Hybrid Sampling with Bagging for Class Imbalance Learning

Yang Lu[1], Yiu-ming Cheung[1(✉)], and Yuan Yan Tang[2]

[1] Department of Computer Science,
Hong Kong Baptist University, Hong Kong, China
{yanglu,ymc}@comp.hkbu.edu.hk
[2] Department of Computer and Information Science,
Faculty of Science and Technology, University of Macau, Macau, China
yytang@umac.mo

Abstract. For class imbalance problem, the integration of sampling and ensemble methods has shown great success among various methods. Nevertheless, as the representatives of sampling methods, undersampling and oversampling cannot outperform each other. That is, undersampling fits some data sets while oversampling fits some other. Besides, the sampling rate also significantly influences the performance of a classifier, while existing methods usually adopt full sampling rate to produce balanced training set. In this paper, we propose a new algorithm that utilizes a new hybrid scheme of undersampling and oversampling with sampling rate selection to preprocess the data in each ensemble iteration. Bagging is adopted as the ensemble framework because the sampling rate selection can benefit from the Out-Of-Bag estimate in bagging. The proposed method features both of undersampling and oversampling, and the specifically selected sampling rate for each data set. The experiments are conducted on 26 data sets from the UCI data repository, in which the proposed method in comparison with the existing counterparts is evaluated by three evaluation metrics. Experiments show that, combined with bagging, the proposed hybrid sampling method significantly outperforms the other state-of-the-art bagging-based methods for class imbalance problem. Meanwhile, the superiority of sampling rate selection is also demonstrated.

Keywords: Class imbalance learning · Hybrid sampling · Sampling method · Ensemble method

1 Introduction

In many classification applications, the problem of learning from imbalanced data is still one of the challenges [22], where the number of data in the minority

Y.M. Cheung is the corresponding author. This work was supported by the Faculty Research Grants of Hong Kong Baptist University (HKBU): FRG2/14-15/075, and by the National Natural Science Foundation of China under Grant Number: 61272366.

J. Bailey et al. (Eds.): PAKDD 2016, Part I, LNAI 9651, pp. 14–26, 2016.
DOI: 10.1007/978-3-319-31753-3_2

class is severely under-represented and overwhelmed by the majority class. In this case, the distribution of the classes is skewed. Subsequently, usual classification methods will generate poor results because the distribution is one of the most important factors that affect the performance [7,15,19].

Usually, standard classification algorithms assume that the class distribution is balanced, and the misclassification cost is equal for both classes. However, there exists some cases that the class distribution is skewed and the misclassification cost is extremely unequal. Further, sometimes people focus more on the minority class because it usually contains more information and interest than the majority class. Let us take cancer diagnosis as an example, the number of patients who have cancer is much less than the number of healthy people in regular checkups. It is obvious that the cost for misdiagnosing a healthy person to be sick, which only brings the person mental stress and more payment to further diagnosis, is much less than diagnosing a patient to be health, which may lead to the loss of the patient's life. Therefore, when dealing with imbalanced data, the misclassification cost is one of the most significant factors that affect the process of learning. In addition to the algorithms, evaluation metrics also play important roles in imbalanced learning. Suppose there are 100 cancer patients out of 10,000 people, the normal classifier will tend to predict "healthy", because even all predictions are "healthy", the accuracy of this classifier is still as high as 99 %. Therefore, simply using the accuracy or error rate is not comprehensive enough to measure the performance of a classifier dealing with imbalanced data. Usually, three evaluation metrics for class imbalance problem, i.e. AUC, F1 and G-mean, will be used. In this paper, we focus on the binary classification problem, and following the convention, we treat samples in the minority class as positive and samples in the majority class as negative.

Among various of methods to tackle the imbalance problem, sampling methods have been proved to be effective. Several studies have shown that training on the balanced data set by sampling methods can achieve better overall classification performance than the original imbalanced one [11,21]. Usually, the sampling methods, such as random undersampling or oversampling, are integrated with ensemble methods, such as bagging or boosting, in order to overcome their drawbacks and provide more diversity to the boosted classifier [12].

However, ensemble-based undersampling and oversampling cannot outperform each other, e.g. see a recent survey [12], in which RUSBoost [17] (undersampling based) wins SMOTEBoost [9] (oversampling based) 22 times, draws 4 times and loses 18 times and UnderBagging [1] (undersampling based) wins SMOTEBagging [20] (oversampling based) 18 times, draws 1 time and loses 25 times (shown in Tables XX and XXI in [12]). It can be seen that the results generated by ensemble-based undersampling or oversampling highly depend on the data. In other words, some data has better performance with undersampling, while the other ones with oversampling. Therefore in terms of the sampling process, it is expected that the hybrid of undersampling and oversampling can take advantage of their properties. That is, the hybrid sampling generally outperforms each individual sampling method because undersampling and oversampling are complementary to each other and cure the skewed distribution of class imbalanced data in different extents.

In addition, no matter undersampling, oversampling or even hybrid sampling is adopted, the sampling rate is one of the key factors that affect the performance of a classifier. Most of the sampling based methods, no matter integrated with ensemble methods or not, tend to make the number of data in both classes balanced after sampling, based on the simple assumption that the balanced training set produces the best result. However, producing poor model caused by training on imbalanced data does not imply that the optimal model is produced by training on totally balanced data by sampling. Estabrooks *et al.* [11] shows that the best results from undersampling and oversampling are not always on the balanced case. It means that conducting sampling to achieve the balanced data for training is not guaranteed to be the best solution. Furthermore, the best sampling rate depends on the distribution and complexity of the data set. That is, the best sampling rate of one dataset would be different from one of another dataset. Therefore, it is necessary to select a proper sampling rate for the sampling methods on each data set. To the best of our knowledge, selection of the proper sampling rate has yet to be studied in the literature.

In this paper, we therefore propose a novel method for the class imbalance problem called Hybrid Sampling with Bagging (HSBagging). It adopts a new hybrid scheme that conducts random undersampling in tandem with oversampling technique SMOTE at a certain sampling rate in each bagging iteration. The sampling rate is selected by Out-Of-Bag (OOB) estimate on a specified metric for each data set. To reduce the computational cost, the sampling rate is only estimated in the first several iterations and the averaged estimated sampling rate will be utilized in the rest iterations then. The major advantages of HSBagging are:

- The new hybrid sampling scheme can take advantage of the merits of both undersampling and oversampling.
- Sampling rate selection can effectively select a proper sampling rate which fits the data to achieve best performance.
- The preferred metric can be selected during OOB estimate according to the application requirement.

To validate the effectiveness of the hybrid sampling scheme and sampling rate selection, four experiments are conducted on 26 UCI data sets with statistical significance tests. The experiments show that the proposed HSBagging significantly outperforms individual sampling method with bagging and verify that both hybrid sampling and sampling rate selection contribute to the superiority of HSBagging.

2 Related Work

Over the past years, much work devoted to solve the class imbalance problem has shown great success in the corresponding application domain, in which sampling methods and ensemble methods are two major branches.

Random oversampling and undersampling are two elementary sampling methods to cure imbalance, by randomly replicating data in minority class and discarding data in majority class, respectively. The drawbacks of them are that oversampling will easily cause overfitting and undersampling may discard useful data that leads to information loss. As an improvement to random oversampling, Synthetic Minority Over-sampling TEchnique (SMOTE) [8] synthesizes artificial data in the minority class instead of replication. Borderline-SMOTE [13] and ADASYN [14] improve SMOTE by assuming that the samples close to the borderline are more important, thus synthesize more data there. The idea of combining undersampling and oversampling has been mentioned in [20]. It combines undersampling with oversampling to create a training set with the same number of positive and negative samples. The number of samples in each class after sampling is determined by a predefined re-sampling rate $a\%$.

Ensemble methods such as bagging [4] and boosting [16] cannot solve the imbalanced problem themselves. Usually, they are combined with sampling methods to utilize the diversity provided by sampling to enhance the ensemble classifier. A comprehensive review of ensemble methods for class imbalance problem can be found in [12]. OverBagging [20] and UnderBagging [1] combine random undersampling and oversampling with bagging, respectively. They adopt oversampling or undersampling after bootstrapping the training data to create a balanced training set. As an improvement of OverBagging, SMOTE-Bagging [20] combines SMOTE with random oversampling and the sampling rate of SMOTE increases in every iteration to provide more diversity. As the counterpart of bagging-based methods, SMOTEBoost [8] and RUSBoost [17] are boosting-based. They created balanced training set by SMOTE and random undersampling in each boosting iteration. After sampling applied, the sample weights are normalized. The following steps are the same as Adaboost [16]. IIVotes [3] combines IVotes ensemble [5] and SPIDER [18] data preprocessing to obtain improved balance between the sensitivity and specificity for the minority class.

3 The Proposed Method

Since training in the balanced data set is not guaranteed to produce the best result [11], the proposed HSBagging does not aim to create the balanced training set, but depending on a specified sampling rate p, which is different from the hybrid scheme in [20]. In HSBagging, the minority class is enlarged by p and meanwhile the majority class is shrank by p. Conducting undersampling and oversampling at sampling rate p at the same time can explore the best sampling rate from severe imbalance, slight imbalance to balance or even reversed imbalance (i.e. the minority class becomes majority after sampling). Since each data set tends to have different best sampling rate, it is necessary to estimate the sampling rate during bagging. HSBagging estimates the best sampling rate by Out-Of-Bag (OOB) estimate, which is used to estimate parameters in the bootstrapped set by leaving the samples not selected by bootstrapping as validation set. There are two advantages of using OOB estimate: (1) it acts as

Algorithm 1. Hybrid Sampling Bagging

```
Require: Training set $S = \{(\mathbf{x}_i, y_i)\}$, $i = 1, ..., n$ and $y_i \in \{+1, -1\}$, weak learner $L$, number of iterations $T$, number of iterations $k$ for sampling rate estimate, sampling rate selection set $I$, evaluation metric $f_m$.
 1: for $t = 1$ to $T$ do
 2:     Create a training set $B$ by bootstrapping each class respectively.
 3:     Create the OOB set $B_o$.
 4:     if $t \leq k$ then
 5:         for each $p$ in $I$ do
 6:             Create the training set $B'$ by both undersampling and SMOTE set $B$ at sampling rate $p$.
 7:             Learn the classifier $h'_p = L(B')$.
 8:         end for
 9:         $p_t^* = \text{argmax}_{p \in I} f_m(h'_p, B_o)$.
10:         $h_t = h'_{p_t^*}$.
11:     else
12:         $p_t^* = \frac{1}{k} \sum_{i=1}^{k} p_i^*$.
13:         Create the training set $B'$ by both undersampling and SMOTE set $B$ at sampling rate $p_t^*$.
14:         Learn the classifier $h_t = L(B')$.
15:     end if
16: end for
17: Output: $H(\mathbf{x}) = sign(\sum_{t=1}^{T} h_t(\mathbf{x}))$
```

validation set, but it needs not separate part of data from training set; (2) the model needs not be trained again on the original training set with the estimated best parameter. Therefore, in HSBagging, the sampling rate is regarded as a parameter to be estimated. The estimate criterion is based on a specified evaluation metric, because commonly used accuracy for classification cannot well assess the class imbalance problem. Usually, it will be computational expensive if the OOB estimate is conducted on every bagging iteration. To save computational cost, we only conduct OOB sampling in the first k iterations, and the rest iterations will use the averaged estimated best sampling rates of the previous iterations.

The proposed HSBagging is shown in Algorithm 1. In each iteration, the training data is bootstrapped on each class, respectively, as shown in Line 2. The bootstrapped training set B keeps the same number of samples for the majority class and minority class as before bootstrapping. The OOB set B_o is then constructed by the samples that are not selected into B. The sampling rate selection is only conducted in the first k iterations in order to save computational cost. In these k iterations, undersampling and SMOTE are used to process B at the same time at sampling rate p in Line 6. The sampling rate $p \in [0, 1]$ is set to each of the values in the set I, in order to find a proper sampling rate for the current data set. For undersampling, it randomly selects $n_{min} + (1 - p)$ $(n_{maj} - n_{min})$ samples from the majority class, and for SMOTE, it synthesizes $p(n_{maj} - n_{min})$ more samples from the minority class and adds them to the original minority class, where n_{maj} and n_{min} represents the number of samples in the majority class and minority class. Therefore, when $p = 0$, the data set B' after sampling is as same as the original data set B and when $p = 1$, the number of samples in the majority class and the minority class gets reversed after sampling. Thus, undersampling and SMOTE are effectively combined. By learning B' by the learner L, a classifier h'_p can be built for the sampling rate p. After that, $f_m(h'_p, B_o)$ estimates the performance of h'_p on the OOB set B_o and metric f_m. The sampling rate p^*_t is set to the p associated with best performance on f_m and $h'_{p^*_t}$ is set to the classifier of the t's iteration h_t. After k iterations of sampling rate selection, the following iterations simply use the averaged value of the first k selected sampling rates to do sampling and train the classifier h_t. At last, each individual classifier is combined into the final boosted classifier $H(x)$.

The computational complexity of HSBagging is $O((T + (k-1)|I|)\mathcal{L}(n))$, where $|\cdot|$ is the cardinality of a set and $\mathcal{L}(n)$ is the computational cost of the weak learning L with n training samples. Compared with SMOTEBagging [20], although HSBagging costs $(k-1)|I|$ more iterations to select the sampling rate, it trains only on n samples in each iteration, while SMOTEBagging trains $2n_{maj}$ samples. If the number of iterations T is relatively large and the imbalance problem of the data set is severe, HSBagging will be computational cheaper than SMOTEBagging.

4 Experiments

In this section, we conducted four experiments. Experiment 1 shows the times of best performance on each sampling rate for each data set. It verifies that the sampling rate corresponding to the best performance varies from data to data. Experiment 2 compares the proposed HSBagging with bagging on original imbalanced data set, UnderBagging [1], SMOTEBagging [20] and IIVotes [3]. We denote SMOTE with bagging by full sampling rate ($p = 1$) as SMOTEBagging-1, and SMOTE with bagging by increasing sampling rate in each iteration, which is proposed in [20], as SMOTEBagging-2. Experiment 3 compares HSBagging with those methods on different sampling rates to verify that the superior performance is not only caused by sampling rate selection, but also effected by the hybrid sampling scheme. Experiment 2 and 3 verify that hybrid sampling is significantly better than individual sampling. Finally, Experiment 4 shows the performance of HSBagging on different number k of iterations for sampling rate estimation.

All experiments were conducted on 26 data sets from UCI data repository [2] summarized in Table 1, which cover a wide range of applications and imbalance ratios. The imbalance ratio (IR) is calculated by the number of data in the majority class divided by the number of data in the minority class. All experiments adopted 5-fold cross validation, where 80 % of the samples in each data were used for training and the rest for testing in each fold. The final results were averaged by 10 runs of experiments. The number of iterations T in bagging was set at 10 for all methods, except IIVotes, whose iteration was automatically determined. CART [6] was adopted as the base learner for all bagging-based methods.

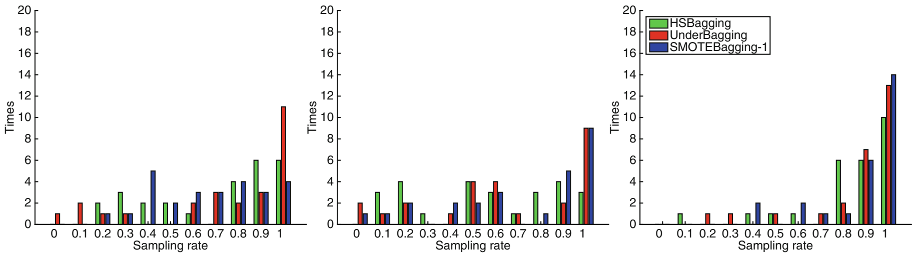
Table 1. Information of 26 UCI data sets.

Data set	#Instance	#Attribute	Minority class	Majority class	IR
glass-2	214	9	bwnfp	remainder	1.8
pima	768	8	positive	negative	1.9
vehicle-2	846	18	saab	remainder	2.9
vehicle-1	846	18	opel	remainder	3.0
glass-123vs567	214	9	non-window	remainder	3.2
wpbc	198	33	recur	nonrecur	3.2
vehicle-4	846	18	van	remainder	3.3
haberman	306	3	within-5-year	5-year-or-longer	2.8
cmc	1473	9	long-term	remainder	3.4
ecoli-2	336	7	im	remainder	3.4
car	1728	6	acc	remainder	3.5
wine-quality	6497	11	score 7	remainder	5.0
segment	2310	19	brickface	remainder	6.0
glass-7	214	9	headlamp	remainder	6.4
yeast-4	1484	8	me3	remainder	8.1
ecoli-4	336	7	imU	remainder	8.6
pageblocks	5473	10	remainder	text	8.8
mf-morph	2000	6	class 10	remainder	9.0
mf-zernike	2000	47	class 10	remainder	9.0
cm1	498	21	defects	no-defects	9.2
satimage	6435	36	class 4	remainder	9.3
yeast-5vs347810	1484	8	me2	mit;me3;exc;vac;erl	9.4
abalone	4177	8	class 7	remainder	9.7
balance	625	4	balanced	remainder	11.8
glass-127vs6	214	9	tableware	bwfp;bwnfp;headlamps	19.4
yeast-6	1484	8	me1	remainder	32.7

The number of nearest neighbor for all kNN related methods was set at 5. In the experiments, three evaluation metrics, i.e. AUC, F1 and G-mean, which are commonly adopted as the benchmark assessment metric for class imbalance learning [15], were used to measure the effectiveness of methods.

4.1 Experiment 1: Sampling Rate Verification

Figure 1 shows the number of data sets with the best performance on different sampling rate from 0 to 1. In this experiment, UnderBagging and SMOTEBagging-1 were set to process the data on a specific sampling rate p instead of producing balanced training set. UnderBagging conducted undersampling by discarding $p(n_{maj} - n_{min})$ samples from the majority class while SMOTEBagging-1 conducted SMOTE by synthesizing $p(n_{maj} - n_{min})$ from the



(a) (b) (c)

Fig. 1. Number of data sets with best (a) AUC, (b) F1, and (c) G-mean performance generated on different sampling rates.

minority class. HSBagging conducted both undersampling and SMOTE at sampling rate p instead of OOB estimate as described in Algorithm 1. From Fig. 1, it can be observed that the best results of all three methods almost appear on all sampling rates on each evaluation metric. Especially, some best sampling rates of HSBagging occur at sampling rate 0 or 1, which means that the original imbalanced data or the reversed imbalanced data may also be able to generate good results. Furthermore, the sampling rate corresponding to the best performance on different evaluation metrics may also be different, e.g. higher sampling rates generate relatively better results on G-mean. Thus, we can argue that, no matter which sampling method is adopted, selecting a proper sampling rate on a specific metric for each data set is effective and necessary.

4.2 Experiment 2: Comparative Studies

Since CART generates discrete outputs, AUC can only be calculated by the ensemble of CART classifiers and is not available for individual CART classifier. Therefore, we use F1 and G-mean as the metric f_m to select the best sampling rate for HSBagging, denoted as HSBagging-F1 and HSBagging-Gmean, respectively. The number of iterations k for sampling rate estimate is set to 3 and sampling rate selection set $I = \{0, 0.2, 0.4, 0.6, 0.8, 1\}$.

The pairwise comparisons by Wilcoxon signed-rank test [10] is provided to show the statistical significance of the compared methods. It measures the difference between two methods and rank their magnitude among data sets. Greater difference will count more in this evaluation. The sum of ranks of each method is calculated by $R^+ = \sum_{d_i>0} rank(|d_i|) + \frac{1}{2}\sum_{d_i=0} rank(|d_i|)$ and $R^- = \sum_{d_i<0} rank(|d_i|) + \frac{1}{2}\sum_{d_i=0} rank(|d_i|)$ where d_i is the difference of the result of the ith data set. If the significance value N with a certain significance level α is greater than $T = min\{R^+, R^-\}$, the null hypothesis is rejected which indicates one method significantly outperforming the other one. In the following Tables 2, 3 and 4, as well as Table 5 in Sect. 4.4, the method shown in the left upper corner is marked as + and the compared methods are marked as −. The sign (+,−) in the T column indicates which method wins more ranks and the symbol • indicates the significance with significance level $\alpha = 0.05$.

Table 2. Wilcoxon signed-rank test for HSBagging-F1 and other methods.

HSBagging-F1 vs.	AUC			F1			G-mean		
	R^+	R^-	T	R^+	R^-	T	R^+	R^-	T
Bagging	335.00	16.00	• **16.00** (+)	331.00	20.00	• **20.00** (+)	341.00	10.00	• **10.00** (+)
UnderBagging	179.00	172.00	172.00(+)	285.00	66.00	• **66.00** (+)	61.00	290.00	• **61.00** (−)
SMOTEBagging-1	205.00	146.00	146.00(+)	279.50	71.50	• **71.50** (+)	264.00	87.00	• **87.00** (+)
SMOTEBagging-2	226.00	125.00	125.00(+)	282.00	69.00	• **69.00** (+)	149.50	201.50	149.50 (−)
IIVotes	350.00	1.00	• **1.00** (+)	350.00	1.00	• **1.00** (+)	328.00	23.00	• **23.00** (+)

Table 3. Wilcoxon signed-rank test for HSBagging-Gmean and other methods.

HSBagging-Gmean vs.	AUC			F1			G-mean		
	R^+	R^-	T	R^+	R^-	T	R^+	R^-	T
Bagging	344.00	7.00	• **7.00** (+)	270.00	81.00	• **81.00** (+)	351.00	0.00	• **0.00** (+)
UnderBagging	221.00	130.00	130.00(+)	296.50	54.50	• **54.50** (+)	177.00	174.00	174.00 (+)
SMOTEBagging-1	269.50	81.50	• **81.50** (+)	210.00	141.00	141.00 (+)	350.00	1.00	• **1.00** (+)
SMOTEBagging-2	257.50	93.50	• **93.50** (+)	171.00	180.00	171.00 (−)	289.00	62.00	• **62.00** (+)
IIVotes	351.00	0.00	• **0.00** (+)	351.00	0.00	• **0.00** (+)	351.00	0.00	• **0.00** (+)

Table 4. Wilcoxon signed-rank test for HSBagging-Gmean and HSBagging-F1.

HSBagging-Gmean vs.	AUC			F1			G-mean		
	R^+	R^-	T	R^+	R^-	T	R^+	R^-	T
HSBagging-F1	276.00	75.00	• **75.00** (+)	117.50	233.50	117.50 (−)	313.00	38.00	• **38.00** (+)

Tables 2 and 3 show the Wilcoxon signed-rank test results of HSBagging-F1 and HSBagging-Gmean compared with the other methods. It can be seen that:

– HSBagging-F1 significantly outperforms all other methods on F1, and HSBagging -Gmean significantly outperforms bagging, SMOTEBagging-1, SMOTEBagging-2 and IIVotes on G-mean.
– Even though the sampling rate is not selected based on AUC, HSBagging-F1 and HSBagging-Gmean also achieve comparable or better performance on AUC. Especially, the performance of HSBagging-Gmean on AUC shows similar significance as its performance on G-mean.
– On G-mean, HSBagging-F1 outperform Bagging, SMOTEBagging-1 and IIVotes, and on F1, HSBagging-Gmean outperform Bagging, UnderBagging and IIVotes.

As a result, it can be observed that no matter the sampling rate is selected on which metric, HSBagging can produce superior results on each metric, especially on its selected metric, i.e. F1 and G-mean.

Table 4 shows the comparison between HSBagging-F1 and HSBagging-Gmean. Both of them have better results on their own selected metrics. However, HSBagging-Gmean significantly outperforms HSBagging-F1 on both AUC and G-mean while HSBagging-F1 is only slightly better than HSBagging-Gmean on F1. Therefore, overall speaking, HSBagging-Gmean performs better than HSBagging-F1.

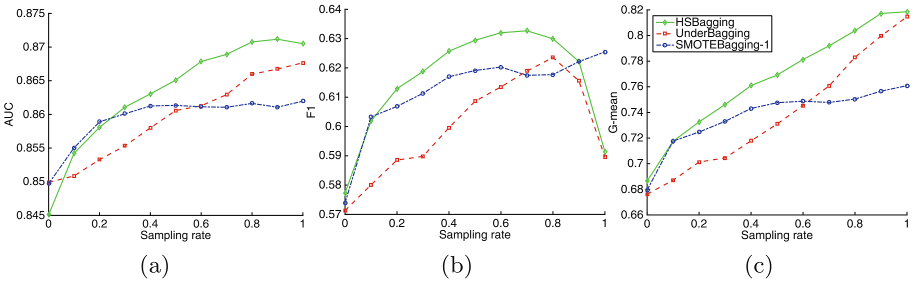


Fig. 2. Average performance of HSBagging, UnderBagging and SMOTEBagging-1 over different sampling rate in terms of (a) AUC, (b) F1, and (c) G-mean.

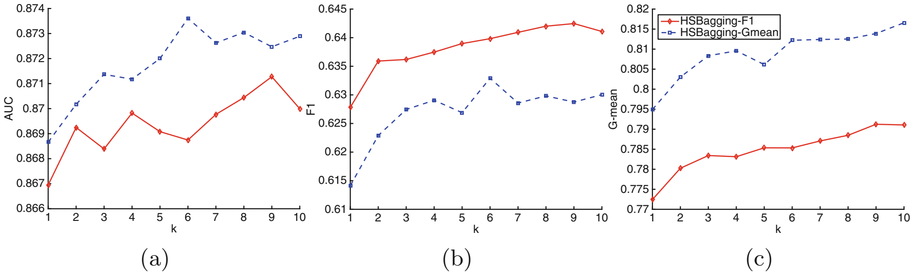


Fig. 3. Average Performance of HSBagging in terms of (a) AUC, (b) F1, and (c) G-mean, respectively, on different number of iterations for sampling rate estimate k.

4.3 Experiment 3: Sampling Rate Comparison

In addition to the sampling rate selection, the new hybrid sampling scheme also plays an important role in terms of the superiority of HSBagging. In this subsection, we show that the effectiveness of the proposed HSBagging depends on not only sampling rate selection, but also the hybrid scheme. The comparison of HSBagging with UnderBagging and SMOTEBagging-1 at different sampling rate is shown in Fig. 2 with the same setting as the experiment in Sect. 4.1. The figures are generated by averaging all 26 UCI data sets. On most of the sampling rates, HSBagging can achieve better results on average than UnderBagging and SMOTEBagging-1. Besides, the best results of HSBagging are better than the best results of UnderBagging and SMOTEBagging among all sampling rates. Figure 2 illustrates that, even sampling rate selection is adopted for UnderBagging and SMOTEBagging, the overall performance cannot be as good as HSBagging. That implies that HSBagging outperforming UnderBagging and SMOTEBagging benefits from not only the choice of a proper sampling rate, but also the hybrid scheme.

Table 5. A comparison of HSBagging with $p = 3$ to HSBagging with $p = 1$ and $p = 10$, respectively, using Wilcoxon signed-rank test.

p = 3 vs.	HSBagging-F1 on F1			HSBagging-Gmean on G-mean		
	R^+	R^-	T	R^+	R^-	T
p = 1	300.50	50.50	• **50.50** (+)	303.50	47.50	• **47.50** (+)
p = 10	165.50	185.50	165.50 (−)	101.50	249.50	101.50 (−)

4.4 Experiment 4: Parameter Selection

The performance of HSBagging-F1 and HSbagging-Gmean over different number k of iterations for sampling rate estimate is shown in Fig. 3. It can be observed that the performance increases from $k = 1$ to 3 for all metrics. After $k = 3$, the increase tends to be modest. Table 5 shows the statistical comparison of the performance of HSBagging on $k = 3$ against $k = 1$ and $k = 10$, respectively. To address the significance of the selected preferred metric, we compare HSBagging-F1 on F1 and HSBagging-Gmean on G-mean only. As shown in Table 5, HSBagging-F1 and HSBagging-Gmean on $k = 3$ significantly outperform the cases on $k = 1$ with the significance level $\alpha = 0.05$ on F1 and G-mean, respectively. Further, they have comparable performance in comparison with the cases on $k = 10$. Therefore, if the longer running time for some certain applications can be tolerated, the selection process is suggested to be conducted in every iteration because it has slightly better performance. Nevertheless, by a rule of thumb, setting $k = 3$ can usually produce significantly better results in comparison with the other bagging-based methods as shown in Tables 2 and 3, meanwhile saving the computational cost compared with $k = 10$.

5 Conclusion

This paper has first investigated the two problems for class imbalance problem. The first is that undersampling and oversampling with ensemble methods have their own irreplaceable property for the imbalanced data. Each of them can only performs well on part of data sets. Second, the sampling rate is crucial to the performance of sampling methods. The sampling rate in regard to the best performance differs from data to data.

A novel method called HSBagging has been proposed to solve the discovered problems. It adopts a new hybrid scheme of undersampling and oversampling integrated with bagging. During the sampling, the sampling rate is selected by OOB estimate on a specified metric. Experiments on 26 UCI data sets have shown that HSBagging can significantly outperform the other related bagging-based methods. The advantages of both the new hybrid sampling scheme and the sampling rate selection are also shown by experiments. Undoubtedly, the hybrid sampling and sampling rate selection are applicable to the other ensemble-based method like boosting as well.

References

1. Barandela, R., Valdovinos, R.M., Sánchez, J.S.: New applications of ensembles of classifiers. Pattern Anal. Appl. **6**(3), 245–256 (2003)
2. Blake, C., Merz, C.J.: UCI repository of machine learning databases (1998)
3. Błaszczyński, J., Deckert, M., Stefanowski, J., Wilk, S.: Integrating selective preprocessing of imbalanced data with ivotes ensemble. In: Szczuka, M., Kryszkiewicz, M., Ramanna, S., Jensen, R., Hu, Q. (eds.) RSCTC 2010. LNCS, vol. 6086, pp. 148–157. Springer, Heidelberg (2010)
4. Breiman, L.: Bagging predictors. Mach. Learn. **24**(2), 123–140 (1996)
5. Breiman, L.: Pasting small votes for classification in large databases and on-line. Mach. Learn. **36**(1–2), 85–103 (1999)
6. Breiman, L., Friedman, J., Stone, C.J., Olshen, R.A.: Classification and Regression Trees. CRC Press, Boca Raton (1984)
7. Chawla, N.V.: Data mining for imbalanced datasets: an overview. In: Maimon, O., Rokach, L. (eds.) Data Mining and Knowledge Discovery Handbook, pp. 853–867. Springer, Heidelberg (2005)
8. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: synthetic minority over-sampling technique. J. Artif. Intell. Res. **16**(1), 321–357 (2002)
9. Chawla, N.V., Lazarevic, A., Hall, L.O., Bowyer, K.W.: SMOTEBoost: improving prediction of the minority class in boosting. In: Lavrač, N., Gamberger, D., Todorovski, L., Blockeel, H. (eds.) PKDD 2003. LNCS (LNAI), vol. 2838, pp. 107–119. Springer, Heidelberg (2003)
10. Demšar, J.: Statistical comparisons of classifiers over multiple data sets. J. Mach. Learn. Res. **7**, 1–30 (2006)
11. Estabrooks, A., Jo, T., Japkowicz, N.: A multiple resampling method for learning from imbalanced data sets. Comput. Intell. **20**(1), 18–36 (2004)
12. Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., Herrera, F.: A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. IEEE Trans. Syst. Man Cybern. Part C Appl. Rev. **42**(4), 463–484 (2012)
13. Han, H., Wang, W.-Y., Mao, B.-H.: Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. In: Huang, D.-S., Zhang, X.-P., Huang, G.-B. (eds.) ICIC 2005. LNCS, vol. 3644, pp. 878–887. Springer, Heidelberg (2005)
14. He, H., Bai, Y., Garcia, E.A., Li, S.: Adasyn: adaptive synthetic sampling approach for imbalanced learning. In: IEEE International Joint Conference on Neural Networks, 2008. IJCNN 2008. (IEEE World Congress on Computational Intelligence), pp. 1322–1328. IEEE (2008)
15. He, H., Garcia, E.A.: Learning from imbalanced data. IEEE Trans. Knowl. Data Eng. **21**(9), 1263–1284 (2009)
16. Schapire, R.E.: The strength of weak learnability. Mach. Learn. **5**(2), 197–227 (1990)
17. Seiffert, C., Khoshgoftaar, T.M., Van Hulse, J., Napolitano, A.: Rusboost: a hybrid approach to alleviating class imbalance. IEEE Trans. Syst. Man Cybern. Part A Syst. Hum. **40**(1), 185–197 (2010)
18. Stefanowski, J., Wilk, S.: Improving rule based classifiers induced by modlem by selective pre-processing of imbalanced data. In: Proceedings of the RSKD Workshop at ECML/PKDD, Warsaw, pp. 54–65. Citeseer (2007)
19. Sun, Y., Wong, A.K., Kamel, M.S.: Classification of imbalanced data: a review. Int. J. Pattern Recognit. Artif. Intell. **23**(04), 687–719 (2009)

20. Wang, S., Yao, X.: Diversity analysis on imbalanced data sets by using ensemble models. In: IEEE Symposium on Computational Intelligence and Data Mining, 2009. CIDM 2009, pp. 324–331. IEEE (2009)
21. Weiss, G.M., Provosti, F.: The effect of class distribution on classifier learning: an empirical study. Rutgers Univ (2001)
22. Yang, Q., Wu, X.: 10 challenging problems in data mining research. Int. J. Inf. Technol. Decis. Mak. **5**(04), 597–604 (2006)