Local Neighbourhood in Generalizing Bagging for Imbalanced Data

Jerzy Błaszczyński, Jerzy Stefanowski and Marcin Szajek

Institute of Computing Sciences, Poznań University of Technology,
60–965 Poznań, Poland
{jerzy.blaszczynski, jerzy.stefanowski, marcin.szajek}@cs.put.poznan.pl

Abstract. Bagging ensembles specialized for class imbalanced data are considered. We show that difficult distributions of the minority class can be handled by analyzing the content of the local neighbourhood of examples. First, we introduce a new generalization of bagging, called Neighbourhood Balanced Bagging, where sampling probabilities of examples are modified according to the class distribution in their neighbourhoods. Experiments show that it is competitive to other extensions of bagging. Finally, we demonstrate that assessing types of the minority examples based on the analysis of their neighbourhoods could help in explaining why some ensembles work better for imbalanced data than others.

1 Introduction

One of the sources of difficulties while constructing accurate classifiers is class imbalance in data. This difficulty manifests itself by the fact that one of the target classes is significantly less numerous than the other classes. This problem often occurs in many applications, and constitutes a difficulty for most learning algorithms. As a results many classifiers are biased toward majority classes and fail to recognize examples from the minority class.

The class imbalance problem has received a growing research interest in the last decade and a number of specialized methods have been proposed, see their review [7]. In general, they are categorized in *data level* and *algorithm level* ones. Methods from the first category are based on pre-processing, and they transform the original data distribution into more balanced one. The simplest methods 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 required degree of balance between classes is reached. The more informative methods, e.g., SMOTE, introduce additional synthetic examples according to an internal characteristics of regions around examples from the minority class [4]. The methods from the second category are classifier dependent ones. They also include ensembles of classifiers. However, the standard methods of ensemble construction are oriented toward improving the overall classification accuracy and they do not solve sufficiently the recognition of the minority class. The new proposed ensembles usually include either integrating pre-processing

methods before learning component classifiers or embedding the cost-sensitive framework in the ensemble learning process, see their review in [5].

Although several specialized ensembles have been presented as adequate to class imbalance, there is still lack of their general comparison or discussion of their competence area. To the best to our knowledge, only two comprehensive studies were carried out in different experimental frameworks [5, 10]. The main conclusions from the comparative study [5] were that simpler versions of using under-sampling or SMOTE inside ensembles worked better than more complex solutions. Then, the experiments with few best boosting and bagging generalizations over noisy and imbalanced data sets showed that bagging outperformed boosting. In our previous study we experimentally compared main bagging variants for class imbalance and also observed that under-sampling bagging performed much better than variants with over-sampling [2]. In particular, the Roughly Balanced Bagging [8] achieved the best results.

We keep our interest in bagging extensions and identify two tasks to be undertaken. The first task is looking for the hypothesis why the under-sampling bagging works better than over-sampling ones. The other task concerns an attempt to construct yet another ensemble, more similar to over-sampling the minority class, leading to performance closer to the Roughly Balanced Bagging. Our new view is to resign from the simple integration of pre-processing with unchanged bagging sampling technique. Unlike using the equal probabilities of each example in bootstrap sampling we want to change probabilities of their drawing and to focus this sampling toward the minority class and additionally more to the examples located in difficult sub-regions of the minority class.

While considering the probability of each example to be drawn we propose to analyze class distributions in the local neighbourhood of the minority example [13]. Depending on the distribution of examples from the majority class in this neighbourhood, we can evaluate whether this example could be safe or unsafe (difficult) to be learned. This approach is inspired by our earlier positive experience with studying the "nature" of imbalanced data where such local characteristics was successfully modeled with the k-nearest neighbourhood [13].

To sum up, the main contributions of our study are the following. The first aim is to introduce a new extension of bagging for class imbalance, where the probability of selecting an example into the bootstrap sample is influenced by the analysis of the class distribution in a local neighbourhood of the example. The new proposal is compared against existing extensions over several data sets. Then, the second aim is to use the same type of analysis to explain how contents of bootstrap samples affect performance of the Roughly Balanced Bagging and the proposed new extension.

2 Related Works

Due to space limits, we briefly discuss the most related works only. The reader is referred to [5] for the most comprehensive review of current ensembles addressing the class imbalance. Below we discuss extensions of bagging.

Recall that original Breiman's bagging [3] is based on the *bootstrap* aggregation, where the training set for each classifier is constructed by random uniformly sampling (with replacement) instances from the original training set (usually keeping the size of the original data). Then, T component classifiers are induced by the same learning algorithm from these T bootstrap samples. Their predictions form the final decision with the equal weight majority voting. However, bootstrap samples are still biased toward the majority class. Most of proposals overcome this drawback by using pre-processing techniques, which change the balance between classes in bootstraps.

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 *Exactly Balanced Bagging* (EBBag), while creating each training bootstrap sample, the entire minority class is just copied and combined with randomly chosen subsets of the majority class to exactly balance cardinalities between classes. The base classifiers and their aggregation are constructed as in the standard bagging.

The *Roughly Balanced Bagging* (RBBag) [8] results from the critique of the EBBag. 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 final bootstrap sample (S_{maj}) is determined probabilistically according to the negative binominal 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 inside the bootstrap samples maybe slightly imbalanced and varies over iterations. According to [8] this approach is more consistent with the nature of the original bagging and performs better than EBBag. In our experiments on larger collection of data [2] both RBBag and EBBag achieved quite similar results for the sensitivity measure while RBBag was slightly better than EBBag for G-mean and F-measure.

Another way to transform bootstrap samples includes over-sampling the minority class before training classifiers. In this way the number of minority examples is increased in each bootstrap sample while the majority class is not reduced as in underbagging. This idea is realized with different over-sampling techniques. We present two approaches further used in experiments.

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

Another approach is used in *SMOTEBagging* to increase diversity of component classifiers. First, SMOTE is used instead of random over-sampling of the minority class. Then, SMOTE resampling rate (α) is stepwise changed in each iteration from small to high values (e.g,. from 10% to 100%). This ratio defines the number if minority examples ($\alpha \times N_{min}$) to be additionally re-sampled in each iteration. Quite similar trick is also used to construct bootstrap samples in "from underbagging to overbagging" ensemble.

3 Neighbourhood Balanced Bagging for Imbalanced Data

The proposed extension of bagging descends from results of studying sources of difficulties in learning from imbalanced classes. The high imbalance ratio between cardinalities of minority and majority class is not the only and not even the main reason of these difficulties. Other, as we call them, data factors, which characterize class distributions, are also influential. The experimental studies, as e.g. [9], demonstrate that the degradation of classification performance is linked to the decomposition of the minority class into many sub-parts containing very few examples. It means 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 (they correspond to the small disjuncts as they are harder to be learned and more contribute to classification errors than larger sub-concepts). Other factors related to the class distribution (occurring together with the class rarity) concern the effect of too strong overlapping between the classes [6] or presence of too many single minority examples inside the majority class regions [12].

We follow studies, as [11, 12], where the data factors are linked to different types of examples creating the minority class distribution. Authors differentiate 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*. The appropriate treatment of these types of minority examples within pre-processing methods should lead to improving learning classifiers, e.g., as it has been done by Stefanowski inside the informed pre-processing method SPIDER [14].

The question is how to identify these types of examples. In [13], it is achieved by analyzing the class distribution inside a *local neighbourhood* of the considered example, which is modeled by *k-nearest neighbour* examples. The distance between the examples is calculated according to the HVDM metric (*Heterogeneous Value Difference Metric*) [16]. Then, the number of neighbours from the opposite class indicates how safe or unsafe is the considered example (see [13] for details).

Inspired by the positive results of [14, 13], we will exploit characteristics of the local neighbourhood in a different quantitative way. The result is a new modification of bagging, which is called *Neighbourhood Balanced Bagging* (NBBag).

The idea behind NBBag is to focus sampling process toward these minority examples, which are hard to be learned (i.e. unsafe ones) while decreasing probabilities of selecting examples from the majority class at the same time. Recall that the idea of changing sampling probabilities has been considered in our previous work with applying bagging to noisy data and improving the overall accuracy [1]. Here, we postulate another strategy to change bootstrap samples. It is carried through a conjunction of sampling modifications at two levels: *global* and *local* ones.

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. We implement it by changing the probability of sampling of majority examples. More precisely, first we set p^1_{min} probability of sampling of each minority example to 1. Then, we downscale p^1_{maj} probability of sampling of a majority example to $\frac{N_{min}}{N_{maj}}$, where N_{min}, N_{maj} are numbers of examples in the minority and majority class 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 factors. Thus, this simple modification of probabilities exploits information about the global between-class imbalance. It should lead to bootstrap samples with approximately globally balanced class cardinalities.

However, the experimental studies [2, 5, 10] show that the global balancing in overbagging (somehow similar to our global level) is not competitive to other extensions of bagging. Moreover, most studied imbalance data sets contain many unsafe minority examples while the majority classes comprise rather safe ones, see results in [12]. However, while more focusing on the local characteristics of the minority class one should differently treat types of unsafe examples, as earlier successful experiments with such pre-processing methods as SPIDER [14] or generalizations of SMOTE as Borderline-SMOTE (see its description in [7]) pointed out that safe minority examples could be less over-sampled than borderline or other unsafe ones.

The second local level is intended to shift sampling of minority examples to these unsafe examples that are harder to be learned – what is identified by analyzing their k-nearest neighbours. This level can be modeled in different ways, having in mind the following rule: the more unsafe example, the more amplified probability of its drawing. This is partly inspired by earlier successful experiences with informed pre-processing methods. The modification rule could be done with either linear or non-linear function. In this study we use the formula L^2_{min}, defined as:

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

where N'_{maj} is the number of examples in the neighbourhood, which belong to the majority class; ψ is an exponential scaling factor, which in default case of a linear modification is set to 1. The value ψ may be increased if one wants to strengthen the role of rare cases and outliers in bootstraps. This increase may correspond to data sets where the minority class distribution in the original data is scattered into many rare cases or outliers, and the number of safe examples is significantly limited (see exemplary data, e.g. `balance-scale`, in the further experiments – section 4).

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 number "1" corresponds to a normalization of sampling probabilities inside the conjunctive combination, if one expects that $p_{min} \in [0, 1]$.

Then, we hypothesize that examples from majority class are, by default, not balanced on the second 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. Even when it is false for some data, it is still quite apparent that amplifying majority rare or outlying examples, at this level, would increase difficulties of learning classifiers from the minority classes interiors disrupted by them.

Finally, local and global levels are combined by a multiplication. This combination could correspond to the independence assumption, i.e. the distribution of examples in the neighbourhood is independent from the global distribution of examples in the whole data set. This leads us to the final formulations of the probability of selecting minority and majority classes, respectively as:

$$p_{min} = p^1_{min} \times \beta(L^2_{min} + 1) = p^1_{min} \times 0.5(L^2_{min} + 1) = 0.5(L^2_{min} + 1), \quad (3)$$

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

resulting from $L^2_{maj} = 0$, and default β set to 0.5.

4 Experiments

The first part experiments is an evaluation of the new proposed NBBag while the others concern using the local neighbourhood analysis to assess types of examples and studying the contents of bootstrap samples.

4.1 Evaluation of Bagging Extensions

First, we compare performance of NBBag with existing extensions of bagging. As a baseline for this comparison we use a balanced bagging (BBag), i.e., a variant which attempts to globally balance cardinalities of majority class and minority class in bootstrap samples (this is achieved by using only the first "global" level of NBBag of decreasing probability of the minority examples according to the imbalance ratio). Following our earlier study [2], we chose Rough Balanced Bagging (RBBag) as the best under-sampling extension. As our approach is more similar to over-sampling, we also consider: Overbagging (OverBag) and SMOTEBagging (SMOBag).

All implementations are done for the WEKA framework. Component classifiers in all bagging variants are learned with C4.5 tree learning algorithm (J4.8), which uses standard parameters except disabling pruning. For all bagging variants, we tested the following numbers T of component classifiers: 20, 50 and 100. Due to space limit, we present detailed results for $T = 50$ only. Results for

other T lead to similar general conclusions. In case of using SMOBag, we used 5 neighbours and the oversampling ratio α was stepwise changed in each sample starting from 10%. In NBBag we tested different sizes of the neighbourhood with $k = 5, 7, 9$ and 11. Their best values depend on a particular data set, however using 7 neighbours is the best in case of the linear amplification ($\psi = 1$). This option is further denoted as 7NBBag. As minority classes in some data sets are composed of mainly rare examples and outliers, we also considered the other variant of increasing a selection of examples with $\psi = 2$. For it the best results are obtained with a slightly smaller number of neighbours equal 5 – so it will be denoted as 5NBBag^2.

We conduct our analysis on 20 real-world data sets representing different domains, sizes and imbalance ratio. Most of data sets come from the UCI repository, and have been used in other works on class imbalance. Two data sets `abominal` 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. Their characteristics are presented in Table 1 where IR is the imbalance ratio defined as $\frac{N_{maj}}{N_{min}}$.

Table 1. Data characteristics

Data set	# examples	# attributes	Minority class	IR
abdominal_pain	723	13	positive	2.58
balance-scale	625	4	B	11.76
breast-cancer	286	9	recurrence-events	2.36
breast-w	699	9	malignant	1.90
car	1728	6	good	24.04
cleveland	303	13	3	7.66
cmc	1473	9	2	3.42
credit-g	1000	20	bad	2.33
ecoli	336	7	imU	8.60
flags	194	29	white	10.41
haberman	306	4	2	2.78
hepatitis	155	19	1	3.84
ionosphere	351	34	b	1.79
new-thyroid	215	5	2	5.14
postoperative	90	8	S	2.75
scrotal_pain	201	13	positive	2.41
solar-flareF	1066	12	F	23.79
transfusion	748	4	1	3.20
vehicle	846	18	van	3.25
yeast-ME2	1484	8	ME2	28.10

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" as-

signed to precision and recall). For their definitions see, e.g. [7]. These measures are estimated with the stratified 10-fold cross-validation repeated several 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 analyzed using either Friedman or Wilcoxon statistical tests (with a standard significance level 0.05). In all these tables the last row contains average ranks calculated as in the Friedman test – the lower average rank, the better classifier.

Table 2. G-mean [%] of compared bagging ensembles

Data set	BBag	SMOBag	OverBag	RBBag	7NBBag	5NBBag^2
abdominal_pain	79.04	80.85	79.44	79.99	80.26	80.82
balance-scale	19.74	0.00	1.40	58.12	47.36	61.07
breast-cancer	60.60	52.57	56.17	58.62	59.32	56.53
breast-w	96.11	95.88	96.23	96.13	96.21	96.14
car	96.21	95.26	95.29	97.09	96.80	96.98
cleveland	51.02	25.03	22.77	72.14	58.06	65.75
cmc	61.12	57.74	59.95	64.86	62.81	64.33
credit-g	65.87	80.68	71.75	86.89	66.48	66.94
ecoli	84.41	58.38	51.42	72.91	86.52	86.74
flags	61.16	62.48	64.30	65.84	61.46	61.46
haberman	61.22	60.02	58.11	65.92	62.24	48.65
hepatitis	74.92	68.47	72.16	80.23	78.19	75.33
ionosphere	90.88	90.30	90.47	90.25	90.76	89.95
new-thyroid	95.35	95.18	95.36	97.15	96.73	97.02
pima	74.51	72.33	73.54	74.59	74.18	72.30
scrotal_pain	73.41	70.42	72.01	74.17	72.29	71.42
solar-flareF	64.97	55.04	58.07	84.91	66.80	71.13
transfusion	67.33	63.96	64.83	67.39	64.98	39.56
vehicle	95.13	94.34	94.61	94.77	95.49	95.91
yeast-ME2	63.59	59.41	59.70	84.37	69.35	74.86
avg. rank	3.65	5.0	4.35	1.95	2.83	3.23

The results of Friedman tests (with CD = 1.69), reveal that, in both cases of G-mean and sensitivity, RBBag, 7NBBag, and 5NBBag^2 are significantly better than the rest of classifiers, without significant difference among them. Still, we can give some more detailed observations.

For G-mean, RBBag is the best classifier according to average ranks (see Table 2). It is also significantly better than several classifiers according to the Wilcoxon test - although the difference is not significant to 7NBBag. The worst classifier with respect to G-mean is SMOBag. Although it is a more complex approach using the informed SMOTE method, one can notice that the much simpler overbagging or BBag give better evaluation measures.

Analyzing the recognition of the minority examples, i.e. the sensitivity measure in Table 3, the best performing is 5NBBag^2 with respect to the average

Table 3. Sensitivity [%] of compared bagging ensembles

Data set	BBag	SMOBag	OverBag	RBBag	7NBBag	5NBBag^2
abdominal_pain	75.54	71.57	74.22	80.99	78.51	80.99
balance-scale	4.90	0.00	0.67	65.67	35.51	72.45
breast-cancer	54.12	34.35	44.91	59.81	59.88	66.71
breast-w	96.02	95.02	95.98	96.41	96.47	96.35
car	94.20	92.54	92.62	100.00	95.36	95.80
cleveland	29.14	17.22	16.11	77.22	39.71	54.57
cmc	50.93	40.05	46.47	66.80	57.12	66.61
credit-g	61.27	71.67	60.83	90.28	67.53	73.93
ecoli	77.14	55.00	66.67	78.33	82.00	84.29
flags	82.35	45.89	52.89	68.56	82.94	82.94
haberman	56.30	49.81	49.86	61.34	69.51	87.28
hepatitis	65.62	54.44	62.78	84.17	73.44	69.38
ionosphere	85.56	83.70	84.70	86.00	87.38	87.94
new-thyroid	92.57	92.22	93.06	97.50	95.43	96.00
pima	74.70	65.13	67.38	78.09	80.56	85.07
scrotal_pain	69.83	58.56	65.89	75.78	70.34	71.86
solar-flareF	47.44	37.33	42.17	87.33	50.70	58.84
transfusion	61.46	51.53	56.54	69.83	72.64	92.08
vehicle	94.77	92.14	93.46	96.48	95.63	96.48
yeast-ME2	41.57	39.11	39.11	91.56	49.80	59.22
avg. rank	4.05	5.78	5.08	1.95	2.52	1.62

ranks. However, according to the Wilcoxon test, its difference to RBBag is not significant. Again, the worst classifier in this comparison is SMOBag.

We do not show values of the F-measure, due to space limits. Nevertheless, these results indicate, similarly to the results for G-mean, that RBBag, 7NBBag, and 5NBBag^2 are better than other classifiers with no significant difference among them (e.g., the Wilcoxon test p value is 0.3 while comparing the pair of best classifiers RBBag and 7NBBag).

Looking more precisely at results in Tables 2 and 3 one can also notice that classifiers leading to high improvements of the sensitivity also strongly deteriorate G-mean at the same time (it means that the recognition of the majority class is much worse). For example see `transfusion` data set, which contains many outliers (see Table 4) and using $\psi = 2$ in the variant 5NBBag^2 leads the highest sensitivity 92.08% and the worst G-mean 39.56% among all compared classifiers. The similar trade off occurs also for, e.g., `haberman` data set and in the case of RBBag for, e.g., `car`. The linear amplification of the local probability of the minority class ($\psi = 1$) is the more conservative approach and it could be used if one wants to improve the sensitivity while still keeping the accuracy of majority classes at the sufficient level. However, tuning other intermediate ψ values between 1 and 2 could be the topic of further experiments.

Finally, notice that using the imbalance ratio to global balancing classes in bootstrap samples is not sufficient. Consider results of BBag which works sim-

ilarly to over-bagging. Taking into account information about the local neighbourhood of minority examples improves classification performance with respect to all evaluation measures. To conclude, the introduction of local modifications of sampling probabilities inside the combination rule of NBBag maybe the crucial element leading to the significantly better performance of the ensembles than all overbagging variants as well as for making it competitive to RBBag.

4.2 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 study class data characteristics in considered data sets and to identify types of examples (recall their distinction in section 3). Following the method introduced in [13] we propose to assign types of examples using information about class labels in their k-nearest local neighbourhood.

In this analysis we will use $k = 5$, because $k = 3$ may poorly distinguish the nature of examples, and $k = 7$ has led to quite similar decisions [13]. This choice is also similar to the size of neighbourhood used in NBBag. 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 analyzed 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 [13]: Proportions 5:0 or 4:1 inside the neighbourhood – the example x is labeled as a safe example (as it is surrounded by examples from the same class); 3:2 or 2:3 – it is a borderline example; 1:4 – it is interpreted as a rare case; 0:5 – it is an outlier. For higher values of k such proportions could be interpreted in a similar way.

The results of such labeling of the minority class examples are presented in Table 4. The first observation is that many data sets contain rather a small number of safe examples. The exceptions are three data sets composed of almost only safe examples: `breast-w`, `car`, and `flags`. On the other hand, there are data sets such as `cleveland`, `balance-scale` or `solar-flareF`, 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-ME2`: 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 [13], 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 2 - 3 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

Table 4. Labeling minority examples expressed as a percentage of each type of examples occurring in this class

Data set	Safe	Border	Rare	Outlier
abdominal_pain	61,39	23,76	6,93	7,92
balance-scale	0,00	0,00	8,16	91,84
breast-cancer	21,18	38,82	27,06	12,94
breast-w	91,29	7,88	0,00	0,83
car	47,83	47,83	0,00	4,35
cleveland	0,00	45,71	8,57	45,71
cmc	13,81	53,15	14,41	18,62
credit-g	15,67	61,33	12,33	10,67
ecoli	28,57	54,29	2,86	14,29
flags	100,00	0,00	0,00	0,00
haberman	4,94	61,73	18,52	14,81
hepatitis	18,75	62,50	6,25	12,50
ionosphere	44,44	30,95	11,90	12,70
new-thyroid	68,57	31,43	0,00	0,00
pima	29,85	56,34	5,22	8,58
scrotal_pain	50,85	33,90	10,17	5,08
solar-flareF	2,33	41,86	16,28	39,53
transfusion	18,54	47,19	11,24	23,03
vehicle	74,37	24,62	0,00	1,01
yeast-ME2	5,88	47,06	7,84	39,22

examples. Furthermore, the best improvements of all evaluation measures for RBBag or NBBag are observed for the unsafe data sets. For instance, consider `cleveland` (no safe examples, nearly 50% of outliers) where RBBag has 72% G-mean comparing to overbagging with 22.7%. Similar highest improvements occur for `balance-scale` (containing the highest number of outliers among all data sets) where NBBag gets 61.07% while OverBag 1.4%. Similar situations also occur for `yeast-ME2`, `ecoli`, `haberman` or `solar-flare`. 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 these improvements achieved by RBBag and NBBag, we perform the same neighbourhood analysis and labeling types of minority examples inside their bootstraps. For each bootstrap sample we label types of minority examples basing on class labels of the k-nearest neighbours. Then, we average results from all bootstraps. The results of this labeling are presented in two rows of Table 5, referring to each classifier. We present result for 7NBBag only, due to space limits and skip some safe data, where there is no big changes of distributions between both variants of NNBag.

In our opinion these results reveal very interesting properties of both ensembles. While comparing Tables 4 and 5 notice that RBBag and NBBag strongly change types of the minority class distributions into safer ones inside their bootstraps. For many data sets which originally contain high numbers of rare cases

Table 5. Distributions of types of the minority examples [in %] inside bootstrap samples for each classifier and data set

Data set	Classifier	Safe	Border	Rare	Outlier
abdominal_pain	NBBag	65.14	21.5	7.41	5.96
	RBBag	72.60	19.15	5.11	3.15
balance-scale	NBBag	59.23	20.52	5.63	14.62
	RBBag	39.68	59.02	0.05	1.25
breast-cancer	NBBag	37.55	43.51	11.39	7.54
	RBBag	35.56	52.82	7.52	4.10
breast-w	NBBag	93.44	6.04	0.40	0.12
	RBBag	93.57	5.60	0.29	0.54
cleveland	NBBag	64.58	17.59	7.07	10.76
	RBBag	42.86	53.33	0.44	3.37
cmc	NBBag	38.33	41.14	10.74	9.79
	RBBag	42.47	50.96	3.13	3.44
credit-g	NBBag	36.07	49.21	7.47	7.26
	RBBag	34.44	59.58	2.79	3.19
ecoli	NBBag	81.61	7.76	3.96	6.67
	RBBag	85.33	11.75	0.00	2.92
flags	NBBag	100.00	0.00	0.00	0.00
	RBBag	100.00	0.00	0.00	0.00
haberman	NBBag	33.37	46.82	9.63	10.19
	RBBag	25.34	66.09	4.07	4.50
hepatitis	NBBag	65.89	23.78	4.38	5.95
	RBBag	67.01	26.60	1.25	5.14
ionosphere	NBBag	94.96	3.62	0.49	0.93
	RBBag	51.98	31.10	6.60	10.32
new-thyroid	BBag	96.54	2.41	0.15	0.90
	RBBag	90.83	9.17	0.00	0.00
scrotal_pain	NBBag	62.92	25.24	6.57	5.27
	RBBag	64.67	29.34	3.50	2.49
solar-flareF	NBBag	84.83	7.67	3.76	3.74
	RBBag	70.52	21.37	2.69	5.43
transfusion	NBBag	35.71	46.43	9.44	8.42
	RBBag	41.00	41.90	3.76	13.33
vehicle	NBBag	86.84	10.4	1.46	1.30
	RBBag	89.80	10.20	0.00	0.00
yeast-ME2	NBBag	86.66	7.51	1.79	4.05
	RBBag	64.31	34.29	0.22	1.18

or outliers, the transformed bootstrap samples contain now more safe examples. For instance, consider the very difficult `balance-scale` data set (containing originally 91.8% outliers), where RBBag creates bootstrap samples with at most 4% outliers and 7.5% rare cases while moving the rest of examples into safe and borderline ones. Similar data type shift could be observed for: `yeast-ME2` (originally 5% safe examples, now over 70%), `solar-flareF`, `ecoli`, `ionosphere`, `hepatitis`, `cleveland`. Finally, one can notice that RBBag usually constructs slightly safer data than NBBag.

Recall that literature known extensions of bagging are based on the simple idea of balancing distributions in bootstrap samples. However, 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 NBBag, which uses larger bootstrap samples.

5 Discussion and Final Remarks

The difficulty of learning classifiers from imbalanced data comes from complex distributions of the minority class. Besides the unequal class cardinalities, the minority class is decomposed into smaller sub-parts, affected by strong overlapping, rare cases or outliers. In our study we attempt to capture these data characteristics by analyzing the local neighbourhood of minority class examples. Our main message it to show that such kind of local information could be useful both for proposing a new type of bagging and to explain why some ensembles work better than others.

Our first contribution includes the introduction of the Nearest Balanced Bagging which is based on different principles than all known bagging extensions for class imbalances. First, instead of integrating bagging with pre-processing, we keep the standard bagging idea but we change radically probabilities of sampling examples by increasing the chance of drawing more difficult minority examples. Furthermore, we promote to amplify the role of difficult examples with respect to their local neighbourhood. The experimental results show that this proposal is significantly better than existing over-sampling generalizations of bagging and

it is competitive to Roughly Balanced Bagging (being the best known undersampling variant).

The other contribution is to use the local neighbourhood analysis to assess the type of examples in data. It comes from the earlier research of Stefanowski and Napierala [13], however it is now applied in the context of ensembles, which uncover new characteristics of studied ensembles.

First, the strongest differences between classifiers have been noticed for data sets containing the most unsafe minority examples. Indeed, both RBBag and NBBag ensembles have strongly outperformed all overbagging variants for such data. Furthermore, the analysis of types of minority examples inside bootstrap samples has clearly showed that RBBag and NBBag strongly changed data characteristics comparing to the original data sets. Many examples from the minority class labeled as unsafe (in particular as rare cases or outliers) have been transformed to more safe ones. This might be more influential for improving the classification performance than the simple global class balancing, which was previously considered in the literature and applied to many of existing approaches to generalize bagging.

References

1. Błaszczyński, J., Słowiński, R., Stefanowski, J.: Feature Set-based Consistency Sampling in Bagging Ensembles. Proc. From Local Patterns To Global Models (LEGO), ECML/PKDD Workshop, 19–35 (2009)
2. Błaszczyński, J., Stefanowski, J., Idkowiak L.: Extending bagging for imbalanced data. Proc. of the 8th CORES 2013, Springer Series on Advances in Intelligent Systems and Computing 226, 269–278 (2013)
3. Breiman, L.: Bagging predictors. Machine Learning, 24 (2), 123–140 (1996)
4. Chawla, N., Bowyer, K., Hall, L., Kegelmeyer, W.: SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artifical Intelligence Research, 16, 341-378 (2002)
5. Galar, M., Fernandez, A., Barrenechea, E., Bustince, H. Herrera, F.: A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches. IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, 99, 1–22 (2011)
6. Garcia, V., Sanchez, J.S., Mollineda, R.A.: An empirical study of the behaviour of classifiers on imbalanced and overlapped data sets. In: Proc. of Progress in Pattern Recognition, Image Analysis and Applications 2007, Springer, LNCS, vol. 4756, 397–406 (2007)
7. He H., Garcia E.: Learning from imbalanced data. IEEE Transactions on Data and Knowledge Engineering, 21 (9), 1263–1284 (2009)
8. Hido S., Kashima H.: Roughly balanced bagging for imbalance data. Statistical Analysis and Data Mining, 2 (5-6), 412–426 (2009)
9. Jo, T., Japkowicz, N.: Class Imbalances versus small disjuncts. ACM SIGKDD Explorations Newsletter, 6 (1), 40–49 (2004)
10. Khoshgoftaar T., Van Hulse J., Napolitano A.: Comparing boosting and bagging techniques with noisy and imbalanced data. IEEE Transactions on Systems, Man, and Cybernetics–Part A, 41 (3), 552–568 (2011)

11. Kubat, M., Matwin, S.: Addresing the curse of imbalanced training sets: one-side selection. In Proc. of Int. Conf. on Machine Learning ICML 97, 179–186 (1997)
12. Napierala, K., Stefanowski, J., Wilk, Sz.: Learning from Imbalanced Data in Presence of Noisy and Borderline Examples. In: Proc. of 7th Int. Conf. RSCTC 2010, Springer, LNAI vol. 6086, pp. 158–167 (2010)
13. Napierała, K., Stefanowski, J.: The influence of minority class distribution on learning from imbalance data. In. Proc. 7th Int. Conference HAIS 2012, Part II, LNAI vol. 7209, Springer, pp. 139-150 (2012)
14. Stefanowski, J., Wilk, Sz.: Selective Pre-processing of Imbalanced Data for Improving Classification Performance. In: Proc. of 10th Int. Conference DaWaK 2008, Springer Verlag, LNCS vol. 5182, 283-292 (2008)
15. Wang, S., Yao, T.: Diversity analysis on imbalanced data sets by using ensemble models. In Proc. IEEE Symp. Comput. Intell. Data Mining, 324-331 (2009).
16. Wilson, D.R., Martinez, T.R.: Improved heterogeneous distance functions. Journal of Artifical Intelligence Research, 6, 1-34 (1997)