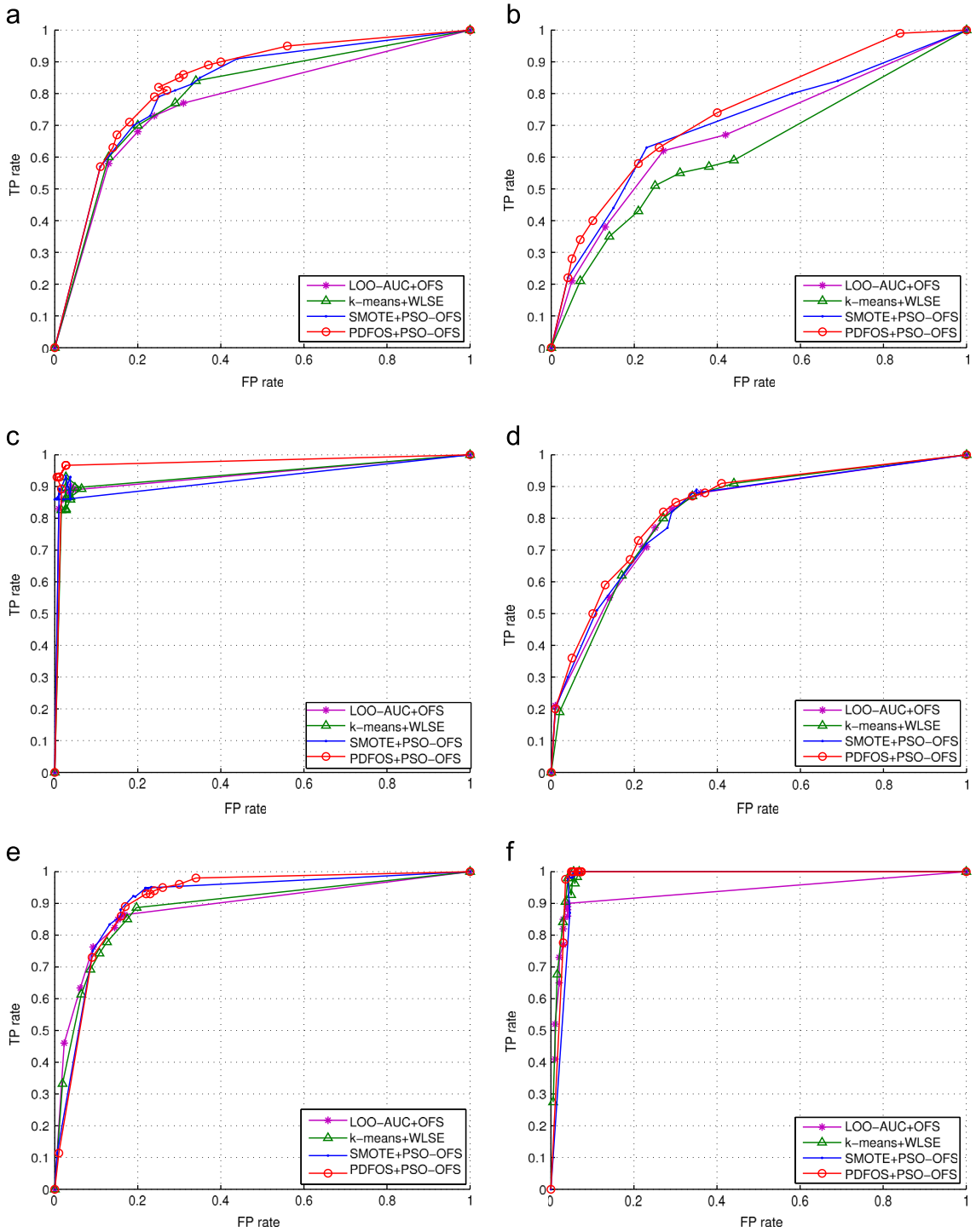

Fig. 3. Comparison of ROC curves on imbalanced data sets: (a) Pima Indians diabetes, (b) Haberman's survival, (c) glass, (d) ADI, (e) satimage, and (f) yeast.

Table 2
Comparison of mean and standard deviation of AUCs.

Data set	LOO-AUC+OFS	κ-means+WLSE	SMOTE+PSO-OFS	PDFOS+PSO-OFS
Pima Diabetes	0.77 ± 0.06	0.80 ± 0.06	0.82 ± 0.06	$\mathbf{0.84 \pm 0.06}$
Haberman's survival	0.68 ± 0.06	0.62 ± 0.06	0.71 ± 0.06	$\mathbf{0.74 \pm 0.06}$
Glass(6)	0.94 ± 0.05	0.93 ± 0.06	0.92 ± 0.06	$\mathbf{0.97 \pm 0.04}$
ADI	0.82 ± 0.03	0.82 ± 0.03	0.82 ± 0.03	$\mathbf{0.83 \pm 0.03}$
Satimage(4)	0.88 ± 0.03	0.88 ± 0.03	$\mathbf{0.91 \pm 0.03}$	$\mathbf{0.91 \pm 0.03}$
Yeast(5)	0.93 ± 0.04	$\mathbf{0.98 \pm 0.02}$	0.97 ± 0.03	$\mathbf{0.98 \pm 0.02}$

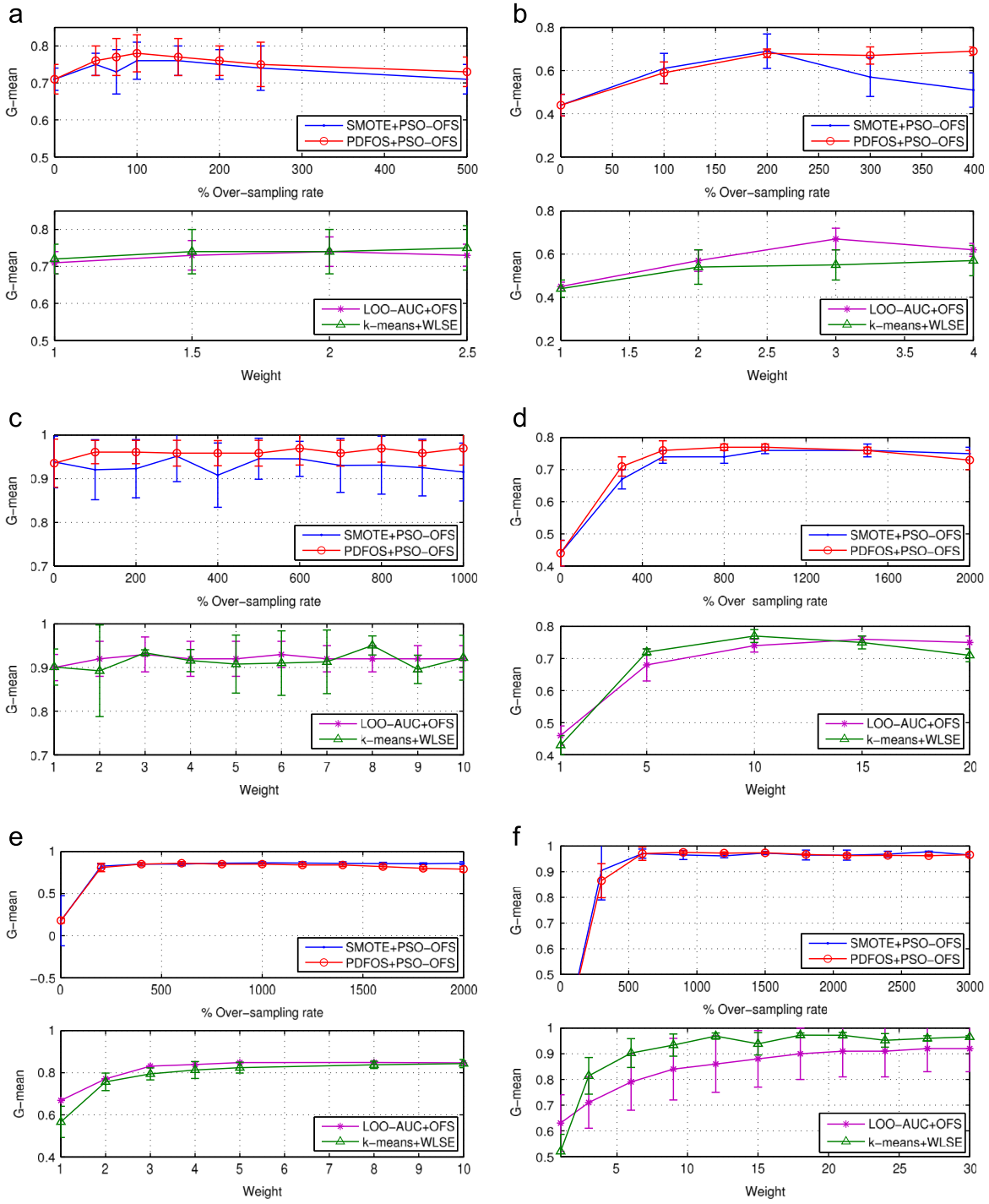

Fig. 4. Comparison of G-mean on imbalanced data sets: (a) Pima Indians diabetes, (b) Haberman's survival, (c) glass, (d) ADI, (e) satimage, and (f) yeast.

of the ascending imbalanced degree (ID), which is defined as $ID = N_- / N_+$. The austempered ductile iron (ADI) material data set came from the study [64], while the other five data sets were from the UCI machine learning repository [65]. Note that the data sets, Glass, Satimage and Yeast, are multiple-class data sets, which were turned into the two-class problems in this study by considering the class with the class label given in the brackets as the chosen positive class and designating the other classes altogether as the negative class. Considering the size of the data set and the sparsity of the positive class, different n-fold cross-validations (CVs) were performed on the different data sets. Each dimension of a feature vector $\mathbf{x}_k = [x_{k,1}\ x_{k,2} \cdots x_{k,m}]^T$ was normalised to the range $[0, 1]$ using the operation

$$\overline{x}_{k,i} = \frac{x_{k,i} - x_{\min,i}}{x_{\max,i} - x_{\min,i}}, \quad 1 \le k \le N, \ \ 1 \le i \le m \tag{25}$$

with

$$\begin{cases} x_{\min,i} = \min\limits_{1 \le k \le N} x_{k,i} \\ x_{\max,i} = \max\limits_{1 \le k \le N} x_{k,i} \end{cases} \tag{26}$$

The mean and standard deviation of σ, determined by the PW density estimator for the positive class of each data set, averaged over the n-fold CV are also reported in the last column of Table 1.

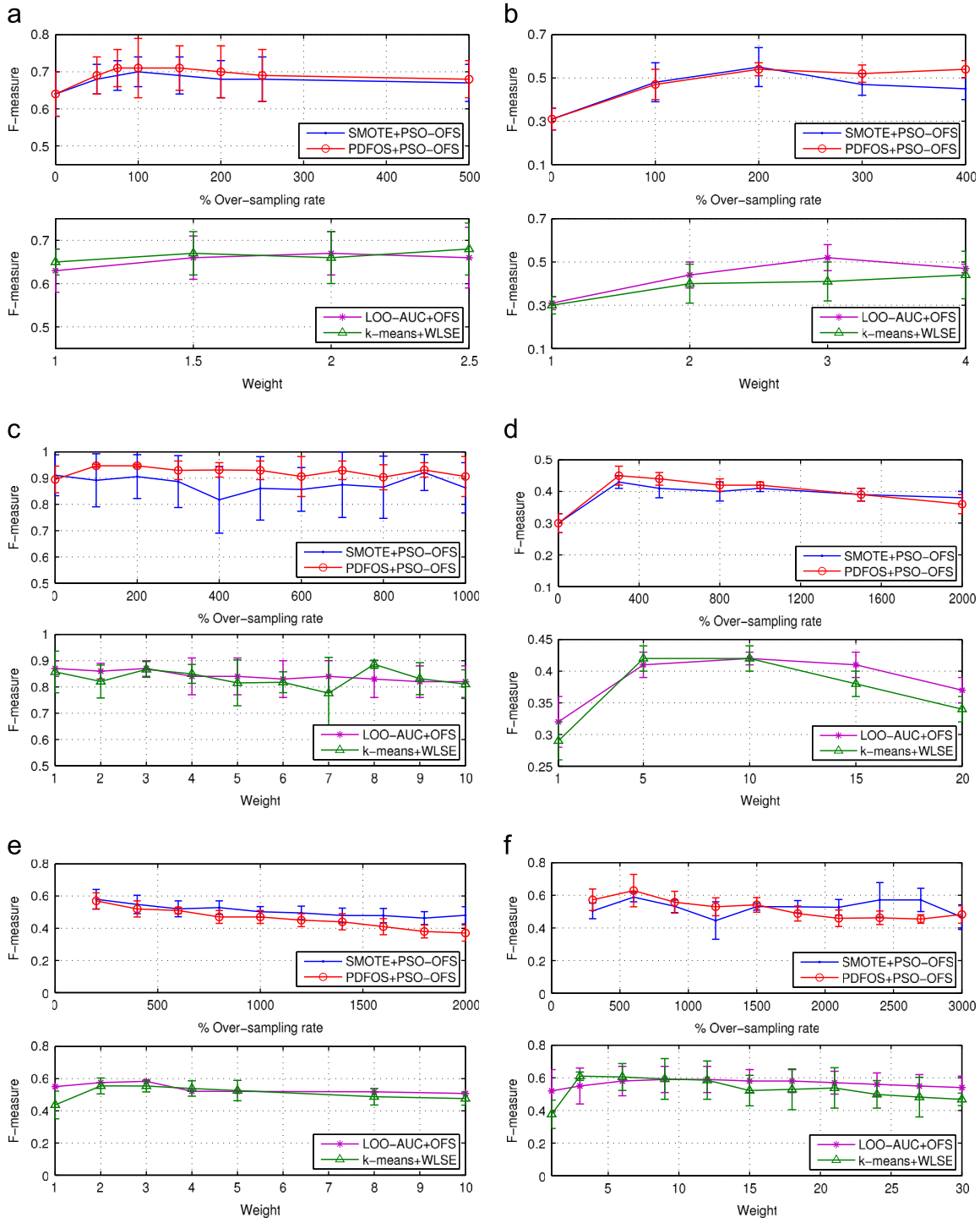

Fig. 5. Comparison of F-measure on imbalanced data sets: (a) Pima Indians diabetes, (b) Haberman's survival, (c) glass, (d) ADI, (e) satimage, and (f) yeast.

As discussed in the introduction section, the data sets were also tested on three other benchmark algorithms, two of which belong to weighted methods, namely, the LOO-AUC+OFS with different weight ρ [13] and the κ-means+WLSE with different weight ρ. More specifically, based on the WLSE, the LOO-AUC+OFS uses the OFS based model selection criterion of maximising the LOO area under the curve (AUC) of receiver operating characteristics (ROC) [13], while the κ-means+WLSE selects the RBF centers via the κ-means clustering. The third algorithm, the SMOTE+PSO-OFS, uses the SMOTE for over-sampling and then applies the same PSO-OFS based RBF classifier, as presented in Section 3, to the rebalanced data set.

Three typical performance metrics for evaluating imbalanced classification performance were involved in the experimental study, and they were the area under the ROC curve (AUC) [66], the G-mean and the F-measure [67]. Receiver operating characteristics (ROC) curves are first presented in Fig. 3, where FP and TP stand for false positive and true positive, respectively. The (FP, TP) pair in the ROC of Fig. 3 is the mean of FP and TP, respectively, averaged over the n-fold CV. Each algorithm is related to one curve formed by the pairs of (FP, TP), obtained for different over-sampling rates r of the SMOTE+PSO-OFS and PDFOS+PSO-OFS or different weights ρ of the LOO-AUC+OFS and κ-means+WLSE. The means and standard deviations of the AUC metric [66] are then listed in Table 2, where the best results are highlighted in boldface. Likewise, the G-mean and F-measure metrics [67] with respect to different r and ρ are reported in Figs. 4 and 5, respectively. Note that, for each data set, the G-mean and F-measure versus r of the SMOTE+PSO-OFS and PDFOS+PSO-OFS and ρ of the LOO-AUC+OFS and κ-means+WLSE are depicted as two separate subplots in the same plot, respectively. The best G-mean and F-measure of each method with the corresponding r or ρ value are listed in the Tables 3 and 4, respectively, where again the best results are highlighted in boldface.

As shown in Fig. 3 and Table 2, the proposed PDF+PSO-OFS algorithm is able to achieve a better performance for the imbalanced data sets with various imbalanced degrees, in terms of the AUC metric, over the three selected competitive and state-of-the-art benchmark methods. The AUC is a metric for evaluating the overall performance across the whole ROC plane. For single operating points, there exist trade-offs between the TP rate and FP rate. As an example, for the satimage data set results depicted in Fig. 3(e), the weighted methods tend to perform better, in terms of TP rate, in the lower FP rate region. The influences of r or ρ to the achievable G-mean and F-measure can be observed in Figs. 4 and 5, respectively. The PDFOS+PSO-OFS shows a very competitive performance, in terms of the best G-mean and F-measure as presented in Tables 3 and 4, respectively. Noticeably, the best G-mean and F-measure tend not to occur at the same r or ρ. Also, the case of fully re-balanced positive and negative classes does not necessarily result in the best G-mean or F-measure. Furthermore, a higher value of r or ρ does not guarantee better G-mean and F-measure, as the FP rate may increase along with the TP rate.

5. Conclusions

Although re-sampling is a straightforward and effective way to deal with imbalanced classification problems, most of the existing methods lack sufficient theoretical insight and justification. This study has followed the principle of over-sampling technique that seeks to re-balance the skewed class distribution, but with the aim of maintaining the true statistical information as manifested in the

Table 3
Comparison of mean and standard deviation of best G-means.

Data set	LOO-AUC+OFS	k-means+WLSE	SMOTE+PSO-OFS	PDFOS+PSO-OFS
Pima Diabetes	0.74 ± 0.04 ($\rho = 2.0$)	0.75 ± 0.06 ($\rho = 2.5$)	0.76 ± 0.05 ($r = 100\%$)	$\mathbf{0.78 \pm 0.05}$ ($r = 100\%$)
Haberman's survival	0.67 ± 0.05 ($\rho = 3.0$)	0.57 ± 0.07 ($\rho = 4.0$)	$\mathbf{0.69 \pm 0.08}$ ($r = 200\%$)	$\mathbf{0.69 \pm 0.02}$ ($r = 400\%$)
Glass(6)	0.93 ± 0.03 ($\rho = 3.0, 6.0$)	0.95 ± 0.02 ($\rho = 8.0$)	0.95 ± 0.06 ($r = 600\%$)	$\mathbf{0.97 \pm 0.04}$ ($r = 600\%$)
ADI	0.76 ± 0.01 ($\rho = 15.0$)	$\mathbf{0.77 \pm 0.02}$ ($\rho = 10.0$)	0.76 ± 0.02 ($r = 1000\%, 1500\%$)	$\mathbf{0.77 \pm 0.01}$ ($r = 800\%, 1000\%$)
Satimage(4)	0.85 ± 0.03 ($\rho = 8.0$)	0.84 ± 0.02 ($\rho = 10.0$)	$\mathbf{0.86 \pm 0.01}$ ($r = 1000\%$)	$\mathbf{0.86 \pm 0.02}$ ($r = 600\%$)
Yeast(5)	0.92 ± 0.09 ($\rho = 27.0, 30.0$)	0.97 ± 0.01 ($\rho = 18.0$)	$\mathbf{0.98 \pm 0.00}$ ($r = 2700\%$)	$\mathbf{0.98 \pm 0.01}$ ($r = 900\%$)

Table 4
Comparison of mean and standard deviation of best F-measures.

Data set	LOO-AUC+OFS	k-means+WLSE	SMOTE+PSO-OFS	PDFOS+PSO-OFS
Pima Diabetes	0.67 ± 0.05 ($\rho = 2.0$)	0.68 ± 0.06 ($\rho = 2.5$)	0.70 ± 0.04 ($r = 100\%$)	$\mathbf{0.71 \pm 0.06}$ ($r = 100\%$)
Haberman's survival	0.52 ± 0.06 ($\rho = 3.0$)	0.44 ± 0.11 ($\rho = 4.0$)	$\mathbf{0.55 \pm 0.09}$ ($r = 200\%$)	0.54 ± 0.03 ($r = 200\%, 400\%$)
Glass(6)	0.87 ± 0.03 ($\rho = 3.0$)	0.89 ± 0.02 ($\rho = 8.0$)	0.92 ± 0.07 ($r = 900\%$)	$\mathbf{0.95 \pm 0.01}$ ($r = 100\%, 200\%$)
ADI	0.42 ± 0.01 ($\rho = 10.0$)	0.42 ± 0.02 ($\rho = 5.0, 10.0$)	0.43 ± 0.02 ($r = 300\%$)	$\mathbf{0.45 \pm 0.03}$ ($r = 300\%$)
Satimage(4)	$\mathbf{0.58 \pm 0.03}$ ($\rho = 3.0$)	0.55 ± 0.05 ($\rho = 2.0$)	$\mathbf{0.58 \pm 0.06}$ ($r = 200\%$)	0.57 ± 0.05 ($r = 200\%$)
Yeast(5)	0.59 ± 0.08 ($\rho = 9.0, 12.0$)	0.61 ± 0.03 ($\rho = 3.0$)	0.59 ± 0.03 ($r = 600\%$)	$\mathbf{0.63 \pm 0.10}$ ($r = 600\%$)

observed data. This has been achieved by a PW based PDF estimator using the positive data samples, followed by drawing data samples according to the estimated PDF in order to re-balance the data. The RBF classifier is then constructed based on the rebalanced data set using the efficient PSO aided OFS procedure. Experimental results have demonstrated that the proposed PDFOS+PSO-OFS approach offers a very competitive method, in comparison with many existing state-of-the-art methods for dealing with imbalanced classification problems.

As reviewed in the introduction section, under-sampling is considered to be able to refine the distribution of the negative class and, as a result, to further improve the overall classification performance. Our future work will investigate how to combine under-sampling with the proposed PDFOS technique. For the challenging class of high-dimensional problems, where the feature space dimension is extremely large, in thousands or even tens of thousands, but the sample size is extremely small by comparison, in hundreds or even in tens, efficient feature selection becomes essential. We are currently investigating suitable feature selection techniques for integrating with the proposed PDFOS+PSO-OFS approach in order to tackle this type of challenging high-dimensional problems effectively.

Acknowledgements

This paper was partly funded by the Deanship of Scientific Research (DSR), King Abdulaziz University, under Grant no. (1-4-1432/HiCi). X. Hong, S. Chen and E. Khalaf acknowledge with thanks DSR technical and financial support.

Appendix A. PSO for optimising a RBF node's parameters

To start the PSO, the search space for the candidate solutions needs to be defined initially. For convenience, define the 2m-dimensional vector $\boldsymbol{\gamma}$ that combines $\boldsymbol{\mu}_n$ and $\boldsymbol{\Sigma}_n$. Then, the optimisation problem defined in (24) can be rewritten as

$$\boldsymbol{\gamma}_{\text{opt}} = \arg\min_{\boldsymbol{\gamma}\in\boldsymbol{\Gamma}} J_{\text{LOO}}^{(n)}(\boldsymbol{\gamma}) \tag{27}$$

where the search space $\boldsymbol{\Gamma}$ is defined by

$$\boldsymbol{\Gamma} \triangleq \prod_{i=1}^{2m} [\Gamma_{i,\min}, \Gamma_{i,\max}] \tag{28}$$

Specifically, the search space for $\boldsymbol{\mu}_n = [\mu_{n,1}\ \mu_{n,2}\cdots\mu_{n,m}]^{\text{T}}$ is specified by the distribution of the training data $\{\mathbf{x}_k\}_{k=1}^N$ according to

$$\mu_{n,i} \in [x_{\min,i}, x_{\max,i}] \triangleq [\Gamma_{i,\min}, \Gamma_{i,\max}], \quad 1 \le i \le m \tag{29}$$

with $x_{\min,i}$ and $x_{\max,i}$ given in (26), while each element of $\boldsymbol{\Sigma}_n$ is limited in the range

$$\sigma_{n,i}^2 \in [\sigma_{\min}^2, \sigma_{\max}^2] \triangleq [\Gamma_{(i+m),\min}, \Gamma_{(i+m),\max}], \quad 1 \le i \le m \tag{30}$$

A swarm of particles, $\{\boldsymbol{\gamma}_s^{[l]}\}_{s=1}^S$, that represents the candidate solutions are initially generated randomly within the search space $\boldsymbol{\Gamma}$, where S is the size of the swarm and $0 \le l \le L$ is the iteration index with L representing the maximum number of iterations. Each particle position $\boldsymbol{\gamma}_s^{[l]}$ has the associated cost $J_{\text{LOO}}^{(n)}(\boldsymbol{\gamma}_s^{[l]})$. For the sth particle, $\boldsymbol{\gamma}_s^{[i]}$ with the best $J_{\text{LOO}}^{(n)}(\boldsymbol{\gamma}_s^{[i]})$ for $0 \le i \le l$ is stored as the cognitive information $\mathbf{CI}_s^{[l]}$. For the entire swarm, $\boldsymbol{\gamma}_s^{[i]}$ with the best $J_{\text{LOO}}^{(n)}(\boldsymbol{\gamma}_s^{[i]})$ for $0 \le i \le l$ and $1 \le s \le S$ is stored as the social information $\mathbf{SI}^{[l]}$. Each particle $\boldsymbol{\gamma}_s^{[l]}$ has a velocity vector $\boldsymbol{\nu}_s^{[l]} = [\nu_{s,1}^{[l]}\nu_{s,2}^{[l]}\cdots\nu_{s,2\ m}^{[l]}]^{\text{T}}$ to direct its "flying" or search, where $\boldsymbol{\nu}_s^{[l]} \in \mathbf{V}$ with the velocity space defined by

$$\mathbf{V} \triangleq \prod_{i=1}^{2m} [-V_{i,\max}, V_{i,\max}] \tag{31}$$

in which $V_{i,\max} = \frac{1}{2}(\Gamma_{i,\max} - \Gamma_{i,\min})$.

The cognitive information $\mathbf{CI}_s^{[l]}$ and the social information $\mathbf{SI}^{[l]}$ are used to update the particle velocities and positions according to

$$\begin{aligned}\boldsymbol{\nu}_s^{[l+1]} = {} & rand() \cdot \boldsymbol{\nu}_s^{[l]} + rand() \cdot c_1 \cdot (\mathbf{CI}_s^{[l]} - \boldsymbol{\gamma}_s^{[l]}) \\ & + rand() \cdot c_2 \cdot (\mathbf{SI}^{[l]} - \boldsymbol{\gamma}_s^{[l]})\end{aligned} \tag{32}$$

$$\boldsymbol{\gamma}_s^{[l+1]} = \boldsymbol{\gamma}_s^{[l]} + \boldsymbol{\nu}_s^{[l+1]} \tag{33}$$

where $rand()$ is the uniform random number in $[0, 1]$, c_1 and c_2 are the two acceleration coefficients. The velocity space $\mathbf{V}$ is applied to confine the updated $\boldsymbol{\nu}_s^{[l+1]}$ according to

$$\begin{cases} \boldsymbol{\nu}_s^{[l+1]} = \pm 0.1 \cdot rand() \cdot \mathbf{V}_{\max}, & \boldsymbol{\nu}_s^{[l+1]} \approx 0 \\ \nu_{s,i}^{[l+1]} = V_{i,\max}, & \nu_{s,i}^{[l+1]} > V_{i,\max} \\ \nu_{s,i}^{[l+1]} = -V_{i,\max}, & \nu_{s,i}^{[l+1]} < -V_{i,\max} \end{cases} \tag{34}$$

Typically, the time varying acceleration coefficients [21]

$$\begin{aligned} c_1 &= 2.5 - (2.5 - 0.5) \cdot l/L \\ c_2 &= 0.5 + (2.5 - 0.5) \cdot l/L \end{aligned} \tag{35}$$

can be adopted for the two coefficients c_1 and c_2 in (32).

The detailed algorithmic steps for applying the PSO algorithm to determine the nth RBF node's parameters can be found in [20,56]. It is notable that at each stage for constructing RBF classifier the computational complexity including one LOO cost evaluation and the associated model column orthogonalisation maintains the order of $O(N)$. Thus, the computational requirements of constructing an M-term RBF model tuned by PSO can be given as $(M+1) \times L \times S \times O(N)$ [20].

References

[1] N. Petrick, H.P. Chan, B. Sahiner, D. Wei, An adaptive density-weighted contrast enhancement filter for mammographic breast mass detection, IEEE Trans. Med. Imaging 15 (1) (1996) 59–67.
[2] T. Fawcett, F. Provost, Adaptive fraud detection, Data Min. Knowl. Discov. 1 (3) (1997) 291–316.
[3] M. Kubat, R.C. Holte, S. Matwin, Machine learning for the detection of oil spills in satellite radar images, Mach. Learn. 30 (2–3) (1998) 195–215.
[4] D.D. Lewis, J. Catlett, Heterogeneous uncertainty sampling for supervised learning, in: Proceedings of the 11th International Conference on Machine Learning, New Brunswick, NJ, USA, July 10–13, 1994, pp. 148–156.
[5] C.X. Ling, C. Li, Data mining for direct marketing: problems and solutions, in: Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining, New York, USA, August 27–31, 1998, pp. 73–79.
[6] E.P.D. Pednault, B.K. Rosen, C. Apte, Handling Imbalanced Data Sets in Insurance Risk Modeling, IBM Research Report RC-21731, 2000.
[7] G.M. Weiss, F. Provost, The Effect of Class Distribution on Classifier Learning: An Empirical Study, Technical Report ML-TR-44, Department of Computer Science, Rutgers University, 2001.
[8] A. Estabrooks, T. Jo, N. Japkowicz, A multiple resampling method for learning from imbalanced data sets, J. Chem. Inf. Model. 20 (1) (2004) 18–36.
[9] N. Japkowicz, S. Stephen, The class imbalance problem: a systematic study, Intell. Data Anal. 6 (5) (2002) 429–449.
[10] R. Akbani, S. Kwek, N. Japkowicz, Applying support vector machines to imbalanced datasets, in: Proceedings of the 15th European Conference on Machine Learning, Pisa, Italy, September 20–24, 2004, pp. 39–50.
[11] G. Wu, E.Y. Chang, KBA: kernel boundary alignment considering imbalanced data distribution, IEEE Trans. Knowl. Data Eng. 17 (6) (2005) 786–795.
[12] H. He, E.A. Garcia, Learning from imbalanced data, IEEE Trans. Knowl. Data Eng. 21 (9) (2009) 1263–1284.
[13] X. Hong, S. Chen, C.J. Harris, A kernel-based two-class classifier for imbalanced data sets, IEEE Trans. Neural Netw. 18 (1) (2007) 28–41.
[14] J. Moody, C.J. Darken, Fast learning in networks of locally-tuned processing units, Neural Comput. 1 (2) (1989) 281–294.
[15] S. Haykin, Neural Networks: A Comprehensive Foundation, 2nd edition, Prentice Hall, Upper Saddle River, NJ, 1998.
[16] Y. Sun, M.S. Kamel, A.K.C. Wong, Y. Wang, Cost-sensitive boosting for classification of imbalanced data, Pattern Recogn. 40 (12) (2007) 3358–3378.

[17] W. Fan, S.J. Stolfo, J. Zhang, P.K. Chan, AdaCost: misclassification cost-sensitive boosting, in: Proceedings of the 16th International Conference on Machine Learning, Bled, Slovenia, June 27–30, 1999, pp. 97–105.
[18] J. Kennedy, R.C. Eberhart, Swarm Intelligence, Morgan Kaufmann, San Francisco, CA, 2001.
[19] S. Chen, X. Hong, C.J. Harris, Radial basis function classifier construction using particle swarm optimisation aided orthogonal forward regression, in: Proceedings of the 2010 International Joint Conference on Neural Networks, Barcelona, Spain, July 18–23, 2010, pp. 3418–3423.
[20] S. Chen, X. Hong, C.J. Harris, Particle swarm optimization aided orthogonal forward regression for unified data modelling, IEEE Trans. Evol. Comput. 14 (4) (2010) 477–499.
[21] A. Ratnaweera, S.K. Halgamuge, H.C. Watson, Self-organizing hierarchical particle swarm optimizer with time-varying acceleration coefficients, IEEE Trans. Evol. Comput. 8 (3) (2004) 240–255.
[22] W.-F. Leong, G.G. Yen, PSO-based multiobjective optimization with dynamic population size and adaptive local archives, IEEE Trans. Syst. Man Cybern. B 38 (5) (2008) 1270–1293.
[23] S. Chen, X. Hong, B.L. Luk, C.J. Harris, Non-linear system identification using particle swarm optimisation tuned radial basis function models, Int. J. Bio-Inspired Comput. 1 (4) (2009) 246–258.
[24] M. Ramezani, M.-R. Haghifam, C. Singh, H. Seifi, M.P. Moghaddam, Determination of capacity benefit margin in multiarea power systems using particle swarm optimization, IEEE Trans. Power Syst. 24 (2) (2009) 631–641.
[25] H.-L. Wei, S.A. Billings, Y. Zhao, L. Guo, Lattice dynamical wavelet neural networks implemented using particle swarm optimization for spatio-temporal system identification, IEEE Trans. Neural Netw. 20 (1) (2009) 181–185.
[26] S. Chen, W. Yao, H.R. Palally, L. Hanzo, Particle swarm optimisation aided MIMO transceiver designs, in: Y. Tenne, C.-K. Goh (Eds.), Computational Intelligence in Expensive Optimization Problems, Springer-Verlag, Berlin, 2010, pp. 487–511 (Chapter 19).
[27] P. Puranik, P. Bajaj, A. Abraham, P. Palsodkar, A. Deshmukh, Human perception-based color image segmentation using comprehensive learning particle swarm optimization, J. Inf. Hiding Multim. Signal Process. 2 (3) (2011) 227–235.
[28] F.-C. Chang, H.-C. Huang, A refactoring method for cache-efficient swarm intelligence algorithms, Inf. Sci. 192 (1) (2012) 39–49.
[29] G.E.A.P.A. Batista, R.C. Prati, M.C. Monard, Study of the behavior of several methods for balancing machine learning training data, ACM SIGKDD Expl. Newsl. 6 (1) (2004) 20–29.
[30] C. Drummond, R.C. Holte, C4.5, class imbalance, and cost sensitivity: Why under-sampling beats over-sampling, in: Proceedings of the 12th International Conference on Machine Learning – Workshop on Learning from Imbalanced Datasets II, Washington DC, USA, August 21, 2003, pp. 1–8.
[31] D.W. Aha, D. Kibler, M.K. Albert, Instance-based learning algorithms, Mach. Learn. 6 (1) (1991) 37–66.
[32] J. Zhang, Selecting typical instances in instance-based learning, in: Proceedings of the 9th International Workshop on Machine learning, Aberdeen, Scotland, July 1–3, 1992, pp. 470–479.
[33] D.B. Skalak, Prototype and feature selection by sampling and random mutation hill climbing algorithms, in: Proceedings of the 11th International Conference on Machine Learning, New Brunswick, USA, July 10–13, 1994, pp. 293–301.
[34] S. Floyd, M. Warmuth, Sample compression, learnability, and the vapnik-chervonenkis dimension, Mach. Learn. 21 (3) (1995) 269–304.
[35] M. Kubat, S. Matwin, Addressing the curse of imbalanced training sets: one-sided selection, in: Proceedings of the 14th International Conference on Machine Learning, Nashville, USA, July 8–12, 1997, pp. 179–186.
[36] J. Zhang, I. Mani, KNN approach to unbalance data distributions: a case study involving information extraction, in: Proceedings of the 12th International Conference on Machine Learning – Workshop on Learning from Imbalanced Datasets II, Washington DC, USA, August 21, 2003, pp. 42–48.
[37] X.Y. Liu, J. Wu, Z.H. Zhou, Exploratory undersampling for class-imbalance learning, IEEE Trans. Syst. Man Cybern. B 39 (2) (2009) 539–550.
[38] R. Barandela, E. Rangel, J.S. Sánchez, F.J. Ferri, Restricted decontamination for the imbalanced training sample problem, in: A. Sanfeliu, J. Ruiz-Shulcloper (Eds.), Progress in Pattern Recognition, Speech and Image Analysis, Lecture Notes in Computer Science, vol. 2905, Springer-Verlag, Berlin, 2003, pp. 424–431.
[39] S. García, J. Cano, A. Fernández, F. Herrera, A proposal of evolutionary prototype selection for class imbalance problems, in: E. Corchado, H. Yin, V. Botti, C. Fyfe (Eds.), Intelligent Data Engineering and Automated Learning, Lecture Notes in Computer Science, vol. 4224, Springer-Verlag, Berlin, 2006, pp. 1415–1423.
[40] R. Barandela, J.K. Hernández, J.S. Sánchez, F.J. Ferri, Imbalanced training set reduction and feature selection through genetic optimization, in: Proceeding of the 2005 Conference on Artificial Intelligence Research and Development, vol. 131, 2005, pp. 215–222.
[41] I. Tomek, Two modifications of CNN, IEEE Trans. Syst. Man Cybern. 6 (11) (1976) 769–772.
[42] D.L. Wilson, Asymptotic properties of nearest neighbor rules using edited data, IEEE Trans. Syst. Man Cybern. 2 (3) (1972) 408–421.
[43] R. Barandela, J.S. Sánchez, V. García, E. Rangel, Strategies for learning in class imbalance problems, Pattern Recogn. 36 (3) (2003) 849–851.
[44] J. Laurikkala, Improving identification of difficult small classes by balancing class distribution, in: Proceedings of the 8th Conference on AI in Medicine in Europe: Artificial Intelligence Medicine, Cascais, Portugal, July 1–4, 2001, pp. 63–66.
[45] P. Hart, The condensed nearest neighbor rule (Corresp.), IEEE Trans. Inf. Theory 14 (3) (1968) 515–516.
[46] R. Barandela, R.M. Valdovinos, J.S. Sánchez, F.J. Ferri, The imbalanced training sample problem: under or over sampling?, in: A. Fred, T. Caelli, R.P.W. Duin, A. Campilho, D.d. Ridder (Eds.), Structural, Syntactic, and Statistical Pattern Recognition, Lectures Notes in Computer Science, vol. 3138, Springer-Verlag, Berlin, 2004, pp. 806–814.
[47] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: Synthetic minority over-sampling technique, J. Artif. Intell. Res. 16 (2002) 321–357.
[48] B.X. Wang, N. Japkowicz, Imbalanced data set learning with synthetic samples, in: Proceedings of IRIS Machine Learning Workshop, Ottawa, Canada, June 9, 2004.
[49] N.V. Chawla, A. Lazarevic, L.O. Hall, K.W. Bowyer, SMOTEBoost: improving prediction of the minority class in boosting, in: Proceedings of the 7th European Conference on Principles and Practice of Knowledge Discovery in Databases, Cavtat-Dubrovnik, Croatia, September 22–26, 2003, pp. 107–119.
[50] H. Han, W.Y. Wang, B.H. Mao, Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning, in: D.-S. Huang, X.-P. Zhang, G.-B. Huang (Eds.), Advances in Intelligent Computing, Lecture Notes in Computer Science, vol. 3644, Springer-Verlag, Berlin, 2005, pp. 878–887.
[51] H. He, Y. Bai, E.A. Garcia, S. Li, ADASYN: adaptive synthetic sampling approach for imbalanced learning, in: Proceedings of the 2008 International Joint Conference on Neural Networks, Hong Kong, China, June 1–8, 2008, pp. 1322–1328.
[52] B.W. Silverman, Density Estimation for Statistics and Data Analysis, Chapman and Hall, London, 1986.
[53] R.O. Duda, P.E. Hart, Pattern Classification and Scene Analysis, John Wiley & Sons Inc., New York, 1973.
[54] C.M. Bishop, Neural Networks for Pattern Recognition, Oxford University Press, New York, 1995.
[55] E. Parzen, On estimation of a probability density function and mode, Ann. Math. Stat. 33 (3) (1962) 1065–1076.
[56] M. Gao, X. Hong, S. Chen, C.J. Harris, On combination of SMOTE and particle swarm optimization based radial basis function classifier for imbalanced problems, in: Proceedings of the 2011 International Joint Conference on Neural Networks, San Jose, USA, July 30–August 5, 2011, pp. 1146–1153.
[57] X. Hong, S. Chen, C.J. Harris, A forward-constrained regression algorithm for sparse kernel density estimation, IEEE Trans. Neural Netw. 19 (1) (2008) 193–198.
[58] S. Chen, X. Hong, C.J. Harris, Sparse kernel density construction using orthogonal forward regression with leave-one-out test score and local regularization, IEEE Trans. Syst. Man Cybern. B 34 (4) (2004) 1708–1717.
[59] X. Hong, S. Chen, C.J. Harris, An orthogonal forward regression technique for sparse kernel density estimation, Neurocomputing 71 (4–6) (2008) 931–943.
[60] K. Fukunaga, Introduction to Statistical Pattern Recognition, 2nd edition, Academic Press, San Diego, CA, 1990.
[61] J.W. Tukey, P.A. Tukey, Graphical display of data sets in 3 or more dimensions, in: V. Barnett (Ed.), Interpreting Multivariate Data, Wiley and Sons, Chichester, UK, 1981, pp. 189–257.
[62] X. Hong, S. Chen, C.J. Harris, A fast linear-in-the-parameters classifier construction algorithm using orthogonal forward selection to minimize leave-one-out misclassification rate, Int. J. Syst. Sci. 39 (2) (2008) 119–125.
[63] R.H. Myers, Classical and Modern Regression with Applications, 2nd edition. Boston: PWS-KENT, 1990.
[64] K.K. Lee, C.J. Harris, S.R. Gunn, P.A.S. Reed, Classification of imbalanced data with transparent kernel, in: Proceedings of the 2001 International Joint Conference on Neural Networks, Washington DC, USA, July 15–19, 2001, pp. 2410–2415.
[65] C.L. Blake, C.J. Merz, UCI repository of machine learning databases, Department of Computer Science, University of California, Department of Computer Science, Irvine, CA, 1998. ⟨http://archive.ics.uci.edu/ml/datasets.html⟩.
[66] A.P. Bradley, The use of the area under the ROC curve in the evaluation of machine learning algorithms, Pattern Recogn. 30 (1997) 1145–1159.
[67] C. van Rijsbergen, Information Retrieval, Butterworths, London, 1979.

Ming Gao received his B.Eng. degree from Northwestern Polytechnical University, Shaanxi, PR China, in 2006, the MEng degree from Beihang University, Beijing, PR China, in 2009, and the PhD from University of Reading, Reading, UK, in 2013.

His research interests are in data modelling, machine learning, pattern recognition, and their applications in imbalanced problems.

Xia Hong received her university education at National University of Defense Technology, PR China (BSc, 1984, M.Sc., 1987), and University of Sheffield, UK (Ph.D., 1998), all in automatic control.

She worked as a research assistant in Beijing Institute of Systems Engineering, Beijing, China from 1987 to 1993. She worked as a research fellow in the Department of Electronics and Computer Science at University of Southampton from 1997 to 2001. She is currently a Professor at School of Systems Engineering, University of Reading. She is actively engaged in research into nonlinear systems identification, data modelling, estimation and intelligent control, neural networks, pattern recognition, learning theory and their applications. She has published over 100 research papers, and coauthored a research book.

Professor Hong was awarded a Donald Julius Groen Prize by IMechE in 1999.

Sheng Chen received his B.Eng. degree from the East China Petroleum Institute, Dongying, China, in January 1982, and his Ph.D. degree from the City University, London, in September 1986, both in control engineering. In 2005, he was awarded the higher doctoral degree, Doctor of Sciences (D.Sc.), from the University of Southampton, Southampton, UK.

From 1986 to 1999, He held research and academic appointments at the Universities of Sheffield, Edinburgh and Portsmouth, all in UK. Since 1999, he has been with Electronics and Computer Science, the University of Southampton, UK, where he currently holds the post of Professor in Intelligent Systems and Signal Processing. Dr. Chen's research interests include adaptive signal processing, wireless communications, modelling and identification of nonlinear systems, neural network and machine learning, intelligent control system design, evolutionary computation methods and optimisation. He has published over 500 research papers.

Dr. Chen is a Fellow of IEEE and a Fellow of IET. He is a Distinguished Adjunct Professor at King Abdulaziz University, Jeddah, Saudi Arabia. Dr. Chen is an ISI highly cited researcher in engineering (March 2004).

Chris J. Harris received his B.Sc. and M.A. degrees from the University of Leicester and the University of Oxford in UK, respectively, and his PhD degree from the University of Southampton, UK, in 1972. He was awarded the higher doctoral degree, the Doctor of Sciences (D.Sc.), by the University of Southampton in 2001.

He is Emeritus Research Professor at the University of Southampton, having previously held senior academic appointments at Imperial College, Oxford and Manchester Universities, as well as Deputy Chief Scientist for the UK Government.

Professor Harris was awarded the IEE senior Achievement Medal for Data Fusion research and the IEE Faraday Medal for distinguished international research in Machine Learning. He was elected to the UK Royal Academy of Engineering in 1996. He is the co-author of over 450 scientific research papers during a 45 year research career.

Emad F. Khalaf received his B.Eng. and M.Eng. degrees in IT from Wroclaw University of Technology in Poland, in 1992, as one certificate, and the PhD degree in Computer networks from Wroclaw University of Technology, in Poland, in 2002.

From 2003 to 2011, he worked as an assistant professor at Computer Engineering Department, Faculty of Engineering, Philadelphia University, in Jordan. Since 2012 he is an assistant professor at Electrical and Computer Engineering Department, Faculty of Engineering, King Abdulaziz University, Jeddah, Saudi Arabia. Dr. Khalaf's research interests are in network security and cryptography, speech classification and recognition.