Provided for non-commercial research and education use.
Not for reproduction, distribution or commercial use.

Volume 20, November 2014
ISSN 1566-2535
ELSEVIER
An International Journal on
INFORMATION FUSION
Multi-Sensor, Multi-Source, Multi-Process Information Fusion
Editor-in-Chief
Francisco Herrera
Available online at www.sciencedirect.com
ScienceDirect
Now indexed by Thomson Scientific

This article appeared in a journal published by Elsevier. The attached copy is furnished to the author for internal non-commercial research and education use, including for instruction at the authors institution and sharing with colleagues.

Other uses, including reproduction and distribution, or selling or licensing copies, or posting to personal, institutional or third party websites are prohibited.

In most cases authors are permitted to post their version of the article (e.g. in Word or Tex form) to their personal website or institutional repository. Authors requiring further information regarding Elsevier's archiving and manuscript policies are encouraged to visit:

http://www.elsevier.com/authorsrights

Author's personal copy

Information Fusion 20 (2014) 99–116

Contents lists available at ScienceDirect

Information Fusion

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

RWO-Sampling: A random walk over-sampling approach to imbalanced data classification

Huaxiang Zhang [*], Mingfang Li

Dept. of Computer Science, Shandong Normal University, Jinan 250014, Shandong, China
Shandong Provincial Key Laboratory for Novel Distributed Computer Software Technology, Jinan 250014, Shandong, China

ARTICLE INFO

Article history:
Received 27 July 2011
Received in revised form 7 December 2013
Accepted 30 December 2013
Available online 10 January 2014

Keywords:
Data fusion
Imbalanced data classification
Over-sampling
Random walk
Probability distribution

ABSTRACT

This study investigates how to alleviate the class imbalance problems for constructing unbiased classifiers when instances in one class are more than that in another. Since keeping the data distribution unchanged and expanding class boundaries after synthetic samples have been added influence the classification performance greatly, we take into account the above two factors, and propose a Random Walk Over-Sampling approach (RWO-Sampling) to balancing different class samples by creating synthetic samples through randomly walking from the real data. When some conditions are satisfied, it can be proved that, both the expected average and the standard deviation of the generated samples equal to that of the original minority class data. RWO-Sampling also expands the minority class boundary after synthetic samples have been generated. In this work, we perform a broad experimental evaluation, and experimental results show that, RWO-Sampling statistically does much better than alternative methods on imbalanced data sets when implementing common baseline algorithms.

© 2014 Elsevier B.V. All rights reserved.

1. Introduction

There are many imbalanced data sets [1] in real problems, and the imbalanced datasets share a characteristic that instances in different classes are imbalanced. Classifiers constructed based on imbalanced data sets usually perform well on one class data but bad on the left class data [2]. There exist imbalanced class distributions in many real world domains, and attentions should be paid to the rare class data in these cases. For example, in medical detection, the number of patients who suffer from cancer is about 2% of all patients in a region. If a disease classifier predicts that all the patients do not suffer from cancer, it will perform well and 98 percent of the patients can be correctly diagnosed. However, the classifier cannot tell us the real cancer patients. therefore, high prediction accuracy of the patients suffering from cancer will be more useful to help patients save their lives. For the above mentioned domains, the decision system aims to detect a rare but important case.

Imbalanced data classification problems have attracted great attention, and the approaches proposed to handle imbalanced data problems can be categorized as algorithmic ones and data processing ones according to the methods used to deal with imbalanced data problems. It is reported [3] that Random Over-Sampling (RO-Sampling) and under-sampling are two simple but effective data-level approaches. A number of proposed data-level solutions to the class imbalance problems generally under-sample the majority negative samples or over-sample the minority positive samples or some combinations of the two to alleviate the skew degree of the class distribution on training data [4–6].

In this study, we present a novel over-sampling approach to tackle the imbalanced data classification problems by generating some synthetic minority class samples. The synthetic samples, which appropriately obey the original minority class distribution and expand the minority class boundaries, are combined with the real samples to become a more useful whole dataset, and the whole is used to construct unbiased classifiers. Here we mean an unbiased classifier is the one that the training data used to construct it and the unlabeled data being recognized by it are dominated by the same probability distribution. If the data distribution is known, and an over-sampling approach generates samples according to the given distribution, we may say the classifier constructed using the original training samples and the generated samples are unbiased for the unseen samples. It is common that the distribution is unknown, but if the distribution can be approximated, and synthetic samples are generated by an over-sampling approach based on the approximated distribution, then the likelihood of obtaining an unbiased classifier can be increased. Although we do not know the real data probability distribution, we can calculate the mean and the variance of the training data, and it has been proved [7] that the training data mean is an unbiased

[*] Corresponding author at: Dept. of Computer Science, Shandong Normal University, Jinan 250014, Shandong, China. Tel.: +86 13583192887.
E-mail address: huaxzhang@hotmail.com (H. Zhang).

1566-2535/$ - see front matter © 2014 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.inffus.2013.12.003

Author's personal copy

estimator for the true mean and a confidence interval can be obtained for the true mean. Since the expected average and the variance of the synthetic data generated by our approach approximately equal to that of the original minority positive samples, the classifier trained using original data together with data generated has more potential to be an unbiased one for the unseen data than the one trained using the original data together with synthetic data generated by other over-sampling approaches. We implement three baseline classifiers to compare the performance of the proposed approach with that of RO-Sampling and SMOTE [8]. The performance is evaluated in terms of common metrics, such as *F-measure*, *G-mean*, *TP rate* and *AUC*. Data sets extracted from several sources with different skew degrees have been used.

We review the related work in Section 2, and present the random walk model together with the description of Random Walk Over-Sampling (RWO-Sampling) in Section 3. The proof of theorems concerning synthetic sample generation is also given in this section. We describe the commonly used evaluation metrics and the experimental methodology in Section 4, and show the rich results on broad data sets in Section 5. Conclusions are made in Section 6.

2. Related work

One solution to imbalanced data classification is artificially balancing the training sets by modifying the distribution of the data sets, and the commonly used methods are known as under-sampling and over-sampling approaches respectively.

Over-sampling approaches balance different class instances through synthetic instance generation for the minority class. RO-Sampling duplicates some randomly selected instances from the minority class, and it increases the risk of overfitting. Chawla et al. [8] proposed a novel approach SMOTE to overcome this issue. SMOTE generates new minority class samples by a linear interpolation between two instance neighbors to make instances in different classes balanced. Experimental results show that, SMOTE improves the predictive accuracy of the minority class while not sacrificing the accuracy of the whole testing set. Since SMOTE does not take into account the original data probability distribution when generating data, which may change the original data distribution, the constructed classifiers may be unbiased ones. Batista et al. [3] evaluated ten different under-sampling and over-sampling methods, and concluded that, SMOTE combined with data cleaning approaches such as Tomek links [9] provided good results. Considering the original data distribution, we proposed an empirical method NDO-Sampling to balance samples in different classes [10]. NDO-Sampling generates synthetic samples around the original samples based on the deviations of a normal distribution by randomly choosing one original sample each time. Although it approximately keeps the minority class data mean and variance unchanged as RO-Sampling does according to the unproved conclusions, which may be an important factor for the success of an over-sampling approach, but it still faces a problem of instability, and some samples may be used more than once, and the others may not be used. Randomly selecting n samples from the minority class data of size n with replacement cause some samples not to be picked, each sample has a $1-\frac{1}{n}$ probability of not being selected, and for the n samples, there will be about $n(1-\frac{1}{n})^n \approx 0.368n$ samples not to be used to generate new samples. Thus, the performance of NDO-Sampling may be unstable, i.e., it may obtain good performance based on the averaged results of several experiments, but the deviations of the results may be large.

Since instances falling in class borders make classification difficult, over-sampling instances near the borderline may be helpful when dealing with imbalanced dataset if it can increase the class borders. Han et al. [11] proposed a Borderline-SMOTE method to over-sample the monitory class instances near the class boundaries. Algorithms that attempt to learn the borderline of each class as exactly as possible may perform well, such as support vector machines (SVMs) whose classification hyperplane is determined by the supported vectors. But for those insensitive to the classification borders, Borderline-SMOTE has no positive effects on the performance of the algorithm, and may have negative effects on the algorithm's generalization, since it changes the data distribution. Tomek links [9] can be used as a borderline-like approach which finds the border instances falling in the overlapped areas from the majority class, and performs under-sampling to discard these instances. Guo and Viktor [12] described an approach DataBoost-IM, which generated new data and classified imbalanced data by an ensemble classifier. In DataBoost-IM, hard instances are identified and are used for synthetic sample generation. Hard instances fall in the overlapped areas are used to sample generation for both classes. Batista et al. [3] made a broad experimental evaluation and concluded that, over-sampling methods helped to construct more accurate classifiers than under-sampling methods did.

Ensemble learning technologies and data processing approaches are used to handle imbalanced data classification problems [12–16]. Peng and Yao [14] used adaptive sampling methods to train different weak classifiers, and constructed a strong classifier using these weak classifiers. The algorithm aims to fully use the whole training data. Guo and Viktor [12] used data over-sampling and boosting approaches together to obtain good accuracy for each class. Sun et al. [16] introduced cost items into the learning framework of AdaBoost to handle class imbalance problems. Woźniak et al. [17] reviewed ensemble classifiers from a new point of view including approaches to imbalanced data classification. Sun et al. [4] comparatively studied imbalanced document classification problems implementing SVM on algorithmic level and data level, and found that classical SVM often performs the best most of the time. Li et al. [18] proposed AdaBoostSVM through adjusting a parameter in different iterations, and it was reported that AdaBoostSVM demonstrated better performance than SVM on imbalanced classification problems. Giacinto et al. [19] proposed a modular multiple classifier system (MCS) for intrusion detection. Since the number of false alarms is relatively large compared to the labeled intrusions, they optimize the overall performance by tuning the false alarm rate in each module. The effects of over-sampling and under-sampling have also been studied by Drummond and Holte [20].

To improve the performance of over-sampling or under-sampling, taking into account both the data distribution and class border expansion may be a feasible way. As addition of new samples generated by over-sampling or exclusion of some training samples by under-sampling keeps the data distribution unchanged, the classifiers constructed using the processed data will have high generalization capability, while class border expansion increases the probability that samples in specific classes can be correctly classified. In real classification problems, the data probability distribution that governs the minority class data is unknown, and each minority class instance in a training dataset is drawn independently from the domain with unknown distribution. We do not know whether the synthetic samples obey the unknown distribution when performing an over-sampling approach. Although the minority class data distribution is unknown, the training instances of the minority class still reveal some information about the unknown distribution. The mean and the variance of a random variable obeying an unknown probability distribution are two very important parameters for the distribution. We can calculate the two parameters using the given training samples of the minority class. These two parameters may not be the true mean and variance of the unknown distribution governing the

Author's personal copy

domain from where the samples are drawn, but may fall in confidence intervals of them. In order to construct unbiased classifiers using the original training data and some ones generated through over-sampling, the most obvious and immediate requirement is that the two parameters of the new samples should approximately equal to that of minority positive data. Based on this idea, we propose the novel approach RWO-Sampling. It generates new samples with approximately the same average and standard deviation as that of minority positive data. Thus, classifiers constructed using the original and newly generated data are approximately unbiased to the newly coming test samples.

As mentioned above, over-sampling instances falling in class borders may expand the class border, thus help the data classification. But if the original sample distribution has been changed after over-sampling, classifiers constructed based on the over-sampled training data are biased for the unlabeled instances. So over-sampling approaches should take into account both the class border change and the instance distribution. RO-Sampling may not change the training data distribution, but it over-samples instances by duplicating the original instances, and newly generated instances fall in the same position of their corresponding original instances. This approach does not expand the minority class border. SMOTE generates new instances on the line segment connected by two neighbors, and it changes the data distribution while having little chance to expand the minority class border. RO and SMOTE data processing approaches change the data distributions and do not expand the minority class border, and some will construct biased classifiers based on the combination of the original and the generated training data. These classifiers are not suitable to classify the unlabeled data sampled from the problem domain. After implementing random walk over-sampling, the data borderline may be expanded since newly generated instances may fall outside the coverage space of the original samples, and newly generated instances may approximately obey the original data probability distribution, thus achieving good classification performance on imbalanced datasets.

In this work, we only discuss over-sampling approaches, and no attribute processing methods are considered [21,22].

3. The random walk over-sampling approach

It is stated by the central limit theorem that, given a distribution with a mean μ and a variance σ, if n approaches infinite is satisfied, the average of n samples from the domain governed by the given distribution obeys a normal distribution with mean μ and variance $\frac{\sigma}{\sqrt{n}}$ respectively. The central limit theorem inspires us to generate synthetic minority class samples for unknown data distribution problems. The original training data obey an unknown probability distribution, and we need to generate synthetic data for the minority class under this case. If the newly generated samples also obey the original data distribution, then combination of the original data and the synthetic data will keep the original distribution unchanged. Over-sampling like this has two merits: high generalization capability and algorithm unrelated. Some over-sampling approaches are effective for some specific algorithms but are harmful to the other ones, their performance is evaluated according to the results of some specific baseline classifiers [11,23–25], such as c4.5, SVM and NN (Neural Networks). We call these over-sampling approaches algorithm-related, since the over-sampling methods employed are not always effective when conducting different algorithms.

3.1. Synthetic data generation

Each training instance is described by a tuple of attribute values, and each attribute is regarded as a random variable, which can be assumed to be independent of each other. Although this assumption may not hold in general for many practical problems as it may not hold for Naïve Bayes classification algorithms, but it is effective in solving many classification problems. For a multiple attribute dataset, we calculate the mean and the standard variance for each attribute using the minority class data, and use μ_i and σ_i to denote the calculated mean and standard variance for the ith attribute a_i. Each attribute can be considered as a random variable, and each value of an attribute can be considered as its one sampling value. We use μ_i' and σ_i' to denote its real mean and standard deviation for random variable a_i, and if the number of the minority class instances n approaches infinite, the following holds

$$\frac{\mu_i - \mu_i'}{\sigma_i'/\sqrt{n}} \to N(0,1) \tag{1}$$

Based on the attribute independent assumption, formula (1) holds for all attributes.

Since (1) is satisfied when n approaches infinite, we use it to approximate the parameters of the unknown distribution, and have the following

$$\mu_i' = \mu_i - r \times \frac{\sigma_i'}{\sqrt{n}} \tag{2}$$

where r is a sampling value of distribution $N(0,1)$. By substituting the mean with an attribute value, then for each instance $j (j \in \{1,2,\ldots,n\})$, given its attribute a_i's value $a_i(j)$, a synthetic value $a_i'(j)$ for a_i can be generated by

$$a_i'(j) = a_i(j) - r_j \times \frac{\sigma_i'}{\sqrt{n}}, \quad i \in \{1,2,\ldots,m\}, \quad j \in \{1,2,\ldots,n\} \tag{3}$$

where r_j is a sampling value of $N(0,1)$. $a_i'(j)$ cannot be obtained unless σ_i' is known in (3), so we approximate it with σ_i, and obtain the following

$$a_i'(j) = a_i(j) - r_j \times \frac{\sigma_i}{\sqrt{n}}, \quad i \in \{1,2,\ldots,m\}, \quad j \in \{1,2,\ldots,n\} \tag{4}$$

We use formula (4) to generate synthetic samples, and call it a random walk model, since each synthetic sample is generated by randomly walking from one real sample. The following conclusions can be made for the generated samples.

Theorem 1. *For each attribute, the expected average of the generated values of attribute $a_i (\forall i \in \{1,2,\ldots,m\})$ obtained using (4) approaches μ_i as n approaches infinite.*

Proof. For each instance and each attribute a_i , we create one synthetic value for it using (4). The mean of a_i' can be gotten by

$$E(a_i') = \frac{1}{n}\sum_{j=1}^{n}\left(a_i(j) - \frac{r_j\sigma_i}{\sqrt{n}}\right) = \mu_i - \frac{\sigma_i}{n\sqrt{n}}\sum_{j=1}^{n} r_j \tag{5}$$

As n approaches infinite, the expected value $\frac{1}{n}\sum_{j=1}^{n} r_j$ approaches zero, so we have $E(a_i')$ approaches μ_i as n approaches ∞. □

Theorem 2. *For $\forall i \in \{1,2,\ldots,m\}$, the expected standard deviation of the synthetic values of attribute a_i obtained using (4) approaches σ_i as n approaches infinite.*

Proof. For each instance and each attribute a_i , we create one synthetic value for it using (4). The squared deviation of a_i' can be gotten by

Author's personal copy

$$\frac{1}{n}\sum_{j=1}^{n}\left(a_i'(j) - E(a_i')\right)^2 \tag{6}$$

By substituting $a_i'(j)$ with (4) and $E(a_i')$ with (5) in (6), and merging similar items, we obtain (7) from (6).

$$\begin{aligned}\frac{1}{n}\sum_{j=1}^{n}&\left(a_i(j) - \frac{\sigma_i}{\sqrt{n}}r_j - \mu_i + \frac{\sigma_i}{n\sqrt{n}}\sum_{j=1}^{n}r_j\right)^2\\ &= \frac{1}{n}\sum_{j=1}^{n}(a_i(j)-\mu_i)^2 - \frac{2\sigma_i}{n\sqrt{n}}\sum_{j=1}^{n}a_i(j)r_j + \frac{2\sigma_i\mu_i}{n\sqrt{n}}\sum_{j=1}^{n}r_j\\ &\quad + \left(\frac{\sigma_i}{n}\right)^2\sum_{j=1}^{n}(\bar{r}_i - r_j)^2\end{aligned} \tag{7}$$

where $\bar{r}_i = \sum_{j=1}^{n} r_j$. Since r_j is a sampling value of standard normal distribution, it is obvious that the expected value of the second and the third term in (7) equal zero. In the following, we calculate value of the fourth term in (7).

$$\sum_{j=1}^{n}(\bar{r}_i - r_j)^2 = \sum_{j=1}^{n}r_j^2 - n\bar{r}_i^2 \tag{8}$$

r_j obeys normal distribution $N(0,1)$, and the sum of such n squared random variables obeys χ^2 distribution with freedom n, so the expected value of $\sum_{j=1}^{n} r_j^2$ is n.

The mean of n sampling values drawn from $N(0,1)$ obeys normal distribution $N(0,\frac{1}{n})$. It is easy to know $\bar{r}_i^2$ obeys χ^2 distribution with expected value $\frac{1}{n}$. According to the above discussion, we obtain the expected value of (8)

$$E\left(\sum_{j=1}^{n}(\bar{r}_i - r_j)^2\right) = n - 1 \tag{9}$$

Based on the above discussion, the expected value of (6) can be calculated by

$$\frac{1}{n}\sum_{j=1}^{n}(a_i - \mu_i)^2 + \left(\frac{\sigma_i}{n}\right)^2(n-1) = \sigma_i^2 - \frac{n-1}{n^2}\sigma_i^2 \tag{10}$$

As n approaches infinite, the value of (10) approaches σ_i^2.

For each instance belonging to the minority class, we use (4) to generate one value for each attribute of it, and use these values to form a vector. The vector is assigned the minority class label, and is considered as a synthetic training sample. We can use the above method to generate required number of synthetic samples. The theorems show that, the expected average and variance of the generated data are the same as that of the original minority positive data respectively under some conditions. Classifiers constructed using both the original and the synthetic training samples may perform well on unlabeled samples from the same domain.

In practical problems, that n approaches infinite does not hold in general, especially for class imbalance problems, since many class imbalance problems have few minority class samples. Experimental results in Section 5 show that, RWO-Sampling is still effective on benchmark datasets when this assumption is violated.

The t-distribution could be another alternative of standard normal distribution. In Eq. (1), when we replace σ_i' with σ_i, the distribution is a t-distribution with $n-1$ degrees of freedom, and in this case, σ_i' is not used. t-distribution is symmetric, and the shape of the distribution resembles the bell shape of $N(0,1)$, except that it is a bit lower and wider. As the degrees of freedom becomes large, t-distribution approaches $N(0,1)$. If the degrees of freedom is greater than a particular integer, their difference will not be obvious. Since the conclusions of the above theorems are obtained based on the assumption that n approaches infinite, when n approaches infinite, t-distribution approaches $N(0,1)$, and there will be no difference between t-distribution and $N(0,1)$. □

3.2. RWO-Sampling algorithm description

We denote the training data set T, and the minority class instance set $P = \{x_1,\ldots,x_n\}$. Each x_j represented by m attributes, is a m-dimensional vector representing a point in the m-dimensional space. We name the attribute set $A = \{a_1,\ldots,a_m\}$, and use $a_i(j)$ to denote the value of attribute a_i of instance x_j. s denotes how many times each instance is used to generate synthetic samples. If $s = 1$, it means n synthetic instances are generated, each real instance in P is used one time, and the minority class is over-sampled at over-sampling rate 100%.

RWO-Sampling can deal with continuous attribute values directly when creating synthetic samples, but cannot deal with discrete attributes. For discrete attributes, we use roulette to generate synthetic values for them. We count the frequency of each value for each discrete attribute using the minority class data, and randomly generate a value for each attribute based on the probability calculated for each attribute value. In SMOTE, discrete attributes are handled according to the approach described in [8].

Algorithm 1. RWO-Sampling (T, k: a positive integer).

```
Input: $T,\ k$
Output: $k \times n$ synthetic instances for minority class
for $i = 1$ to $m$
  if $a_i$ is a continuous attribute
    calculating the mean $\mu_i = \frac{\sum_{j=1}^{n} a_i(j)}{n}$
    and the variance $\sigma_i^2 = \frac{1}{n}\sum_{j=1}^{n}(a_i(j) - \mu_i)^2$
  if $a_i$ is a discrete attribute
    calculating the occurrence probability for each value of $a_i$
while($k > 0$)
  for each $x_j \in P$
    for each $a_i \in A$
      if $a_i$ is a continuous attribute
        generating a random value $a_i'(j) = a_i(j) - \frac{\sigma_i}{\sqrt{n}}N(0,1)$ based on (4)
      if $a_i$ is a hybrid attribute
        generating a random value for attribute $a_i$ using roulette
    forming a synthetic instance $(a_1'(j), a_2'(j), \ldots, a_m'(j))$
  $k = k - 1$
return the $k \times n$ instances for the minority class
```

We use points in a two dimensional space to illustrate the differences among RWO-Sampling, SMOTE and RO-Sampling. As shown in Fig. 1, all the synthetic data generated by RO-Sampling fall at the positions of the original data, and the data generated by SMOTE fall on the dashed line segments between pairs of two neighbors. The data introduced by RWO-Sampling are around their original data, and the distances between the synthetic data and their corresponding original data vary from one datum to another. The dashed closed line around the original data represents the boundary of the minority class data. Some generated points, such as 5.1 and 2.1, fall outside the boundary. Since instances generated by RO-sampling are duplications of real instances, they do not expand the minority class coverage space. SMOTE generates synthetic instances through a linear interpolation between two

Author's personal copy

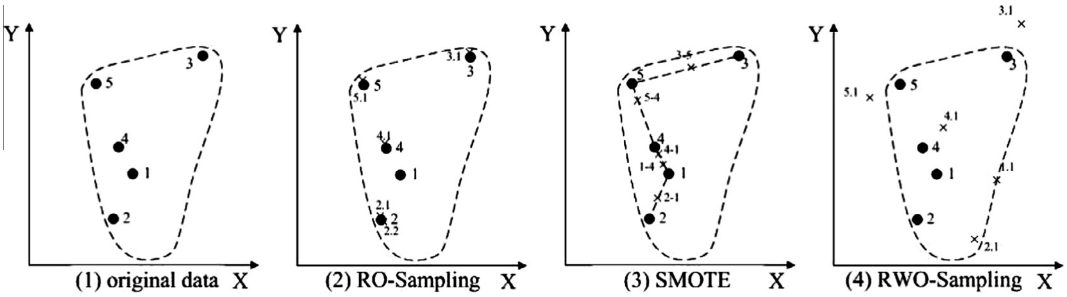

Fig. 1. Illustrations of different over-sampling approaches. Black dots represent the original data, and crosses represent the synthetic data. (1) shows the original data. (2) shows data after RO-Sampling. Point numbered 2 is duplicated 2 times and two synthetic points labeled 2.1 and 2.2 are created using it. (3) synthetic points labeled 2–4 are created using instance 2 with its neighbor point 4. (4) point labeled 5.1 is generated using point 5 based on RWO-Sampling.

near instances, and it still has little chance to expand the minority class coverage space. RWO-Sampling uses random walk to generate synthetic instances, thus it has more opportunities to expand the positive class border than SMOTE and RO-Sampling, and increases the positive class classification accuracy.

4. Evaluation metrics and experimental methodology

The performance of a classifier is commonly evaluated based on its global accuracy on an independent test dataset. Since the overall classification accuracy on an imbalanced dataset is mainly dominated by the majority class, researchers use different metrics to evaluate the performance of imbalanced data classification approaches. The metrics are the accuracy rate [26], F-measure [27], Geometric Mean (G-mean) [28], the AUC and the overall accuracy. In this paper, the positive and negative class refer to the minority and majority class respectively. *TP* and *FP* are the number of true positive and false positive respectively, and *TN* and *FN* are the number of true negative and false negative respectively. For the minority class data, its $precision = \frac{TP}{TP+FP}$, and $recall = TPrate = \frac{TP}{TP+FN}$. For the majority class data, its $precision = \frac{TN}{TN+FN}$, and $recall = TNrate = \frac{TN}{TN+FP}$. Positive accuracy $pa = \frac{TP}{TP+FN}$ and negative accuracy $na = \frac{TN}{TN+FP}$ are used to calculate $G-Mean = \sqrt{pa * na}$. $F-Measure = \frac{(1+\beta^2)\times recall \times precision}{\beta^2 \times recall + precision}$, where β is set to 1 in this paper. *F-Measure* for each class is calculated based on its corresponding *precision* and *recall*. We also describe the overall accuracy $\left(\frac{TP+TN}{TP+FN+FP+TN}\right)$ in the paper. Since each metric represents only one aspect of a learning algorithm, we calculate the *F-measure* for both classes, and obtain the overall classification accuracy together with *G-mean* and *TP rate* to evaluate different over-sampling approaches. AUCs are also calculated for different classifiers under different over-sampling approaches.

Extensive experiments are done on selected benchmark datasets[1] by implementing three baseline classifier algorithms C4.5, NB and KNN($k = 3$). The datasets characterized in Table 1 are extensively used for evaluating classification approaches. We do experiments to compare RWO-Sampling with the others, and obtain *F-measure*, *TP rate*, *G-mean*, *overall accuracy* and AUC for each over-sampling approach under different imbalance ratios by implementing the baseline classifier algorithms. We implement a ten-fold cross validation when constructing a classifier, each experiment is repeated ten times, and both the average and the variance of the results are obtained. We also evaluate each approach on imbalanced multi-class problems, that means the number of instances in one class is the smallest. We also evaluate different over-sampling approaches together with randomly under-sampling approaches in a separate group of experiments, and random under-sampling is performed on the majority class by randomly eliminating instances from the majority class.

Data sets are divided into three categories: data with all continuous attributes, data with all discrete attributes and data with both continuous and discrete attributes (hybrid attributes). The over-sampling rate is set to 100%, 200%, 300%, 400%, and 500% for data with only continuous attributes, and the over-sampling rates 200% and 300% are considered for data with all discrete and hybrid attributes. The nearest neighbor number is set to 5 for SMOTE. If an attribute is discrete, the missing value of an attribute is assigned a value with the most occurrence frequency of that attribute. If an attribute is continuous, its missing value is assigned the mean of that attribute.

We do differently to transform a multi-class dataset into an imbalanced binary class dataset. The smallest class of Glass is considered as the positive class, and all the other classes are considered as the negative class. A similar transformation is made on Satimage, Segment, Vehicle and Vowel. For Hypothyroid, Postoperative-patient-data, and Primary-tumor, we eliminate all classes but the biggest and the smallest class.

We implement the selected baseline algorithms under different over-sampling rates on Weka platform[2]. The following abbreviations are used in tables and figures of section five: *F-measure* for the minority and majority class are abbreviated as F-min and F-maj respectively, and the overall accuracy and algorithm are abbreviated as O-acc and Alg respectively; RWO-Sampling, SMOTE and RO-Sampling are abbreviated as RWO, SMO and RO respectively, and Over-Sampling strategy is abbreviated as OS.

5. Experimental results

Classifiers are trained using the synthetic data together with the original data. That means classifiers are trained on balanced dataset. A ten-fold cross validation scheme is used to evaluate the performance of each over-sampling approach under different over-sampling rates when different classical classifiers are conducted, and the values in boldface in the following tables indicate the optimal ones. The results described in tables from 2 to 7 show that, each over-sampling strategy performs differently under different over-sampling rates and when different classification algorithms are implemented. We use Diabets to explain the results. Table 2 show that, RWO-Sampling performs the best in all evaluation metrics when C4.5 is implemented, but SMOTE outperforms the others in four metrics when implementing Naïve Bayes classifier. This does not keep unchanged under different over-sampling rates. From the results shown in Table 3 we know that, on the same

[1] http://www.ics.uci.edu/mlearn/MLRepository.html

[2] http://www.cs.waikato.ac.nz/ml/weka

Author's personal copy

Table 1
Dataset description.

Dataset	# Of instances	# Of positive instances (%)	# Of attributes	# Of discrete attributes	Positive class label
Breast-cancer	286	85	9	9	Recurrence-events
Breast-w	699	241	9	0	Malignant
Colic	368	136	24	15	Surgical_lesioin = no
Credit-g	1000	300	20	13	Bad
Diabetes	768	268	8	0	Tested_positive
Glass	214	9	9	0	Type = tableware
Haberman	306	81	3	1	Survival_status = 2
Hepatitis	155	32	19	13	DIE
Hypothyroid	3675	194	29	22	Compensated_hypothyroid
Ionosphere	351	126	34	0	b
Monk2	169	64	6	6	1
Postoperative	88	24	8	8	Decision = S
Primary-tumor	104	20	17	17	Esophagus
Satimage	6430	625	36	0	2
Segment	1500	205	19	0	brickface
Sick	3772	231	29	22	Sick
Sonar	208	97	60	0	Rock
Splice-i.e.	2423	768	60	60	IE
Tic-tac-toe	958	332	9	9	Negative
Vehicle	846	199	18	0	Van
Vowel	990	90	13	3	Hid

Table 2
Averaged results and standard deviations on 8 continuous attribute datasets (over-sampling rate equals 100%).

Dataset	Alg	OS	(%)				
			F-min	F-maj	O-acc	G-mean	TP rate
Breast-w	C4.5	RWO	**93.12** ± 2.28	**96.24** ± 1.76	**95.14** ± 2.09	**95.23** ± 1.93	**95.52** ± 2.17
		SMO	92.80 ± 2.28	96.11 ± 2.14	94.95 ± 2.20	94.83 ± 2.05	94.47 ± 2.15
		RO	92.39 ± 2.13	95.94 ± 2.34	94.71 ± 2,14	94.35 ± 1.85	93.22 ± 2.15
	NB	RWO	94.38 ± 1.63	96.89 ± 1.86	95.99 ± 1.45	96.35 ± 1.86	97.51 ± 1.79
		SMO	94.38 ± 1.52	96.89 ± 1.64	95.99 ± 1.77	96.35 ± 1.86	97.51 ± 1.76
		RO	94.38 ± 2.08	96.89 ± 2.07	95.99 ± 1.75	96.35 ± 1.62	97.51 ± 1.92
	3-NN	RWO	95.94 ± 1.71	97.77 ± 1.51	97.12 ± 1.49	97.45 ± 1.57	98.51 ± 1.28
		SMO	**96.36** ± 1.72	**98.01** ± 1.45	**97.42** ± 1.80	**97.73** ± 1.45	**98.76** ± 1.50
		RO	95.56 ± 1.62	97.62 ± 1.47	96.90 ± 1.54	96.88 ± 1.62	96.82 ± 1.61
Diabetes	C4.5	RWO	**67.31** ± 4.26	**78.52** ± 4.04	**74.08** ± 4.25	**74.61** ± 4.07	**76.49** ± 4.05
		SMO	66.30 ± 4.04	77.78 ± 4.05	73.22 ± 4.02	73.73 ± 4.03	75.50 ± 4.00
		RO	60.52 ± 4.08	76.88 ± 4.07	70.83 ± 4.09	69.06 ± 4.07	64.05 ± 4.07
	NB	RWO	65.93 ± 4.07	79.73 ± 4.02	74.58 ± 4.08	73.57 ± 4.00	**70.49** ± 4.09
		SMO	**66.27** ± 4.04	**80.25** ± 4.09	**75.09** ± 4.05	**73.84** ± 4.04	70.15 ± 4.00
		RO	65.57 ± 4.00	79.95 ± 4.01	74.65 ± 4.09	73.26 ± 4.01	69.15 ± 4.07
	3-NN	RWO	65.03 ± 4.40	75.61 ± 4.19	71.26 ± 4.21	72.38 ± 4.13	**76.57** ± 4.43
		SMO	**65.14** ± 4.43	**76.12** ± 4.19	**71.66** ± 4.45	**72.56** ± 4.07	75.87 ± 4.28
		RO	60.07 ± 4.04	74.21 ± 4.40	68.66 ± 4.09	68.40 ± 4.06	67.54 ± 4.21
Glass	C4.5	RWO	94.74 ± 2.09	99.76 ± 2.41	99.53 ± 1.97	99.76 ± 2.02	100.00 ± 1.89
		SMO	94.74 ± 2.83	99.76 ± 2.50	99.53 ± 2.41	99.76 ± 2.15	100.00 ± 2.60
		RO	94.74 ± 2.60	99.76 ± 2.42	99.53 ± 2.25	99.76 ± 2.33	100.00 ± 2.65
	NB	RWO	68.44 ± 2.99	97.93 ± 3.12	96.12 ± 3.04	**97.95** ± 2.90	**100.00** ± 3.17
		SMO	**69.44** ± 2.98	**98.18** ± 2.96	**96.57** ± 2.93	94.65 ± 2.74	92.59 ± 3.49
		RO	64.86 ± 2.76	97.85 ± 2.78	95.95 ± 2.82	92.50 ± 2.66	88.89 ± 3.09
	3-NN	RWO	73.30 ± 1.82	98.55 ± 2.42	97.24 ± 1.76	**93.70** ± 2.33	**90.00** ± 2.05
		SMO	**73.85** ± 1.94	**98.61** ± 1.80	**97.35** ± 2.30	93.20 ± 2.16	88.89 ± 2.41
		RO	71.64 ± 2.14	98.44 ± 1.90	97.04 ± 2.31	93.05 ± 2.25	88.89 ± 2.19
Ionosphere	C4.5	RWO	86.71 ± 2.73	92.72 ± 2.76	90.60 ± 2.47	89.38 ± 2.54	85.48 ± 2.60
		SMO	85.45 ± 2.35	91.73 ± 2.91	89.46 ± 2.79	88.72 ± 2.51	**86.24** ± 2.61
		RO	**87.60** ± 2.49	**93.26** ± 2.26	**91.26** ± 2.57	**90.01** ± 2.50	85.98 ± 2.84
	NB	RWO	**82.34** ± 3.58	**89.62** ± 3.66	**86.92** ± 3.68	**86.47** ± 3.75	84.92 ± 3.55
		SMO	80.54 ± 3.74	87.79 ± 3.65	85.00 ± 3.69	85.32 ± 3.66	86.51 ± 3.67
		RO	77.09 ± 3.54	84.46 ± 3.29	81.48 ± 3.64	82.54 ± 3.31	**86.77** ± 3.45
	3-NN	RWO	85.79 ± 3.04	92.78 ± 2.78	90.43 ± 2.91	87.90 ± 2.68	80.48 ± 2.93
		SMO	**90.61** ± 2.73	**94.97** ± 2.74	**93.45** ± 3.05	**92.18** ± 3.08	**88.10** ± 2.55
		RO	82.11 ± 2.85	91.43 ± 2.94	88.41 ± 2.93	84.52 ± 3.37	74.07 ± 2.86
Satimage	C4.5	RWO	63.38 ± 2.35	95.56 ± 2.33	92.08 ± 2.11	**81.61** ± 2.30	**70.56** ± 2.26
		SMO	**63.39** ± 2.48	**95.64** ± 2.51	**92.20** ± 2.17	81.07 ± 2.40	69.44 ± 2.57
		RO	55.57 ± 2.32	95.18 ± 2.42	91.30 ± 2.24	72.98 ± 2.49	56.00 ± 2.38
	NB	RWO	47.98 ± 3.98	88.78 ± 3.58	81.54 ± 3.92	**84.18** ± 3.57	**87.60** ± 3.72
		SMO	**48.87** ± 3.90	**89.40** ± 3.33	**82.44** ± 3.82	84.15 ± 3.79	86.35 ± 3.68
		RO	47.96 ± 3.79	88.81 ± 3.77	81.58 ± 3.81	84.10 ± 3.86	87.36 ± 3.72
	3-NN	RWO	68.90 ± 2.35	95.74 ± 2.00	92.50 ± 1.72	89.27 ± 1.39	85.44 ± 1.91

Author's personal copy

Table 2 (*continued*)

Dataset	Alg	OS	(%)				
			F-min	F-maj	O-acc	G-mean	TP rate
		SMO	**72.17** ± 2.42	**96.08** ± 2.20	**93.12** ± 1.81	**92.52** ± 2.04	**91.79** ± 2.09
		RO	67.18 ± 2.46	95.75 ± 2.10	92.47 ± 2.02	86.26 ± 1.53	79.25 ± 1.99
Segment	C4.5	RWO	**97.23** ± 0.76	**99.56** ± 0.57	**99.24** ± 1.10	98.59 ± 0.88	97.71 ± 0.46
		SMO	96.92 ± 0.76	99.51 ± 0.89	99.16 ± 0.98	98.27 ± 1.07	97.07 ± 0.90
		RO	96.86 ± 1.10	99.50 ± 0.61	99.13 ± 0.92	**98.61** ± 1.01	**97.89** ± 0.91
	NB	RWO	**65.47** ± 1.61	**91.79** ± 2.15	**86.73** ± 1.93	**89.55** ± 2.16	**98.37** ± 1.98
		SMO	62.44 ± 2.02	89.69 ± 1.69	83.82 ± 2.14	88.91 ± 1.86	92.03 ± 1.66
		RO	61.24 ± 2.10	89.08 ± 2.03	82.96 ± 2.06	89.06 ± 1.85	98.04 ± 1.74
	3-NN	RWO	**97.91** ± 0.61	**99.66** ± 0.84	**99.42** ± 1.04	**99.35** ± 0.92	**99.24** ± 0.96
		SMO	97.44 ± 0.88	99.59 ± 0.89	99.29 ± 0.91	99.11 ± 0.92	98.86 ± 1.01
		RO	96.90 ± 1.10	99.50 ± 0.88	99.13 ± 1.09	99.16 ± 0.94	99.19 ± 0.83
Sonar	C4.5	RWO	**82.90** ± 4.31	**80.62** ± 4.79	**81.83** ± 4.25	**81.77** ± 4.24	**94.43** ± 5.09
		SMO	78.15 ± 4.69	77.94 ± 4.57	78.04 ± 4.80	78.22 ± 4.69	84.19 ± 4.69
		RO	70.63 ± 4.22	76.48 ± 4.61	73.88 ± 4.97	73.21 ± 4.86	67.35 ± 5.03
	NB	RWO	**71.85** ± 4.85	63.77 ± 5.02	**68.32** ± 4.76	67.31 ± 4.81	**86.70** ± 5.03
		SMO	70.40 ± 5.09	64.67 ± 5.26	67.79 ± 4.45	67.37 ± 5.03	82.13 ± 4.72
		RO	70.86 ± 5.19	**64.78** ± 5.09	68.11 ± 5.55	**67.60** ± 5.16	83.16 ± 4.86
	3-NN	RWO	85.96 ± 2.76	85.96 ± 3.03	85.96 ± 2.48	86.16 ± 2.35	92.16 ± 2.77
		SMO	**88.31** ± 3.24	**88.61** ± 2.89	**88.46** ± 2.71	**88.65** ± 2.85	**93.47** ± 2.90
		RO	87.31 ± 2.54	87.99 ± 2.72	87.66 ± 2.77	87.82 ± 2.55	91.07 ± 2.78
Vehicle	C4.5	RWO	**88.47** ± 2.91	**96.25** ± 3.31	**94.34** ± 2.90	**93.64** ± 3.36	**92.36** ± 2.73
		SMO	86.80 ± 3.10	95.72 ± 3.72	93.54 ± 3.12	92.39 ± 3.28	90.28 ± 3.09
		RO	86.76 ± 2.89	95.79 ± 3.28	93.62 ± 2.68	91.95 ± 3.23	88.94 ± 2.83
	NB	RWO	55.66 ± 4.07	72.06 ± 3.83	65.72 ± 4.44	72.71 ± 4.30	91.46 ± 3.79
		SMO	56.26 ± 3.97	72.35 ± 4.11	66.12 ± 4.15	73.27 ± 4.39	92.63 ± 4.02
		RO	**54.52** ± 3.78	**71.93** ± 3.52	**65.29** ± 4.47	**71.72** ± 3.60	**88.44** ± 4.24
	3-NN	RWO	86.40 ± 2.74	95.26 ± 2.98	92.97 ± 2.89	**93.64** ± 2.95	**94.92** ± 2.97
		SMO	**87.00** ± 2.81	**95.56** ± 2.83	**93.38** ± 3.34	93.64 ± 2.72	94.14 ± 2.44
		RO	85.29 ± 2.78	94.98 ± 3.34	92.51 ± 2.58	92.44 ± 2.98	92.29 ± 2.40

Table 3
Averaged results and standard deviations on 8 continuous attribute datasets (over-sampling rate equals 200%).

dataset	Alg	OS	(%)				
			F-min	F-maj	O-acc	G-mean	TP rate
Breast-w	C4.5	RWO	**92.95** ± 1.96	**96.10** ± 2.15	**94.98** ± 1.76	**95.22** ± 1.73	**96.02** ± 1.93
		SMO	92.71 ± 2.45	96.02 ± 1.99	94.85 ± 2.00	94.89 ± 1.89	95.02 ± 1.88
		RO	92.14 ± 1.82	95.73 ± 1.87	94.47 ± 2.10	94.37 ± 2.01	94.05 ± 2.11
	NB	RWO	94.38 ± 1.64	96.89 ± 1.83	95.99 ± 1.61	96.35 ± 1.80	97.51 ± 1.71
		SMO	94.38 ± 1.72	96.89 ± 1.88	95.99 ± 1.85	96.35 ± 1.82	97.51 ± 1.79
		RO	94.38 ± 1.87	96.89 ± 1.79	95.99 ± 1.88	96.35 ± 1.82	97.51 ± 1.75
	3-NN	RWO	96.39 ± 1.48	98.00 ± 1.21	97.42 ± 1.47	**97.96** ± 1.21	**99.75** ± 1.50
		SMO	**96.50** ± 1.21	**98.08** ± 1.23	**97.52** ± 1.32	97.90 ± 1.37	99.17 ± 1.51
		RO	95.21 ± 1.31	97.38 ± 1.52	96.61 ± 1.58	96.82 ± 1.37	97.51 ± 1.47
Diabetes	C4.5	RWO	**66.43** ± 3.98	**77.04** ± 4.43	**72.73** ± 4.03	**73.71** ± 3.99	77.31 ± 4.14
		SMO	64.84 ± 3.93	74.50 ± 4.44	70.44 ± 3.51	71.98 ± 4.61	**78.11** ± 3.90
		RO	60.18 ± 4.25	75.66 ± 3.65	69.79 ± 3.55	68.70 ± 3.69	65.42 ± 4.44
	NB	RWO	**66.74** ± 3.92	**77.78** ± 3.73	**73.36** ± 3.27	**74.07** ± 3.99	**76.60** ± 3.39
		SMO	65.59 ± 3.71	76.71 ± 4.31	72.22 ± 3.40	73.01 ± 3.45	75.87 ± 4.05
		RO	65.35 ± 3.86	77.36 ± 4.14	72.61 ± 3.60	72.93 ± 4.64	74.00 ± 3.62
	3-NN	RWO	**74.42** ± 4.09	**78.38** ± 4.23	**76.56** ± 4.29	**79.83** ± 4.14	**97.69** ± 4.19
		SMO	65.73 ± 4.12	73.68 ± 4.17	70.23 ± 4.27	72.37 ± 4.37	81.84 ± 4.28
		RO	60.54 ± 4.30	69.69 ± 4.41	65.71 ± 4.45	67.55 ± 4.37	75.37 ± 4.03
Glass	C4.5	RWO	94.74 ± 2.15	99.76 ± 2.27	99.53 ± 2.47	99.76 ± 1.96	100.00 ± 2.34
		SMO	94.74 ± 2.53	99.76 ± 2.29	99.53 ± 1.92	99.76 ± 2.75	100.00 ± 2.33
		RO	94.74 ± 2.71	99.76 ± 2.55	99.53 ± 2.36	99.76 ± 2.23	100.00 ± 2.59
	NB	RWO	66.91 ± 2.69	97.78 ± 2.58	95.84 ± 2.56	**97.81** ± 2.68	**100.00** ± 2.11
		SMO	**71.64** ± 3.09	**98.44** ± 2.88	**97.04** ± 2.44	93.05 ± 2.94	88.89 ± 2.82
		RO	60.76 ± 3.19	97.43 ± 2.04	95.17 ± 2.47	92.11 ± 3.47	88.89 ± 2.41
	3-NN	RWO	**74.69** ± 2.26	**98.49** ± 1.67	**97.15** ± 2.53	**98.50** ± 1.27	**100.00** ± 1.46
		SMO	69.57 ± 2.43	98.27 ± 1.91	96.73 ± 1.98	92.89 ± 1.95	88.89 ± 2.30
		RO	62.16 ± 2.65	97.69 ± 1.99	95.64 ± 2.38	90.48 ± 1.77	85.19 ± 2.21
Ionosphere	C4.5	RWO	**86.40** ± 1.96	**92.82** ± 2.38	**90.60** ± 1.79	88.78 ± 2.10	83.17 ± 1.69
		SMO	86.43 ± 2.51	91.76 ± 2.64	89.74 ± 2.39	**90.02** ± 2.16	**91.01** ± 2.28
		RO	83.99 ± 2.95	90.92 ± 2.59	88.41 ± 2.38	87.54 ± 2.03	84.66 ± 2.88
	NB	RWO	**82.63** ± 3.70	**89.84** ± 3.18	**87.18** ± 3.21	**86.66** ± 3.26	84.92 ± 3.05
		SMO	81.44 ± 3.48	88.56 ± 3.44	85.85 ± 3.41	85.99 ± 3.16	86.51 ± 4.16
		RO	76.74 ± 3.49	83.95 ± 3.90	81.01 ± 4.04	82.25 ± 3.38	**87.30** ± 4.51

(*continued on next page*)

Author's personal copy

Table 3 (*continued*)

dataset	Alg	OS	(%)				
			F-min	F-maj	O-acc	G-mean	TP rate
	3-NN	RWO	**95.57** ± 2.39	**97.40** ± 2.21	**96.72** ± 2.32	**97.10** ± 2.88	**98.49** ± 2.70
		SMO	95.02 ± 2.55	97.05 ± 2.36	96.30 ± 3.03	96.75 ± 2.84	98.41 ± 3.24
		RO	84.00 ± 3.02	92.03 ± 2.82	89.36 ± 1.96	86.34 ± 3.05	77.78 ± 2.97
Satimage	C4.5	RWO	**68.70** ± 1.75	**96.05** ± 2.16	**92.99** ± 1.85	**86.50** ± 2.27	**79.20** ± 2.17
		SMO	63.91 ± 2.48	95.58 ± 2.52	92.13 ± 2.80	82.23 ± 2.22	71.68 ± 2.30
		RO	57.25 ± 1.96	95.25 ± 2.71	91.45 ± 2.56	74.77 ± 2.21	58.88 ± 2.32
	NB	RWO	47.59 ± 3.88	88.57 ± 3.77	81.23 ± 3.82	84.03 ± 3.80	87.68 ± 3.34
		SMO	**48.76** ± 4.31	**89.32** ± 3.95	**82.33** ± 4.07	**84.16** ± 3.90	86.51 ± 3.63
		RO	47.52 ± 4.03	88.50 ± 3.61	81.13 ± 3.28	84.06 ± 3.75	**87.89** ± 3.32
	3-NN	RWO	**74.67** ± 1.34	**96.21** ± 1.23	**93.41** ± 1.31	**96.28** ± 1.34	**100.00** ± 1.09
		SMO	71.40 ± 2.37	95.58 ± 1.27	92.34 ± 1.41	94.98 ± 1.59	98.40 ± 1.60
		RO	65.55 ± 1.48	95.08 ± 1.50	91.39 ± 1.62	88.12 ± 1.60	84.27 ± 2.19
Segment	C4.5	RWO	97.82 ± 0.61	99.65 ± 0.86	99.40 ± 0.62	**98.97** ± 0.80	**98.39** ± 0.63
		SMO	**97.96** ± 0.89	**99.68** ± 0.87	**99.44** ± 0.67	98.65 ± 0.64	97.56 ± 0.79
		RO	96.84 ± 0.61	99.50 ± 0.83	99.13 ± 0.70	98.26 ± 0.81	97.07 ± 0.74
	NB	RWO	**66.55** ± 1.93	**92.22** ± 1.53	**87.38** ± 1.27	**89.23** ± 1.77	**99.02** ± 1.94
		SMO	61.90 ± 2.06	89.41 ± 2.20	83.42 ± 1.42	89.36 ± 1.80	98.54 ± 2.03
		RO	61.67 ± 2.23	89.22 ± 1.90	83.18 ± 2.16	89.38 ± 2.16	98.87 ± 2.30
	3-NN	RWO	**97.82** ± 0.67	**99.65** ± .061	**99.39** ± 0.99	**99.62** ± 0.78	**99.95** ± 0.61
		SMO	97.53 ± 0.83	99.60 ± 0.95	99.31 ± 0.66	99.33 ± 0.80	99.35 ± 0.87
		RO	96.54 ± 0.92	99.43 ± 0.69	99.02 ± 0.79	99.30 ± 0.88	99.67 ± 0.62
Sonar	C4.5	RWO	**84.40** ± 4.73	**82.07** ± 5.10	**83.32** ± 4.77	**83.21** ± 5.18	**96.80** ± 4.92
		SMO	81.75 ± 4.96	80.72 ± 5.10	81.25 ± 5.02	81.39 ± 5.02	90.03 ± 4.91
		RO	72.70 ± 4.74	76.67 ± 5.04	74.84 ± 4.95	74.60 ± 5.10	71.82 ± 5.11
	NB	RWO	**73.90** ± 4.75	**65.13** ± 4.93	**70.14** ± 4.73	**68.81** ± 4.95	**90.62** ± 4.87
		SMO	70.41 ± 4.77	65.03 ± 5.19	67.95 ± 4.93	67.59 ± 5.09	81.79 ± 4.95
		RO	71.14 ± 4.80	64.77 ± 5.18	68.27 ± 5.01	67.70 ± 5.07	83.85 ± 4.70
	3-NN	RWO	**89.23** ± 2.09	**88.43** ± 2.84	**88.85** ± 2.83	**88.98** ± 2.72	**99.07** ± 2.86
		SMO	87.66 ± 2.76	87.34 ± 2.77	87.50 ± 3.08	87.69 ± 2.99	95.19 ± 2.91
		RO	86.54 ± 2.74	86.54 ± 2.89	86.54 ± 3.06	86.74 ± 2.80	92.78 ± 3.07
Vehicle	C4.5	RWO	**87.17** ± 2.95	**95.70** ± 3.12	**93.56** ± 3.11	**93.39** ± 2.77	**93.07** ± 2.71
		SMO	86.40 ± 3.28	95.56 ± 3.09	93.30 ± 3.09	92.30 ± 3.29	90.45 ± 3.03
		RO	86.96 ± 2.88	95.91 ± 3.02	93.77 ± 2.80	91.80 ± 2.80	88.27 ± 3.22
	NB	RWO	**56.63** ± 4.40	**71.98** ± 4.27	**65.96** ± 4.13	**73.50** ± 4.11	**94.47** ± 4.07
		SMO	56.27 ± 4.11	71.87 ± 4.58	65.76 ± 4.38	73.18 ± 4.25	93.63 ± 4.35
		RO	55.20 ± 4.22	71.82 ± 4.09	65.41 ± 4.22	72.28 ± 4.28	90.62 ± 4.13
	3-NN	RWO	**88.36** ± 2.87	**95.78** ± 2.87	**93.80** ± 3.01	**95.86** ± 2.83	**100.00** ± 3.04
		SMO	86.97 ± 2.97	95.42 ± 2.98	93.22 ± 3.09	94.22 ± 2.97	96.15 ± 3.06
		RO	84.48 ± 2.91	94.43 ± 2.88	91.80 ± 3.04	92.82 ± 2.90	94.81 ± 3.07

dataset, when Naïve Bayes classifier is implemented, SMOTE loses in the same four metrics to RWO-Sampling with over-sampling rate 200%. When implementing 3NN classifier under different over-sampling rates, no over-sampling approach performs the best all the time. So we say that, whether an over-sampling approach performs well or not on a given dataset, is determined by both the classifier and the number of synthetic samples.

The results on the large scale and highly imbalanced data set Setimage show that, most of the time, if C4.5 and 3NN are conducted, RWO-Sampling performs the best in all metrics when 200%, 300%, 400% and 500% are assigned to the over-sampling rate, and it does not perform well only on part of metrics when 100% is assigned to the over-sampling rate. When implementing Naïve Bayes classifier, RWO-Sampling loses to SMOTE in some metrics. Sonar is the highest-dimensional sample dataset among the continuous attribute datasets, and RWO-Sampling always performs the best under all over-sampling rates when conducting C4.5 on Sonar. When implementing Naïve Bayes classifier and 3NN on Sonar, RWO-Sampling outperforms the other two approaches most of the time. The time of wins and losses on each dataset under different cases is summarized in Table 7.

In order to compare the performance conveniently, we count the number of wins under all cases for each over-sampling approach, and describe the results in Table 8. The results show that, RWO-Sampling outperforms the other two approaches in all five metrics when conducting C4.5 and 3NN, it loses to SMOTE only one time in F-maj. The results also reveal that, random over-sampling loses most of the time, and it never wins when implementing 3NN, since random over-sampling just copies the original samples to over-sample the minority class data. New samples around their original samples are generated by both SMOTE and RWO-Sampling, thus the performance of KNN is improved statistically.

AUC (Area Under the Curve) is also an important metric to evaluate the performance of classifiers. Based on the above experiments, we calculate the mean of ten experiments, and describe the results on 8 continuous attribute datasets in Table 7. For each specific classifier, we count the number of wins under different over-sampling rate, and also show the win time in the table. For example, when conducting C4.5 on Breast-w under over-sampling rates from 100% to 500%, RWO wins 3 times in terms of AUC, and both SMO and RO win once each. The results summarized in Table 8 show that, RWO outperforms the others statistically in terms of AUC.

We also describe the overall classification accuracies and standard deviations on the above mentioned datasets in Table 9. "The original" denotes the classification accuracies on the original datasets without data over-sampling. The results show that, the overall classification accuracies have no obvious differences before and after data over-sampling, and whether the overall classification accuracy is improved or not, depends on the over-sampling approaches and the classifiers.

Results on the other thirteen datasets with all discrete attributes or hybrid attributes are shown in Tables 10 and 11. As

Author's personal copy

Table 4
Averaged results and standard deviations on 8 continuous attribute datasets (over-sampling rate equals 300%).

dataset	Alg	OS	(%)				
			F-min	F-maj	O-acc	G-mean	TP rate
Breast-w	C4.5	RWO	93.04 ± 1.73	96.13 ± 1.59	95.02 ± 1.39	95.36 ± 2.01	**96.47** ± 1.44
		SMO	**93.52** ± 1.96	**96.46** ± 1.72	**95.42** ± 1.83	**95.52** ± 1.86	95.85 ± 2.04
		RO	92.40 ± 1.63	95.83 ± 1.80	94.61 ± 1.62	94.71 ± 2.08	95.02 ± 1.98
	NB	RWO	94.38 ± 1.75	96.89 ± 1.84	95.99 ± 1.75	96.35 ± 1.78	97.51 ± 1.62
		SMO	94.38 ± 1.69	96.89 ± 1.72	95.99 ± 1.89	96.35 ± 1.63	97.51 ± 1.72
		RO	94.38 ± 1.62	96.89 ± 1.79	95.99 ± 1.81	96.35 ± 1.69	97.51 ± 1.83
	3-NN	RWO	96.39 ± 1.30	98.00 ± 1.37	97.42 ± 1.45	**97.96** ± 1.45	**99.75** ± 1.47
		SMO	96.39 ± 1.26	98.00 ± 1.41	97.42 ± 1.46	97.92 ± 1.58	99.59 ± 1.31
		RO	95.44 ± 1.31	97.48 ± 1.38	96.76 ± 1.47	97.16 ± 1.29	98.48 ± 1.46
Diabetes	C4.5	RWO	**66.86** ± 3.27	**77.76** ± 3.42	**73.39** ± 3.86	**74.16** ± 3.57	76.94 ± 3.6
		SMO	66.34 ± 3.12	73.72 ± 4.16	70.49 ± 4.32	72.80 ± 4.15	**83.33** ± 3.98
		RO	60.84 ± 3.46	76.26 ± 4.58	70.44 ± 4.21	69.27 ± 3.69	65.80 ± 4.46
	NB	RWO	**66.44** ± 3.53	74.95 ± 3.89	71.32 ± 3.67	73.24 ± 3.29	81.38 ± 3.44
		SMO	66.33 ± 3.93	74.48 ± 3.30	70.96 ± 3.25	73.03 ± 3.30	**81.97** ± 4.42
		RO	66.39 ± 3.67	**76.14** ± 3.94	**72.09** ± 3.44	**73.50** ± 4.09	78.98 ± 3.26
	3-NN	RWO	**73.53** ± 4.09	**76.92** ± 4.28	**75.34** ± 4.21	**78.70** ± 4.33	**98.13** ± 4.10
		SMO	67.82 ± 4.38	73.55 ± 4.04	70.96 ± 4.27	73.73 ± 4.32	87.69 ± 4.27
		RO	61.43 ± 4.14	68.79 ± 4.34	65.49 ± 4.32	67.81 ± 4.31	78.73 ± 4.35
Glass	C4.5	RWO	94.74 ± 2.30	91.53 ± 2.40	99.59 ± 2.41	99.22 ± 2.42	99.59 ± 2.31
		SMO	94.74 ± 2.31	94.74 ± 2.46	99.76 ± 2.44	99.53 ± 2.56	99.76 ± 2.58
		RO	94.74 ± 2.55	94.74 ± 2.67	99.76 ± 2.59	99.53 ± 2.63	99.76 ± 2.68
	NB	RWO	68.44 ± 2.93	64.98 ± 2.60	97.58 ± 3.24	95.47 ± 2.97	**97.61** ± 2.97
		SMO	**69.44** ± 3.59	**71.64** ± 2.20	**98.44** ± 2.86	**97.04** ± 3.01	93.05 ± 2.94
		RO	64.86 ± 2.65	64.00 ± 3.26	97.77 ± 3.25	95.79 ± 2.94	92.42 ± 4.01
	3-NN	RWO	73.30 ± 1.63	**73.77** ± 1.66	**98.41** ± 2.01	**97.01** ± 2.40	**98.43** ± 1.58
		SMO	**73.85** ± 1.35	67.57 ± 1.92	98.02 ± 2.31	96.26 ± 2.81	94.49 ± 1.46
		RO	71.64 ± 1.64	65.75 ± 2.03	97.94 ± 2.25	96.11 ± 2.58	92.58 ± 2.46
Ionosphere	C4.5	RWO	**86.22** ± 1.76	**92.80** ± 2.39	**90.54** ± 2.59	88.54 ± 2.25	82.46 ± 1.96
		SMO	85.50 ± 2.52	90.87 ± 2.47	88.79 ± 2.59	**89.48** ± 2.01	**92.06** ± 2.08
		RO	83.22 ± 2.51	90.82 ± 2.49	88.13 ± 2.52	86.65 ± 2.26	82.01 ± 3.03
	NB	RWO	**82.91** ± 3.33	**90.07** ± 3.38	**87.44** ± 3.90	86.86 ± 3.15	84.92 ± 3.32
		SMO	82.71 ± 3.44	89.45 ± 3.91	86.89 ± 3.75	**86.98** ± 3.08	**87.30** ± 3.33
		RO	76.62 ± 3.85	83.74 ± 3.36	80.82 ± 3.41	82.13 ± 3.39	87.57 ± 3.89
	3-NN	RWO	**95.42** ± 1.57	**97.31** ± 1.89	**96.61** ± 1.94	**97.01** ± 2.48	98.49 ± 1.95
		SMO	95.04 ± 1.98	97.04 ± 2.48	96.30 ± 2.48	96.86 ± 2.90	**98.94** ± 2.35
		RO	84.99 ± 1.88	92.43 ± 2.17	89.93 ± 1.92	87.22 ± 1.92	79.37 ± 2.38
Satimage	C4.5	RWO	**67.20** ± 1.65	**95.70** ± 2.31	**92.40** ± 1.80	**86.67** ± 2.18	**80.16** ± 2.06
		SMO	65.13 ± 2.71	95.59 ± 2.99	92.17 ± 1.42	84.08 ± 2.42	75.20 ± 2.35
		RO	57.04 ± 2.86	95.30 ± 2.88	91.53 ± 2.36	74.17 ± 2.16	57.81 ± 2.46
	NB	RWO	47.15 ± 3.39	88.34 ± 3.69	80.89 ± 3.42	83.84 ± 3.74	87.68 ± 3.77
		SMO	**48.82** ± 3.37	**89.31** ± 3.23	**82.32** ± 3.43	**84.27** ± 3.85	86.77 ± 3.92
		RO	47.49 ± 3.04	88.48 ± 3.90	81.10 ± 3.69	84.05 ± 4.04	**87.89** ± 4.05
	3-NN	RWO	**74.18** ± 1.89	**96.11** ± 1.88	**93.23** ± 1.69	**96.18** ± 2.00	**100.00** ± 1.38
		SMO	69.51 ± 2.16	95.06 ± 1.70	91.50 ± 1.49	95.02 ± 1.76	99.63 ± 1.81
		RO	64.87 ± 1.97	94.80 ± 2.26	90.94 ± 2.38	88.73 ± 2.06	86.08 ± 2.11
Segment	C4.5	RWO	**97.75** ± 0.69	**99.64** ± 0.79	**99.38** ± 0.89	**99.07** ± 0.87	**98.63** ± 0.63
		SMO	97.18 ± 0.62	99.55 ± 0.81	99.22 ± 0.83	98.66 ± 0.77	97.89 ± 0.85
		RO	97.24 ± 0.78	99.56 ± 0.86	99.24 ± 0.77	98.53 ± 0.60	97.56 ± 0.74
	NB	RWO	**67.19** ± 1.39	**92.47** ± 1.25	**87.75** ± 1.42	**89.41** ± 0.94	91.76 ± 1.33
		SMO	61.15 ± 1.82	89.03 ± 1.39	82.89 ± 1.78	89.01 ± 1.32	98.54 ± 1.10
		RO	60.81 ± 2.03	88.78 ± 2.04	82.56 ± 1.72	88.98 ± 1.04	**99.02** ± 1.40
	3-NN	RWO	**97.59** ± 0.81	**99.61** ± 0.87	**99.33** ± 0.77	**99.59** ± .094	**99.95** ± 0.68
		SMO	97.45 ± 0.71	99.59 ± 0.79	99.29 ± 0.64	99.38 ± 0.94	99.51 ± 0.82
		RO	96.23 ± 0.81	99.38 ± 0.81	98.93 ± 0.79	99.24 ± 0.70	99.67 ± 0.85
Sonar	C4.5	RWO	**84.17** ± 4.80	**81.14** ± 4.75	**82.79** ± 5.01	**82.51** ± 5.05	**98.14** ± 4.73
		SMO	82.12 ± 4.90	79.93 ± 4.74	81.09 ± 4.78	81.07 ± 4.78	93.13 ± 5.06
		RO	69.86 ± 5.17	75.15 ± 4.77	72.76 ± 4.98	72.28 ± 5.06	67.70 ± 5.13
	NB	RWO	**73.89** ± 5.26	65.02 ± 5.17	**70.10** ± 4.93	**68.73** ± 5.26	**90.72** ± 4.89
		SMO	70.43 ± 5.29	**65.39** ± 5.00	68.11 ± 4.78	67.81 ± 4.88	81.44 ± 4.98
		RO	70.33 ± 5.21	63.20 ± 4.80	67.15 ± 4.71	66.43 ± 4.87	83.51 ± 5.08
	3-NN	RWO	**89.27** ± 2.71	**88.49** ± 2.64	**88.89** ± 2.72	**89.03** ± 2.83	**99.07** ± 2.89
		SMO	88.65 ± 3.03	87.93 ± 2.83	88.30 ± 2.77	88.45 ± 3.05	97.94 ± 2.91
		RO	87.94 ± 2.92	87.70 ± 2.87	87.82 ± 2.96	88.01 ± 2.74	95.19 ± 2.98
Vehicle	C4.5	RWO	**86.70** ± 2.89	**95.57** ± 2.97	**93.36** ± 2.76	**92.90** ± 2.73	**92.06** ± 2.80
		SMO	85.69 ± 2.87	95.35 ± 3.15	92.99 ± 2.80	91.67 ± 3.03	89.28 ± 2.82
		RO	85.55 ± 2.94	95.33 ± 3.19	92.95 ± 2.91	91.46 ± 2.90	88.78 ± 3.14
	NB	RWO	**56.46** ± 4.28	71.73 ± 4.44	65.72 ± 4.11	73.30 ± 4.08	**94.47** ± 4.04
		SMO	56.45 ± 4.40	**71.90** ± 4.06	**65.84** ± 4.33	**73.34** ± 4.18	94.14 ± 4.53
		RO	56.38 ± 4.54	71.71 ± 4.44	65.68 ± 4.35	73.24 ± 4.12	94.30 ± 4.14
	3-NN	RWO	**88.12** ± 2.81	**95.67** ± 2.82	**93.66** ± 2.89	**95.76** ± 3.09	**100.00** ± 2.81

(*continued on next page*)

Author's personal copy

Table 4 (*continued*)

dataset	Alg	OS	(%)				
			F-min	F-maj	O-acc	G-mean	TP rate
		SMO	87.29 ± 2.93	95.45 ± 2.85	93.30 ± 2.93	94.82 ± 2.89	97.82 ± 2.88
		RO	84.87 ± 2.80	94.49 ± 2.82	91.92 ± 2.83	93.40 ± 2.88	96.31 ± 2.81

Table 5
Averaged results and standard deviations on 8 continuous attribute datasets (over-sampling rate equals 400%).

dataset	Alg	OS	(%)				
			F-min	F-maj	O-acc	G-mean	TP rate
Breast-w	C4.5	RWO	**93.44** ± 1.22	**96.35** ± 1.50	**95.31** ± 1.97	**95.69** ± 1.81	**96.97** ± 2.11
		SMO	92.61 ± 2.45	95.99 ± 1.99	94.80 ± 2.00	94.72 ± 1.49	94.47 ± 1.88
		RO	92.64 ± 2.06	95.98 ± 2.15	94.80 ± 1.76	94.82 ± 2.73	94.88 ± 2.11
	NB	RWO	94.38 ± 1.64	96.89 ± 1.83	95.99 ± 1.61	96.35 ± 1.80	97.51 ± 1.71
		SMO	94.38 ± 1.72	96.89 ± 1.88	95.99 ± 1.85	96.35 ± 1.82	97.51 ± 1.79
		RO	94.38 ± 1.87	96.89 ± 1.79	95.99 ± 1.88	96.35 ± 1.82	97.51 ± 1.65
	3-NN	RWO	96.29 ± 1.48	97.94 ± 1.21	97.35 ± 1.47	97.90 ± 1.21	**99.75** ± 1.31
		SMO	**96.51** ± 1.21	**98.08** ± 1.23	**97.52** ± 1.32	**98.00** ± 1.37	99.59 ± 1.51
		RO	95.25 ± 1.31	97.37 ± 1.52	96.61 ± 1.58	97.05 ± 1.35	98.48 ± 1.28
Diabetes	C4.5	RWO	**65.85** ± 3.93	**77.33** ± 4.44	**72.75** ± 3.51	**73.31** ± 3.61	75.30 ± 3.90
		SMO	64.15 ± 3.98	70.91 ± 4.43	67.88 ± 4.03	70.37 ± 3.99	**82.34** ± 4.14
		RO	59.34 ± 4.25	75.58 ± 3.65	69.49 ± 3.55	68.03 ± 3.69	63.81 ± 4.44
	NB	RWO	**66.71** ± 3.26	**73.26** ± 4.14	**70.34** ± 3.10	**72.89** ± 4.24	**85.15** ± 3.62
		SMO	66.09 ± 3.71	72.34 ± 4.31	69.53 ± 3.40	72.16 ± 3.45	85.07 ± 4.05
		RO	65.95 ± 3.92	73.07 ± 3.73	69.92 ± 3.27	72.32 ± 3.99	83.46 ± 3.39
	3-NN	RWO	**72.96** ± 4.09	**75.89** ± 4.23	**74.51** ± 4.29	**77.93** ± 4.14	**98.54** ± 4.19
		SMO	67.83 ± 4.12	72.37 ± 4.17	70.27 ± 4.27	73.28 ± 4.37	89.80 ± 4.28
		RO	61.69 ± 4.30	68.18 ± 4.14	65.23 ± 4.45	67.74 ± 4.37	80.22 ± 4.03
Glass	C4.5	RWO	100. ± 2.53	91.53 ± 2.29	99.59 ± 1.92	99.22 ± 2.75	99.59 ± 2.33
		SMO	100. ± 2.15	94.74 ± 3.27	99.76 ± 2.47	99.53 ± 1.96	99.76 ± 2.53
		RO	100. ± 2.71	94.74 ± 2.55	99.76 ± 2.36	99.53 ± 2.23	99.76 ± 2.59
	NB	RWO	**100.** ± 2.65	64.98 ± 2.58	97.58 ± 2.56	95.47 ± 3.28	**97.61** ± 2.11
		SMO	88.89 ± 3.09	**76.19** ± 2.82	**98.77** ± 3.44	**97.66** ± 2.94	93.36 ± 2.82
		RO	88.89 ± 3.19	61.54 ± 2.04	97.51 ± 2.47	95.33 ± 3.47	92.19 ± 2.41
	3-NN	RWO	**100.** ± 2.26	**71.71** ± 1.67	**98.24** ± 2.53	**96.68** ± 1.27	**98.25** ± 1.46
		SMO	92.59 ± 2.43	68.49 ± 1.91	98.10 ± 1.98	96.42 ± 1.95	94.57 ± 2.30
		RO	88.89 ± 2.65	64.00 ± 1.99	97.77 ± 2.38	95.79 ± 1.77	92.42 ± 2.21
Ionosphere	C4.5	RWO	**86.26** ± 1.96	**92.82** ± 2.38	**90.57** ± 1.79	88.56 ± 2.10	82.46 ± 1.69
		SMO	85.50 ± 2.51	90.87 ± 2.64	88.79 ± 2.39	**89.48** ± 2.16	**92.06** ± 2.28
		RO	81.70 ± 2.95	89.79 ± 2.50	86.89 ± 2.38	85.60 ± 2.03	81.48 ± 2.88
	NB	RWO	**82.95** ± 3.70	**90.09** ± 3.18	**87.46** ± 3.21	86.88 ± 3.26	84.92 ± 3.05
		SMO	82.71 ± 3.48	89.45 ± 3.44	86.89 ± 3.41	**86.98** ± 3.16	**87.30** ± 4.16
		RO	76.39 ± 3.49	83.57 ± 3.90	80.63 ± 4.04	81.93 ± 3.38	87.30 ± 4.51
	3-NN	RWO	**95.28** ± 1.59	**97.21** ± 1.44	**96.50** ± 1.52	96.92 ± 2.28	98.49 ± 1.90
		SMO	94.97 ± 2.22	96.95 ± 2.02	96.20 ± 1.16	**96.99** ± 2.25	**100.00** ± 2.17
		RO	85.71 ± 1.75	92.67 ± 1.56	90.31 ± 2.23	87.95 ± 2.04	80.95 ± 2.44
Satimage	C4.5	RWO	**66.71** ± 1.75	**95.69** ± 2.16	**92.36** ± 1.85	**85.94** ± 2.27	**78.72** ± 2.17
		SMO	65.49 ± 2.48	95.54 ± 2.52	92.10 ± 2.80	84.99 ± 2.22	77.07 ± 2.30
		RO	56.03 ± 1.96	95.07 ± 2.71	91.13 ± 2.56	74.19 ± 2.21	58.13 ± 2.32
	NB	RWO	47.18 ± 3.88	88.25 ± 3.77	80.78 ± 3.82	84.04 ± 3.80	**88.32** ± 3.64
		SMO	**48.84** ± 4.31	**89.32** ± 3.95	**82.32** ± 4.07	**84.30** ± 3.90	86.83 ± 3.63
		RO	47.07 ± 4.03	88.26 ± 3.61	80.79 ± 3.28	83.87 ± 3.75	87.89 ± 3.32
	3-NN	RWO	**73.57** ± 1.48	**95.98** ± 1.50	**93.02** ± 1.62	**96.05** ± 1.60	**100.00** ± 1.59
		SMO	68.43 ± 2.37	94.78 ± 1.41	91.04 ± 1.59	94.88 ± 1.27	99.95 ± 1.60
		RO	64.54 ± 1.39	94.62 ± 1.23	90.66 ± 2.31	89.22 ± 2.34	87.47 ± 1.69
Segment	C4.5	RWO	97.68 ± 0.67	99.63 ± 0.66	99.36 ± 0.66	**98.97** ± 0.69	98.44 ± 0.62
		SMO	**97.89** ± 0.72	**99.67** ± 0.78	**99.42** ± 0.63	98.84 ± 0.72	**98.05** ± 0.67
		RO	96.83 ± 0.77	99.50 ± 0.81	99.13 ± 0.68	98.19 ± 0.75	96.91 ± 0.84
	NB	RWO	**65.19** ± 1.94	**91.55** ± 1.17	**86.40** ± 1.83	**89.16** ± 1.78	93.17 ± 1.65
		SMO	61.00 ± 1.37	88.95 ± 1.78	82.78 ± 1.76	88.94 ± 1.81	98.54 ± 1.89
		RO	60.00 ± 2.17	88.42 ± 1.87	82.04 ± 2.44	88.47 ± 1.50	**98.54** ± 1.97
	3-NN	RWO	**97.57** ± 0.74	**99.60** ± 0.95	**99.32** ± 0.63	**99.60** ± 0.87	**100.00** ± 0.64
		SMO	97.30 ± 0.99	99.56 ± 0.96	99.24 ± 0.70	99.43 ± 0.65	99.67 ± 0.86
		RO	96.31 ± 0.61	99.39 ± 0.91	98.96 ± 0.73	99.33 ± 0.88	99.84 ± 0.79
Sonar	C4.5	RWO	**84.62** ± 5.08	**81.66** ± 5.14	**83.27** ± 4.79	**83.00** ± 4.95	**98.66** ± 5.00
		SMO	80.36 ± 5.05	77.82 ± 4.86	79.17 ± 4.71	79.11 ± 4.94	91.41 ± 5.00
		RO	71.36 ± 5.15	73.73 ± 5.04	72.60 ± 5.07	72.63 ± 5.15	73.20 ± 5.12
	NB	RWO	**73.52** ± 4.80	64.10 ± 4.71	**69.52** ± 5.28	68.01 ± 4.98	**90.72** ± 4.72
		SMO	69.73 ± 5.18	64.46 ± 4.84	67.31 ± 4.99	66.98 ± 5.12	80.76 ± 4.73

Author's personal copy

Table 5 (*continued*)

dataset	Alg	OS	(%)				
			F-min	F-maj	O-acc	G-mean	TP rate
		RO	71.70 ± 5.04	**65.12** ± 5.23	68.75 ± 4.80	**68.11** ± 5.00	84.88 ± 5.10
	3-NN	RWO	**89.15** ± 2.72	**88.32** ± 3.02	**88.75** ± 2.75	**88.88** ± 3.08	**99.07** ± 2.88
		SMO	88.99 ± 2.90	88.23 ± 3.02	88.62 ± 2.96	88.76 ± 2.95	98.63 ± 2.87
		RO	87.44 ± 2.73	86.57 ± 2.98	87.02 ± 2.90	87.15 ± 3.02	96.91 ± 3.03
Vehicle	C4.5	RWO	87.60 ± 2.75	95.83 ± 2.93	93.76 ± 2.73	**93.74** ± 2.95	**93.72** ± 2.87
		SMO	**87.73** ± 2.77	95.93 ± 3.19	**93.89** ± 2.93	93.51 ± 3.09	92.80 ± 2.95
		RO	87.21 ± 2.80	**95.95** ± 3.18	93.85 ± 3.06	92.16 ± 3.07	89.11 ± 2.70
	NB	RWO	**56.80** ± 4.59	71.75 ± 4.23	65.84 ± 4.20	**73.59** ± 4.03	**95.48** ± 4.25
		SMO	56.54 ± 4.10	71.70 ± 4.11	65.72 ± 4.57	73.37 ± 4.44	94.81 ± 4.32
		RO	56.79 ± 4.06	**71.81** ± 4.29	**65.88** ± 4.55	73.59 ± 4.16	95.31 ± 4.56
	3-NN	RWO	**87.92** ± 2.92	**95.59** ± 3.01	**93.54** ± 3.00	**95.68** ± 2.83	**100.00** ± 2.84
		SMO	86.26 ± 3.09	95.00 ± 2.99	92.67 ± 3.00	94.39 ± 3.09	97.82 ± 2.96
		RO	83.85 ± 2.89	93.96 ± 2.96	91.21 ± 2.85	93.14 ± 2.85	96.98 ± 3.06

Table 6
Averaged results and standard deviations on 8 continuous attribute datasets (over-sampling rate equals 500%).

dataset	Alg	OS	(%)				
			F-min	F-maj	O-acc	G-mean	TP rate
Breast-w	C4.5	RWO	**93.59** ± 1.73	**96.44** ± 1.86	**95.42** ± 2.13	**95.79** ± 1.58	**97.01** ± 1.45
		SMO	92.84 ± 1.62	96.09 ± 2.34	94.95 ± 1.99	94.96 ± 1.81	95.02 ± 1.65
		RO	92.31 ± 1.75	95.80 ± 1.64	94.56 ± 1.82	94.57 ± 1.87	94.61 ± 1.65
	NB	RWO	94.38 ± 1.71	96.89 ± 1.68	95.99 ± 1.84	96.35 ± 1.73	97.51 ± 1.80
		SMO	94.38 ± 1.74	96.89 ± 1.78	95.99 ± 1.89	96.35 ± 1.77	97.51 ± 1.86
		RO	94.38 ± 1.67	96.89 ± 1.67	95.99 ± 1.81	96.35 ± 1.63	97.51 ± 1.84
	3-NN	RWO	96.26 ± 1.30	97.92 ± 1.37	97.32 ± 1.27	97.88 ± 1.43	**99.75** ± 1.48
		SMO	**96.39** ± 1.20	**98.00** ± 1.26	**97.42** ± 1.36	**97.92** ± 1.38	99.59 ± 1.45
		RO	95.45 ± 1.43	97.48 ± 1.32	96.76 ± 1.36	97.22 ± 1.38	98.76 ± 1.45
Diabetes	C4.5	RWO	**65.37** ± 3.59	**76.80** ± 4.45	**72.21** ± 4.19	**72.86** ± 3.73	75.15 ± 3.82
		SMO	64.67 ± 3.82	70.41 ± 3.88	67.80 ± 4.46	70.51 ± 3.40	**84.45** ± 3.88
		RO	59.30 ± 4.04	75.15 ± 4.50	69.14 ± 4.05	67.95 ± 4.06	64.43 ± 3.78
	NB	RWO	66.25 ± 3.57	70.99 ± 3.62	68.80 ± 3.94	71.74 ± 3.82	87.76 ± 3.61
		SMO	66.04 ± 3.26	70.61 ± 3.71	68.49 ± 3.74	71.45 ± 4.42	**87.81** ± 3.59
		RO	**66.48** ± 3.50	**72.08** ± 3.67	**69.53** ± 4.42	**72.31** ± 3.41	86.57 ± 3.94
	3-NN	RWO	**72.43** ± 4.34	**74.97** ± 4.11	**73.76** ± 4.17	**77.21** ± 4.00	**98.77** ± 4.05
		SMO	68.23 ± 4.03	71.67 ± 4.33	70.05 ± 4.39	73.24 ± 4.30	92.16 ± 4.00
		RO	61.97 ± 4.12	67.76 ± 4.42	65.10 ± 4.33	67.74 ± 4.19	81.47 ± 4.23
Glass	C4.5	RWO	90.45 ± 2.49	99.53 ± 2.67	99.11 ± 2.20	99.54 ± 2.16	100.00 ± 2.20
		SMO	**94.74** ± 2.99	**99.76** ± 2.48	**99.53** ± 2.68	**99.76** ± 2.31	100.00 ± 2.79
		RO	94.74 ± 2.37	99.76 ± 1.89	99.53 ± 2.48	99.76 ± 2.57	100.00 ± 2.30
	NB	RWO	64.52 ± 2.62	97.53 ± 2.29	95.37 ± 3.06	**97.56** ± 2.75	**100.00** ± 2.94
		SMO	**76.19** ± 2.46	**98.77** ± 2.61	**97.66** ± 3.27	93.36 ± 3.22	88.89 ± 3.12
		RO	62.34 ± 2.97	97.60 ± 2.66	95.48 ± 3.02	92.27 ± 3.21	88.89 ± 3.06
	3-NN	RWO	**72.29** ± 1.83	**98.29** ± 1.97	**96.78** ± 2.51	**98.30** ± 2.00	**100.00** ± 1.85
		SMO	67.57 ± 1.70	98.02 ± 1.85	96.26 ± 2.04	94.49 ± 2.02	92.59 ± 2.19
		RO	64.00 ± 2.40	97.77 ± 2.17	95.79 ± 2.33	92.42 ± 2.23	88.89 ± 2.27
Ionosphere	C4.5	RWO	**86.25** ± 2.60	**92.79** ± 2.19	**90.54** ± 1.80	88.58 ± 2.46	82.62 ± 2.61
		SMO	84.77 ± 2.74	90.27 ± 2.25	88.13 ± 2.86	**88.94** ± 2.85	**92.06** ± 2.54
		RO	80.90 ± 2.53	89.24 ± 2.20	86.23 ± 2.84	85.04 ± 2.57	81.22 ± 2.67
	NB	RWO	82.95 ± 3.54	**90.09** ± 2.75	87.46 ± 3.32	86.88 ± 3.71	84.92 ± 3.21
		SMO	**83.29** ± 3.61	89.97 ± 3.04	87.46 ± 3.43	**87.37** ± 3.65	87.04 ± 3.98
		RO	76.43 ± 3.03	83.28 ± 3.06	80.44 ± 3.41	81.95 ± 3.80	**88.36** ± 2.98
	3-NN	RWO	**95.21** ± 1.16	**97.17** ± 2.12	**96.44** ± 2.07	96.88 ± 2.19	98.49 ± 1.61
		SMO	94.96 ± 2.46	96.95 ± 2.11	96.20 ± 1.89	**96.94** ± 2.12	**99.74** ± 2.22
		RO	86.19 ± 1.36	92.87 ± 2.99	90.60 ± 2.19	88.38 ± 1.85	81.75 ± 1.65
Satimage	C4.5	RWO	**66.51** ± 1.88	**95.61** ± 2.24	**92.24** ± 2.71	**86.16** ± 2.55	**79.28** ± 1.88
		SMO	65.59 ± 2.83	95.50 ± 2.65	92.04 ± 2.16	85.44 ± 2.21	78.03 ± 2.19
		RO	55.67 ± 2.13	95.05 ± 2.18	91.10 ± 2.70	73.79 ± 1.95	57.49 ± 2.12
	NB	RWO	46.93 ± 3.53	88.09 ± 4.15	80.55 ± 3.49	83.97 ± 3.53	**88.48** ± 3.58
		SMO	**48.71** ± 4.09	**89.25** ± 3.58	**82.23** ± 3.77	**84.24** ± 3.52	86.83 ± 3.52
		RO	47.27 ± 3.58	88.38 ± 4.08	80.95 ± 4.08	83.94 ± 4.50	87.84 ± 3.98
	3-NN	RWO	**73.49** ± 1.55	**95.96** ± 1.05	**92.99** ± 1.71	**96.04** ± 1.20	**100.00** ± 1.20
		SMO	67.45 ± 1.98	94.53 ± 1.20	90.63 ± 2.39	94.60 ± 0.88	99.84 ± 1.82
		RO	64.43 ± 2.26	94.56 ± 2.28	90.57 ± 1.76	89.36 ± 1.56	87.89 ± 1.86
Segment	C4.5	RWO	97.92 ± 0.62	99.67 ± 0.80	99.43 ± 0.74	**99.15** ± 0.85	**98.78** ± 0.80
		SMO	**97.97** ± 0.69	**99.68** ± 0.63	**99.44** ± 0.74	98.92 ± 0.64	98.21 ± 0.76
		RO	97.18 ± 0.76	99.55 ± 0.64	99.22 ± 0.65	98.73 ± 0.61	98.05 ± 0.87
	NB	RWO	**72.90** ± 1.86	**94.64** ± 2.00	**91.04** ± 1.81	**89.80** ± 1.60	88.13 ± 1.80

(*continued on next page*)

Author's personal copy

Table 6 (*continued*)

dataset	Alg	OS	(%)				
			F-min	F-maj	O-acc	G-mean	TP rate
		SMO	60.60 ± 1.87	88.74 ± 2.06	82.49 ± 1.80	88.76 ± 2.59	98.54 ± 1.85
		RO	60.02 ± 2.02	88.39 ± 1.99	82.00 ± 1.78	88.56 ± 1.79	**98.86** ± 2.36
	3-NN	RWO	**97.62** ± 0.83	**99.61** ± 0.63	**99.33** ± 0.89	**99.61** ± 0.80	**100.00** ± 0.63
		SMO	97.37 ± 0.69	99.57 ± 0.85	99.27 ± 0.96	99.37 ± 0.83	99.51 ± 0.60
		RO	96.09 ± 0.86	99.35 ± 0.86	98.89 ± 0.99	99.35 ± 0.97	100.00 ± 0.64
Sonar	C4.5	RWO	**84.20** ± 5.13	**80.52** ± 4.80	**82.55** ± 4.71	**82.07** ± 4.72	**99.69** ± 4.76
		SMO	82.28 ± 4.94	79.73 ± 4.97	81.09 ± 5.00	80.99 ± 4.94	94.16 ± 4.80
		RO	67.47 ± 5.12	71.94 ± 5.01	69.87 ± 4.88	69.64 ± 4.79	67.01 ± 4.77
	NB	RWO	**73.06** ± 4.81	62.94 ± 4.86	**68.80** ± 4.99	67.11 ± 4.77	**90.72** ± 4.72
		SMO	70.12 ± 4.72	**64.69** ± 5.02	67.63 ± 5.02	**67.27** ± 4.99	81.44 ± 4.86
		RO	71.77 ± 5.08	63.35 ± 5.11	68.11 ± 4.96	67.01 ± 5.21	86.94 ± 4.82
	3-NN	RWO	**89.06** ± 2.92	**88.21** ± 2.78	**88.65** ± 2.74	**88.78** ± 2.94	**99.07** ± 3.07
		SMO	87.29 ± 2.96	86.05 ± 3.07	86.70 ± 2.75	86.77 ± 2.92	97.94 ± 2.99
		RO	87.09 ± 2.86	86.28 ± 2.73	86.70 ± 2.76	86.84 ± 2.72	96.22 ± 2.99
Vehicle	C4.5	RWO	87.60 ± 2.73	95.87 ± 3.29	93.81 ± 3.00	93.53 ± 2.78	**93.02** ± 2.88
		SMO	**87.80** ± 3.21	**95.96** ± 3.21	**93.93** ± 2.80	**93.54** ± 2.71	92.80 ± 2.87
		RO	87.13 ± 3.26	95.89 ± 3.17	93.77 ± 2.93	92.29 ± 3.26	89.61 ± 2.89
	NB	RWO	56.63 ± 4.28	71.50 ± 4.50	65.60 ± 4.10	73.39 ± 4.20	95.48 ± 4.03
		SMO	**56.94** ± 4.38	**71.63** ± 4.33	**65.80** ± 4.26	**73.68** ± 4.39	**96.15** ± 4.07
		RO	56.80 ± 4.01	71.42 ± 4.51	65.60 ± 4.03	73.51 ± 4.19	96.15 ± 4.59
	3-NN	RWO	**87.82** ± 2.96	**95.54** ± 2.88	**93.47** ± 3.02	**95.64** ± 2.85	**100.00** ± 2.80
		SMO	86.83 ± 3.01	95.18 ± 2.92	92.95 ± 3.04	94.91 ± 2.90	98.83 ± 2.90
		RO	83.77 ± 3.09	93.94 ± 2.93	91.17 ± 2.83	93.05 ± 2.81	96.82 ± 2.85

Table 7
Times of win for each metric when different over-sampling approach and baseline algorithm are executed on each data set under five different over-sampling rates ("*n* + –" means winning n times and drawing 5 – *n* times) together with the averaged AUCs under five different over-sampling rates.

Dataset	Alg	OS	Times of win					Averaged AUC value					
			F-min	F-maj	O-acc	G-mean	TP rate	100%	200%	300%	400%	500%	Win time
Breast-w	C4.5	RWO	4	4	4	4	5	0.946828	0.940532	0.945481	0.945445	0.9448	3
		SMO	1	1	1	1	0	0.941404	0.944753	0.942898	0.942242	0.946392	1
		RO	0	0	0	0	0	0.945845	0.940532	0.942934	0.939512	0.946647	1
	NB	RWO	–	–	–	–	–	0.963534	0.963534	0.963534	0.963534	0.963534	
		SMO	–	–	–	–	–	0.963534	0.963534	0.963534	0.963534	0.963534	
		RO	–	–	–	–	–	0.963534	0.963534	0.963534	0.963534	0.963534	
	3-NN	RWO	–	–	–	2	4	0.973542	0.974198	0.973798	0.976272	0.973798	0
		SMO	4 + –	4 + –	4 + –	3	1	0.9774	0.979111	0.979366	0.980094	0.979366	5
		RO	0	0	0	0	0	0.96881	0.968265	0.97165	0.970559	0.972306	0
Diabetes	C4.5	RWO	5	5	5	5	1	0.737488	0.722214	0.734667	0.712358	0.716597	5
		SMO	0	0	0	0	4	0.697199	0.684408	0.687109	0.680313	0.687109	0
		RO	0	0	0	0	0	0.695229	0.694891	0.688358	0.70795	0.682716	0
	NB	RWO	3	2	2	2	3	0.727662	0.733602	0.738144	0.734821	0.73997	3
		SMO	1	1	1	1	2	0.739413	0.730687	0.735159	0.731373	0.729721	1
		RO	1	2	2	2	0	0.729284	0.734358	0.732035	0.734488	0.73597	1
	3-NN	RWO	4	4	4	4	5	0.726353	0.729204	0.748433	0.748005	0.751821	5
		SMO	1	1	1	1	0	0.694856	0.690473	0.68807	0.684224	0.685333	0
		RO	0	0	0	0	0	0.695527	0.68809	0.687035	0.687896	0.690517	0
Glass	C4.5	RWO	–	–	–	–	–	0.997561	0.997561	0.995935	0.995935	0.995122	
		SMO	1 + –	2 + –	1 + –	1 + 1	–	0.997561	0.997561	0.997561	0.997561	0.997561	
		RO	–	–	–	–	–	0.997561	0.997561	0.997561	0.997561	0.997561	
	NB	RWO	1	0	1	3	5	0.923306	0.942638	0.939386	0.940199	0.938573	4
		SMO	4	5	4	2	0	0.926558	0.931436	0.931436	0.934688	0.934688	0
		RO	0	0	0	0	0	0.928184	0.924932	0.924932	0.924932	0.922493	1
	3-NN	RWO	3	4	4	5	5	0.933062	0.930623	0.911292	0.92981	0.928997	2
		SMO	2	1	1	0	0	0.91617	0.92981	0.928184	0.946703	0.945077	3
		RO	0	0	0	0	0	0.92981	0.924932	0.926558	0.924932	0.924932	0
Ionosphere	C4.5	RWO	4	4	4	0	0	0.887513	0.900212	0.895132	0.895132	0.889947	4
		SMO	0	0	0	4	5	0.893492	0.891693	0.872698	0.858466	0.866878	1
		RO	1	1	1	1	0	0.890159	0.889841	0.888254	0.890159	0.888254	0
	NB	RWO	4	5	4	2	0	0.865344	0.864339	0.868042	0.869524	0.869524	4
		SMO	1	0	–	3	2	0.85254	0.859947	0.863069	0.869418	0.870899	1
		RO	0	0	–	0	3	0.824392	0.824074	0.820952	0.824762	0.823439	0
	3-NN	RWO	4	4	4	2	1	0.860265	0.858466	0.869048	0.868889	0.870794	0
		SMO	1	1	1	3	4	0.887143	0.926349	0.942222	0.948783	0.945979	5
		RO	0	0	0	0	0	0.847619	0.866085	0.878571	0.88328	0.885185	0
Satimage	C4.5	RWO	4	4	4	5	5	0.782989	0.796905	0.811658	0.815786	0.821497	5
		SMO	1	1	1	0	0	0.761964	0.77348	0.771552	0.769591	0.768746	0

Author's personal copy

Table 7 (*continued*)

Dataset	Alg	OS	Times of win					Averaged AUC value					
			F-min	F-maj	O-acc	G-mean	TP rate	100%	200%	300%	400%	500%	Win time
		RO	0	0	0	0	0	0.751321	0.756773	0.766587	0.762156	0.763469	0
	NB	RWO	0	0	0	1	3	0.843453	0.841041	0.839807	0.83943	0.838597	1
		SMO	5	5	5	4	0	0.84099	0.840913	0.840818	0.841778	0.841155	2
		RO	0	0	0	0	2	0.841948	0.841513	0.841312	0.839914	0.840184	2
	3-NN	RWO	4	4	4	4	4	0.896214	0.914279	0.918091	0.92288	0.923175	5
		SMO	1	1	1	1	1	0.868574	0.880161	0.888913	0.892856	0.893168	0
		RO	0	0	0	0	0	0.874515	0.876439	0.878942	0.879947	0.884382	0
Segment	C4.5	RWO	2	2	2	4	3	0.983476	0.987284	0.983476	0.986986	0.990752	2
		SMO	3	3	3	0	1	0.982792	0.986518	0.986599	0.988442	0.989255	2
		RO	0	0	0	1	1	0.984804	0.98643	0.98416	0.985488	0.983862	1
	NB	RWO	5	5	5	5	2	0.889519	0.889635	0.892683	0.892495	0.898179	1
		SMO	0	0	0	0	0	0.898305	0.896456	0.894613	0.892812	0.889296	3
		RO	0	0	0	0	3	0.897316	0.894912	0.892209	0.898333	0.89225	1
	3-NN	RWO	5	5	5	5	5	0.991647	0.99353	0.995156	0.993957	0.994642	4
		SMO	0	0	0	0	0	0.991091	0.993273	0.993829	0.994256	0.9937	1
		RO	0	0	0	0	0	0.991004	0.992284	0.992454	0.993822	0.993436	0
Sonar	C4.5	RWO	5	5	5	5	5	0.798892	0.777221	0.773815	0.757732	0.812018	5
		SMO	0	0	0	0	0	0.764775	0.729126	0.74072	0.729374	0.732795	0
		RO	0	0	0	0	0	0.740503	0.708291	0.71086	0.740271	0.738336	0
	NB	RWO	5	1	5	2	5	0.689948	0.686945	0.689096	0.684592	0.685242	2
		SMO	0	2	0	1	0	0.681558	0.686496	0.681558	0.683911	0.683059	0
		RO	0	2	0	2	0	0.684994	0.684576	0.689514	0.695969	0.696186	3
	3-NN	RWO	4	4	4	4	5	0.887774	0.879849	0.88909	0.892527	0.874075	4
		SMO	1	1	1	1	0	0.867806	0.874679	0.868023	0.876614	0.878549	1
		RO	0	0	0	0	0	0.877883	0.866537	0.869555	0.877496	0.872775	0
Vehicle	C4.5	RWO	3	3	3	4	5	0.924118	0.923153	0.917032	0.935135	0.935393	5
		SMO	2	1	1	1	0	0.919414	0.91046	0.906532	0.902537	0.904534	0
		RO	0	1	0	0	0	0.915355	0.912264	0.915743	0.918964	0.928368	0
	NB	RWO	3	1	1	2	3	0.745798	0.754947	0.757524	0.757524	0.756945	1
		SMO	2	3	3	3	2	0.752948	0.754109	0.756364	0.757911	0.763066	2
		RO	0	1	1	0	0	0.744895	0.754109	0.760166	0.760811	0.759523	2
	3-NN	RWO	4	4	4	5	5	0.936426	0.942355	0.948669	0.944548	0.94983	5
		SMO	1	1	1	0	0	0.928567	0.932434	0.931854	0.933014	0.932371	0
		RO	0	0	0	0	0	0.926633	0.929537	0.932308	0.930119	0.932374	0

to make the work compacted, we just show the results at over-sampling rate 200% and 300%. Since the table will be too wide, we do not show the standard deviations.

We summarize in Table 12 the win times in each metric for each over-sampling approach based on the results shown in Tables 10 and 11, and results show that, on datasets with hybrid attributes, RWO-Sampling performs well statistically in terms of *F-min*, *G-mean* and TP rate if the three baseline classifiers are conducted except that it draws with RO-Sampling in terms of TP rate when 3NN is conducted. SMOTE wins in terms of the *F-maj*, the overall accuracy and AUC. On datasets with only discrete attributes, the results show that our approach performs well in *F-min*, *G-mean* and *TP rate* if C4.5 and NB are implemented. When 3NN is implemented, SMOTE performs well in terms of both *F-measure*, overall accuracy, *G-mean* and AUC, 3NN gets good results in term of *TP rate*.

We just show the results for over-sampling rate 200% and 300%, and the above conclusion still holds based on results obtained at all five over-sampling rates.

From the results shown in Tables 2–12, we may say that RWO-Sampling performs potentially well on imbalanced datasets no matter what classification algorithms are conducted.

Table 8
Five evaluation metric win summary on 8 continuous attribute datasets when implementing three baseline classifiers together with AUC win summary.

Alg	OS	F-min	F-maj	O-acc	G-mean	TPrate	AUC
C4.5	RWO	**27**	**27**	**27**	**27**	**24**	**29**
	SMO	8	8	7	7	5	4
	RO	1	2	1	2	1	2
NB	RWO	**21**	14	**18**	**17**	**21**	**16**
	SMO	13	**15**	13	14	6	9
	RO	1	5	3	4	8	10
3-NN	RWO	**28**	**29**	**29**	**31**	**34**	**25**
	SMO	11	10	10	9	6	15
	RO	0	0	0	0	0	0

Over-sampling the minority class and under-sampling the majority class at the same time is also evaluated in this work. 200% is assigned to the over-sampling rate, and under this condition, under-sampling is executed on the majority class at different rates, such as 75%, 50% and 25%. 75% means 75 percent of the majority class samples are randomly chosen without replication. Ten independent experiments have been done for each baseline classifier and under-sampling rate. we report the win times for each metric when implementing different over-sampling approaches and baseline classifiers in Table 13, and show the averaged results in Fig. 2. Results in Table 13 show that, when implementing 3NN, RWO performs the best in terms of the six metrics statistically, it is slightly better than SMO in terms of *F-min*, *O-acc* and *TP rate* statistically when implementing C4.5, and is slightly worse than SMO in terms of *F-maj* and *O-acc* statistically when implementing NB. Regardless of the classifiers implemented, RO is statistically the worst one.

Fig. 2 indicates that, the RWO-Sampling performs well in many cases when combined with under-sampling on the majority class. The results also reveal that, as the under-sampling rate changes from 75% to 25%, *F-min*, *G-mean* and *TP rate* increase gradually, and there is no obvious change in *F-maj*. That means as fewer majority class instances are used when constructing a classifier, RWO-Sampling does not favor the minority class at the expense

Author's personal copy

Table 9
Overall classification accuracies on 8 continuous attribute datasets under 5 data over-sampling rates and on the original datasets. "↑" denotes that in each row, no less than 3 accuracies under different over-sampling rates are larger than the original accuracy; "↓" denotes that in each row, no less than 3 accuracies under different over-sampling rates are less than the original accuracy.

Dataset	Alg	OS	100%	200%	300%	400%	500%	Original	
Breast-w	C4.5	RWO	95.14 ± 2.09	94.98 ± 1.76	95.02 ± 1.39	95.31 ± 1.97	95.42 ± 2.13		↑
		SMO	94.95 ± 2.20	94.85 ± 2.00	95.42 ± 1.83	94.80 ± 2.00	94.95 ± 1.99	94.56 ± 2.22	↑
		RO	92.39 ± 2.13	94.47 ± 2.10	94.61 ± 1.62	94.80 ± 1.76	94.56 ± 1.82		
	NB	RWO	95.99 ± 1.45	95.99 ± 1.61	95.99 ± 1.75	95.99 ± 1.61	95.99 ± 1.84		
		SMO	95.99 ± 1.77	95.99 ± 1.85	95.99 ± 1.89	95.99 ± 1.85	95.99 ± 1.89	95.99 ± 1.99	
		RO	95.99 ± 1.75	95.99 ± 1.88	95.99 ± 1.81	95.99 ± 1.88	95.99 ± 1.81		
	3-NN	RWO	97.12 ± 1.49	97.42 ± 1.47	97.42 ± 1.45	97.35 ± 1.47	97.32 ± 1.27		↑
		SMO	97.42 ± 1.80	97.52 ± 1.32	97.42 ± 1.46	97.52 ± 1.32	97.42 ± 1.36	96.85 ± 1.65	↑
		RO	96.90 ± 1.54	96.61 ± 1.58	96.76 ± 1.47	96.61 ± 1.58	96.76 ± 1.36		↓
Diabetes	C4.5	RWO	74.08 ± 4.25	72.73 ± 4.03	73.39 ± 3.86	72.75 ± 3.51	72.21 ± 4.19		↓
		SMO	73.22 ± 4.02	70.44 ± 3.51	70.49 ± 4.32	67.88 ± 4.03	67.80 ± 4.46	73.82 ± 4.46	↓
		RO	70.83 ± 4.09	69.79 ± 3.55	70.44 ± 4.21	69.49 ± 3.55	69.14 ± 4.05		↓
	NB	RWO	74.58 ± 4.08	73.36 ± 3.27	71.32 ± 3.67	70.34 ± 3.10	68.80 ± 3.94		↓
		SMO	75.09 ± 4.05	72.22 ± 3.40	70.96 ± 3.25	69.53 ± 3.40	68.49 ± 3.74	76.30 ± 4.16	↓
		RO	74.65 ± 4.09	72.61 ± 3.60	72.09 ± 3.44	69.92 ± 3.27	69.53 ± 4.42		↓
	3-NN	RWO	71.26 ± 4.21	76.56 ± 4.29	75.34 ± 4.21	74.51 ± 4.29	73.76 ± 4.17		↑
		SMO	71.66 ± 4.45	70.23 ± 4.27	70.96 ± 4.27	70.27 ± 4.27	70.05 ± 4.39	72.65 ± 4.52	↓
		RO	68.66 ± 4.09	65.71 ± 4.45	65.49 ± 4.32	65.23 ± 4.45	65.10 ± 4.33		↓
Glass	C4.5	RWO	99.53 ± 1.97	99.53 ± 2.47	99.59 ± 2.41	99.59 ± 1.92	99.11 ± 2.20		↑
		SMO	99.53 ± 2.41	99.53 ± 1.92	99.76 ± 2.44	99.76 ± 2.47	99.53 ± 2.68	97.19 ± 1.47	↑
		RO	99.53 ± 2.25	99.53 ± 2.36	99.76 ± 2.59	99.76 ± 2.36	99.53 ± 2.48		↑
	NB	RWO	96.12 ± 3.04	95.84 ± 2.56	97.58 ± 3.24	97.58 ± 2.56	95.37 ± 3.06		↓
		SMO	96.57 ± 2.93	97.04 ± 2.44	98.44 ± 2.86	98.77 ± 3.44	97.66 ± 3.27	96.26 ± 1.85	↑
		RO	95.95 ± 2.82	95.17 ± 2.47	97.77 ± 3.25	97.51 ± 2.47	95.48 ± 3.02		↓
	3-NN	RWO	97.24 ± 1.76	97.15 ± 2.53	98.41 ± 2.01	98.24 ± 2.53	96.78 ± 2.51		↓
		SMO	97.35 ± 2.30	96.73 ± 1.98	98.02 ± 2.31	98.10 ± 1.98	96.26 ± 2.04	97.66 ± 1.42	↓
		RO	97.04 ± 2.31	95.64 ± 2.38	97.94 ± 2.25	97.77 ± 2.38	95.79 ± 2.33		↓
Ionosphere	C4.5	RWO	90.60 ± 2.47	90.60 ± 1.79	90.54 ± 2.59	90.57 ± 1.79	90.54 ± 1.80		↓
		SMO	89.46 ± 2.79	89.74 ± 2.39	88.79 ± 2.59	88.79 ± 2.39	88.13 ± 2.86	91.45 ± 2.90	↓
		RO	91.26 ± 2.57	88.41 ± 2.38	88.13 ± 2.52	86.89 ± 2.38	86.23 ± 2.84		↓
	NB	RWO	86.92 ± 3.68	87.18 ± 3.21	87.44 ± 3.90	87.46 ± 3.21	87.46 ± 3.32		↑
		SMO	85.00 ± 3.69	85.85 ± 3.41	86.89 ± 3.75	86.89 ± 3.41	87.46 ± 3.43	82.62 ± 3.93	↑
		RO	81.48 ± 3.64	81.01 ± 4.04	80.82 ± 3.41	80.63 ± 4.04	80.44 ± 3.41		↓
	3-NN	RWO	90.43 ± 2.91	96.72 ± 2.32	96.61 ± 1.94	96.50 ± 1.52	96.44 ± 2.07		↓
		SMO	93.45 ± 3.05	96.30 ± 3.03	96.30 ± 2.48	96.20 ± 1.16	96.20 ± 1.89	86.60 ± 3.32	↓
		RO	88.41 ± 2.93	89.36 ± 1.96	89.93 ± 1.92	90.31 ± 2.23	90.60 ± 2.19		↓
Satimage	C4.5	RWO	92.08 ± 2.11	92.99 ± 1.85	92.40 ± 1.80	92.36 ± 1.85	92.24 ± 2.71		↑
		SMO	92.20 ± 2.17	92.13 ± 2.80	92.17 ± 1.42	92.10 ± 2.80	92.04 ± 2.16	91.88 ± 2.77	↑
		RO	91.30 ± 2.24	91.45 ± 2.56	91.53 ± 2.36	91.13 ± 2.56	91.10 ± 2.70		↓
	NB	RWO	81.54 ± 3.92	81.23 ± 3.82	80.89 ± 3.42	80.78 ± 3.82	80.55 ± 3.49		↓
		SMO	82.44 ± 3.82	82.33 ± 4.07	82.32 ± 3.43	82.32 ± 4.07	82.23 ± 3.77	82.19 ± 4.13	↑
		RO	81.58 ± 3.81	81.13 ± 3.28	81.10 ± 3.69	80.79 ± 3.28	80.95 ± 4.08		↓
	3-NN	RWO	92.50 ± 1.72	93.41 ± 1.31	93.23 ± 1.69	93.02 ± 1.62	92.99 ± 1.71		↓
		SMO	93.12 ± 1.81	92.34 ± 1.41	91.50 ± 1.49	91.04 ± 1.59	90.63 ± 2.39	94.03 ± 2.12	↓
		RO	92.47 ± 2.02	91.39 ± 1.62	90.94 ± 2.38	90.66 ± 2.31	90.57 ± 1.76		↓
Segment	C4.5	RWO	99.24 ± 1.10	99.40 ± 0.62	99.38 ± 0.89	99.36 ± 0.66	99.43 ± 0.74		↑
		SMO	99.16 ± 0.98	99.44 ± 0.67	99.22 ± 0.83	99.42 ± 0.63	99.44 ± 0.74	99.13 ± 0.92	↑
		RO	99.13 ± 0.92	99.13 ± 0.70	99.24 ± 0.77	99.13 ± 0.68	99.22 ± 0.65		↑
	NB	RWO	86.73 ± 1.93	87.38 ± 1.27	87.75 ± 1.42	86.40 ± 1.83	91.04 ± 1.81		↑
		SMO	83.82 ± 2.14	83.42 ± 1.42	82.89 ± 1.78	82.78 ± 1.76	82.49 ± 1.80	83.93 ± 3.83	↑
		RO	82.96 ± 2.06	83.18 ± 2.16	82.56 ± 1.72	82.04 ± 2.44	82.00 ± 1.78		↓
	3-NN	RWO	99.42 ± 1.04	99.39 ± 0.99	99.33 ± 0.77	99.32 ± 0.63	99.33 ± 0.89		↓
		SMO	99.29 ± 0.91	99.31 ± 0.66	99.29 ± 0.64	99.24 ± 0.70	99.27 ± 0.96	99.44 ± 0.63	↓
		RO	99.13 ± 1.09	99.02 ± 0.79	98.93 ± 0.79	98.96 ± 0.73	98.89 ± 0.99		↓
Sonar	C4.5	RWO	81.83 ± 4.25	83.32 ± 4.77	82.79 ± 5.01	83.27 ± 4.79	82.55 ± 4.71		↑
		SMO	78.04 ± 4.80	81.25 ± 5.02	81.09 ± 4.78	79.17 ± 4.71	81.09 ± 5.00	71.15 ± 5.20	↑
		RO	73.88 ± 4.97	74.84 ± 4.95	72.76 ± 4.98	72.60 ± 5.07	69.87 ± 4.88		↑
	NB	RWO	68.32 ± 4.76	70.14 ± 4.73	70.10 ± 4.93	69.52 ± 5.28	68.80 ± 4.99		↑
		SMO	67.79 ± 4.45	67.95 ± 4.93	68.11 ± 4.78	67.31 ± 4.99	67.63 ± 5.02	67.78 ± 5.31	↑
		RO	68.11 ± 5.55	68.27 ± 5.01	67.15 ± 4.71	68.75 ± 4.80	68.11 ± 4.96		↑
	3-NN	RWO	85.96 ± 2.48	88.85 ± 2.83	88.89 ± 2.72	88.75 ± 2.75	88.65 ± 2.74		↑
		SMO	88.46 ± 2.71	87.50 ± 3.08	88.30 ± 2.77	86.62 ± 2.96	86.70 ± 2.75	86.05 ± 3.13	↑
		RO	87.66 ± 2.77	86.54 ± 3.06	87.82 ± 2.96	87.02 ± 2.90	86.70 ± 2.76		↑
Vehicle	C4.5	RWO	94.34 ± 2.90	93.56 ± 3.11	93.36 ± 2.76	93.76 ± 2.73	93.81 ± 3.00		↑
		SMO	93.54 ± 3.12	93.30 ± 3.09	92.99 ± 2.80	93.89 ± 2.93	93.93 ± 2.80	93.26 ± 2.50	↑
		RO	93.62 ± 2.68	93.77 ± 2.80	92.95 ± 2.91	93.85 ± 3.06	93.77 ± 2.93		↑
	NB	RWO	65.72 ± 4.44	65.96 ± 4.13	65.72 ± 4.11	65.84 ± 4.20	65.60 ± 4.10		↓

Author's personal copy

Table 9 (*continued*)

Dataset	Alg	OS	100%	200%	300%	400%	500%	Original	
		SMO	66.12 ± 4.15	65.76 ± 4.38	65.84 ± 4.33	65.72 ± 4.57	65.80 ± 4.26	66.08 ± 5.63	↓
		RO	65.29 ± 4.47	65.41 ± 4.22	65.68 ± 4.35	65.88 ± 4.55	65.60 ± 4.03		↓
	3-NN	RWO	92.97 ± 2.89	93.80 ± 3.01	93.66 ± 2.89	93.54 ± 3.00	93.47 ± 3.02		↑
		SMO	93.38 ± 3.34	93.22 ± 3.09	93.30 ± 2.93	92.67 ± 3.00	92.95 ± 3.04	93.14 ± 2.21	↑
		RO	92.51 ± 2.58	91.80 ± 3.04	91.92 ± 2.83	91.21 ± 2.85	91.17 ± 2.83		↓

Table 10
Averaged results on 6 discrete attribute dataset.

Dataset	Alg	OS	200% (%)						300% (%)					
			F-min	F-maj	O-acc	G-mean	TPrate	AUC	F-min	F-maj	O-acc	G-mean	TPrate	AUC
Breast-cancer	C4.5	RWO	**50.95**	74.89	66.78	**63.96**	**58.04**	0.626430	**53.33**	72.93	65.73	**65.78**	**65.88**	0.633938
		SMO	44.27	**83.45**	**74.48**	55.89	34.12	**0.698551**	44.27	**83.45**	**74.48**	55.89	34.12	**0.698551**
		RO	44.37	69.78	60.84	58.15	52.55	0.573117	43.32	68.42	59.44	57.10	52.16	0.562992
	NB	RWO	**54.40**	73.15	66.20	**66.66**	**67.84**	**0.641080**	**54.52**	72.06	65.38	**66.59**	**69.80**	**0.639922**
		SMO	32.84	**79.45**	**68.53**	47.33	25.88	0.591578	32.56	**80.36**	**69.58**	46.77	24.71	0.606405
		RO	52.73	71.81	64.69	65.13	66.27	0.627810	53.09	71.54	64.57	65.37	67.45	0.629582
	3-NN	RWO	**55.41**	80.24	72.61	**67.30**	57.25	**0.675379**	**54.33**	76.80	69.23	**66.80**	**65.49**	0.651445
		SMO	41.79	**82.19**	**72.73**	54.31	32.94	0.665461	41.79	**82.19**	**72.73**	54.31	32.94	**0.665461**
		RO	45.58	67.23	59.09	58.67	**57.65**	0.573077	46.98	62.49	56.06	58.40	61.57	0.573661
Monk2	C4.5	RWO	**53.78**	34.37	45.76	43.64	**83.33**	0.544665	51.67	21.71	40.24	33.54	**84.38**	0.477874
		SMO	40.71	**70.22**	**60.36**	52.00	35.94	**0.563861**	40.71	**70.22**	**60.36**	52.00	35.94	**0.563861**
		RO	50.21	54.89	52.66	**54.05**	63.02	0.545026	**52.49**	53.23	52.86	**54.48**	68.75	0.559157
	NB	RWO	**53.66**	28.57	43.79	39.43	**85.94**	0.534321	**54.57**	12.68	40.24	25.73	**94.79**	0.535329
		SMO	34.34	**72.80**	**61.54**	**46.91**	26.56	**0.567484**	31.91	**73.77**	**62.13**	**44.82**	23.44	**0.573741**
		RO	51.42	29.88	42.60	39.73	80.21	0.499189	52.47	15.85	39.25	28.55	88.54	0.470717
	3-NN	RWO	49.46	38.16	44.38	44.55	71.88	0.497035	51.85	31.90	43.59	41.30	80.21	0.510590
		SMO	29.63	**66.96**	**55.03**	42.82	25.00	0.489818	29.63	**66.96**	**55.03**	42.82	25.00	0.489818
		RO	**54.41**	49.48	52.07	**53.41**	**75.52**	**0.571044**	**58.21**	48.46	53.85	**54.45**	**84.90**	**0.617151**
Postoperative	C4.5	RWO	**42.86**	57.89	51.52	**55.28**	**66.67**	**0.550752**	**42.19**	45.59	43.94	**49.21**	**75.00**	**0.534239**
		SMO	15.00	**75.00**	**61.36**	31.56	12.50	0.447917	15.79	**76.81**	**63.64**	32.17	12.50	0.465251
		RO	21.98	58.96	46.21	38.41	27.78	0.422078	26.14	63.07	50.76	42.97	31.94	0.457452
	NB	RWO	37.38	57.32	49.24	51.03	**55.56**	0.509698	**44.62**	46.27	45.45	51.00	**80.56**	**0.562150**
		SMO	**38.30**	**77.52**	**67.05**	**54.13**	37.50	**0.580268**	34.78	**76.92**	**65.91**	**51.03**	33.33	0.560606
		RO	31.79	60.06	49.62	47.35	43.06	0.480626	36.12	51.83	45.08	48.10	56.94	0.490056
	3-NN	RWO	**28.87**	58.68	47.73	**44.55**	**38.89**	0.459825	**35.96**	51.33	44.70	**47.79**	**56.94**	0.487892
		SMO	23.81	**76.12**	**63.64**	40.75	20.83	**0.503175**	23.81	**76.12**	**63.64**	40.75	20.83	**0.503175**
		RO	19.15	55.29	42.42	34.99	25.00	0.395154	26.32	44.00	36.36	37.85	41.67	0.401709
Primary-tumor	C4.5	RWO	90.48	97.59	96.15	95.71	95.00	0.925721	**91.34**	97.79	**96.47**	**96.55**	**96.67**	**0.928754**
		SMO	90.48	97.59	96.15	95.71	95.00	0.925721	90.48	97.59	96.15	95.71	95.00	0.925721
		RO	87.80	97.01	95.19	93.16	90.00	0.916523	88.71	97.20	95.51	94.02	91.67	0.919607
	NB	RWO	89.55	97.14	95.51	**97.18**	**100.00**	0.905405	**92.31**	**97.98**	**96.79**	**98.00**	**100.00**	**0.928571**
		SMO	**90.48**	**97.59**	**96.15**	95.71	95.00	**0.925721**	90.48	97.59	96.15	95.71	95.00	0.925721
		RO	89.06	97.18	95.51	95.32	95.00	0.912970	88.89	97.19	95.51	94.67	93.33	0.916112
	3-NN	RWO	81.08	94.12	91.03	94.28	100.00	0.840909	76.43	92.08	88.14	92.37	100.00	0.809278
		SMO	**85.11**	**95.65**	**93.27**	**95.74**	100.00	**0.870370**	**85.11**	**95.65**	**93.27**	**95.74**	100.00	**0.870370**
		RO	77.42	92.54	88.78	92.80	100.00	0.815789	72.73	90.20	85.58	90.63	100.00	0.785714
Splice-i.e.	C4.5	RWO	95.39	97.76	96.99	**97.37**	**98.44**	0.958932	94.75	97.41	96.53	97.08	**98.61**	0.952509
		SMO	**95.77**	**97.99**	**97.28**	97.27	97.27	**0.965153**	**95.77**	**97.99**	**97.28**	**97.27**	97.27	**0.965153**
		RO	95.03	97.60	96.77	96.99	97.61	0.957266	94.83	97.50	96.63	96.89	97.61	0.955356
	NB	RWO	**96.23**	**98.22**	**97.58**	**97.53**	**97.40**	**0.969313**	**96.19**	**98.19**	**97.55**	**97.58**	**97.66**	**0.968368**
		SMO	8.24	81.83	69.67	20.73	4.30	0.846234	0.26	81.19	68.35	3.61	0.13	0.841660
		RO	95.98	98.11	97.43	97.27	96.83	0.968280	96.05	98.14	97.47	97.37	97.09	0.968333
	3-NN	RWO	75.06	82.18	79.21	83.22	98.70	0.798528	71.81	78.01	75.29	79.81	99.26	0.778581
		SMO	**84.80**	**91.41**	**89.02**	**90.89**	96.61	**0.868779**	**84.80**	**91.41**	**89.02**	**90.89**	96.61	**0.868779**
		RO	69.37	74.64	72.25	76.98	**99.13**	0.763413	66.96	70.64	68.91	73.78	**99.39**	0.749864
Tic-tac-toe	C4.5	RWO	73.77	81.78	78.50	80.27	**87.25**	0.777543	70.72	77.51	74.57	77.12	**88.65**	0.752986
		SMO	76.95	88.38	84.55	81.80	74.40	0.832801	76.95	**88.38**	84.55	81.80	74.40	0.832801
		RO	**80.21**	**88.42**	**85.39**	**85.40**	85.44	**0.836410**	**79.89**	88.09	**85.04**	**85.20**	85.74	**0.832912**
	NB	RWO	**59.74**	63.20	61.55	**64.50**	**82.33**	0.656197	**60.76**	56.01	58.52	**61.20**	**92.67**	**0.682134**
		SMO	47.06	**81.66**	**72.76**	56.95	34.94	**0.724740**	29.84	**79.76**	**68.58**	42.73	19.28	0.674264
		RO	59.05	63.83	61.59	64.38	79.92	0.648939	60.29	56.29	58.39	61.14	91.16	0.673916
	3-NN	RWO	85.32	90.09	88.17	90.37	99.20	0.871668	79.49	84.20	82.15	85.23	99.80	0.829502
		SMO	**98.01**	**98.97**	**98.64**	**98.10**	96.39	**0.989023**	**98.01**	**98.97**	**98.64**	**98.10**	96.39	**0.989023**
		RO	89.57	93.42	91.93	93.62	**100.00**	0.905537	80.98	85.77	83.72	86.65	**100.00**	0.840164

Author's personal copy

Table 11
Averaged results on 7 hybrid attribute datasets.

Dataset	Alg	OS	200% (%)						300% (%)					
			F-min	F-maj	O-acc	G-mean	TPrate	AUC	F-min	F-maj	O-acc	G-mean	TPrate	AUC
Credit-g	C4.5	RWO	42.05	78.96	69.13	55.59	37.33	0.618186	40.79	79.13	69.13	54.43	35.44	0.615853
		SMO	50.00	**81.32**	**72.80**	61.92	45.33	**0.670223**	48.92	**81.19**	**72.50**	60.99	43.89	**0.665720**
		RO	**51.03**	74.89	66.80	**63.86**	**57.67**	0.626746	**51.28**	74.24	66.30	**64.04**	**59.11**	0.625561
	NB	RWO	**61.92**	77.90	72.03	**73.05**	**75.78**	**0.697473**	**62.11**	76.25	70.80	**73.08**	**79.78**	**0.696945**
		SMO	40.35	**82.55**	**73.00**	52.70	30.44	0.675995	35.87	**82.66**	**72.70**	48.63	25.44	0.675807
		RO	61.12	78.08	71.97	72.38	73.44	0.692882	61.36	75.97	70.37	72.45	78.44	0.691300
	3-NN	RWO	**54.09**	76.98	69.33	**66.41**	60.22	0.651062	**54.22**	74.57	67.30	**66.49**	64.56	0.642929
		SMO	51.72	**81.96**	**73.73**	63.22	46.89	**0.682872**	51.41	**81.81**	**73.53**	63.00	46.67	**0.680190**
		RO	51.90	68.68	62.07	63.67	**68.22**	0.616193	52.30	65.81	60.17	63.13	**72.78**	0.616236
Haberrman	C4.5	RWO	**51.30**	77.68	69.39	**66.42**	**60.91**	**0.640221**	**49.42**	75.26	66.78	**64.92**	**61.32**	**0.622715**
		SMO	35.21	**81.43**	**71.13**	50.50	29.63	0.603171	34.70	**80.93**	**70.48**	50.24	29.63	0.594691
		RO	42.71	72.87	63.18	59.06	51.85	0.579104	42.91	72.83	63.18	59.22	52.26	0.580015
	NB	RWO	**44.90**	79.94	70.59	**60.07**	**45.27**	0.623566	**47.10**	71.89	63.29	**62.78**	**61.73**	0.601615
		SMO	42.29	**83.82**	**74.73**	55.81	34.98	**0.663211**	42.96	**83.86**	**74.84**	56.41	35.80	**0.665344**
		RO	42.32	79.47	69.72	57.84	41.98	0.609560	43.68	69.79	60.68	59.66	57.61	0.576841
	3-NN	RWO	**46.79**	73.72	64.81	**62.62**	58.44	0.603900	**47.26**	67.95	60.13	**62.28**	**67.49**	0.597236
		SMO	40.72	**81.21**	**71.46**	55.73	37.04	**0.619733**	40.54	**81.03**	**71.24**	55.63	37.04	**0.617186**
		RO	45.11	71.05	62.09	61.01	**58.85**	0.587988	45.89	67.19	59.15	61.01	65.43	0.586923
Hepatitis	C4.5	RWO	50.91	84.79	76.77	68.98	**58.33**	0.667155	**58.63**	84.88	77.85	**77.17**	50.91	0.701703
		SMO	**55.32**	**88.68**	**81.94**	**69.49**	54.17	**0.723627**	56.25	**88.62**	**81.94**	70.60	**55.32**	**0.724339**
		RO	50.00	84.87	76.77	67.96	56.25	0.664130	50.00	85.24	77.20	67.66	50.00	0.666844
	NB	RWO	65.14	89.33	83.66	79.84	**73.96**	0.754540	64.25	88.86	83.01	79.46	65.14	0.747235
		SMO	**70.77**	**92.24**	**87.74**	**81.26**	71.88	**0.811600**	**71.88**	**92.68**	**88.39**	**81.62**	**70.77**	**0.822790**
		RO	61.24	88.77	82.58	76.04	66.67	0.737731	62.33	88.67	82.58	77.43	61.24	0.739605
	3-NN	RWO	58.67	86.81	80.00	**75.51**	**68.75**	0.711171	**60.76**	86.58	80.00	**78.09**	**58.67**	0.718282
		SMO	**59.70**	**88.89**	**82.58**	74.08	62.50	**0.735714**	59.70	**88.89**	**82.58**	74.08	59.70	**0.735714**
		RO	52.21	82.53	74.41	71.81	67.71	0.662739	50.19	80.36	71.83	70.66	52.21	0.647269
Colic	C4.5	RWO	**78.28**	86.72	83.51	**82.83**	**80.39**	0.822048	**76.89**	85.59	82.25	**81.74**	**78.28**	0.808708
		SMO	76.68	**87.78**	**83.97**	80.73	71.32	**0.836841**	76.68	**87.78**	**83.97**	80.73	76.68	**0.836841**
		RO	74.26	83.01	79.53	79.61	79.90	0.782140	75.40	83.76	80.43	80.58	74.26	0.791402
	NB	RWO	69.74	77.90	74.46	75.42	**79.66**	0.738563	**69.34**	76.89	73.64	74.88	69.74	0.733907
		SMO	**70.02**	**81.80**	**77.36**	**76.02**	71.57	**0.757177**	68.95	**81.73**	**76.99**	**75.10**	**70.02**	**0.753124**
		RO	68.73	76.95	73.46	74.46	78.92	0.729565	68.64	76.58	73.19	74.31	68.73	0.728295
	3-NN	RWO	68.22	78.13	74.09	74.33	**75.25**	0.729476	68.39	77.00	73.37	74.23	68.22	0.727278
		SMO	**71.53**	**83.12**	**78.80**	**77.22**	72.06	**0.772464**	**71.53**	**83.12**	**78.80**	**77.22**	**71.53**	**0.772464**
		RO	67.38	71.48	69.57	71.73	85.05	0.715661	68.06	70.77	69.47	71.82	67.38	0.723824
Hypothyroid	C4.5	RWO	96.60	99.80	99.63	99.80	100.	0.967095	95.17	99.72	99.46	99.64	96.60	0.954568
		SMO	98.73	99.93	99.86	99.93	100.	0.987437	98.73	99.93	99.86	99.93	98.73	0.987437
		RO	98.73	99.93	99.86	99.93	100.	0.987437	98.73	99.93	99.86	99.93	98.73	0.987437
	NB	RWO	**61.75**	**98.25**	**96.66**	**71.15**	**51.03**	**0.877403**	61.85	98.12	96.41	**73.78**	**61.75**	0.839626
		SMO	48.91	97.08	94.47	69.74	50.17	0.724637	43.36	96.18	92.84	70.26	48.91	0.672483
		RO	58.60	98.12	96.40	69.16	48.28	0.858545	**63.04**	**98.22**	**96.61**	73.64	58.60	**0.858519**
	3-NN	RWO	**49.32**	97.70	95.60	63.25	40.55	0.798423	**50.38**	97.54	95.32	66.46	**49.32**	0.770885
		SMO	46.40	**97.75**	**95.68**	59.21	35.40	**0.819062**	46.85	**97.73**	**95.64**	60.03	46.40	**0.810887**
		RO	46.80	96.92	94.19	**68.46**	**48.45**	0.711904	47.59	96.70	93.79	**71.64**	46.80	0.701327
Sick	C4.5	RWO	88.52	99.22	98.53	95.59	**92.35**	0.922458	87.11	99.10	98.32	95.62	88.52	0.908591
		SMO	89.05	99.27	98.64	94.65	90.33	0.935831	89.77	99.32	98.73	95.12	89.05	0.939081
		RO	**91.85**	**99.47**	**99.00**	**95.62**	91.92	**0.956298**	**92.35**	**99.50**	**99.05**	**96.28**	**91.85**	**0.955292**
	NB	RWO	**49.85**	94.10	89.44	**87.68**	**85.71**	0.670597	48.01	93.57	88.56	**87.49**	**49.85**	0.661305
		SMO	49.17	**94.44**	**89.97**	84.75	79.22	**0.670882**	**48.17**	**94.24**	**89.63**	84.29	49.17	**0.665974**
		RO	49.23	94.01	89.28	87.18	84.85	0.667892	47.21	93.32	88.14	87.41	49.23	0.657338
	3-NN	RWO	**70.57**	**97.88**	**96.04**	**86.81**	**77.49**	0.816446	**69.45**	97.66	95.66	**88.22**	**70.57**	0.798763
		SMO	65.72	97.85	95.96	78.74	63.20	**0.830245**	65.24	**97.78**	**95.83**	79.11	65.72	**0.821347**
		RO	65.09	97.34	95.05	85.19	75.32	0.778281	64.27	97.15	94.72	86.18	65.09	0.766995
Vowel	C4.5	RWO	**94.01**	**99.39**	**98.89**	**97.54**	**95.93**	**0.958809**	94.47	99.42	98.96	**98.59**	**94.01**	0.954393
		SMO	88.36	98.73	97.71	96.73	95.56	0.908569	87.48	98.64	97.54	96.13	88.36	0.904526
		RO	93.58	99.35	98.82	96.82	94.44	0.960853	**95.20**	**99.52**	**99.12**	97.50	93.58	**0.972041**
	NB	RWO	64.31	94.72	90.81	90.94	91.11	0.743636	61.89	94.11	89.80	**90.39**	64.31	0.729378
		SMO	**70.26**	**96.35**	**93.50**	89.29	84.44	**0.792687**	**70.75**	**96.49**	**93.74**	88.87	**70.26**	**0.798736**
		RO	63.99	94.62	90.64	**91.02**	**91.48**	0.741356	61.58	94.03	89.66	90.31	63.99	0.727598
	3-NN	RWO	100.	100.	100.	100.	100.	1	100.	100.	100.	100.	100.	1
		SMO	100.	100.	100.	100.	100.	1	100.	100.	100.	100.	100.	1
		RO	99.63	99.96	99.93	99.96	100.	0.996324	99.63	99.96	99.93	99.96	99.63	0.996324

of the majority class. This is because that, both the majority class distribution and the minority class distribution are not obviously changed after random under-sampling and random walk over-sampling, thus leading to no obvious change of *F-maj*. We perform the pairwise two-tailed t-tests for assessing the statistical differences between two sets of *F-maj* values when implementing under-sampling under different rates together with RWO-sampling, and show the *P-values*. If *P-value*$<$ 0.05, it indicates that there is

Author's personal copy

Table 12
Number of wins on 13 datasets when implementing baseline classifiers.

		F-min	F-maj	O-acc	G-mean	TPrate	AUC
Discrete attribute							
C4.5	RWO	**6**	0	1	**6**	**11**	3
	SMO	2	**9**	**7**	1	0	**6**
	RO	3	1	3	4	0	2
NB	RWO	**10**	3	4	**8**	**12**	**7**
	SMO	2	**9**	**8**	4	0	5
	RO	0	0	0	0	0	0
3-NN	RWO	4	0	0	4	4	1
	SMO	**6**	**12**	**12**	**6**	0	**9**
	RO	2	0	0	2	**6**	2
Hybrid attribute							
C4.5	RWO	**6**	1	1	**7**	**8**	2
	SMO	1	**8**	**9**	1	0	**6**
	RO	5	3	2	4	4	4
NB	RWO	**8**	1	1	**9**	**11**	4
	SMO	5	**12**	**12**	4	0	**10**
	RO	1	1	1	1	2	0
3-NN	RWO	**9**	1	1	**8**	**6**	0
	SMO	3	**10**	**11**	2	0	**12**
	RO	0	1	0	2	**6**	0

obvious difference between two sets of *F-maj* values. The large the *P-value*, the small the difference of two sets of *F-maj* values is. When implementing C4.5, NB and 3NN for comparing under-sampling rate 25% to 50%, the *P*-value is 0.87447, 0.99924 and 0.99795 respectively; When implementing C4.5, NB and 3NN for comparing under-sampling rate 25–75%, the *P*-value is 0.77365, 0.99032 and 0.99217 respectively; When implementing C4.5, NB and 3NN for comparing under-sampling rate 50–75%, the *P*-value is 0.74835, 0.99011 and 0.99807 respectively. The above *P*-values support the conclusion that there is no significant difference between the *F-maj* values when implementing under-sampling under different rates together with RWO-sampling.

Table 13
Number of wins on Breast-w when simultaneously implementing over-sampling and under-sampling.

OS		F-min	F-maj	O-acc	G-mean	TPrate	AUC
C4.5	RWO	**18**	10	**15**	14	**18**	11
	SMO	8	**15**	14	**15**	9	**15**
	RO	4	5	1	1	1	4
NB	RWO	13	11	13	**13**	12	13
	SMO	13	**15**	**15**	8	12	13
	RO	4	6	2	7	6	4
3-NN	RWO	**18**	**19**	**18**	**17**	**26**	**20**
	SMO	6	6	8	9	4	6
	RO	4	3	4	4	0	2

6. Conclusions

We have evaluated three classical classification algorithms including C4.5, Naïve Bayes classifier and *k*-nearest neighbors on the balanced datasets. Over-sampling often has impacts on the performance of C4.5 in response to the modification of data distribution, since it influences the measure called information gain used for choosing the best attribute in C4.5, thus leads to the modification of the constructed decision tree. Modifying the structure of the decision tree influences pruning and overfitting avoidance, thus influences the performance of C4.5. Over-sampling also influences the performance of *k* nearest neighbors (*k*NN), since after over-sampling, the number of the minority class instances in a fixed volume may be increased, and instances belonging to the

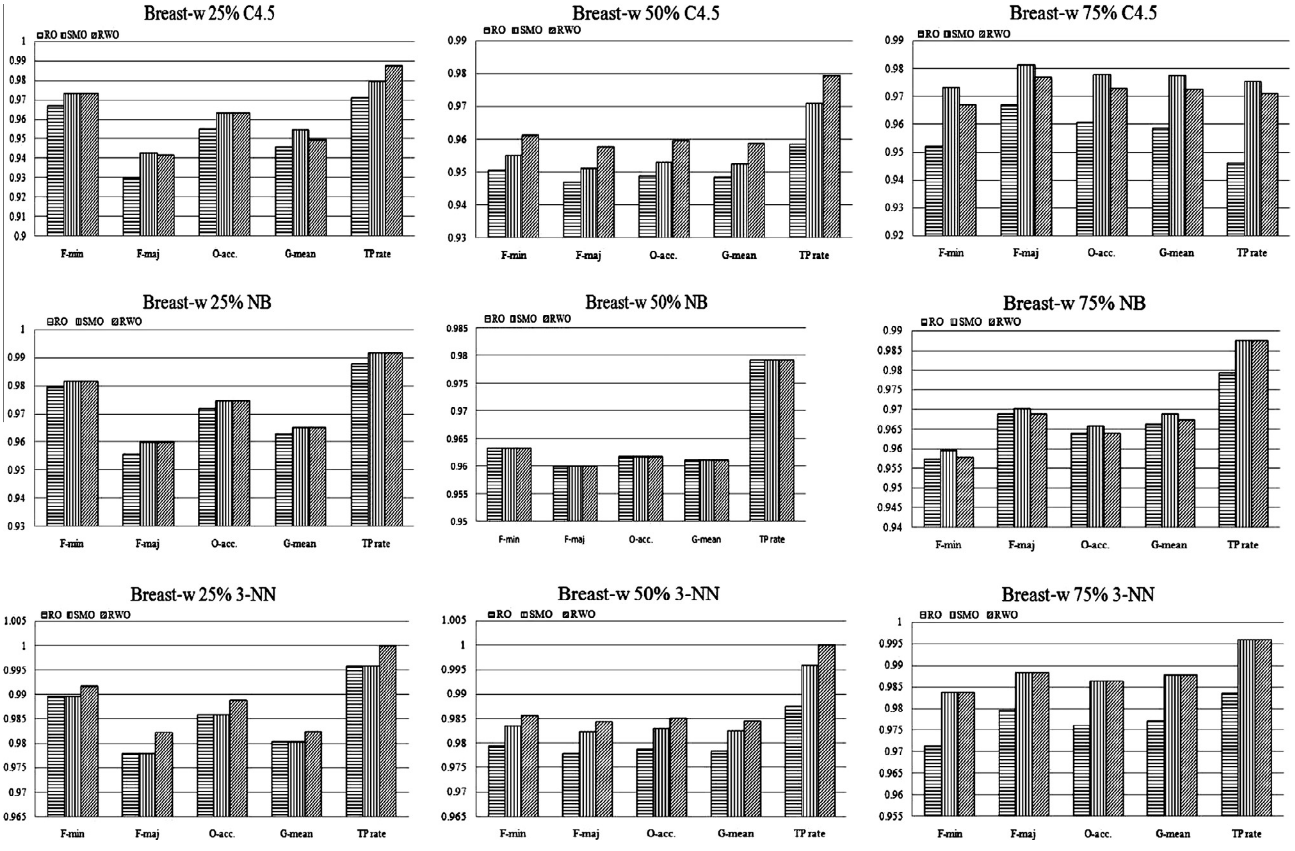

Fig. 2. Metric values when combining under-sampling and over-sampling.

Author's personal copy

minority class will increase their chance to become neighbors of a test sample, thus *k*NN could increase the classification performance of the minority class. A Naïve Bayes classifier calculates the posterior probability for a test sample, and the posterior probability represents the degree to which the sample is a member of a class. The calculation is based on the priori probabilities of the training data, and over-sampling changes the distribution of the training data by increasing the priori probability of the minority class data, thus over-sampling influences the classification performance of Naïve Bayes classifiers.

It is obvious that random over-sampling takes the least time among the three approaches to generate synthetic samples. SMOTE needs to conduct *k*NN to obtain the *k* neighbors for a chosen sample before generating synthetic samples, and RWO-Sampling needs to calculate the mean and standard deviation for all attributes using the minority class data. Since *k*-nearest neighbor algorithm is time consuming compared with the calculation of the mean and standard deviation conducted by RWO-Sampling, we say that RWO-Sampling takes less time than SMOTE for generating synthetic samples. Different methods are employed by the three over-sampling approaches. RO-Sampling is simple, but it increases the possibility of over-fitting. SMOTE uses linear interpolation for sample generation, and new samples fall on the line segment connected by two neighbors. It does not expand the space occupied by the minority class data, and also changes the original data distribution of the minority class. When generating synthetic samples, RWO-Sampling tries to keep the minority class data distribution unchanged while probably expands the space occupied by the minority class data, thus it has high generalization capability and does well on imbalanced data classification.

In order to meet the conclusions of the above mentioned theorems, each original minority class instance should be used to generate the same number of synthetic instances when RWO-Sampling is performed, thus the minority class must be over-sampled at rate of 100s% of its original size, where *s* is a positive integer. So over-sampling rate cannot be a float. That is the one drawback of RWO-Sampling. We do not generate synthetic samples around the mean point of the real minority class data through random walk model, since it also tends to increase the likelihood of over-fitting.

On datasets with all continuous attributes, RWO-Sampling perform well in many cases. On datasets with all discrete attributes or hybrid attributes, even though RWO-Sampling does not perform as well as it does on datasets with all continuous attributes, it still yields promising results. When generating synthetic values for a discrete attribute, roulette may not be the optimal approach. It is worth studying the methods used to generate values for discrete attributes while implementing random walk sampling approach on continuous attributes, and evaluating their effects on the classification performance of imbalanced problems.

The over-sampling approach is based on the strong assumption that attributes are independent. This assumption does not hold in many practical problems, and may have impact on the performance of the proposed algorithm. Some methods may be possible ways to solve this problem, such as using a multivariate normal distribution to generate multivariate normal random values for the synthetic data, but this needs further research.

Acknowledgements

We would like to thank the anonymous reviewers and the journal editors for their valuable comments and suggestions, which have significantly improved this paper. This research is partially supported by the National Natural Science Foundation of China (61170145 and 61373081), the Specialized Research Fund for the Doctoral Program of Higher Education of China (20113704110001), the Natural Science Foundation of Shandong (ZR2010FM021), the Scientific Technology and Development Project of Shandong (2013GGX10125) and the Taishan Scholar Project of Shandong, China.

References

[1] R. Barandela, J.S. Sanchez, V. Garcia, E. Rangel, Strategies for learning in class imbalance problems, Pattern Recogn. 36 (2003) 849–851.

[2] Z.-H. Zhou, X.-Y. Liu, Training cost-sensitive neural networks with methods addressing the class imbalance problem, IEEE Trans. Knowl. Data Eng. 18 (2006) 63–77.

[3] G.E.A.P.A. Batista, R.C. Prati, M.C. Monard, A study of the behavior of several methods for balancing machine learning training data, ACM SIGKDD Explor. Newsl. 6 (1) (2004) 20–29.

[4] A. Sun, E.-P. Lim, Y. Liu, On strategies for imbalanced text classification using SVM: a comparative study, Decis. Support Syst. 48 (1) (2009) 191–201.

[5] S. Kotsiantis, D. Kanellopoulos, P. Pintelas, Handling imbalanced datasets: a review, GESTS Int. Trans. Comput. Sci. Eng. 30 (2006) 1–12.

[6] S.-J. Yen, Y.-S. Lee, Cluster-based under-sampling approaches for imbalanced data distributions, Expert Syst. Appl. 36 (2009) 5718–5727.

[7] T.M. Mitchell, Machine Learning, McGraw Hill, 1997.

[8] N. Chawla, K. Bowyer, L. Hall, W.P. Kegelmeyer, SMOTE: synthetic minority over-sampling technique, J. Artif. Intell. Res. 16 (2002) 321–357.

[9] I. Tomek, Two modifications of CNN, IEEE Trans. Syst. Man Commun. SMC-6 (1976) 769–772.

[10] H. Zhang, Z. Wang, A normal distribution-based over-sampling approach to imbalanced data classification, Lect. Notes Artif. Intell. 7120 (2011) 83–96.

[11] H. Han, W.-Y. Wang, B.-H. Mao, Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning, Lect. Notes Comput. Sci. 3644 (2005) 878–887.

[12] H. Guo, H.L. Viktor, Learning from imbalanced data sets with boosting and data generation: the DataBoost-IM approach, SIGKDD Explor. Newsl. 6 (2004) 30–39.

[13] A. Estabrooks, T. Jo, N. Japkowicz, A multiple resampling method for learning from imbalanced data sets, Comput. Intell. 20 (1) (2004) 18–36.

[14] Y. Peng, J. Yao, AdaOUBoost: adaptive over-sampling and under-sampling to boost the concept learning in large scale imbalanced data sets, in: MIR'10, Philadelphia, Pennsylvania, USA, March, 2010, pp. 111–118.

[15] X. Zhu, Lazy bagging for classifying imbalanced data, in: 17th IEEE International Conference on Data Mining, 2007, pp. 763–768.

[16] Y. Sun, M.S. Kamel, A.K.C. Wong, Y. Wang, Cost-sensitive boosting for classification of imbalanced data, Pattern Recogn. 40 (2007) 3358–3378.

[17] M. Woźniak, M. Grana, E. Corchado, A survey of multiple classifier systems as hybrid systems, Inf. Fusion 16 (2014) 3–17.

[18] X. Li, L. Wang, E. Sung, AdaBoost with SVM-based component classifiers, Eng. Appl. Artif. Intell. 21 (2008) 785–795.

[19] G. Giacinto, R. Perdisci, M.D. Rio, F. Roli, Intrusion detection in computer networks by a modular ensemble of one-class classifiers, Inf. Fusion 9 (2008) 69–82.

[20] C. Drummond, R.C. Holte, C4.5, Class imbalance and cost sensitivity: why under-sampling beats over-sampling, in: Proceedings of the ICML'03 Workshop on Learning from Imbalanced Data Sets, 2003.

[21] Z. Zheng, X. Wu, R. Srihari, Feature selection for text categorization on imbalanced data, SIGKDD Explor. Newsl. 6 (1) (2004) 80–89.

[22] M.-C. Chen, L.-S. Chen, C.-C. Hsu, W.-R. Zeng, An information granulation based data mining approach for classifying imbalanced data, Inf. Sci. 178 (2008) 3214–3227.

[23] Y. Tang, Y. Zhang, N. Chawla, S. Krasser, SVMs modeling for highly imbalanced classification, IEEE Trans. Syst. Man Cybern. Part B 39 (1) (2009) 281–288.

[24] M.A. Mazurowski, P.A. Habas, J.M. Zurada, J.Y. Lob, J.A. Baker, Training neural network classifiers for medical decision making: the effects of imbalanced datasets on classification performance, Neural Networks 21 (2008) 427–436.

[25] A. Fernández, M.J.d. Jesus, F. Herrera. Hierarchical fuzzy rule based classification systems with genetic rule selection for imbalanced data-sets, doi: 10.1016/j.ijar.2008.11.004.

[26] F. Provost, T. Fawcett, R. Kohavi, The case against accuracy estimation for comparing induction algorithms, in: Proceedings of the Fifteenth International Conference on Machine Learning, Morgan Kaufmann, San Francisco, CA, 1998, pp. 445–453.

[27] G. Wu, E.Y. Chang, KBA: kernel boundary alignment considering imbalanced data distribution, IEEE Trans. Knowl. Data Eng. 17 (6) (2005) 786–795.

[28] M. Kubat, S. Matwin, Addressing the curse of imbalanced training sets: one-sided selection, in: Proceedings of the Fourteenth International Conference on Machine Learning, Morgan Kaufmann, San Francisco, CA, 1997, pp. 179–186.