Neurocomputing ■ (■■■■) ■■■–■■■

Contents lists available at ScienceDirect

Neurocomputing

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

Neighbourhood sampling in bagging for imbalanced data

Jerzy Błaszczyński *, Jerzy Stefanowski **

Institute of Computing Sciences, Poznań University of Technology, 60-965 Poznań, Poland

ARTICLE INFO

Article history:
Received 21 December 2013
Received in revised form
27 April 2014
Accepted 26 July 2014

Keywords:
Class imbalance
Ensemble classifiers
Bagging

ABSTRACT

Various approaches to extend bagging ensembles for class imbalanced data are considered. First, we review known extensions and compare them in a comprehensive experimental study. The results show that integrating bagging with under-sampling is more powerful than over-sampling. They also allow to distinguish Roughly Balanced Bagging as the most accurate extension. Then, we point out that complex and difficult distribution of the minority class can be handled by analyzing the content of a neighbourhood of examples. In our study we show that taking into account such local characteristics of the minority class distribution can be useful both for analyzing performance of ensembles with respect to data difficulty factors and for proposing new generalizations of bagging. We demonstrate it by proposing Neighbourhood Balanced Bagging, where sampling probabilities of examples are modified according to the class distribution in their neighbourhood. Two of its versions are considered: the first one keeping a larger size of bootstrap samples by hybrid over-sampling and the other reducing this size with stronger under-sampling. Experiments prove that the first version is significantly better than existing over-sampling bagging extensions while the other version is competitive to Roughly Balanced Bagging. Finally, we demonstrate that detecting types of minority examples depending on their neighbourhood may help explain why some ensembles work better for imbalanced data than others.

1. Introduction

An analysis of challenging real-world classification problems still reveals difficulties in finding accurate classifiers. One of the sources of these difficulties is class imbalance in data, where at least one of the target classes contains a much smaller number of examples than the other classes. For instance, in medical problems the number of patients requiring special attention (e.g., therapy or treatment) is usually much smaller than the number of patients who do not need it. Similar situations occur in other problems, such as fraud detection, risk management, technical diagnostics, image recognition, text categorization or information filtering. In all those problems, the correct recognition of the minority class is of key importance. Nevertheless, class imbalance constitutes a great difficulty for most learning algorithms. Often the resulting classifiers are biased toward the majority classes and fail to recognize examples from the minority class. As it turns out, even ensemble methods, where multiple classifiers are trained to deal with complex classification tasks are not particularly well suited to this problem.

Although the difficulty with learning classifiers from imbalanced data has been known earlier from applications, this challenging problem has received a growing research interest in the last decade and a number of specialized methods have already been proposed, for their review see, e.g., [11,17,18,40]. In general, they may be categorized into *data level* and *algorithm level* ones. Methods within the first category try to re-balance the class distribution inside the training data by either adding examples to the minority class (*over-sampling*) or removing examples from the majority class (*under-sampling*). They also include informed pre-processing methods as, e.g., SMOTE [10] or SPIDER [37].

The other category of algorithm level methods involves specific solutions dedicated to improving a given classifier. They usually include modifications of the learning algorithm, its classification strategy or adaptation to the cost sensitive framework. Within the algorithm level approaches, *ensembles* are also quite often applied. However, as the standard techniques for constructing ensembles are rather too overall accuracy oriented they do not sufficiently recognize the minority class and new extensions of standard techniques have been introduced. These new proposed solutions usually either employ pre-processing methods before learning component classifiers or embed the cost-sensitive framework in the ensemble learning process; see their review in [13,29]. Most of these ensembles are based on known strategies from bagging, boosting or random forests.

* Principal corresponding author.
** Corresponding author.
E-mail addresses: jerzy.blaszczynski@cs.put.poznan.pl (J. Błaszczyński), jerzy.stefanowski@cs.put.poznan.pl (J. Stefanowski).

http://dx.doi.org/10.1016/j.neucom.2014.07.064
0925-2312/

Although the ensemble classifiers are recognized as a remedy to imbalanced problems, there is still a lack of a wider study of their properties. Authors often compare their proposals against the basic versions of other methods or compare over a too limited collection of data sets. Up to now, only two quite comprehensive studies were carried out in different experimental frameworks [13,24]. The first study [13] covers comparison of 20 different ensembles from simple modifications of bagging or boosting to complex cost or hybrid approaches. The main conclusion from this study is that simple versions of under-sampling or SMOTE re-sampling combined with bagging works better than more complex solutions. In the second study [24], two best boosting and bagging ensembles are compared over noisy and imbalanced data. The experimental results show that bagging significantly outperforms boosting. The difference is more significant when data are more noisy. The similar observations on good performance of under-sampling generalizations of bagging vs. cost like generalization of boosting have been recently reported in [2]. Furthermore, the most recent chapter of [29] includes a limited experimental study showing that new ensembles specialized for class imbalance should work better than an approach consisting of first pre-processing data and then using standard ensembles.

Following these related works which show good performance of bagging extensions for class imbalance vs. other boosting like or cost sensitive proposals, we have decided to focus our interest in this paper on studying more deeply bagging ensembles and to look for possible other directions of their generalizations. First, we want to study behaviour of bagging extensions more thoroughly than it was done in [13,24]. In particular, Roughly Balanced Bagging [19] was missed in [13], although it is appreciated in the literature. On the other hand, the study presented in [24] was too much oriented on the noise level and only two versions of random under-sampling in bagging were considered. Therefore, we will consider a larger family of known extensions of bagging. Our comparison will include Exactly Balanced Bagging, Roughly Balanced Bagging, and more variants of using over-sampling in bagging, in particular, a new type of integrating SMOTE.

While analyzing existing extensions of bagging one can also notice that most of them employ the simplest random re-sampling technique and, what is even more important, they modify bootstraps to simply balance the cardinalities of minority and majority classes. So, they represent a kind of a *global* point of view on handling the *imbalance ratio* between classes.

Recent studies on class imbalances have shown that this global ratio between imbalanced classes is not a problem itself. For some data sets with high imbalance ratio, the minority class can still be sufficiently recognized even by standard classifiers. The degradation of classification performance is often linked to other *difficulty factors* related to data distribution, such as decomposition of the minority class into many rare sub-concepts [23], the effect of too strong overlapping between the classes [36,16] or the presence of too many minority examples inside the majority class regions [32]. When these factors occur together with class imbalance, they seriously hinder the recognition of the minority class. In earlier research of Napierala and Stefanowski on single classifiers [33] it has been shown that these data difficulty factors could be at least partly approximated by analyzing the *local characteristics* of learning examples from the minority class. Depending on the distribution of examples from the majority class in the local neighbourhood of the given minority example, we can evaluate whether this example could be safe or unsafe (difficult) to be learned. This local view on distributions of imbalanced classes leads us to main aims of this paper.

The main aim of our paper is to study usefulness of incorporating the information about the results of analyzing the local neighbourhood of minority examples into two directions: proposing new generalizations of bagging for class imbalance and extending analysis of classifier performance over different imbalanced data sets.

Following the first direction our aim is to propose extensions of bagging specialized for imbalanced data, which are based on a different principle than the existing ones. Our new approach is to resign from simple integration of pre-processing with unchanged bootstrap sampling technique. Unlike standard bootstrap sampling, we want to change probability of drawing different types of examples. We would like to focus the sampling toward the minority class and even more to the examples located in the most difficult sub-regions of the minority class. The probability of each minority example to be drawn will depend on the class distribution in the neighbourhood of the example [33]. We plan to consider this modification of sampling in two versions of generalizing bagging: (1) over-sampling one, which replicates the minority examples and filters some majority examples to keep the size of a bootstrap sample larger, similar to the size of the original data set; (2) under-sampling one, which is following the idea of explored in Rough Balanced Bagging, and Exactly Balanced Bagging. The under-sampling modification constructs a smaller bootstrap with the size equal to the double the size of the minority class. We plan to evaluate usefulness of both versions in comparative experiments.

The next aim is to better explain differences in performance of various generalizations of bagging ensemble. Current, related studies on this subject are based on a global view on selected evaluations measures over many imbalanced data sets. We hypothesize that it could be beneficial to differentiate between groups of data sets with respect to their underlying data difficulty factors and to study differences in performance of classifiers within these groups. We will show that it could be done by analyzing contents of the neighbourhood of the examples as it leads to an identification of dominating types of difficulty for minority examples. Furthermore, we plan to study more thoroughly contents of bootstrap samples generated by the best performing extensions of bagging. This examination will also be based on analyzing neighbourhood of the minority examples. We will identify differences between bootstrap samples and the original data, and we will try to find a new view on learning of these generalized ensembles.

To sum up, the main contributions of our study are the following. The first one is to study more closely the best known extensions of bagging over a representative collection of imbalanced data sets. Then, we will present a method for analyzing contents of the neighbourhood of the examples and to discuss its consequences. The next methodological contribution is to introduce a new extension of bagging for imbalanced data based on this analysis of a neighbourhood of each example, which affects the probability of its selection into a bootstrap sample. The new proposal will be compared against the best identified extensions. Finally, we will use the same type of the local analysis to explain differences in performance of bagging classifier and to answer a question why contents of bootstrap samples in particular extension of bagging may lead to its good performance.

2. Related works on ensembles for imbalanced data

Several studies have already investigated the problem of class imbalance. The reader is referred to the recent book [18] for a comprehensive overview of several methods and the current state of the art in the literature. Below we very briefly summarize these methods only, which are most relevant to our paper.

First, we describe data *pre-processing methods* as they are often integrated with many ensembles. The simplest data pre-processing re-sampling techniques are *random over-sampling*, which replicates examples from the minority class, and *random under-sampling*, which randomly eliminates examples from the majority classes until a

Please cite this article as: J. Błaszczyński, J. Stefanowski, Neighbourhood sampling in bagging for imbalanced data, Neurocomputing (2014), http://dx.doi.org/10.1016/j.neucom.2014.07.064

required degree of balance between classes is reached. However, random under-sampling may potentially remove some important examples and simple over-sampling may also lead to overfitting. Thus, focused (also called *informed*) *methods*, which attempt to take into account internal characteristics of regions around minority class examples, were introduced. Popular representatives of such methods are OSS [25], NCR [27] for filtering difficult examples from the majority class, as well as SMOTE [10] for introducing additional minority examples. SMOTE considers each example from the minority class and generates new synthetic examples along the lines between the selected example and some of its randomly selected k-nearest neighbours from the minority class. The number of generated examples depends on the main parameter of this method – an over-sampling ratio α. Although its usefulness is experimentally confirmed [4], and SMOTE is the most popular informed pre-processing method, some of the assumptions behind this technique are questioned and authors still work on its extensions, see, e.g., [31]. There also exist hybrid informed methods which integrate over-sampling of selected minority class examples with removing the most harmful majority class examples, e.g., SPIDER [37].

The proposed extensions of ensembles for imbalanced data may be categorized differently. The taxonomy proposed by Galar et al. in [13] distinguishes between *cost-sensitive* approaches vs. integrations with *data pre-processing*. The first group covers mainly cost-minimizing techniques combined with boosting ensembles, e.g., like AdaCost, AdaC or RareBoost. The second group of approaches is divided into three sub-categories: Boosting-based, Bagging-based or Hybrid depending on the type of classical ensemble technique which is integrated into the schema for learning component classifiers and their aggregation. Liu et al. categorize the ensembles for class imbalance into bagging-like, boosting-based methods or hybrid ensembles depending on their relation to standard approaches [29].

As the most of related works [2,7,13,24,29] indicate good performance of bagging extensions versus the other ensembles, below we focus on the bagging based ensembles and they are further considered in our study.

Recall that original Breiman's bagging [8] is an ensemble of T base (component) classifiers induced by the same learning algorithm from T *bootstrap samples* drawn from the original training set. The predictions of component classifiers form the final decision as the result of equal weight majority voting. The key concept is a *bootstrap* aggregation, where the training set for each classifier is constructed by random uniform sampling (with replacement) instances from the original training set (usually keeping the size of the original set).

As the bootstrap sampling will not change drastically the class distribution in the final training sample, it will be still biased toward the majority class. Most of proposals overcome this drawback by applying pre-processing techniques, which change the balance between classes in each bootstrap samples – usually leading to the same, or similar, cardinalities of the minority and majority classes.

In *Underbagging* approaches the number of the majority class examples in each bootstrap sample is randomly reduced to the cardinality of the minority class (N_{min}). In the simplest proposal, called *Exactly Balanced Bagging* (EBBag), while constructing training bootstrap sample, the entire minority class is copied and combined with randomly chosen subsets of the majority class to exactly balance cardinalities between classes.

Another proposal *Roughly Balanced Bagging* (RBBag) results from the critique of the EBBag and other its variants, which use exactly the same numbers of majority and minority examples in each bootstrap [19]. Instead of fixing the constant sample size, it equalizes the sampling probability of each class. For each of T iterations the size of the majority class in the bootstrap (S_{maj}) is set according to the negative binomial distribution. Then, N_{min} examples are drawn from the minority class and S_{maj} examples are drawn from the entire majority class using bootstrap sampling as in the standard bagging (with or without replacement). The class distribution of the bootstrap samples may be slightly imbalanced and varies over iterations. According to [19], this approach is more consistent with the nature of the original bagging, better uses information about the minority examples and performs better than EBBag.

There are also other variants of underbagging (see Section 3 in [13] or Section 4 in [29]), but we focus on the above ones as they have performed better in related works. Another way to overcome class imbalance in a bootstrap sample consists in performing over-sampling the minority class before training a component classifier. In this way, the number of minority examples is increased in each sample (e.g., by a random replication), while the majority class is not reduced as in underbagging. Note that in overbagging more examples will take part in at least one bootstrap sample but, due to their replication, the size of bootstrap samples will be larger than in the standard bagging. This idea was realized in many ways as authors considered integration with different over-sampling techniques. Some of these ways are also focused on increasing diversity of bootstrap samples. We present two approaches further used in experiments.

OverBagging is the simplest version which applies a simplest random over-sampling to transform each training bootstrap sample. S_{maj} of minority class examples is sampled with replacement to exactly balance the cardinality of the minority and the majority class in each sample. Majority examples are sampled with replacement as in the original bagging.

Another approach is used in *SMOTEBagging* to increase diversity of component classifiers [39]. First, SMOTE is used instead of the random over-sampling of the minority class. Then, SMOTE resampling rate (α) is stepwise changed in each iteration from smaller to higher values (e.g., from 10% to 100%). The ratio defines the number of minority examples ($\alpha \times N_{min}$) to be additionally re-sampled in each iteration. Quite similar way of varying ratio α to construct bootstrap samples is also used in "from underbagging to overbagging" ensemble also mentioned in [39]. According to [13], SMOTEBagging gives slightly better results than other good random re-sampling ensembles. However, our preliminary experiments in [7] have already shown that it is not as accurate and works similar to basic OverBagging. Now we want to check it more precisely in experiment presented in Section 3.

Finally, there exist two other variations of underbagging. The method proposed by Chan and Stolfo partitions the majority class into a set of non-overlapping subsets, with each subset having approximately N_{min} examples [9]. Then, each of these majority subsets and all examples from the minority class form a bag for building component classifiers. The predictions of these classifiers were originally combined by stacking although Liu et al. argued for switching to the majority voting [29]. The other option is to construct *Balanced Random Forests* as a extension of classical Random Forests [12]. This algorithm first draws with replacement a bootstrap sample containing N_{min} from the minority class and the same number of the majority class examples. Then, the random tree procedure originating from CART with random feature subset selection is used at each tree split (it is the same solution as in the original Random Forest). Liu et al. in their experiments have noticed that it works not as good as Chan and Stolfo's method or Balance Cascade [29].

3. Comparison of known bagging extension

In the first experiments we compare known best extensions of bagging. All their implementations are done[1] in Java for WEKA framework. The following bagging variants are considered: Exactly

[1] We are grateful to our Master students Lukasz Idkowiak and Marcin Szajek for their help in implementing these algorithms.

Please cite this article as: J. Błaszczyński, J. Stefanowski, Neighbourhood sampling in bagging for imbalanced data, Neurocomputing (2014), http://dx.doi.org/10.1016/j.neucom.2014.07.064

Balanced Bagging (denoted further as EBBag), Roughly Balanced Bagging (RBBag) as the best representatives of under-sampling extensions, OverBagging (abbreviated as OvBag) and SMOTEBagging (abbreviated as SmBag) for over-sampling perspectives. In case of using SMOTE with Bagging, following literature recommendations we choose 5 neighbours and oversampling ratio α was stepwise changed in each sample starting from 10%. Moreover, we decide to use SMOTE in yet another way. In the new ensemble, called BaggingSMOTE (abbreviated BagSm), the bootstrap samples are drawn in a standard way, and than SMOTE is applied to balance majority and minority class distribution in each bootstrap sample (but with the same α ratio). We also include standard bagging (abbreviated as Bag) as a baseline for the comparison.

Component classifiers in all ensembles are learned with C4.5 tree learning algorithm (J4.8), which uses standard parameters except disabling pruning (following experiences from earlier experiments as [37]). For all bagging variants, we test the following numbers T of component classifiers: 20, 50 and 100. The results for $T=50$ are slightly better than for $T=20$, while increasing T leads to similar general conclusions but introduces additional computational costs. Thus is why we present detailed results for $T=50$ only, due to space limit.

We choose 23 real-world data sets representing different domains, sizes and imbalance ratios and because they have been used in most related experimental studies [4,20,24,30]. Most of them come from the UCI repository [3]. Three data sets `abdominal`, `hsv` and `scrotal-pain` come from our medical applications. For data sets with more than two classes, we chose the smallest one as a minority class and combined other classes into one majority class. The characteristics of data sets are presented in Table 1, where IR is the imbalance ratio defined as N_{maj}/N_{min}. The data sets were ordered from the safest one, at the top of Table 1, to the most unsafe at the bottom. This ordering results from the analysis of data set types presented in Section 6.2.

The performance of bagging ensembles is measured using *sensitivity* of the minority class (the minority class accuracy), its *specificity* (an accuracy of recognizing majority classes), their aggregation to the *geometric mean* (*G*-mean) and *F-measure* (referring to the minority class, and used with equal weights 1 assigned to precision and recall). For their definitions see, e.g., [18,17,22]. These measures are estimated with the stratified 10-fold cross-validation repeated ten times to reduce the variance. The average values of *G*-mean and sensitivity are presented in Tables 2 and 3, respectively. The differences between classifier average results will be also analysed using Friedman and Wilcoxon statistical tests. For their description see, e.g., [22]. In all these tables the last row contains average ranks calculated as in the Friedman test – the lower the average rank, the better the classifier.

Let us analyse first values of *G*-mean presented in Table 2. In the Friedman test we reject the null hypothesis (p-value in this case is smaller than 0.00001). Carrying out the Nemenyi post hoc analysis (critical difference CD=1.61) shows that all extensions, except SmBag, are significantly better than the standard version. Then both under-sampling extensions EBBag and RBBag are significantly better than all over-sampling variants. According to average ranks RBBag seems to be slightly better than EBBag and this trend is even more visible for a higher number of component classifiers, and using bootstrap sampling with replacement. However, according to the paired Wilcoxon test the null hypothesis on no significant difference between results of both ensembles cannot be rejected (p-value=0.24). While using SMOTE to over-sample the minority class, the new integration BagSm performs better than the previously known SmBag and OvBag (this is reflected by average ranks). However according to the Wilcoxon test BagSm is not so strongly outperforming OvBag (p-value=0.53) but it is significantly better than SmBag (p-value=0.009).

Table 1
Data characteristics.

Data set	# examples	# attributes	Minority class	IR
breast-w	699	9	Malignant	1.90
abdominal-pain	723	13	Positive	2.58
acl	140	6	1	2.5
new-thyroid	215	5	2	5.14
vehicle	846	18	Van	3.25
car	1728	6	Good	24.04
scrotal-pain	201	13	Positive	2.41
ionosphere	351	34	b	1.79
pima	768	8	1	1.87
credit-g	1000	20	Bad	2.33
ecoli	336	7	imU	8.60
hepatitis	155	19	1	3.84
haberman	306	4	2	2.78
breast-cancer	286	9	Recurrence-events	2.36
cmc	1473	9	2	3.42
cleveland	303	13	3	7.66
hsv	122	11	4.0	7.71
abalone	4177	8	0–4 16–29	11.47
postoperative	90	8	S	2.75
solar-flareF	1066	12	F	23.79
transfusion	748	4	1	3.20
yeast	1484	8	ME2	28.10
balance-scale	625	4	B	11.76

Table 2
G-mean (%) for known bagging extensions.

Data set	Bag	EBBag	RBBag	OvBag	SmBag	BagSm
breast-w	95.88	96.03	96.37	96.23	95.88	96.77
abdominal-pain	78.95	80.65	80.35	79.44	80.85	79.86
acl	88.18	90.71	89.35	88.35	88.64	87.81
new-thyroid	92.41	96.91	96.58	95.36	95.18	92.89
vehicle	93.91	94.58	95.44	94.61	94.34	94.20
car	84.53	96.73	96.58	95.29	95.26	95.18
scrotal-pain	70.75	73.18	75.65	72.01	70.42	70.68
ionosphere	88.96	90.44	90.67	90.47	90.30	90.26
pima	71.54	74.22	75.64	73.54	72.33	71.38
credit-g	63.98	65.82	67.82	71.75	80.68	66.11
ecoli	68.67	72.24	88.85	51.42	58.38	80.11
hepatitis	62.81	78.93	78.66	72.16	68.47	74.29
haberman	43.11	65.41	63.43	58.11	60.02	62.82
breast-cancer	54.30	58.82	59.37	56.17	52.57	57.25
cmc	52.76	64.61	65.27	59.95	57.74	62.77
cleveland	12.61	72.32	71.02	22.77	25.03	50.96
hsv	0.00	36.27	35.74	2.84	5.37	16.61
abalone	49.58	78.93	79.32	61.95	63.67	69.65
postoperative	1.99	24.97	34.03	15.01	1.57	11.55
solar-flare	13.70	85.39	83.21	58.07	55.04	54.40
transfusion	55.72	66.75	67.32	64.83	63.96	65.76
yeast	51.48	84.55	84.68	59.70	59.41	57.94
balance-scale	0.00	59.07	54.23	1.40	0.00	0.67
Average rank	5.61	1.96	1.61	3.65	4.26	3.91

The similar analysis is carried out for the sensitivity measure, which are presented in Table 3. The Friedman test allows us to claim significance of differences between compared classifiers (again with p-value, which is smaller than 0.00001). Nemenyi post hoc analysis (with the same critical difference CD=1.61) shows that both EBBag and RBBag lead to significantly better sensitivity than all other bagging variants. According to average ranks EBBag is only very slightly better than RBBag but the paired Wilcoxon test indicates that differences between these two classifiers are not significant (p-value=0.24), while they are both significantly better than all other variants. Again while considering over-sampling generalization, the new integration BagSm performs better than the previously known SmBag and OvBag (this is reflected by average ranks and also the Wilcoxon test BagSm vs OvBag (p-value=0.023) and BagSm vs SmBag (p-value=0.002).

Please cite this article as: J. Błaszczyński, J. Stefanowski, Neighbourhood sampling in bagging for imbalanced data, Neurocomputing (2014), http://dx.doi.org/10.1016/j.neucom.2014.07.064

Table 3
Sensitivity (%) for known bagging extensions.

Data set	Bag	EBBag	RBBag	OvBag	SmBag	BagSm
breast-w	94.88	96.01	96.98	95.98	95.02	95.17
abdominal-pain	72.05	81.65	79.16	74.22	71.57	76.86
acl	83.33	93.33	89.00	85.00	85.00	85.83
new-thyroid	87.50	95.50	95.71	93.06	92.22	93.89
vehicle	91.29	91.16	97.04	93.46	92.14	94.97
car	73.97	100.00	100.00	92.62	92.54	92.13
scrotal-pain	58.11	73.78	75.59	65.89	58.56	58.56
ionosphere	81.79	85.73	85.24	84.70	83.70	83.76
pima	61.28	76.70	78.54	67.38	65.13	63.38
credit-g	48.89	72.50	68.13	60.83	71.67	63.11
ecoli	56.67	78.20	91.14	66.67	55.00	77.11
hepatitis	49.44	81.00	76.56	62.78	54.44	67.25
haberman	26.38	60.56	55.68	49.86	49.81	66.25
breast-cancer	35.93	56.06	57.41	44.91	34.35	50.05
cmc	36.67	66.61	64.50	46.47	40.05	53.10
cleveland	9.72	77.22	69.43	16.11	17.22	36.11
hsv	0.00	55.00	23.48	3.33	5.00	21.67
abalone	25.47	79.98	77.58	40.51	42.98	54.99
postoperative	1.67	27.22	22.08	11.67	1.11	8.89
solar-flare	7.00	86.00	85.12	42.17	37.33	34.40
transfusion	34.62	65.45	66.69	56.54	51.53	68.66
yeast	32.22	90.22	87.65	39.11	39.11	57.94
balance-scale	0.00	49.33	60.00	0.67	0.00	0.67
Average rank	5.76	1.67	1.72	3.85	4.70	3.30

We also analysed sampling with or without replacement. Conclusions are not univocal. For the best under-sampling variants like EBBag differences are insignificant while for over-sampling standard replacement sampling works much better.

We skipped the presentation of *F*-measure due to space limits. The results are quite similar to analyzing the sensitivity, i.e. the ranking of the methods is nearly the same (the only difference is that RBBag is now better than EBBag). In this case, RBBag with replacement is better than EBBag in the Wilcoxon test (p-value$=0.038$). Again, underbagging generalizations are better than all overbagging (for instance, according to the Wilcoxon test EBBag is better than BagSm with p-value$=0.04$).

To sum up these experiments we can conclude that under-sampling bagging extensions such as EBBag and RBBag have outperformed all over-sampling ensembles. The difference between them and the best oversampling bagging is much higher than we could expect from the literature survey. Moreover, a new over-sampling bagging variant, where SMOTE is applied with the same over-sampling ratio, works better than the previously promoted SmBag applying different ratios [39].

If one should choose between under-sampling variants EBBag and RBBag, we will rather promote Roughly Balanced Bagging as its experimental evaluation is slightly better (in particular for the most important measure in our study – *G*-mean) and its methodological principle are more consistent with the bagging sampling paradigm. This is why, we will choose it for further experiments in Section 6.

4. Studying local characteristics of minority examples

The further proposed extensions of bagging and method for analyzing distributions of minority examples in data sets descend from results of studying sources of difficulties in learning classifiers from imbalanced data. Notice first that although many authors have experimentally shown that standard classifiers met difficulties while recognizing the minority class, it has also been observed that in some problems characterized by strong class imbalance (e.g., `new-thyroid` data set from [3]) standard classifiers are capable to be sufficiently accurate. Therefore, the discussion of data difficulty in imbalanced data still goes on, for its current review see, e.g., [30,35,38].

Several researchers have already hypothesized that the *class imbalance ratio* (i.e. cardinality of the majority class referred to the total number of minority class examples) is not necessarily the only, or even the main, problem causing the decrease of classification performance and focusing only on this ratio may be insufficient for improving classification performance. In other words, besides the imbalanced ratio other data difficulty factors may cause a severe deterioration of classification performance.

The experimental studies by Japkowicz et al. on large collection of artificial data sets have clearly demonstrated that degradation of classification performance is linked to the decomposition of the minority class into many sub-parts containing very few examples [21,23]. They have shown that the minority class does not form a homogeneous, compact distribution of the target concept but it is scattered into many smaller sub-clusters surrounded by majority examples. In other words, minority examples form, the so-called, *small disjuncts*, which are harder to learn and cause more classification errors than larger sub-concepts.

Other data factors related to the class distribution are linked to the effect of too strong *overlapping* between minority and majority classes. Strong overlapping occurs frequently together with class rarity. In [36], authors have generated many artificial, numerical, data sets and based on them they have shown that increasing overlapping has been more influential than changing the class imbalance ratio. An analogous experiment, but concerning six classifiers compared with more evaluation measures, has been carried out in [16] leading to similar conclusions. However, these authors have also noticed that the *local imbalance* inside overlapping area is more influential than changing the global imbalance ratio. Finally, few researchers have claimed that another data factor, which influences degradation of classifiers performance on imbalanced data, is noisy examples [1]. Experiments presented in [32] have shown that single minority examples located inside the majority class regions cannot be treated as noise since their proper treatment by informed pre-processing may improve classifiers. In most of these experiments researchers focused on studying a single data difficulty factor only. Studies as [38] emphasize that several data factors usually occur together for imbalanced data sets.

Although all of these studies give an insight into the important aspects of imbalanced data distribution and sources of difficulties in learning classifiers in this setting, their conclusions might not be easy to apply in the real-world settings. The main problem is that it is not easy to identify different data factors in the real-world data sets.

In our opinion one of the main conclusions from the studies is that the global information about the data sets (mainly the global imbalance ratio) is not so important as considering *local characteristics* of the class distribution. Local characteristics of learning examples could be modeled in different ways. Here, we follow earlier works on specialized informed pre-processing methods [25,27,37] and on other studies on the nature of imbalanced data [32,35]. We link data factors to *different types of examples* forming the minority class distribution. What follows is a differentiation between safe and unsafe examples.

Safe examples are ones located in the homogeneous regions populated by examples from one class only. Other examples are *unsafe* and more difficult for learning. Unsafe examples are categorized into *borderline* (placed close to the decision boundary between classes), *rare cases* (isolated groups of few examples located deeper inside the opposite class), or *outliers*. As the minority class can be highly under-represented in the data, we claim that the rare examples or outliers could represent a very small but valid sub-concepts of which no other representatives could be collected for training. Therefore they cannot be considered as noise examples which typically are then removed or

Please cite this article as: J. Błaszczyński, J. Stefanowski, Neighbourhood sampling in bagging for imbalanced data, Neurocomputing (2014), http://dx.doi.org/10.1016/j.neucom.2014.07.064

re-labeled. A similar opinion was also expressed in [25], where authors suggested that minority examples should not be removed as they are too rare to be wasted while majority examples could be removed. Moreover, earlier works of Napierala with graphical visualizations of real-world imbalanced data sets [33,35] have confirmed usefulness of such a classification of example types.

The next question is how to automatically and possibly simply *identify these types of examples*. We keep the hypotheses [33] on role of the mutual positions of the learning examples in the attribute space and the idea of assessing the type of example by analyzing class labels of the other examples in its *local neighbourhood*. Such a local neighbourhood of the minority class example could be modeled in different ways. In further considerations we will use an analysis of the class labels among *k-nearest neighbours* following positive experiences with single classifiers and pre-processing methods [33,35]. Depending on the number of examples from the majority class in the local neighbourhood of the given minority class example, we can evaluate whether this example could be safe or unsafe (difficult) to be learned. If its all, or nearly all, neighbours belong the minority class, this example is treated as the safe example. On the other hand, a minority example with all neighbours from the majority class is clearly an outlier. Then, when the numbers of neighbours from both classes are approximately the same, we assume that this example could be located close to the decision boundary between the classes. Finally, an example having one minority neighbour and other majority ones is a candidate for a rare case.

In general, constructing this type of the neighbourhood is related with choosing the value of k and the *distance function*. In further considerations we follow results of analyzing different distance metrics [35] in the method considered here and also more general experimental comparisons of several heterogeneous distances applied to k-NN classifier [28]. Following these recommendations we choose the HVDM metric (*Heterogeneous Value Difference Metric*) [41]. It aggregates normalized distances for qualitative and quantitative attributes. Compared to other metrics it provides more appropriate handling of qualitative attributes. Instead of simple value matching, HVDM makes use of the class information to compute attribute value conditional probabilities by using a Stanfil and Valtz value difference metric for nominal attributes [41]. For numeric attributes, it uses a standardized Euclidean distance.

Considering the value of k, different values could be used with respect to particular data set characteristics. We will check several values during further experiments to see their impact on the types of minority examples and Neighbourhood Balanced Bagging ensemble. However, as the distribution of the minority class is "difficult", this class is often decomposed into smaller sub-parts, and as our assumptions focus on quite local neighbourhood for minority class example we claim that it is reasonable to choose rather small values of k. Moreover one can refer to some related experimental studies as, e.g., [5,14] containing systematic examinations of different values k over many UCI imbalanced data sets, which concluded that for difficult data distributions and using HVDM, more local classifiers (with smaller k values from 5 till 11) were recommended. Finally, following earlier experimental studies of Napierala [35] we will start modeling the neighbourhood with $k=5$, and additionally examine higher values as 7 and 9.

Finally, we will repeat our hypothesis that the appropriate treatment of these types of minority examples within new proposal of classifiers should lead to improving classification performance. Recall that it has been earlier observed by Stefanowski for the informed pre-processing method SPIDER [37] and in BRACID a novel rule induction algorithm [34] specialized for imbalanced data. Now, we want to introduce this way of thinking on the local characteristics into designing new extensions of bagging ensemble.

5. Neighbourhood balanced bagging for imbalanced data

5.1. Motivations

Our aim is to show that the analysis of class distribution in the neighbourhood of examples can be applied to propose a new kind of generalizing bagging ensembles for imbalanced data. Recall that existing approaches to generalize bagging treat all learning examples in the same way while constructing bootstrap samples. It results from the fact that these generalizations do not change the standard bootstrap sampling technique. They rather offer different ways to integrate bootstrap sampling with various pre-processing techniques applied on constructed bootstraps. For instance, over-sampling extensions rely on coping randomly selected examples from the minority class. In such a case, due to the global imbalance ratio, the amount of replication of minority examples may be quite large. One can ask whether each minority example is equally important. Moreover, one can ask whether drawing of minority examples should be done in a blind way or whether it should be directed depending on the difficulty type of example. Earlier related works on pre-processing methods for single classifiers have already showed that plain random sampling is less efficient than informed methods as, e.g., NCR [27], SMOTE [10] or SPIDER [37]. Moreover focusing transformations around more unsafe examples has been usually more beneficial than amplifying safe minority examples, see, e.g., the discussion in [17] or recent extensions of SMOTE [15]. Similar experiences with differentiated role of learning examples have been reported as to edited k-nearest neighbour classifier, and specialized methods integrating rule and instances representations for class imbalanced as, e.g., BRACID [34].

Following these motivations, we present new generalizations of bagging. In these propositions, we resign from treating all minority examples in the same way. We focus bootstrap sampling toward more difficult sub-regions of the minority class. Our hypothesis is that by increasing probabilities of drawing less safe types of the minority class examples and by decreasing, at the same time, probabilities of drawing majority class examples, we can modify the local characteristics of examples in resulting bootstrap samples. This modification should lead to bootstrap samples with more safe distribution of minority class examples as compared to original learning set. As a result, we expect component classifiers in constructed bagging ensembles to be more likely to better learn the minority class.

Referring to experimental studies on the characteristics of often tested UCI imbalanced data sets, see, e.g., [35], and also some results presented in Section 6.2, one may notice that the minority class distributions are generally quite unsafe with many borderline examples or even outliers. Therefore, we think that treating all minority examples in the same way and using only the global between class ratio to simply balance class cardinalities inside bootstrap samples is less realistic and more limited than applying local approaches presented in the previous section. We plan to consider both options of modifying bagging which follows either increasing the cardinality of the minority class or reducing the number of majority examples in the bootstraps.

The first option is more similar to over-sampling minority class inside the bootstraps, however, since it also decreases chance for sampling majority examples it can also be seen as a kind of hybrid approach. Within this proposal we would like to keep the final size of the bootstrap similar to the cardinality of the original data set. We expect that this generalization could be more accurate than existing over-sampling extensions of bagging ensemble.

Considering the other option comes mainly from experimental studies, as presented in Section 3 or [24], which show that generalizations with under-sampling of majority classes are more accurate than over-sampling based bagging ensembles. This is why

Please cite this article as: J. Błaszczyński, J. Stefanowski, Neighbourhood sampling in bagging for imbalanced data, Neurocomputing (2014), http://dx.doi.org/10.1016/j.neucom.2014.07.064

we would like to construct the bootstraps with the size equal to double cardinality of the minority class inside the original data. However, we think that for such bootstraps, being much smaller than in other generalizations of bagging, it is particularly interesting to check which minority examples should be sampled. Recall that EBBag just copies all the content of the minority class inside each bootstrap and even RBBag selects around 66% examples from this class and randomly amplifies some of these examples. Here we want to put a question on the usefulness of a more informed sampling process which takes into account the local characteristics of these examples.

5.2. Modification of sampling technique

The idea behind a new extension called *Neighbourhood Balanced Bagging* (NBBag) is to focus sampling of bootstraps toward these minority examples, which are hard to learn (i.e. unsafe ones) while decreasing probabilities of selecting examples from the majority class at the same time. The idea of changing sampling probabilities has been considered in our previous work with applying bagging to noisy data and improving the overall accuracy [6]. Here, we postulate another strategy to change bootstrap samples, which is carried through a conjunction of modifications at two levels: *global level* (the whole data set level) and *local level* (the example neighbourhood level).

At the first, global level, we attempt at increasing the chance of drawing the minority examples with respect to the imbalance ratio in the original data set. We implement it by changing the probability of sampling of majority examples. More precisely, probability of sampling is, in our setting, proportional to the weight that we associate with each learning example. First we set weight p^1_{min} for each minority example to 1. Then, we downscale weight p^1_{maj} associated with sampling of each majority example to N_{min}/N_{maj}, where N_{min} and N_{maj} are numbers of examples in the minority and majority classes in the original data, respectively. Intuitively, it could refer to the situation, where minority and majority classes contain examples of the same type, e.g., safe ones, and the class distributions are not affected by other data difficulty factors. Thus, this modification of probabilities exploits information about the global between-class imbalance.

Recall that such a global balancing of bootstraps is not the sufficient technique according to the experimental studies as [7,13,24]. Moreover, most studied imbalance data sets contain many unsafe minority examples while the majority classes comprise rather safe ones, see, e.g., [32]. This leads us to consider an additional local level of modifying probabilities, which is based on the analysis of the local characteristics of examples.

This local level of modifying probabilities is intended to shift sampling of minority examples to these unsafe examples that are harder to learn. The extent to which a minority example is unsafe may be quantified by analyzing its k-nearest neighbours (using HVDM distance metric as described in Section 4). We have decided to take a rather simple approach and to only count the number of majority examples in the neighbourhood. Then, partly inspired by earlier successful experiences with informed pre-processing methods, we use a simple rule: the more the unsafe example, the more should be amplified probability of its drawing. We also decide that the probability should be monotonic with respect to the number of majority examples in the neighbourhood. This leads to the following formula L^2_{min}, which defined as below, is either linear or exponential:

$$L^2_{min} = \frac{(N'_{maj})^{\psi}}{k}, \tag{1}$$

where N'_{maj} is the number of examples in the neighbourhood, which belong to the majority class; ψ is a scaling factor, which in case of a linear amplification is set to 1. Although this factor introduces a problem of parametrization, our intuition is that it can be optimized depending on results of analyzing characteristics of particular data set (see further analysis presented in Section 6.2). So, the value ψ may be increased, resulting in an exponential amplification, if one wants to strengthen the role of rare cases and outliers in bootstraps. We claim that this exponential amplification may be beneficial for such data sets where the analysis of types of examples indicates that the minority class distribution is scattered into many rare cases or outliers, and the number of safe examples is significantly limited. In Fig. 1 we present an illustration of different profiles representing amplifications of probability of selecting the minority class with respect to a few selected values of ψ, which will be further considered in experimental studies.

The formula L^2_{min} requires re-scaling as it may lead to the probability equal to 0 for completely safe examples, i.e., for $N'_{maj} = 0$. We propose to re-formulate it as

$$\beta \times (L^2_{min} + 1) \tag{2}$$

where β is a technical coefficient referring to drawing a completely safe example. Intuitively, safe examples from both minority and majority classes should have the same probability of being selecting to bootstraps. Setting β to 0.5 keeps this intuition. Adding the 1 corresponds to a normalization of sampling probabilities inside the conjunctive combination, if one expects that for linear amplification $p_{min} \in [0, 1]$ (p_{min} is the weight of minority examples – see definition (3)).

Then, we hypothesize that examples from majority class are, by default, not exactly balanced on the second, local level, which is reflected by $L^2_{maj} = 0$. The intuition behind this hypothesis is that examples from majority class are more likely to be safe (see the results of such analysis further presented in Section 6.2). Even when the hypothesis is false for some data, it is still quite apparent that amplifying majority rare or outlying examples, at this level, would interact with the amplification of minority examples and increase difficulties of learning classifiers from the minority classes.

Finally, local and global levels are combined by multiplication. This leads us to the final formulation of weights associated with probability of selecting examples from minority and majority classes, respectively, as

$$\begin{aligned} p_{min} &= p^1_{min} \times \beta(L^2_{min} + 1) \\ &= p^1_{min} \times 0.5(L^2_{min} + 1) = 0.5(L^2_{min} + 1), \end{aligned} \tag{3}$$

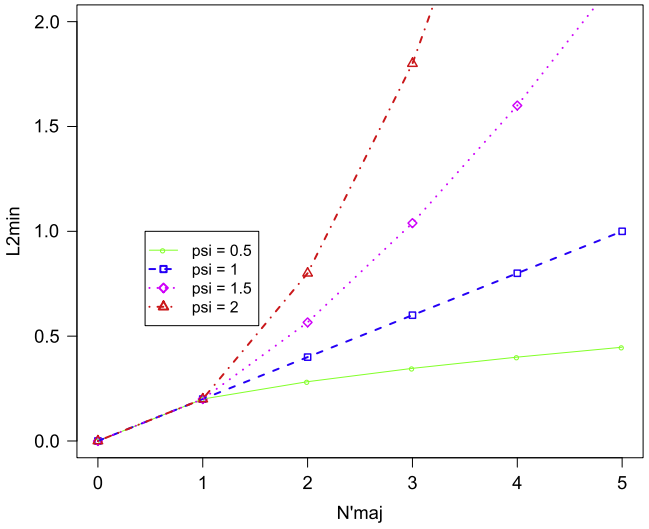

Fig. 1. L^2_{min} weights depending on ψ.

Please cite this article as: J. Błaszczyński, J. Stefanowski, Neighbourhood sampling in bagging for imbalanced data, Neurocomputing (2014), http://dx.doi.org/10.1016/j.neucom.2014.07.064

$$p_{maj} = p^1_{maj} \times \beta(L^2_{maj}+1) = p^1_{maj} \times 0.5 = \frac{N_{min}}{N_{maj}} \times 0.5, \quad (4)$$

resulting from $L^2_{maj}=0$, and default β set to 0.5. Such a formulation may be interpreted as amplification of chances to select minority examples according to parameterized local factor L^2_{min} in combination with lowering chances to select majority examples according to imbalance rate in the whole data set.

Finally, we present the general schema of using these modifications of probability sampling in both types of Neighbourhood Balanced Bagging, i.e., following the ideas of under-sampling the majority class and the other, similar to over-sampling the minority class (see Algorithm 1).

6. Experimental evaluation of NBBag

The first part of experiments is focused on an evaluation of classification performance of Neighbourhood Balanced Bagging (NBBag), and its comparison to known extensions of bagging. The second part concerns an analysis of local characteristics of different types of minority class examples in the bootstrap samples produced by these extensions.

6.1. Evaluation of bagging extensions

We compare performance of NBBag with the best previously proposed extensions of bagging. Following our earlier study [7], we choose Rough Balanced Bagging (RBBag) as the best under-sampling extension. Since NBBag is considered in two variants, under-sampling and more following over-sampling, we also include Overbagging (OvBag) and SMOTEBagging (SmBag) in the comparison. All experiments have been performed in the same setting as the ones presented in Section 3.

We tested different sizes of neighbourhood for NBBag: $k=5$, 7, 9 and 11. Their best performance depends on data set. However in general, we have noticed that good performance can be achieved for small neighbourhoods for under-sampling, and for over-sampling, regardless of the amount of amplification applied to the weights of minority class examples (i.e., value of ψ scaling parameter). Thus, we present results only for $k=5$ – which is also consistent with a discussion from Section 4.

We also checked the values of scaling factor ψ responsible for amplification of weights of minority class examples in NBBag bootstrap sampling. More precisely, we applied $\psi=0.5$, 1, 1.5, 2, 4. The best value depends on data set. However, on the average the best results for over-sampling were achieved for $\psi=2$, and the best result for under-sampling was achieved for, considerably lighter amplification, $\psi=0.5$.

This is why due to space limits we present only results of the best performing over-sampling NBBag: oNBBag^2 ($k=5$, $\psi=2$), and the best performing under-sampling NBBag: uNBBag$^{0.5}$ ($k=5$, $\psi=0.5$).

The results of G-mean, sensitivity, and F-measure are presented in Tables 4, 5, and 6, respectively. Note that, as it was already done in Section 3, data sets in the analysed tables are ordered from the safest one to the most unsafe one. In general, RBBag and uNBBag$^{0.5}$ stand out as the best classifiers in comparison on each of the presented measures. However, comparison on F-measure does not show significant difference between compared classifiers (p-value in Friedman test in this case is 0.21). On the other hand, comparison on G-mean and sensitivity leads to significant differences discovered by Friedman test (p-values in both cases smaller than 0.00001). In further analysis we focus more on G-mean (as this

Algorithm 1. Neighbourhood Balanced Bagging Algorithm.

Input: LS training set; TS testing set; CLA base classifier learning algorithm; m number of bootstrap samples; N_{min}, N_{maj} size of minority and majority class (respectively); L^2_{min} minority class local balancing weights;
Output: C^* ensemble classifier

```
Learning phase;
if under-sampling then
    n = 2 × N_min;
else
    n = N_min + N_maj;
foreach x ∈ LS do
    If x ∈ minority class then
        w(x) = p_min = 0.5(L^2_min + 1);
    else
        w(x) = p_maj = (N_min / N_maj) × 0.5
for i:=1 to m do
    S_i = bootstrap sample of n examples from LS sampled
    according to weights w;
    C_i:=CLA(S_i) {generate a base classifier};
Classification phase;
foreach x in TS do
    C*(x):=majority vote of C_i(x), where i = 1, ..., m {the suggestion of the classifier for object
    x is a combination of suggestions of component
    classifiers C_i};
```

Please cite this article as: J. Błaszczyński, J. Stefanowski, Neighbourhood sampling in bagging for imbalanced data, Neurocomputing (2014), http://dx.doi.org/10.1016/j.neucom.2014.07.064

measure takes into account classifier performance on both minority and majority classes, i.e., an increase of recognition of the minority examples cannot be achieved at cost of a deterioration of the majority class), and on sensitivity – which, on the other hand, is the accuracy of minority class. In the following, we present some more detailed observations from the experimental comparison.

For *G*-mean, uNBBag$^{0.5}$ is the best classifier according to average ranks (see Table 4). It is also significantly better than all other classifiers except RBBag according to Nemenyi post hoc test (CD = 1.33). This result is confirmed by Wilcoxon test (with *p*-values smaller than 0.01 in each case except comparison between uNBBag$^{0.5}$ and RBBag). RBBag is better than OvBag and SmBag according to Nemenyi, and better than OvBag, SmBag and oNBBag^2

Table 4
G-mean (%) of NBBag and other compared bagging ensembles.

Data set	RBBag	OvBag	SmBag	oNBBag^2	uNBBag$^{0.5}$
breast-w	96.37	96.23	95.88	96.14	96.32
abdominal-pain	80.35	79.44	80.85	80.82	81.11
acl	89.35	88.35	88.64	88.20	89.37
new-thyroid	96.58	95.36	95.18	97.02	96.49
vehicle	95.44	94.61	94.34	95.91	95.53
car	96.58	95.29	95.26	96.98	96.47
scrotal-pain	75.65	72.01	70.42	71.42	74.26
ionosphere	90.67	90.47	90.30	89.95	90.71
pima	75.64	73.54	72.33	72.30	74.80
credit-g	67.82	64.30	62.48	66.94	67.68
ecoli	88.85	71.75	80.68	86.74	88.44
hepatitis	78.66	72.16	68.47	75.33	79.81
haberman	63.43	58.11	60.02	48.65	64.28
breast-cancer	59.37	56.17	52.57	56.53	59.99
cmc	65.27	59.95	57.74	64.33	65.54
cleveland	71.02	22.77	25.03	65.75	74.29
hsv	35.74	2.84	5.37	30.43	43.62
abalone	79.32	61.95	63.67	76.25	79.59
postoperative	34.03	15.01	1.57	41.43	40.22
solar-flare	83.21	58.07	55.04	71.13	83.32
transfusion	67.32	64.83	63.96	39.56	66.60
yeast	84.68	59.70	59.41	74.86	84.57
balance-scale	54.23	1.40	0.00	61.07	32.76
Average rank	1.87	4.00	4.39	3.09	1.65

Table 5
Sensitivity (%) of NBBag and other compared bagging ensembles.

Data set	RBBag	OvBag	SmBag	oNBBag^2	uNBBag$^{0.5}$
breast-w	96.68	95.98	95.02	96.35	96.72
abdominal-pain	79.16	74.22	71.57	80.99	82.08
acl	89.00	85.00	85.00	89.00	90.25
new-thyroid	95.71	93.06	92.22	96.00	95.43
vehicle	97.04	93.46	92.14	96.48	97.29
car	100.00	92.62	92.54	95.80	100.00
scrotal-pain	75.59	65.89	58.56	71.86	76.10
ionosphere	85.24	84.70	83.70	87.94	86.03
pima	78.54	67.38	65.13	85.07	81.53
credit-g	68.13	52.89	45.89	73.93	72.67
ecoli	91.14	60.83	71.67	84.29	91.71
hepatitis	76.56	62.78	54.44	69.38	79.06
haberman	55.68	49.86	49.81	87.28	62.22
breast-cancer	57.41	44.91	34.35	66.71	65.53
cmc	64.50	46.47	40.05	66.61	68.35
cleveland	69.43	16.11	17.22	54.57	76.29
hsv	23.48	3.33	5.00	13.57	39.29
abalone	77.58	40.51	42.98	65.70	79.76
postoperative	22.08	11.67	1.11	42.92	35.83
solar-flare	85.12	42.17	37.33	58.84	86.51
transfusion	66.69	56.54	51.53	92.08	74.33
yeast	87.65	39.11	39.11	59.22	88.63
balance-scale	60.00	0.67	0.00	72.45	98.78
Average rank	2.43	4.22	4.78	2.15	1.41

Table 6
F-measure (%) of NBBag and other compared bagging ensembles.

Data set	RBBag	OvBag	SmBag	oNBBag^2	uNBBag$^{0.5}$
breast-w	94.72	94.83	94.56	94.43	94.60
abdominal-pain	69.83	70.20	74.23	70.16	70.38
acl	82.92	84.04	84.62	80.75	82.45
new-thyroid	91.70	92.03	92.71	93.26	91.82
vehicle	89.44	90.38	90.84	91.19	89.49
car	55.32	80.60	81.28	79.91	54.58
scrotal-pain	64.64	62.16	62.68	59.49	62.83
ionosphere	89.00	88.84	88.95	86.99	88.79
pima	68.54	66.24	64.75	66.21	67.93
credit-g	55.87	52.07	51.06	55.62	56.14
ecoli	59.56	56.64	64.70	60.96	57.64
hepatitis	61.24	59.19	56.52	57.98	62.33
haberman	47.86	42.51	44.95	44.82	48.72
breast-cancer	46.17	43.75	41.54	46.02	48.17
cmc	45.96	41.90	41.05	44.97	46.25
cleveland	36.66	15.21	17.35	34.79	39.33
hsv	11.13	1.33	3.89	7.61	14.70
abalone	39.34	42.34	43.55	44.15	38.37
postoperative	17.49	10.96	1.11	27.81	24.92
solar-flare	27.10	21.34	20.68	23.89	26.37
transfusion	49.54	47.81	47.89	40.24	48.99
yeast	25.08	38.70	37.06	38.24	24.25
balance-scale	15.80	0.51	0.00	19.52	15.86
Average	2.61	3.43	3.30	3.09	2.57

in paired Wilcoxon test (*p*-values smaller than 0.001 in this case). OvBag, SmBag, and oNBBag^2 are not significantly different with respect to Nemenyi test but Wilcoxon test shows significant difference in pairs between oNBBag^2 and OvBag, as well as SmBag. The worst classifier is SmBag, which is consistent with conclusions from experiments in Section 3. Some of the results on *G*-mean require distinguishing since they are much better than the results achieved by the other compared classifiers. These are oNBBag^2 on `postoperative`, and `balance-scale`, and uNBBag$^{0.5}$ on `cleveland`, and `hsv`. It is also worth noting that higher differences between classifiers are more visible for more difficult (unsafe) data sets. This effect is observable as one moves from the top of the tables to the bottom, since, as it was mentioned earlier, data sets are ordered according to their difficulty (which is explained in more detail in Section 6.2).

Analyzing the recognition of the minority examples, i.e., the sensitivity measure in Table 5, the best performing with respect to the average ranks is again uNBBag$^{0.5}$. Post hoc Nemenyi test divides classifiers into two groups: RBBag, oNBBag^2, and uNBBag$^{0.5}$ are better than OvBag and SmBag. uNBBag$^{0.5}$ is significantly better than all classifiers except oNBBag^2 in paired Wilcoxon test (*p*-values lower than 0.001). It is also worth noting that all the best results on sensitivity are achieved by either oNBBag^2 or uNBBag$^{0.5}$ (with one shared best result between RBBag and uNBBag$^{0.5}$ for `car`).

For *F*-measure results, we can observe that also in this case, the best average rank is achieved by uNBBag$^{0.5}$. However, we need to take into account that the observed differences in average ranks between classifiers are not significant according to Friedman test. We also failed to find significant differences between pairs of classifiers with respect to Wilcoxon test.

Looking more precisely at results in Tables 4 and 5, one can notice that some classifiers showing high improvements of the sensitivity also show strong deterioration on *G*-mean (it means that the recognition of the majority class is much worse). Such effect is visible for oNBBag^2 on `pima`, `haberman`, `breast-cancer`, and `transfusion`. Similar effect, but less evident, is visible in case of `yeast` for uNBBag$^{0.5}$. Performance on `balance-scale`, which is the most difficult data set in our comparison, illustrates perfectly the effect of high sensitivity on *G*-mean. In this case, the second best result on sensitivity achieved by oNBBag^2 leads to the best result on *G*-mean. At

Please cite this article as: J. Błaszczyński, J. Stefanowski, Neighbourhood sampling in bagging for imbalanced data, Neurocomputing (2014), http://dx.doi.org/10.1016/j.neucom.2014.07.064

the same time, the best result on sensitivity achieved by uNBBag$^{0.5}$ leads to a result on *G*-mean which is not only worse than oNBBag^{2} but also worse than RBBag. On the other hand, we can also show data sets, for which the best result on sensitivity translates into the best result on *G*-mean. These are `postoperative` for oNBBag^{2}, and `cleveland`, as well as `hsv` for uNBBag$^{0.5}$.

Finally, we can observe that simple use of the imbalance ratio in global balancing of classes in bootstraps is not sufficient. It is apparent when we consider results of OvBag. Taking into account information about the neighbourhood of minority examples improves classification performance with respect to *G*-mean, and sensitivity evaluation measures. This hypothesis is supported by results of both oNBBag^{2} and uNBBag$^{0.5}$. To conclude, the introduction of local modifications of sampling probabilities inside the combination rule of NBBag may be the crucial element leading to the significantly better performance than all over-sampling variants as well as for making it competitive to RBBag.

When we analyse which parameters lead to the best *G*-mean, we have noticed that, in most of the cases, neighbourhood composed of $k=5$ examples is sufficient. Larger neighbourhood may lead to slightly better results in under-sampling NBBag for only small fraction of the data sets, which are averagely difficult to more difficult: `credit-g`, `ecoli`, `haberman`, `breast-cancer`, and `solar-flare`. This is an important observation from the effectiveness of learning point of view. Larger neighbourhoods may lead to more computational effort during learning. When we look for the best values of ψ, the choice clearly depends on whether over-sampling NBBag or under-sampling NBBag is applied. For over-sampling higher $\psi=2$ is often the best choice for unsafe data sets but also lower values are desirable for more safe data sets. In under-sampling NBBag the best value of ψ is almost always 0.5, higher value 1 leads to small improvement for safe data sets. In both cases, over-sampling and under-sampling NBBag, ψ higher than 2 may lead to slightly better result in the safest data sets (only `breast-w` in our comparison) it is, however, followed with high deterioration of results on other types of data sets.

6.2. Analyzing data characteristics and bootstrap samples

The aim of this part of experiments is to learn more about the nature of the best bagging extensions. First, we want to identify proportion of different types of examples in the minority class of considered data sets (recall their distinction in Section 3). Following the method introduced in [33], we propose to assign types of examples using information about class labels in their *k*-nearest local neighbourhood.

In this analysis we will again use $k=5$ mainly because $k=3$ may poorly distinguish the nature of examples, and in earlier experiments [35], as well as in the current ones, examining higher values as $k=7$ has led to quite similar decisions as to identification of types examples in the data sets. This choice is also similar to the size of neighbourhood used in NBBag and in main pre-processing methods such as SMOTE or SPIDER.

For the considered example *x* and $k=5$, the proportion of the number of neighbours from the same class as *x* against neighbours from the opposite class can range from 5:0 (all neighbours are from the same class as the analysed example *x*) to 0:5 (all neighbours belong to the opposite class). Depending on this proportion, we assign the type labels to the example *x* in the following way [33]: Proportions 5:0 or 4:1 inside the neighbourhood – the example *x* is labeled as a *safe example* (as it is surrounded mainly by examples from the same class); 3:2 or 2:3 – it is a *borderline example* (the explanation is that the number of neighbours from both classes is approximately the same, so it refers to class overlapping near the decision boundary. Notice that within this interpretation the examples with the proportion 3:2 although still correctly classified by its neighbours, this example could be located close to the decision boundary between the classes); 1:4 – it is interpreted as a *rare case* (as explained in Section 4); 0:5 – it is an *outlier*. For higher values of *k* such proportions could be interpreted in a similar way – see their definitions in [35].

Although this categorization could be seen as based on intuitive thresholding, its results are consistent with a more probabilistic analysis of the neighbourhood, modeled with kernel functions, as it is shown in [35]. Knowing also that higher values *k* have led to identification of similar distributions of minority class examples in considered UCI data sets we will stay with presenting results for $k=5$.

The results of such labeling of the minority class examples are presented in Table 7. The first observation is that many data sets contain rather a small number of safe minority examples. The exceptions are three data sets composed of almost only safe examples: `breast-w`, `car`. On the other hand, there are data sets such as `cleveland`, `balance-scale` or `solar-flare`, which do not contain any safe examples. We carried out the similar neighbourhood analysis for the majority classes and make a contrary observation – nearly all data sets contain mainly safe majority examples (e.g., `yeast`: 98.5%, `ecoli`: 91.7%) and sometimes a limited number of borderline examples (e.g., `balance-scale`: 84.5% safe and 15.6% borderline examples). What is even more important, nearly all data sets do not contain any majority outliers and at most 2% of rare examples. Thus, we can repeat similar conclusions to [33], saying that in most data sets the minority class includes mainly difficult unsafe examples.

Then, one can observe that for safe data sets nearly all bagging extensions achieve similar high performance (see Tables 4 and 5 for `breast-w`, `new-thyroid`). A quite similar observation concerns data sets with still high number of safe examples, limited borderline ones and no/or nearly no rare cases or outliers – see, e.g., `vehicle`. One the other hand, the strong differences between classifiers occur for the most difficult data distributions with a limited number of safe minority examples. Furthermore, the best improvements of all evaluation measures for RBBag and NBBag are observed for the unsafe data sets. For instance, consider `cleveland` (no safe examples, nearly 50% of outliers) where uNBBag$^{0.5}$ has 74.3% *G*-mean compared to OvBag with 22.7%. Similar highest improvements occur for

Table 7
Labeling minority class examples expressed as a percentage of each type of examples occurring in this class.

Data set	Safe (%)	Border (%)	Rare (%)	Outlier (%)
breast-w	91.29	7.88	0.00	0.83
abdominal-pain	61.39	23.76	6.93	7.92
acl	67.5	30.00	0.00	2.5
new-thyroid	68.57	31.43	0.00	0.00
vehicle	74.37	24.62	0.00	1.01
car	47.83	47.83	0.00	4.35
scrotal-pain	50.85	33.90	10.17	5.08
ionosphere	44.44	30.95	11.90	12.70
pima	29.85	56.34	5.22	8.58
credit-g	15.67	61.33	12.33	10.67
ecoli	28.57	54.29	2.86	14.29
hepatitis	18.75	62.50	6.25	12.50
haberman	4.94	61.73	18.52	14.81
breast-cancer	21.18	38.82	27.06	12.94
cmc	13.81	53.15	14.41	18.62
cleveland	0.00	45.71	8.57	45.71
hsv	0.00	0.00	28.57	71.43
abalone	8.36	20.6	20.6	50.45
postoperative	0.00	41.67	29.17	29.17
solar-flare	2.33	41.86	16.28	39.53
transfusion	18.54	47.19	11.24	23.03
yeast	5.88	47.06	7.84	39.22
balance-scale	0.00	0.00	8.16	91.84

Please cite this article as: J. Błaszczyński, J. Stefanowski, Neighbourhood sampling in bagging for imbalanced data, Neurocomputing (2014), http://dx.doi.org/10.1016/j.neucom.2014.07.064

balance-scale (containing the highest number of outliers among all data sets) where oNBBag^2 gets 61.07% while OvBag 1.4%, and SmBag 0%. Analogous situations also occur for yeast, solar-flare, postoperative, hsv, and cleveland. We can conclude that RBBag and NBBag strongly outperform other bagging extensions for the most difficult data sets with large numbers of outliers or rare cases – sometimes occurring with borderline examples.

In order to better understand the improvements achieved by RBBag and NBBag, we perform a similar, but more detailed, neighbourhood analysis of minority examples inside their bootstraps. For each bootstrap sample constructed by standard bagging, NBBag and RBBag, we calculate distribution of N'_{maj}, which are numbers of examples from majority class belonging to k-nearest neighbourhood of minority class example present in the sample. More precisely, we take an average of proportion of a number of examples having a specific N'_{maj} to the number of all minority examples in the original data set (not the number of minority class examples in the bootstrap sample). We consider standard bagging bootstrap samples, as well as, RBBag samples and samples obtained by oNBBag^2, and uNBBag$^{0.5}$. The results of the averaging are presented in Fig. 2. The results for standard bootstrap sampling reflect the distribution of labels presented in Table 7.

In our opinion these results reveal very interesting properties of RBBag and NBBag ensembles. Both ensembles strongly change types of the minority class distributions into much safer ones inside their bootstraps. This result is visible by relatively more examples with lower N'_{maj} than in standard bootstrap samples. For many data sets, which originally contain high numbers of rare cases or outliers, the transformed bootstrap samples contain now more safe examples. For instance, consider very difficult balance-scale data set (containing 91.8% of outliers in the whole data set, and a significant amount of examples with high value of N'_{maj} in standard bootstrap), where oNBBag^2 creates bootstrap samples with majority of safe examples characterized by low value of N'_{maj}. Similar example type shift can be observed for yeast (only 5% safe examples in the whole data set), abalone, hsv, cleveland, and ecoli. The distribution of weights produced by NBBag shows that increase of ψ affects strongly the amount of safe examples in samples. The result is that oNBBag^2 classifier is, in most cases, trained on bootstraps composed of safe examples. Also note that, oNBBag^2 is producing bootstraps with monotonic distribution of N'_{maj}, which results from choosing the formula L^2_{min}. Another aspect of oNBBag^2 sampling is decrease of diversity of samples. When we compare the sums of average proportions of examples having different N'_{maj} it is apparent that in some cases almost all of examples from minority class are present in each constructed sample (abdominal-pain, balance-scale, car, ecoli, new-thyroid, vehicle, yeast). This is not the case for uNBBag$^{0.5}$ classifier, whose distributions are affected by lighter amplification by ψ parameter, as well as by the under-sampling effect. The resulting distributions produced by uNBBag$^{0.5}$ are often characterized by smaller representation of minority class examples than in oNBBag^2 and RBBag. It is especially true for unsafe data sets with high IR (balance-scale, yeast, solar-flare, abalone, cleveland). The same effect can be, however, also noticed for large, safe data sets having high IR (car).

RBBag and uNBBag$^{0.5}$ are less aggressive than oNBBag^2 in converting outliers to safe examples. In fact, we have not noticed a situation where RBBag or uNBBag$^{0.5}$ is able to construct completely safe bootstraps from unsafe data set. Distribution of weights in RBBag bootstraps is not monotonic, since the change of this distribution results from random omission of majority examples regardless of their type.

Data set	Type	N'maj 0	1	2	3	4	5
breast-w	standard	0.62	0.07	0.03	0.01	0.01	0.00
	RBBag	0.66	0.06	0.02	0.01	0.00	0.00
	oNBBag2	0.84	0.06	0.01	0.00	0.00	0.00
	uNBBag0.5	0.59	0.03	0.01	0.01	0.01	0.00
abdominal-pain	standard	0.37	0.13	0.11	0.08	0.07	0.00
	RBBag	0.46	0.13	0.08	0.05	0.04	0.00
	oNBBag2	0.71	0.17	0.05	0.01	0.00	0.00
	uNBBag0.5	0.34	0.10	0.06	0.04	0.02	0.00
acl	standard	0.62	0.07	0.03	0.01	0.01	0.00
	RBBag	0.66	0.06	0.02	0.01	0.00	0.00
	oNBBag2	0.84	0.06	0.01	0.00	0.00	0.00
	uNBBag0.5	0.34	0.08	0.04	0.02	0.01	0.00
new-thyroid	standard	0.52	0.09	0.09	0.04	0.01	0.00
	RBBag	0.64	0.08	0.03	0.01	0.00	0.00
	oNBBag2	1.00	0.00	0.00	0.00	0.00	0.00
	uNBBag0.5	0.15	0.01	0.00	0.00	0.00	0.00
vehicle	standard	0.47	0.13	0.08	0.04	0.02	0.00
	RBBag	0.57	0.12	0.04	0.01	0.00	0.00
	oNBBag2	0.93	0.05	0.01	0.00	0.00	0.00
	uNBBag0.5	0.37	0.06	0.03	0.02	0.01	0.00
car	standard	0.17	0.18	0.20	0.15	0.05	0.00
	RBBag	0.64	0.10	0.01	0.00	0.00	0.00
	oNBBag2	1.00	0.00	0.00	0.00	0.00	0.00
	uNBBag0.5	0.01	0.00	0.00	0.00	0.00	0.00
scrotal-pain	standard	0.19	0.18	0.17	0.12	0.09	0.00
	RBBag	0.37	0.17	0.11	0.07	0.04	0.00
	oNBBag2	0.62	0.21	0.09	0.02	0.00	0.00
	uNBBag0.5	0.27	0.11	0.08	0.05	0.02	0.00
ionosphere	standard	0.29	0.12	0.12	0.12	0.11	0.00
	RBBag	0.33	0.14	0.11	0.09	0.08	0.00
	oNBBag2	0.54	0.22	0.10	0.02	0.00	0.00
	uNBBag0.5	0.45	0.10	0.08	0.06	0.03	0.00
pima	standard	0.16	0.18	0.18	0.13	0.09	0.00
	RBBag	0.25	0.21	0.15	0.09	0.05	0.00
	oNBBag2	0.46	0.31	0.12	0.02	0.00	0.00
	uNBBag0.5	0.25	0.17	0.13	0.08	0.03	0.00
credit-g	standard	0.07	0.14	0.18	0.20	0.16	0.00
	RBBag	0.18	0.22	0.19	0.12	0.05	0.00
	oNBBag2	0.49	0.32	0.12	0.02	0.00	0.00
	uNBBag0.5	0.17	0.17	0.14	0.09	0.03	0.00
ecoli	standard	0.17	0.15	0.17	0.15	0.10	0.00
	RBBag	0.49	0.17	0.06	0.02	0.02	0.00
	oNBBag2	0.98	0.02	0.00	0.00	0.00	0.00
	uNBBag0.5	0.06	0.02	0.01	0.00	0.00	0.00
hepatitis	standard	0.13	0.16	0.20	0.16	0.14	0.00
	RBBag	0.35	0.16	0.12	0.07	0.05	0.00
	oNBBag2	0.84	0.13	0.02	0.00	0.00	0.00
	uNBBag0.5	0.20	0.07	0.04	0.02	0.02	0.00
haberman	standard	0.06	0.13	0.20	0.21	0.15	0.00
	RBBag	0.13	0.21	0.21	0.15	0.05	0.00
	oNBBag2	0.54	0.33	0.10	0.01	0.00	0.00
	uNBBag0.5	0.13	0.15	0.13	0.08	0.03	0.00
breast-cancer	standard	0.12	0.13	0.16	0.20	0.15	0.00
	RBBag	0.21	0.17	0.19	0.13	0.06	0.00
	oNBBag2	0.52	0.27	0.11	0.04	0.01	0.00
	uNBBag0.5	0.19	0.15	0.13	0.09	0.03	0.00
cmc	standard	0.08	0.13	0.17	0.20	0.18	0.00
	RBBag	0.20	0.21	0.18	0.11	0.05	0.00
	oNBBag2	0.65	0.25	0.07	0.01	0.00	0.00
	uNBBag0.5	0.15	0.13	0.10	0.07	0.03	0.00
cleveland	standard	0.02	0.05	0.12	0.21	0.34	0.00
	RBBag	0.23	0.27	0.15	0.07	0.03	0.00
	oNBBag2	0.94	0.05	0.01	0.00	0.00	0.00
	uNBBag0.5	0.11	0.06	0.04	0.02	0.01	0.00
hsv	standard	0.01	0.03	0.09	0.25	0.39	0.00
	RBBag	0.04	0.14	0.27	0.24	0.08	0.00
	oNBBag2	0.99	0.01	0.00	0.00	0.00	0.00
	uNBBag0.5	0.01	0.03	0.02	0.02	0.00	0.00
abalone	standard	0.06	0.08	0.12	0.20	0.30	0.00
	RBBag	0.25	0.19	0.15	0.10	0.05	0.00
	oNBBag2	0.98	0.02	0.00	0.00	0.00	0.00
	uNBBag0.5	0.09	0.05	0.03	0.02	0.01	0.00
postoperative	standard	0.02	0.09	0.18	0.24	0.24	0.00
	RBBag	0.11	0.17	0.21	0.19	0.08	0.00
	oNBBag2	0.44	0.32	0.16	0.05	0.01	0.00
	uNBBag0.5	0.08	0.17	0.16	0.10	0.04	0.00
solar-flare	standard	0.02	0.06	0.14	0.22	0.25	0.07
	RBBag	0.39	0.19	0.08	0.05	0.04	0.00
	oNBBag2	0.80	0.13	0.04	0.02	0.01	0.00
	uNBBag0.5	0.06	0.01	0.01	0.00	0.00	0.00
transfusion	standard	0.09	0.11	0.15	0.17	0.18	0.04
	RBBag	0.19	0.19	0.13	0.10	0.07	0.06
	oNBBag2	0.55	0.28	0.11	0.02	0.00	0.00
	uNBBag0.5	0.15	0.14	0.11	0.08	0.04	0.00
yeast	standard	0.04	0.09	0.16	0.20	0.27	0.00
	RBBag	0.35	0.20	0.11	0.07	0.02	0.00
	oNBBag2	1.00	0.00	0.00	0.00	0.00	0.00
	uNBBag0.5	0.04	0.01	0.01	0.01	0.00	0.00
balance-scale	standard	0.00	0.01	0.04	0.18	0.51	0.00
	RBBag	0.14	0.29	0.22	0.08	0.02	0.00
	oNBBag2	1.00	0.00	0.00	0.00	0.00	0.00
	uNBBag0.5	0.05	0.04	0.03	0.02	0.01	0.00

Fig. 2. Average distribution of N'_{maj} in bootstraps: standard bagging, RBBag, over-sampling NBBag with $\psi = 2$, and under-sampling NBBag with $\psi = 0.5$.

Please cite this article as: J. Błaszczyński, J. Stefanowski, Neighbourhood sampling in bagging for imbalanced data, Neurocomputing (2014), http://dx.doi.org/10.1016/j.neucom.2014.07.064

Recall that extensions of bagging known from the literature are based on the simple idea of balancing distributions in bootstrap samples. Our results indicate that transforming distributions of examples into safer ones can be more influential. In case of RBBag it could be connected with strong filtering majority class examples in each bootstrap sample. Notice that many data sets contain nearly 1000 examples with around 50 minority ones. For instance, the number of all examples in `solar-flare` is 1066 while the minority class contains 43 examples only. The new created bootstrap samples include only 43 safe majority examples and as a result most of the majority class examples (also reflecting their original distribution) disappear. It can be interpreted as a kind of cleaning around the minority class examples, so they become safer in their local neighbourhood. Having such a transformed distribution in each sample can help construct base classifiers, which are more biased toward the minority class. On the other hand, the size of the learning set can be dramatically reduced. As a result, some bootstrap samples may lead to weak classifiers, and this type of ensemble may need more component classifiers than over-sampling NBBag, which uses larger bootstrap samples.

Now, when we compare the change in distribution of examples in bootstrap samples produced by different parameters of NBBag to the classification performance resulting from learning on these samples it is quite apparent that the favourable type of change is dependent on the data set. This allows us to draw a conclusion that one should look for the best parameters with respect to a particular data set.

7. Discussion and final remarks

Our paper is devoted to various extensions of bagging ensembles dedicated to improve classification of imbalanced data. We have focused our attention on the most studied approaches, which integrate pre-processing methods specialized for class imbalance within bootstrap construction. The first contribution of our paper is a comprehensive experimental comparison of the well known bagging extensions over a large collection of diversified imbalanced data sets. Results of these experiments have clearly shown that all under-sampling extensions of bagging have achieved much better classification performance than all over-sampling variants. Definitely, for all considered evaluation measures both Exactly Balanced Bagging (EBBag) and Roughly Balanced Bagging (RBBag) produce the best results. Still, the difference between them and the best over-sampling bagging is much higher than we have expected and what has been previously shown in [13]. The superiority of using underbagging is also consistent with some of opinions from [19,24], although these experiments were done in other frameworks.

What is also worth noting is that, according to our results, SMOTEBagging is not as accurate as it has been postulated by its authors in [39]. Quite often it is the worst classifier and it is also worse than much simpler random over-sampling. Moreover, a new over-sampling bagging variant, where SMOTE is applied with the same oversampling ratio, works better than this previously promoted more diversified SMOTEBagging.

Although EBBag and RBBag performs similar to each other with respect to the sensitivity and *G*-mean, RBBag seems to be slightly better than EBBag for *F*-measure, in particular when sampling is done with replacement. Similar performance of these both classifiers was also observed in [24], however their experiments were carried out in another framework with artificially modified noisy data. However, authors of RBBag have already shown its slightly better performance over EBBag [19] but over 9 data sets only (4 of them was also used in our experiments). Yet another novel observation is that sampling with replacement may be profitable for RBBag unlike EBBag, where our results show no differences between sampling with or without replacement. If one expects a single strong recommendation from these parts of experiments, we will suggest to use Roughly Balanced Bagging as the most efficient, simplest to use (tune only the number of component classifiers) and the most consistent with the original Breiman's idea of bootstrap aggregation.

Analyzing known extensions of bagging one can also notice that they usually use the simplest random re-sampling since more complex SMOTE inside overbagging does not work. This a contradictory situation to single classifiers, where SMOTE usually is one of the best options see, e.g., [4]. However, considering future research, one could still ask a question how BorderlineSMOTE, LN-SMOTE and other SMOTE extensions [31] work in this setting. Yet another future issue could be studying diversity in a more advanced way than presented in [39]. We would also consider looking for new diversity measures more specialized for class imbalance than the previous ones, which are oriented at total accuracy [26].

Our next methodological contributions result from the hypothesis saying that the difficulty of learning classifiers from imbalanced data comes from complex distributions of the minority class [33]. Besides the unequal class cardinalities, the minority class is decomposed into smaller sub-parts, affected by strong over-lapping, rare cases and/or outliers. In our study we attempt to capture these data characteristics by analyzing the neighbourhood of minority class examples. The main message of our study is that such a kind of local information can be useful both for proposing a new direction of generalizing bagging and for analyzing more deeply data conditions which may provide explanation why some ensembles work better than others.

We have proposed a new type of bagging, called Neighbourhood Balanced Bagging (NBBag), which is based on a different principle than all known bagging extensions for class imbalance. First, instead of integrating bagging with pre-processing, we keep the standard bagging idea but we change, sometimes radically, probabilities of sampling examples to bootstraps by increasing the chance of drawing minority examples. Furthermore, we amplify the role of difficult minority examples with respect to the type of their neighbourhood. The strength of amplification can be parameterized in our setting. We have given some indications how the choice of values of this parameter affects learning. We have also identified the values of the parameter that work the best on the average. The experiments have proven that the choice of parameter values that lead to satisfactory results is rather limited.

We have shown that our proposition can be applied in both types of bagging generalizations: over-sampling and under-sampling. In the first type of generalization, our proposition is similar to over-sampling minority class examples into bootstraps, however, at the same time, the probabilities of drawing majority class examples are decreased. The size of bootstrap is kept the same as the size of the original learning set. The second type is inspired by under-sampling generalizations, which are proven to work better than over-sampling generalizations. We construct bootstraps of double size of the minority class. The probabilities of drawing minority class examples are increased, while probabilities of drawing majority class examples are decreased. The experimental results show that under-sampling Neighbourhood Balanced Bagging is at least competitive to Roughly Balanced Bagging, which is considered as the best known bagging generalization. They also show that our proposal is significantly better than existing over-sampling extensions of bagging regardless whether over-sampling variant or under-sampling one is considered. The strongest differences between classifiers performance have been noticed for data sets containing the most unsafe minority examples. Indeed, both NBBag and RBBag ensembles

Please cite this article as: J. Błaszczyński, J. Stefanowski, Neighbourhood sampling in bagging for imbalanced data, Neurocomputing (2014), http://dx.doi.org/10.1016/j.neucom.2014.07.064

have strongly outperformed all over-sampling bagging variants for such data.

We have also shown that the source of this difference in performance can be linked to the change of distribution of unsafe (difficult to learn) minority class examples into safe ones introduced directly by NBBag and indirectly by RBBag. The analysis of types of minority examples inside bootstrap samples has clearly shown that NBBag and RBBag strongly changed data characteristics compared to the original data sets. This analysis follows the earlier research by Stefanowski and Napierala [33], however, its application in the context of ensembles uncovers new interesting characteristics of studied ensembles. Many examples from the minority class labeled as unsafe in the original learning set (in particular as rare cases or outliers) have been transformed to more safe ones. We have demonstrated that over-sampling NBBag, in the best performing variant, learns, in most of the cases, from almost safe type bootstrap samples. Under-sampling NBBag, in the best performing variant, is relatively less altering the distribution in this direction. RBBag is closer to the under-sampling variant of NBBag but it is less adapting to the type of distribution of minority class examples in the data set. This change of the local characteristics of learning examples may be more influential for improving the classification performance than the simple global class balancing, which has previously been considered in the literature and applied in many of known approaches to extend bagging. It introduces additional computational cost associated with detection of how safe (or unsafe) learning examples are. This cost is, however, introduced only for minority class examples. Moreover, our experimental evaluation shows that relatively small neighbourhoods are sufficient to achieve satisfactory performance. This is confirmed by our observation that even though under-sampling Neighbourhood Balanced Bagging is computationally more costly than Roughly Balanced Bagging, the differences in cost of learning are not significant.

Acknowledgements

The authors' research was funded by the Polish National Science Center, Grant no. DEC-2013/11/B/ST6/00963.

References

[1] D. Anyfantis, M. Karagiannopoulos, S. Kotsiantis, P. Pintelas, Robustness of learning techniques in handling class noise in imbalanced datasets, in: Proceedings of the IFIP International Federation for Information Processing Conference, AIAI 2007, 2007, pp. 21–28.

[2] D. Anyfantis, M. Karagiannopoulos, S. Kotsiantis, P. Pintelas, Creating ensembles of classifiers by distributing an imbalance dataset to reach balance in each resulting training set, in: Proceedings of the IEEE International Conference on Distributed Human-Machine Systems Conference, DHMS, 2008.

[3] A. Asuncion, D.J. Newman, UCI Machine Learning Repository. ⟨http://www.ics.uci.edu/~mlearn/MLRepository.html⟩. University of California, School of Information and Computer Science, Irvine, CA, 2007.

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

[5] G. Batista, D.F. Silva, How k-nearest neighbor parameters affect its performance, in: Proceedings of Argentine Symposium on Artificial Intelligence, Mar del Plata, Argentina, 2009, pp. 1–12.

[6] J. Błaszczyński, R. Słowiński, J. Stefanowski, Feature set-based consistency sampling in bagging ensembles, in: Proceedings From Local Patterns To Global Models (LEGO), ECML/PKDD Workshop, 2009, pp. 19–35.

[7] J. Błaszczyński, J. Stefanowski, L. Idkowiak, Extending bagging for imbalanced data, in: Proceedings of the eighth CORES 2013, Springer Series on Advances in Intelligent Systems and Computing, vol. 226, 2013, pp. 269–278.

[8] L. Breiman, Bagging predictors, Mach. Learn. 24 (2) (1996) 123–140.

[9] P.K. Chan, S. Stolfo, Toward scalable learning with non-uniform class and cost distributions: a case study in credit fraud detection, in: Proceedings of ACM SIGKD'98, 1998, pp. 164–168.

[10] N. Chawla, K. Bowyer, L. Hall, W. Kegelmeyer, SMOTE: synthetic minority over-sampling technique, J. Artif. Intell. Res. 16 (2002) 341–378.

[11] N. Chawla, Data mining for imbalanced datasets: an overview, in: O. Maimon, L. Rokach (Eds.), The Data Mining and Knowledge Discovery Handbook, 2005, pp. 853–867.

[12] C. Chen, A. Liaw, L. Breiman, Using Random Forest to Learn Imbalanced Data, Technical Report, University of California, Berkley, 2004.

[13] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, F. Herrera, A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches, IEEE Trans. Syst. Man Cybern. Part C: Appl. Rev. 99 (2011) 1–22.

[14] V. Garcia, R.A. Mollineda, J.S. Sanchez, On the k-NN performance in a challenging scenario of imbalance and overlapping, Pattern Anal. Appl. 11 (3–4) (2008) 269–280.

[15] H. Han, W. Wang, B. Mao, Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. In Proceedings of ICIC, Lecture Notes in Computer Science, vol. 3644, Springer, Berlin, Heidelberg, 2005, pp. 878–887.

[16] V. Garcia, J.S. Sanchez, R.A. Mollineda, An empirical study of the behaviour of classifiers on imbalanced and overlapped data sets, in: Proceedings of Progress in Pattern Recognition, Image Analysis and Applications 2007, Lecture Notes in Computer Science, vol. 4756, Springer, Berlin, Heidelberg, 2007, pp. 397–406.

[17] H. He, E. Garcia, Learning from imbalanced data, IEEE Trans. Data Knowl. Eng. 21 (9) (2009) 1263–1284.

[18] H. He, Ma Yungian, Imbalanced Learning. Foundations, Algorithms and Applications, Wiley-IEEE Press, 2013.

[19] S. Hido, H. Kashima, Roughly balanced bagging for imbalance data, Stat. Anal. Data Min. 2 (5–6) (2009) 412–426.

[20] J. Hulse, T.M. Khoshgoftaar, A. Napolitano, Experimental perspectives on learning from imbalanced data, in: Proceedings of the 24th International Conference on Machine Learning (ICML), 2007, pp. 17–23.

[21] N. Japkowicz, Class imbalance: are we focusing on the right issue?, in: Proceedings of Second Workshop on Learning from Imbalanced Data Sets, ICML Conference, 2003, pp. 17–23.

[22] N. Japkowicz, Mohak Shah, Evaluating Learning Algorithms: A Classification Perspective, Cambridge University Press, 2011.

[23] T. Jo, N. Japkowicz, Class Imbalances versus small disjuncts, ACM SIGKDD Explor. Newslett. 6 (1) (2004) 40–49.

[24] T. Khoshgoftaar, J. Van Hulse, A. Napolitano, Comparing boosting and bagging techniques with noisy and imbalanced data, IEEE Trans. Syst. Man Cybern. Part A 41 (3) (2011) 552–568.

[25] M. Kubat, S. Matwin, Addresing the curse of imbalanced training sets: one-side selection, in: Proceedings of International Conference on Machine Learning (ICML), vol. 97, 1997, pp. 179–186.

[26] L. Kuncheva, Combining Pattern Classifiers: Methods and Algorithms, Wiley-Interscience, 2004.

[27] J. Laurikkala, Improving Identification of Difficult Small Classes by Balancing Class Distribution. Technical Report A-2001-2, University of Tampere, 2001. Another version was published in: The Eighth Conference on AI in Medicine in Europe (AIME01), Lecture Notes in Computer Science, vol. 2001, Springer, Berlin, Heidelberg, 2001, pp. 63–66.

[28] J. Lumijarvi, J. Laurikkala, M. Juhola, A comparison of different heterogeneous proximity functions and Euclidean distance, Stud. Health Technol. Informatics 107 (Pt 2) (2004) 1362–1366.

[29] A. Liu, Zh Zhu, Ensemble methods for class imbalance learning, in: H. He, Ma Yungian (Eds.), Imbalanced Learning. Foundations, Algorithms and Applications, Wiley-IEEE Press, 2013, pp. 61–82.

[30] V. Lopez, A. Fernandez, S. Garcia, V. Palade, F. Herrera, An insight into classification with imbalanced data: empirical results and current trends on using data intrinsic characteristics, Inf. Sci. 257 (2014) 113–141.

[31] T. Maciejewski, J. Stefanowski, Local neighnourhood extension of SMOTE for mining imbalanced data, in: Procedings of IEEE Symposium on Computational Intelligence and Data Mining, 2011, pp. 104–111.

[32] K. Napierala, J. Stefanowski, Sz. Wilk, Learning from imbalanced data in presence of noisy and borderline examples, in: Proceedings of Seventh International Conference RSCTC 2010, Lecture Notes in Artificial Intelligence, vol. 6086, Springer, Berlin, Heidelberg, 2010, pp. 158–167.

[33] K. Napierała, J. Stefanowski, Identification of different types of minority class examples in imbalanced data, in: Proceedings of Seventh International Conference HAIS 2012, Part II, Lecture Notes in Computer Science, vol. 7209, Springer, Berlin, Heidelberg, 2012, pp. 139–150.

[34] K. Napierala, J. Stefanowski, BRACID: a comprehensive approach to learning rules from imbalanced data, J. Intell. Inf. Syst. 39 (2) (2012) 335–373.

[35] K. Napierala, Improving rule classifiers for imbalanced data (Ph.D. thesis), Poznan University of Technology, 2013.

[36] R. Prati, G. Batista, M. Monard, Class imbalance versus class overlapping: an analysis of a learning system behavior, in: Proceedings of the Third Mexican International Conference on Artificial Intelligence, 2004, pp. 312–321.

[37] J. Stefanowski, Sz. Wilk, Selective pre-processing of imbalanced data for improving classification performance, in: Proceedings of 10th International Conference DaWaK 2008, Lecture Notes in Computer Science, vol. 5182, Springer, Berlin, Heidelberg, 2008, pp. 283-292.

[38] J. Stefanowski, Overlapping, rare examples and class decomposition in learning classifiers from imbalanced data, in: S. Ramanna, L.C. Jain, R.J. Howlett (Eds.), Emerging Paradigms in Machine Learning, 2013, pp. 277–306.

[39] S. Wang, T. Yao, Diversity analysis on imbalanced data sets by using ensemble models, in: Proceedings of IEEE Symposium on Computational Intelligence and Data Mining, 2009, pp. 324–331.

[40] G.M. Weiss, Mining with rarity: a unifying framework, ACM SIGKDD Explor. Newslett. 6 (1) (2004) 7–19.

[41] D.R. Wilson, T.R. Martinez, Improved heterogeneous distance functions, J. Artif. Intell. Res. 6 (1997) 1–34.

Please cite this article as: J. Błaszczyński, J. Stefanowski, Neighbourhood sampling in bagging for imbalanced data, Neurocomputing (2014), http://dx.doi.org/10.1016/j.neucom.2014.07.064

Jerzy Błaszczyński is an Assistant Professor in the Institute of Computing Science, Poznan University of Technology. He obtained his Ph.D. from this university. In 2010 he received scholarship for outstanding young researchers from Polish Ministry of Education. His research interests focus on decision support, including multiple criteria decision aiding, preference modeling and on machine learning.

Jerzy Stefanowski is an Associate Professor in the Institute of Computing Science, Poznan University of Technology. He received the Ph.D. and Habilitation degrees in computer science from this university. His research interests include machine learning, data mining and intelligent decision support – in particular, rule induction, ensemble classifiers, class imbalance, data preprocessing, handling uncertainty in data, and medical applications.

Please cite this article as: J. Błaszczyński, J. Stefanowski, Neighbourhood sampling in bagging for imbalanced data, Neurocomputing (2014), http://dx.doi.org/10.1016/j.neucom.2014.07.064