Probability Density Function Estimation Based Over-Sampling for Imbalanced Two-Class Problems

Ming Gao, Xia Hong, Sheng Chen and Chris J. Harris

***Abstract*— A novel probability density function (PDF) estimation based over-sampling approach is proposed for two-class imbalanced classification problems. The Parzen-window kernel function is applied to estimate the PDF of the positive class, from which synthetic instances are generated as additional training data to re-balance the class distribution. Utilising the re-balanced over-sampled training data, a radial basis function (RBF) classifier is constructed by applying an orthogonal forward regression, 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 approach is demonstrated by an empirical study on several imbalanced data sets.**

I. Introduction

Two-class imbalanced classification problems, in which the instances of one class outnumbers the instances of the other class, widely arise in life threatening or safety critical and many other real-world applications [1]–[6]. The imbalance between two classes is problematic for many standard classification algorithms [7]–[11], whose performance deteriorate as class imbalance degree increases, or equivalently as the data samples of minority or positive class become sparser [9]. For example, kernel-based methods, which are regarded as robust classifiers [12], construct a decision hyperplane separating the two classes. Without special countermeasure for imbalance in the training data, the resultant hyperplane will tend to be placed in favour of classification performance for the majority or negative class, but the classification performance for the positive class becomes unsatisfactory. Techniques for tackling the imbalanced problem can be categorised into two categories: resampling methods and imbalanced learning algorithms.

Imbalanced learning algorithms are obtained by modifying 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 can be modified, in order to cope with the imbalanced distribution during the classifier construction process [11], [13]. A well-known radial basis function (RBF) modelling approach is the two staged procedure [14], in which the RBF centres are first determined using 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.

Resampling methods are *external* as they operate on the original imbalanced data set, aiming to provide a re-balanced input to train a conventional classifier. There have been many studies [16]–[22] 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 [23] has been applied to minimise the leave-one-out (LOO) misclassification rate in the orthogonal forward selection (OFS) construction of a tunable RBF classifier [24]. The tunable RBF modelling advocated in [24] 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 the learning process. Resampling methods can be divided into the two basic categories, under-sampling and over-sampling.

Various under-sampling techniques have been proposed in the literature [3], [18]–[20], [25]–[31]. Under-sampling tends to be an ideal option when the imbalance degree is not very severe. However, as pointed out in [32], the use of over-sampling is necessary when the imbalance degree is high. Random over-sampling is a simple yet competitive method [9], [25], but it suffers from a serious problem of over-fitting. The study [21] proposed a synthetic minority over-sampling technique (SMOTE), which enhances the significance of some specific regions in the feature space by over-sampling the positive class. Although SMOTE is a well acknowledged technique, it has some drawbacks, including over generalisation and high variance [33]. Some improved SMOTE methods, such as SMOTEBoost [22], were proposed to alleviate the limitations of SMOTE. Despite the empirical evidences that the foregoing methods have been effective in improving the classification performance for positive class, the reason behind the success of the oversampling approaches, such as SMOTE, is not fully understood, as there are little theoretical

M. Gao and X. Hong is with School of Systems Engineering, University of Reading, Reading RG6 6AY, U.K. (E-mails: ming.gao@pgr.reading.ac.uk, x.hong@reading.ac.uk).

S. Chen and C.J. Harris are with Electronics and Computer Science, Faculty of Physical and Applied Sciences, University of Southampton, Southampton SO17 1BJ, U.K. (E-mails: sqc@ecs.soton.ac.uk, cjh@ecs.soton.ac.uk). S. Chen is also with Faculty of Engineering, King Abdulaziz University, Jeddah 21589, Saudi Arabia.

This work was partly supported by UK EPSRC.

studies that justify most of the oversampling methods.

An ideal oversampling technique should generate synthetic data according to the same probability distribution which produces the observed positive-class data samples. We propose an oversampling approach based on the Parzen window (PW) or kernel density estimation [34], [35] from positive-class data samples. According to the estimated probability density function (PDF), synthetic instances are generated as the additional training data. Th RBF classifier proposed in [24] is then applied to the rebalanced data set, to complete the classification process. The significance of the proposed method is twofold. Firstly, the proposed oversampling 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 classifiers [24].

II. PDF ESTIMATION BASED OVER-SAMPLING

Consider the two-class data set given as

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

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, with $N_+ \ll N_-$. The samples $\mathbf{x}_k$ are generated independently and identically from the unknown underlying PDF.

Kernel density estimation for positive class: Denote the unknown PDF that generates the positive-class sample set D_{N_+} by $p(\mathbf{x})$. A 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 smoothing parameter, and $\Phi_\sigma(\mathbf{x} - \mathbf{x}_i)$ is the kernel function with the training instance $\mathbf{x}_i$ as its centre, scaled by σ. The normal kernel scaled by a single σ is often chosen as kernel function [36]

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

which implies that all the dimensions of the feature space are uncorrelated and have the same spread. To obtain a better PDF estimate 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

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

where

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

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

$$\mathbf{S} = \frac{1}{N_+ - 1} \sum_{i=1}^{N_+} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^{\mathrm{T}} \tag{5}$$

with $\bar{\mathbf{x}} = \frac{1}{N_+} \sum_{i=1}^{N_+} \mathbf{x}_i$ being the mean vector of the positive class. The inclusion of $\mathbf{S}$ in (4) is to account for the coordinates of the feature space being correlated and the spreads of the coordinates being different.

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), based on which the value of σ can be found by minimising the score function $M(\sigma)$ [35]

$$M(\sigma) = N_+^{-2} \sum_i \sum_j \Phi_\sigma^*\left(\mathbf{S}^{-1/2}(\mathbf{x}_j - \mathbf{x}_i)\right) + 2N_+^{-1}\Phi_\sigma(\mathbf{0}) \tag{6}$$

where $\Phi_\sigma^*\left(\mathbf{S}^{-1/2}(\mathbf{x}_j - \mathbf{x}_i)\right) \approx \Phi_\sigma^{(2)}\left(\mathbf{S}^{-1/2}(\mathbf{x}_j - \mathbf{x}_i)\right) - 2\Phi_\sigma\left(\mathbf{S}^{-1/2}(\mathbf{x}_j - \mathbf{x}_i)\right)$, in which $\Phi_\sigma^{(2)}\left(\mathbf{S}^{-1/2}(\mathbf{x}_j - \mathbf{x}_i)\right)$ is given by $\Phi_\sigma^{(2)}\left(\mathbf{S}^{-1/2}(\mathbf{x}_j - \mathbf{x}_i)\right) = \frac{(\sqrt{2}\sigma)^{-m}}{(2\pi)^{m/2}} e^{-\frac{1}{2}(\sqrt{2}\sigma)^{-2}(\mathbf{x}_j-\mathbf{x}_i)^{\mathrm{T}}\mathbf{S}^{-1}(\mathbf{x}_j-\mathbf{x}_i)}$. The optimal σ can be found by a grid search.

Over-sampling based on a 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 is generated by the two following steps:

1) Based on the discrete uniform distribution, randomly draw a data sample, $\mathbf{x}_o$, from the positive-class data set.

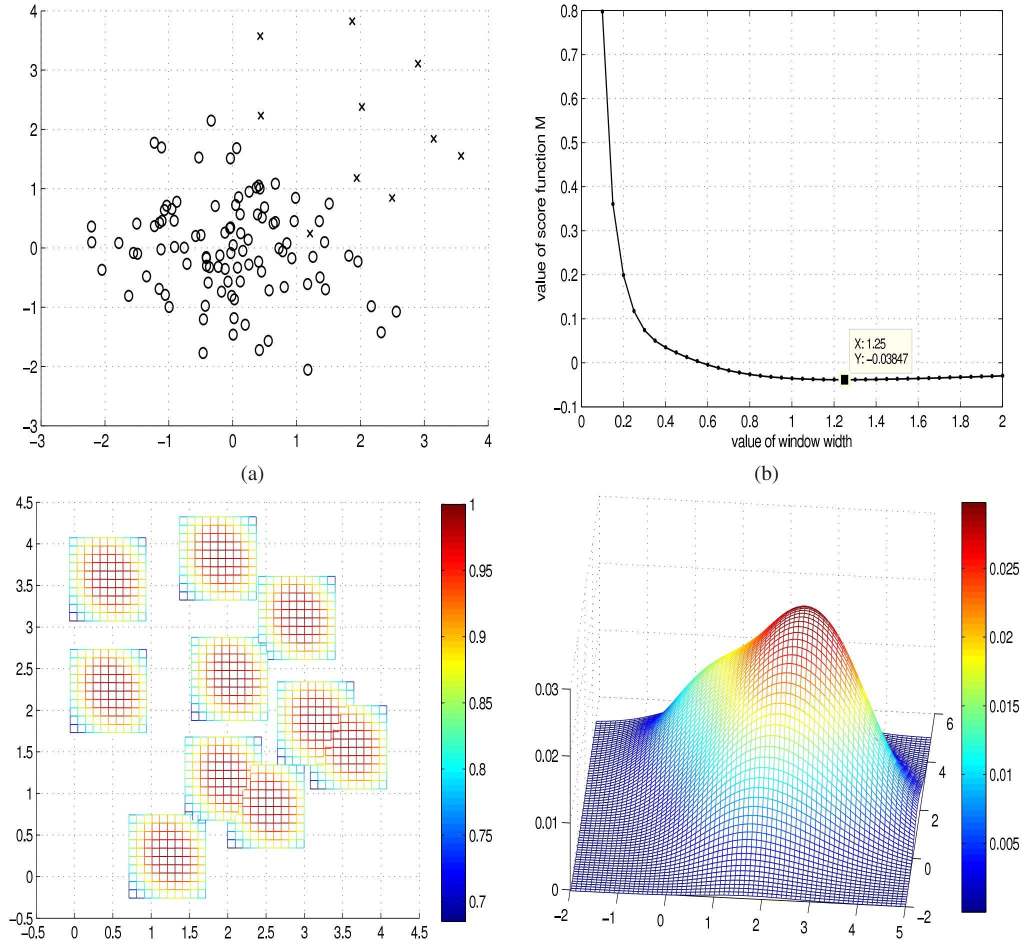
2) Generate a synthetic data sample, $\mathbf{x}_n$, using the Gaussian distribution with $\mathbf{x}_o$ as the mean and $\sigma^2\mathbf{S}$ as the covariance matrix.

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

$$\mathbf{x}_n = \mathbf{x}_o + \sigma\mathbf{R} \cdot \mathbf{randn}() \tag{7}$$

where $\mathbf{R}$ is the upper triangular matrix that is the Cholesky decomposition of $\mathbf{S}$, 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 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. 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. In Fig. 1 (c), the kernel function placed at each positive-class instance is constructed according to $\sigma^2\mathbf{S}$. In this example, $\mathbf{S} \approx \mathbf{I}_2$. Fig. 1 (d) presents the density estimate for the positive class, which is the mixture of all the kernels in Fig. 1 (c) with an equal weighting for each component.



(a)

(b)

(c)

(d)

Fig. 1. PDF estimation for the synthetic imbalanced data set: (a) the imbalanced data set with x denoting positive-class instance and ∘ 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.

The over-sampled data distributions for the imbalanced data set of Fig. 1 (a), obtained by the proposed method and the SMOTE 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. Both the proposed and SMOTE methods increase the positive-class instances, particularly in the decision region. 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, because the SMOTE generates the synthetic instances in the line linking the original instance to its k-NN neighbours [37]. As a result, increasing the oversampling rate r only leads to a higher density in this region. By contrast, the over-sampled positive class generated by the proposed method expands along the direction of the ideal decision boundary, as can be seen from Fig. 2 (a).

III. Tunable RBF classifier

After oversampling the positive class with a required oversampling rate r, a tunable RBF classifier is constructed based on the rebalanced training set using the algorithm proposed in [24]. For notational simplicity, the oversampled training data set is still denoted as $D_N = \{\mathbf{x}_k, y_k\}_{k=1}^{N}$. The RBF classifier to be constructed takes the form

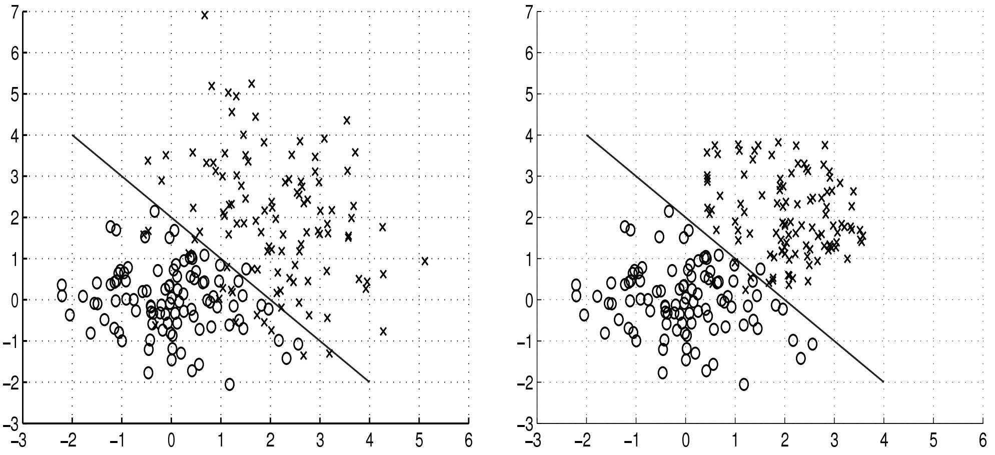
$$\hat{y}_k^{(M)} = \sum_{i=1}^{M} w_i g_i(\mathbf{x}_k) = \mathbf{g}_M^{\mathrm{T}}(k)\mathbf{w}_M,\ \tilde{y}_k^{(M)} = \mathrm{sgn}\big(\hat{y}_k^{(M)}\big) \quad (8)$$

where M is the number of RBF kernels, $\hat{y}_k^{(M)}$ is the output of the classifier with the M kernels $g_i(\bullet)$ for $1 \le i \le M$, $\mathbf{w}_M = \big[w_1\ w_2 \cdots w_M\big]^{\mathrm{T}}$ the weight vector and $\mathbf{g}_M^{\mathrm{T}}(k) = \big[g_1(\mathbf{x}_k)\ g_2(\mathbf{x}_k) \cdots g_M(\mathbf{x}_k)\big]$, while $\tilde{y}_k^{(M)}$ denotes the estimated class label for $\mathbf{x}_k$ with $\mathrm{sgn}(y) = -1$ if $y \le 0$ and $\mathrm{sgn}(y) = 1$ if $y > 0$. The Gaussian kernel $g_i(\mathbf{x}) = e^{-(\mathbf{x}-\boldsymbol{\mu}_i)^{\mathrm{T}}\boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)}$ 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 $\boldsymbol{\Sigma}_i = \mathrm{diag}\{\sigma_{i,1}^2, \sigma_{i,2}^2, \cdots, \sigma_{i,m}^2\}$. 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 (8), the RBF classifier over D_N can be written as

$$\mathbf{y} = \mathbf{G}_M\mathbf{w}_M + \boldsymbol{\varepsilon}^{(M)} \quad (9)$$

where $\boldsymbol{\varepsilon}^{(M)} = \big[\varepsilon_1^{(M)}\ \varepsilon_2^{(M)} \cdots \varepsilon_N^{(M)}\big]^{\mathrm{T}}$ with $\varepsilon_k^{(M)} = y_k - \hat{y}_k^{(M)}$, $\mathbf{y} = \big[y_1\ y_2 \cdots y_N\big]^{\mathrm{T}}$, and $\mathbf{G}_M = \big[\mathbf{g}_1\ \mathbf{g}_2 \cdots \mathbf{g}_M\big]$ with



(a) (b)

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

$\mathbf{g}_l = \left[g_l(\mathbf{x}_1)\ g_l(\mathbf{x}_2)\cdots g_l(\mathbf{x}_N)\right]^{\mathrm{T}}$ for $1 \leq l \leq M$. 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

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

and $\mathbf{P}_M = \left[\mathbf{p}_1\ \mathbf{p}_2\cdots\mathbf{p}_M\right]$ with $\mathbf{p}_i^{\mathrm{T}}\mathbf{p}_j = 0$ for $i \neq j$. The RBF classifier (9) can alternatively be represented as:

$$\mathbf{y} = \mathbf{P}_M\boldsymbol{\theta}_M + \boldsymbol{\varepsilon}^{(M)} \tag{10}$$

where $\boldsymbol{\theta}_M = \left[\theta_1\ \theta_2\cdots\theta_M\right]^{\mathrm{T}}$ satisfies $\boldsymbol{\theta}_M = \mathbf{A}_{\mathbf{M}}\mathbf{w}_M$.

The OFS procedure constructs the RBF kernels one by one by minimising the LOO misclassification rate [24]. 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 $D_N \setminus (\mathbf{x}_k, y_k)$, calculated at $\mathbf{x}_k$, as $\hat{y}_k^{(n,-k)}$. Further define the associated LOO decision variable as

$$s_k^{(n,-k)} = \mathrm{sgn}(y_k)\hat{y}_k^{(n,-k)} = y_k\hat{y}_k^{(n,-k)} \tag{11}$$

Then the LOO misclassification rate is defined by

$$J_{\mathrm{LOO}}^{(n)} = \frac{1}{N}\sum_{k=1}^{N}\mathcal{I}_d\big(s_k^{(n,-k)}\big) \tag{12}$$

in which the indicator function $\mathcal{I}_d(s)$ is given by $\mathcal{I}_d(s) = 1$ if $s \leq 0$ and $\mathcal{I}_d(s) = 0$ if $s > 0$. The LOO decision variable can be efficiently calculated according to [24]

$$s_k^{(n,-k)} = \frac{\psi_k^{(n)}}{\eta_k^{(n)}} \tag{13}$$

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

$$\psi_k^{(n)} = \psi_k^{(n-1)} + y_k\theta_n p_n(k) - \frac{p_n^2(k)}{\mathbf{p}_n^{\mathrm{T}}\mathbf{p}_n + \lambda} \tag{14}$$

$$\eta_k^{(n)} = \eta_k^{(n-1)} - \frac{p_n^2(k)}{\mathbf{p}_n^{\mathrm{T}}\mathbf{p}_n + \lambda} \tag{15}$$

where $p_n(k)$ is the kth element of $\mathbf{p}_n$ and $\lambda \geq 0$ 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

$$\{\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n\}_{\mathrm{opt}} = \arg\min_{\boldsymbol{\mu},\boldsymbol{\Sigma}} J_{\mathrm{LOO}}^{(n)}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{16}$$

The PSO algorithm is used to solve this optimisation problem, and the detailed algorithmic steps to determine the nth RBF node's parameters can be found in [24], [37]. The construction of the RBF classifier automatically terminates at the size of M when $J_{\mathrm{LOO}}^{(M+1)} \geq J_{\mathrm{LOO}}^{(M)}$.

IV. Experimental results

The proposed PDFOS+PSO-OFS method was examined on the six data sets summarised in Table I in the order of the ascending imbalanced degree (ID), defined as ID $= N_-/N_+$. The austempered ductile iron (ADI) data set came from [38], while the other five data sets were from [39]. The multiple-class data sets, Glass, Satimage and Yeast, were turned into the two-class problems 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. Different n-fold cross-validations (CVs) were performed on the different data sets. Each dimension of a feature vector $\mathbf{x}_k = \left[x_{k,1}\ x_{k,2}\cdots x_{k,m}\right]^{\mathrm{T}}$ was normalised using

$$\bar{x}_{k,i} = \frac{x_{k,i} - x_{\min,i}}{x_{\max,i} - x_{\min,i}},\ 1 \leq k \leq N, 1 \leq i \leq m \tag{17}$$

with $x_{\min,i} = \min\limits_{1\leq k\leq N} x_{k,i}$ and $x_{\max,i} = \max\limits_{1\leq k\leq N} x_{k,i}$. The mean and standard deviation σ, determined by the PW estimator for the positive class of each data set, averaged over the n-fold CV are also reported in Table I.

Three benchmark algorithms were used. The first benchmark used the same PSO-OFS based RBF classifier applied

TABLE I

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

TABLE II

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	$\mathbf{0.98 \pm 0.02}$

TABLE III

COMPARISON OF MEAN AND STANDARD DEVIATION OF BEST G-MEANS

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

TABLE IV

COMPARISON OF MEAN AND STANDARD DEVIATION OF BEST F-MEASURES

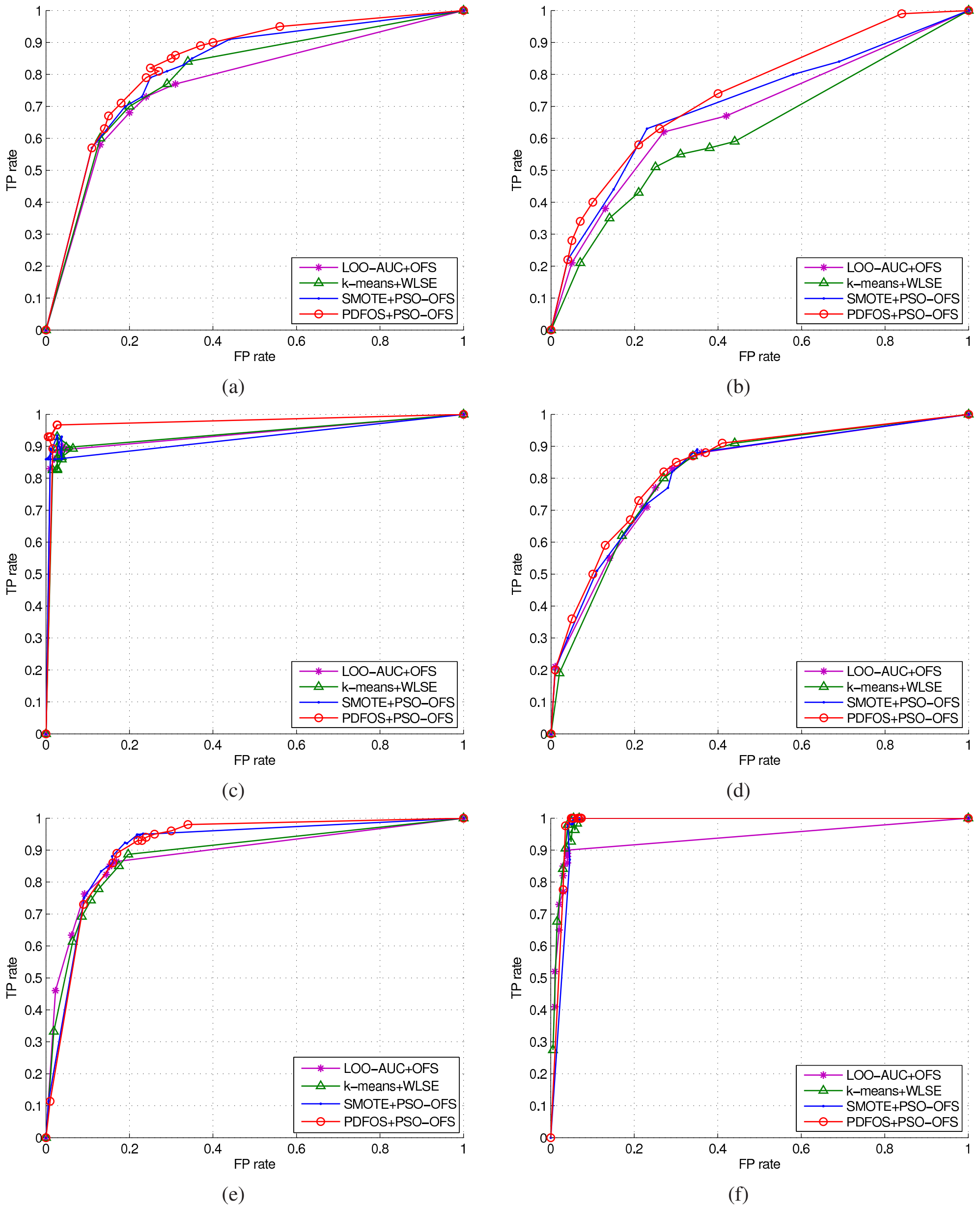
Data set	LOO-AUC+OFS ($\rho =$)	k-means+WLSE ($\rho =$)	SMOTE+PSO-OFS ($r =$)	PDFOS+PSO-OFS ($r =$)
Pima Diabetes	0.67 ± 0.05 (2.0)	0.68 ± 0.06 (2.5)	0.70 ± 0.04 (100%)	$\mathbf{0.71 \pm 0.06}$ (100%)
Haberman's survival	0.52 ± 0.06 (3.0)	0.44 ± 0.11 (4.0)	$\mathbf{0.55 \pm 0.09}$ (200%)	0.54 ± 0.03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

to the SMOTE oversampling data set [37], denoted by the SMOTE+PSO-OFS. The second benchmark [13], denoted by the LOO-AUC+OFS, is a state-of-the-art weighted method. The third benchmark, the κ-means+WLSE algorithm, was also an imbalanced learning method.

Three performance metrics were utilised, and they were the area under the ROC curve (AUC) [40], the G-mean and the F-measure [41]. Receiver operating characteristics (ROC) curves are first presented in Fig. 3, where FP rate and TP rate stand for false positive rate and true positive rate, respectively. The (FP rate, TP rate) pair in the ROC of Fig. 3 is the mean of FP rate and TP rate, respectively, averaged over the n-fold CV. Each algorithm is related to one curve formed by the pairs of (FP rate, TP rate), 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 [40] are then listed in Table II, where the best results are highlighted in boldface. Likewise, the G-mean and F-measure metrics [41] with respect to different r and ρ are reported in Figs. 4 and 5, respectively. 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 III and IV, respectively, where again the best results are highlighted in boldface.

V. CONCLUSIONS

This study has proposed an over-sampling technique that seeks to re-balance the skewed class distribution according to the original statistical information as manifested in the observed data. This has been achieved by a PW PDF estimator using the positive data samples, followed by drawing data samples according to the estimated PDF 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 approach offers a very competitive method, in comparison with many existing state-of-the-art methods for dealing with imbalanced classification problems.



(a)

(b)

(c)

(d)

(e)

(f)

Fig. 3. ROC curves of imbalanced data sets: (a) Pima Indians diabetes, (b) Haberman's survival, (c) Glass, (d) ADI, (e) Satimage, and (f) Yeast.

REFERENCES

[1] N. Petrick, H. P. Chan, B. Sahiner, and D. Wei, "An adaptive density-weighted contrast enhancement filter for mammographic breast mass detection," *IEEE Trans. Medical Imaging*, vol. 15, no. 1, pp. 59–67, 1996.

[2] T. Fawcett and F. Provost, "Adaptive fraud detection," *Data Mining and Knowledge Discovery*, vol. 1, no. 3, pp. 291–316, 1997.

[3] M. Kubat, R. C. Holte, and S. Matwin, "Machine learning for the detection of oil spills in satellite radar images," *Machine Learning*, vol. 30, no. 2-3, pp. 195–215, 1998.

[4] D. D. Lewis and J. Catlett, "Heterogeneous uncertainty sampling for supervised learning," in *Proc. 11th Int. Conf. Machine Learning* (New Brunswick, USA), July 10-13, 1994, pp. 148–156.

[5] C. X. Ling and C. Li, "Data mining for direct marketing: Problems and solutions," in *Proc. 4th Int. Conf. Knowledge Discovery and Data Mining* (New York, USA), August 27-31, 1998, pp. 73–79.

[6] E. P. D. Pednault, B. K. Rosen, and C. Apte, "Handling imbalanced data sets in insurance risk modeling," *IBM Research Report RC-21731*, 2000.

[7] G. M. Weiss and 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, and N. Japkowicz, "A multiple resampling method for learning from imbalanced data sets," *J. Chemical Information and Modeling*, vol. 20, no. 1, pp. 18–36, 2004.

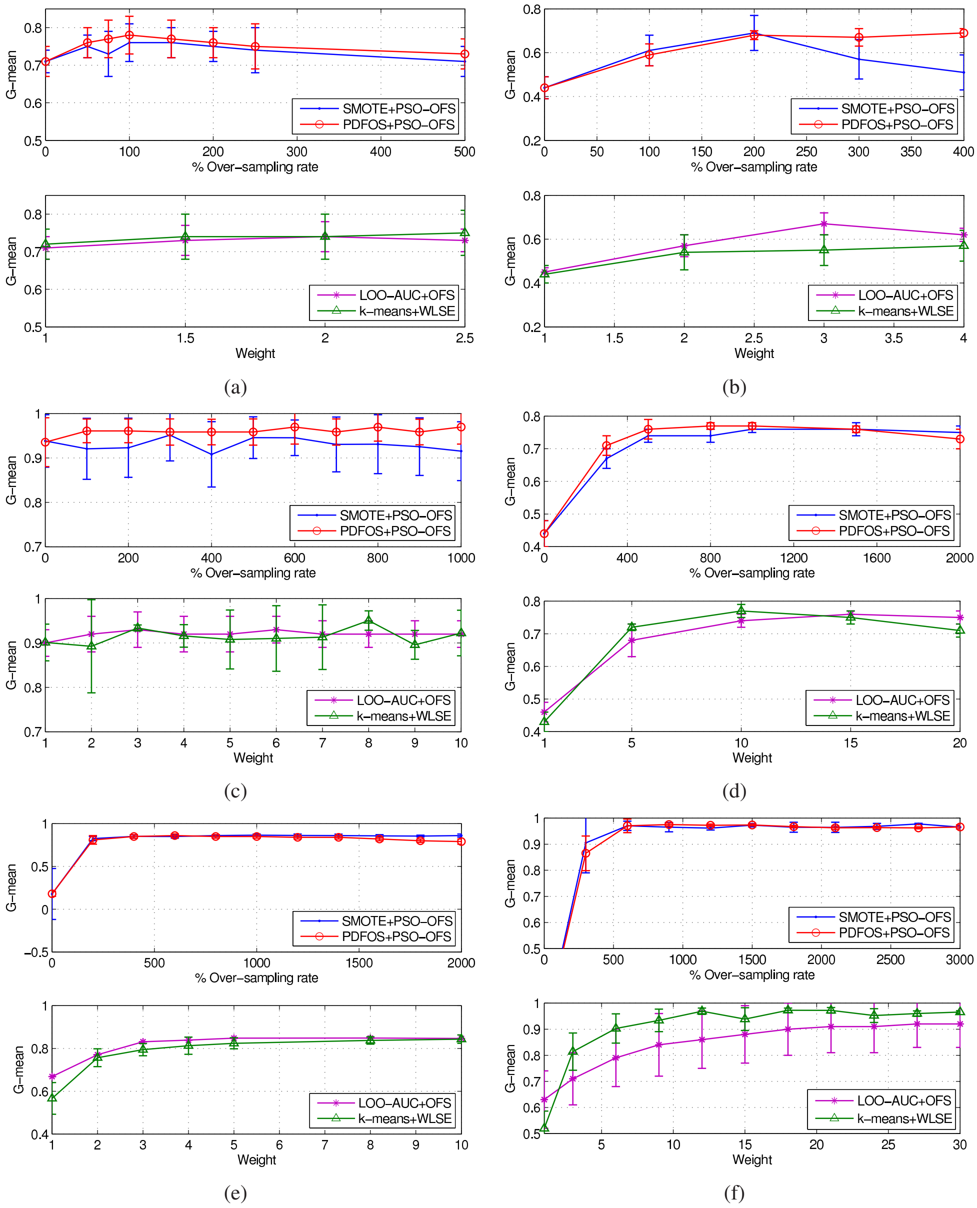
[9] N. Japkowicz and S. Stephen, "The class imbalance problem: A systematic study," *Intelligence Data Analysis*, vol. 6, no. 5, pp. 429–449, 2002.

[10] R. Akbani, S. Kwek, and N. Japkowicz, "Applying support vector machines to imbalanced datasets," in *Proc. 15th European Conf. Machine Learning* (Pisa, Italy), Sept. 20-24, 2004, pp. 39–50.

[11] G. Wu and E. Y. Chang, "KBA: Kernel boundary alignment considering imbalanced data distribution," *IEEE Trans. Knowledge and Data Engineering*, vol. 17, no. 6, pp. 786–795, 2005.

[12] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, 2009.

[13] X. Hong, S. Chen, and C. J. Harris, "A kernel-based two-class classifier for imbalanced data sets," *IEEE Trans. Neural Networks*, vol. 18, no. 1, pp. 28–41, 2007.



(a) (b)

(c) (d)

(e) (f)

Fig. 4. G-means of imbalanced data sets: (a) Pima Indians diabetes, (b) Haberman's survival, (c) Glass, (d) ADI, (e) Satimage, and (f) Yeast.

[14] J. Moody and C. J. Darken, "Fast learning in networks of locally-tuned processing units," *Neural Computation*, vol. 1, No. 2, pp. 281–294, 1989.

[15] S. Haykin, *Neural Networks: A Comprehensive Foundation*, 2nd Edition. Upper Saddle River, NJ: Prentice Hall, 1998.

[16] Y. Sun, M. S. Kamel, A. K. C. Wong, and Y. Wang, "Cost-sensitive boosting for classification of imbalanced data," *Pattern Recognition*, vol. 40, no. 12, pp. 3358–3378, 2007.

[17] W. Fan, S. J. Stolfo, J. Zhang, and P. K. Chan, "AdaCost: Misclassification cost-sensitive boosting," in *Proc. 16th Int. Conf. Machine Learning* (Bled, Slovenia), June 27-30, 1999, pp. 97–105.

[18] D. W. Aha, D. Kibler, and M. K. Albert, "Instance-based learning algorithms," *Machine Learning*, vol. 6, no. 1, pp. 37–66, 1991.

[19] J. Zhang, "Selecting typical instances in instance-based learning," in *Proc. 9th Int. Workshop Machine learning* (Aberdeen, Scotland), July 1-3, 1992, pp. 470–479.

[20] D. B. Skalak, "Prototype and feature selection by sampling and random mutation hill climbing algorithms," in *Proc. 11th Int. Conf. Machine Learning* (New Brunswick, USA), July 10-13, 1994, pp. 293–301.

[21] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[22] N. V. Chawla, A. Lazarevic, L. O. Hall, and K. W. Bowyer, "SMOTEBoost: Improving prediction of the minority class in boosting," in *Proc. 7th European Conf. Principles and Practice of Knowledge Discovery in Databases* (Cavtat-Dubrovnik), Sept. 22-26, 2003, pp. 107–119.

[23] J. Kennedy and R. C. Eberhart, *Swarm Intelligence*. Morgan Kaufmann, 2001.

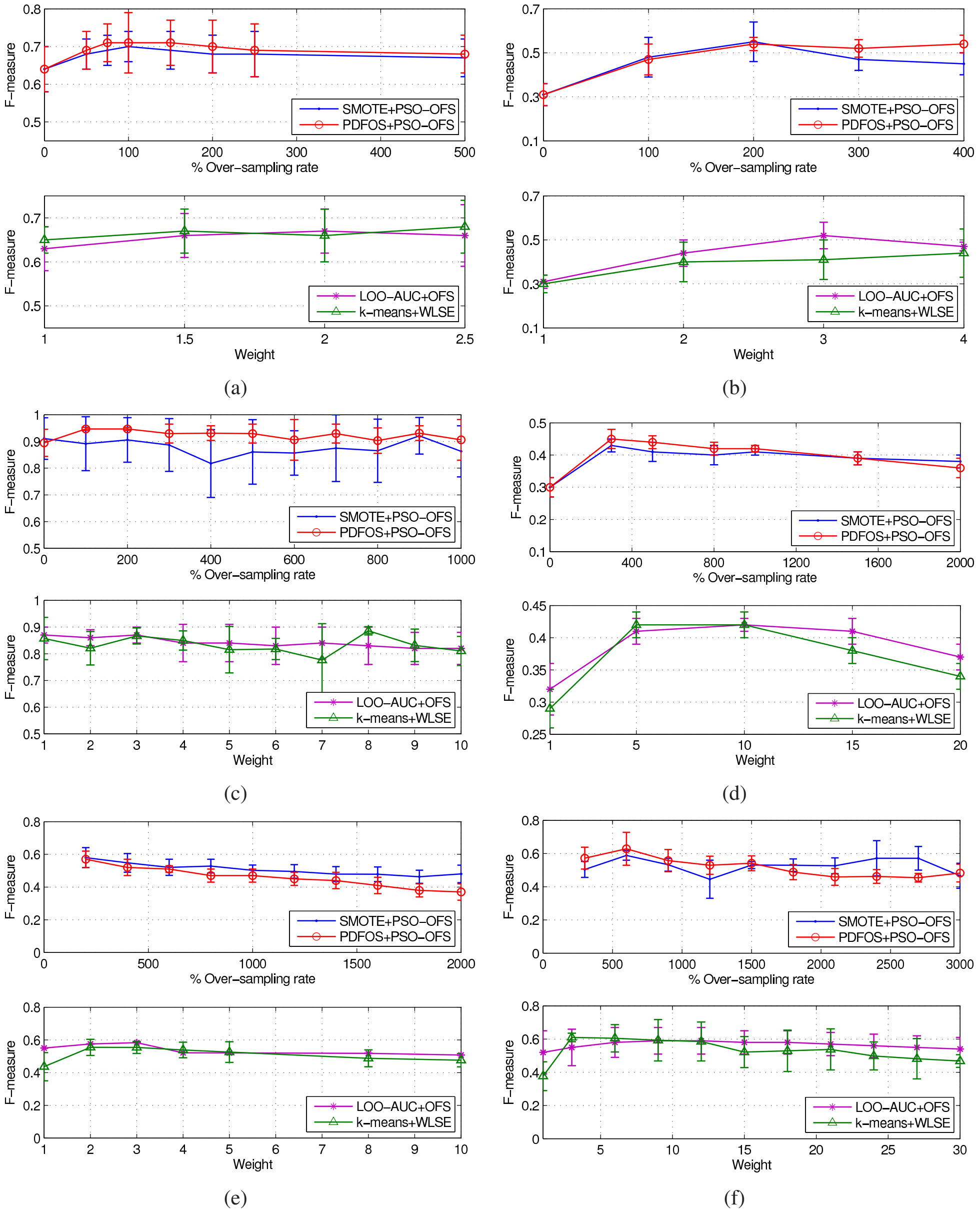
[24] S. Chen, X. Hong, and C. J. Harris, "Particle swarm optimization aided orthogonal forward regression for unified data modelling," *IEEE Trans. Evolutionary Computation*, vol. 14, no. 4, pp. 477–499, 2010.

[25] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," *ACM SIGKDD Exploration Newsletter*, vol. 6, no. 1, pp. 20–29, 2004.

[26] C. Drummond and R. C. Holte, "C4.5, class imbalance, and cost sensitivity: Why under-sampling beats over-sampling," in *Proc. 12th Int. Conf. Machine Learning – Workshop on Learning from Imbalanced Datasets II* (Washington DC, USA), Aug. 21, 2003, pp. 1–8.

[27] S. Floyd and M. Warmuth, "Sample compression, learnability, and the vapnik-chervonenkis dimension." *Machine Learning*, vol. 21, no. 3, pp. 269–304, 1995.

[28] M. Kubat and S. Matwin, "Addressing the curse of imbalanced training



(a)

(b)

(c)

(d)

(e)

(f)

Fig. 5. F-measures of imbalanced data sets: (a) Pima Indians diabetes, (b) Haberman's survival, (c) Glass, (d) ADI, (e) Satimage, and (f) Yeast.

sets: One-sided selection," in *Proc. 14th Int. Conf. Machine Learning* (Nashville, USA), July 8-12, 1997, pp. 179–186.

[29] J. Zhang and I. Mani, "KNN approach to unbalance data distributions: A case study involving information extraction," in *Proc. 12th Int. Conf. Machine Learning - Workshop on Learning from Imbalanced Datasets II* (Washington DC, USA), Aug. 21, 2003, pp. 42–48.

[30] X. Y. Liu, J. Wu, and Z. H. Zhou, "Exploratory undersampling for class-imbalance learning," *IEEE Trans. Systems, Man, and Cybernetics, Part B*, vol. 39, no. 2, pp. 539–550, 2009.

[31] R. Barandela, E. Rangel, J. S. Sánchez, and F. J. Ferri, "Restricted decontamination for the imbalanced training sample problem," in: A. Sanfeliu and J. Ruiz-Shulcloper, Eds., *Progress in Pattern Recognition, Speech and Image Analysis*, Berlin: Springer-Verlag, 2003, pp. 424–431.

[32] R. Barandela, R. M. Valdovinos, J. S. Sánchez, and F. J. Ferri, "The imbalanced training sample problem: Under or over sampling?" in: A. Fred, T. Caelli, R. P. W. Duin, A. Campilho, and D. d. Ridder, Eds., *Structural, Syntactic, and Statistical Pattern Recognition*, Berlin: Springer-Verlag, 2004, pp. 806–814.

[33] B. X. Wang and N. Japkowicz, "Imbalanced data set learning with synthetic samples," in *Proc. IRIS Machine Learning Workshop* (Ottawa, Canada), June 9, 2004.

[34] E. Parzen, "On estimation of a probability density function and mode," *Annals of Mathematical Statistics*, vol. 33, no. 3, pp. 1065–1076, 1962.

[35] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall, 1986.

[36] X. Hong, S. Chen, and C. J. Harris, "An orthogonal forward regression technique for sparse kernel density estimation," *Neurocomputing*, vol. 71, no. 4-6, pp. 931–943, 2008.

[37] M. Gao, X. Hong, S. Chen, and C. J. Harris, "A combined SMOTE and PSO based RBF classifier for two-class imbalanced problems," *Neurocomputing*, vol. 74, no. 17, pp. 3456–3466, 2011.

[38] K. K. Lee, C. J. Harris, S. R. Gunn, and P. A. S. Reed, "Classification of imbalanced data with transparent kernel," in *Proc. 2001 IJCNN* (Washington DC, USA), July 15-19, 2001, pp. 2410–2415.

[39] C. L. Blake and C. J. Merz, "UCI repository of machine learning databases," Department of Computer Science, University of California, 1998. `http://archive.ics.uci.edu/ml/datasets.html`

[40] A. P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," *Pattern Recognition*, vol. 30, pp. 1145–1159, 1997.

[41] C. van Rijsbergen, *Information Retrieval*. London: Butterworths, 1979.