J Intell Inf Syst (2018) 50:97–127
DOI 10.1007/s10844-017-0446-7

Multi-class and feature selection extensions of Roughly Balanced Bagging for imbalanced data

Mateusz Lango[1] · **Jerzy Stefanowski**[1]

Received: 23 June 2016 / Revised: 7 November 2016 / Accepted: 13 January 2017 /
Published online: 10 February 2017

Abstract Roughly Balanced Bagging is one of the most efficient ensembles specialized for class imbalanced data. In this paper, we study its basic properties that may influence its good classification performance. We experimentally analyze them with respect to bootstrap construction, deciding on the number of component classifiers, their diversity, and ability to deal with the most difficult types of the minority examples. Then, we introduce two generalizations of this ensemble for dealing with a higher number of attributes and for adapting it to handle multiple minority classes. Experiments with synthetic and real life data confirm usefulness of both proposals.

Keywords Class imbalance · Roughly balanced bagging · Types of minority examples · Feature selection · Multiple imbalanced classes

1 Introduction

Many real–life problems involve learning classifiers from *imbalanced data*, where one of the classes (further called a *minority class*) includes much smaller number of examples than the other *majority classes*. For instance, in medical problems the number of patients requiring special attention is usually much smaller than the number of patients who do not need it. The correct recognition of the minority class is of key importance in these problems. Similar challenges with imbalanced classes have been also observed in many other application domains such as fraud detection in telephone calls or credit cards transactions, bank risk analysis, technical diagnostics, network intrusion detection, recognition of oil spills in

✉ Mateusz Lango
mateusz.lango@cs.put.poznan.pl

Jerzy Stefanowski
jerzy.stefanowski@cs.put.poznan.pl

[1] Institute of Computing Science, Poznań University of Technology, 60-965 Poznań, Poland

images, detecting specific astronomical objects in sky surveys, text categorization, information filtering; for some reviews see, e.g., (Chawla 2005; Fernandez et al. 2011; He and Garcia 2009; He and Ma 2013; Weiss 2004).

The standard learning algorithms usually do not work properly on imbalanced data as they are biased toward better recognition of the majority classes and they have difficulties (or even they are unable) to classify correctly new objects from the minority class.

Learning from imbalanced data has received a growing research interest in the last decade and several specialized methods have already been proposed. For their review see, e.g., (Branco et al. 2016; He and Garcia 2009; He and Ma 2013; Krawczyk 2016). These methods are usually categorized in *data level* and *algorithm level* approaches. The first category includes classifier-independent methods that rely on transforming the original data to change the distribution into a more appropriate one (usually being more balanced). The most simple representatives of these preprocessing approaches include *random undersampling* (i.e., removing some examples from the majority classes) or *oversampling* (i.e., adding minority class examples). As they may not be sufficiently effective, more sophisticated *informed pre-processing methods* that use the characteristics of the *local data distribution*, such as one-side sampling (Kubat and Matwin 1997), NCR (Laurikkala 2001), SMOTE (Chawla et al. 2002) or SPIDER (Stefanowski and Wilk 2008), are often considered.

The other category of the specialized approaches involves modifications of either a learning phase of the algorithm, classification strategies, optimizing appropriate evaluation criteria inside the algorithm, construction of specialized ensembles or adaptation of cost-sensitive learning. New types of ensemble classifiers are also visible among these methods; see their review in Galar et al. (2011), Liu and Zhu (2013). Most of them are modifications of bagging or boosting schemes known from the typical approaches to improve the total predictive accuracy (Kuncheva 2014). These modifications usually employ either pre-processing methods before learning component classifiers or embed the cost-sensitive framework in the learning process.

The number of comprehensive comparative studies of the new ensembles dedicated to imbalanced data is still quite limited. However, experimental comparisons (Galar et al. 2011; Khoshgoftaar et al. 2011) have shown that extensions of bagging work better than generalizations of boosting and other, more complex solutions. Moreover, other experiments from Błaszczyński et al. (2013) have shown that undersampling modifications of bagging are significantly better than oversampling alternatives. Our recent experimental studies (Błaszczyński et al. 2013; Błaszczyński and Stefanowski 2015; Stefanowski 2016b) have also demonstrated that the undersampling bagging called Roughly Balanced Bagging (RBBag) (Hido and Kashima 2009) has achieved the best prediction results comparing to other extensions of bagging as well as to modified boosting ensembles (Khoshgoftaar et al. 2011).

The key idea behind Roughly Balanced Bagging is a random undersampling before generating component classifiers, which reduces the presence of the majority class examples inside each bootstrap sample. Although this ensemble has been successfully used in several studies, there are not enough attempts to check which of its properties are the most crucial for improving classification of complex imbalanced data. In our opinion, they should be examined more precisely.

Therefore, the first aim of this paper is to experimentally study the following aspects of constructing Roughly Balanced Bagging:

- Studying the influence of choosing algorithms for learning component classifiers.
- Deciding on the number of component classifiers in this ensemble.

– Examine diversity of the component classifier predictions and its relation to the final accuracy of the ensemble.
– Abilities of the ensemble to deal with data difficulty factors – which will be modeled by distinguishing the different types of minority classes following the methodology from Napierala and Stefanowski (2012).

To the best of our knowledge, such a detailed analysis of Roughly Balanced Bagging has not been carried out yet. Here, we significantly extend our previous conference paper (Lango and Stefanowski 2015). On the other hand, one could still ask research questions with respect to other directions for further extensions and improvements of this ensemble. In this paper, we consider two kinds of problem: dealing with a higher number of attributes and adapting this ensemble to handle multiple minority classes.

Imbalanced datasets characterized by relatively many attributes often occur in image recognition, fraud detection, and genetic data analysis (Pio et al. 2014) and also need more specialized approaches. We will introduce an extension of Roughly Balanced Bagging – which integrates its specific bootstrap sampling with a random selection of attributes.

The other extension concerns imbalanced multiple classes as it is a more complex task than standard binary imbalanced problems. Although considering single minority class versus the majority class (which could also result from aggregating other classes) is often justified in several domains, sometimes it is necessary to distinguish additional classes with low cardinality. Dealing with several minority classes is a more complicated scenario and approaches for binary imbalanced data are not applicable to it. In particular, it concerns Roughly Balanced Bagging where the modification of bootstrap sampling is defined with binary probability distribution only. The recent research on using ensembles over such multi-class problems usually include various decompositions of the original data into binary ones (Fernandez et al. 2013). In this paper, we have decided to follow a quite different perspective where all classes are handled simultaneously.

The rest of the paper is organized as follows. In the next section, we recall the bagging scheme, discuss the basics of Roughly Balanced Bagging and show some other related works. In Section 3, we experimentally study basic properties of constructing Roughly Balanced Bagging. Then, in Section 4, we introduce an attribute selection generalization of this ensemble and evaluate its usefulness. In the following Section 5, two kinds of its extension for dealing with multiple minority classes are put forward and compared in experiments with special synthetic and real datasets. In Section 6 we draw conclusions and discuss lines of future research.

2 Related works

2.1 Preliminaries

Here, we do not intend to provide a comprehensive review of methods for dealing with class imbalances and we will briefly present the selected ensemble methods only. For more details, the reader is referred to a recently published monograph (He and Ma 2013) covering the most representative issues and to the earlier systematic surveys, such as Chawla (2005), He and Garcia (2009), Sun et al. (2009). A recent, comprehensive review of pre-processing methods could be found in Branco et al. (2016) and their comparative studies are provided by Napierala and Stefanowski (2016), Van Hulse et al. (2007).

Another issue concerns the nature of imbalance data which poses challenges for learning accurate classifiers. Although the global ratio between cardinalities of both minority and majority classes (called an *imbalance ratio*) is a main characteristic of the imbalanced data, it may not sufficiently explain differences between classification performance of various methods. Some researchers have already shown that the global imbalance ratio is not a problem itself and it may not be the main source of difficulties (Japkowicz 2003). The degradation of classification performance is linked to other factors related to data distribution, such as the decomposition of the minority class into many rare sub-concepts playing a role of small disjuncts (Jo and Japkowicz 2004), the effect of too strong overlapping between the classes (Garcia et al. 2007) or a presence of too many minority examples inside the majority class regions (Napierala and Stefanowski 2012). It has been shown that when these *data difficulty factors* occur *together* with class imbalance, they seriously hinder the recognition of the minority class (Lopez et al. 2014; Napierala et al. 2010; Napierala and Stefanowski 2012; Stefanowski 2013, 2016a). In the experimental analysis of Roughly Balanced Bagging (see Section 3) we will refer to some difficulty factors by analysing types of unsafe examples in the distribution of the minority class following the methodology presented in Napierala and Stefanowski (2012, 2016).

Most of the research on imbalance concerns binary (two-class) problems. It is justified by a semantic importance of the rare class versus other classes. The multi-class formulation of the imbalanced problem will be discussed in Section 5.

2.2 Ensembles dedicated to imbalanced data

Several ensembles dedicated to class imbalance have been proposed in the recent decades. Their most comprehensive surveys are provided in Galar et al. (2011), Liu and Zhu (2013). The authors of these papers categorized these ensembles slightly differently.

The taxonomy proposed by Galar et al. in (2011) distinguishes between *cost-sensitive* approaches vs. integrations with *data pre-processing*. The first group covers mainly cost-minimizing techniques combined with boosting ensemble, e.g., AdaCost, AdaC or RareBoost. The second group of approaches is divided into three sub-categories: Boosting-based, Bagging-based and Hybrid. A classifier ensemble is assigned to them depending on the type of classical ensemble technique which is integrated into the schema for learning component classifiers and their aggregation. In their view, most of these proposals integrate some pre-processing techniques. For instance, the majority of Bagging-based ensembles apply a kind of random undersampling, or oversampling to change class distribution inside the bootstrap sampling. Few authors refer to the use of the SMOTE method (Chawla et al. 2002) or playing with different oversampling ratios in each bootstrap to increase ensemble diversity (Wang and Yao 2009). Similar modifications of examples are proposed inside each iteration of AdaBoost, see SMOTEBoost, RUSBoost or DataBoost. For these and other ensembles descriptions please refer, e.g., to Galar et al. (2011). The authors of (Galar et al. 2011) also distinguish approaches which exploit kinds of re-sampling in more untypical adaptive ensembles, like IIVotes (Błaszczyński et al. 2010). Then, different combinations of boosting and bagging extensions into one complex ensemble, such as EasyEnsemble or BalancedCasade, are assigned to the hybrid generalizations.

Liu and Zhu (2013) also categorizes the ensembles for class imbalance into bagging-like, boosting-based methods and hybrid ensembles, depending on their relation to standard approaches. All other proposals, like cost embedding, are put in a simple category "others". Note, that the authors of Liu and Zhu (2013) have paid more attention to extensions of

Random Forest to imbalanced data and showed a good performance of Balanced Random Forest (Chen et al. 2004) in their experiments.

Although these ensembles are promoted as a remedy to imbalanced data, 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 datasets. Up to now, only two comprehensive studies were carried out in different experimental frameworks (Galar et al. 2011; Khoshgoftaar et al. 2011). The first study (Galar et al. 2011) 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 undersampling or SMOTE re-sampling combined with bagging work better than more complex solutions. In the study (Khoshgoftaar et al. 2011), two best boosting and bagging ensembles are compared over noisy and imbalanced data. The experimental results show that bagging significantly outperforms boosting and the difference is more significant when the data is noisier. The similar observations on the good performance of undersampling generalizations of bagging vs. cost like generalizations of boosting have been recently reported in Anyfantis et al. (2008). Furthermore, the most recent chapter of Liu and Zhu (2013) includes another 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.

There are also two theoretical papers on properties of random re-sampling with respects to probability distributions. The research of Wallace et al. (2011) provides a probabilistic theory of imbalance and its reference to undersampling ensembles. Another recent work (Dal Pozzolo et al. 2015) proposes a theoretical analysis specifying under which conditions undersampling could be effective in pre-processing for a single classifier.

Following these experimental and other motivations, we have decided to consider the under-bagging extensions and Roughly Balanced Bagging as a basis for our study. Below we briefly describe them.

2.3 Bagging and its re-sampling generalizations

The *Bagging* approach (its name is shortening from **B**ootstrap **agg**regat**ing**) was introduced by Breiman (1996). It aggregates by voting classifiers generated from different bootstrap samples. Its main element is adopting bootstrap sampling to inject some random perturbation into parallel training sets that allows learning more diverse component classifiers in this ensemble. The *bootstrap sample* is obtained by uniformly sampling with replacement examples from the original training dataset. Each sample has usually the same size as the original set, however, some examples do not appear in it, while others may appear more than once. For a training set with N examples, the probability of an example being selected at least once is $1-(1-1/N)^N$. For a large N, this is about $1-1/e$. Each bootstrap sample contains, on the average, 63.2% unique examples from the training set.

The generic schema of bagging is presented in Algorithm 1. Given a parameter k, which is a number of component classifiers, the algorithm draws with replacement k bootstrap samples from training data D of size N. Then, the same learning algorithm LA is applied to generate component classifiers C_i, which are aggregated to create the final ensemble C^*. The classification decision of the ensemble is a result of the simple majority voting (see (1)) – a new instance is assigned to a class predicted by the most of the component classifiers. In the case of using classifiers with probabilistic outputs, this prediction formula aggregates probabilities of classes from component classifiers and may use different operators as sum,

median or product of probabilities. The reader is referred to such books as Kuncheva (2014) for more information on the details of bagging, an explanation why it works and its popular generalizations.

Algorithm 1 Standard bagging algorithm

Input: D: original training set of examples of size N, k: number of bootstrap samples, LA: learning algorithm;
Output: C^* bagging ensemble with k component classifiers
Learning phase:

1: **for** $i = 1 \rightarrow k$ **do**
2: $S_i \leftarrow$ bootstrap sample from D;
3: generate classifier $C_i \leftarrow LA(S_i)$
4: **end for**
Predicting class label for new instance x:
5:

$$C^*(x) = \arg\max_y \sum_{i=1}^{k} [C_i(x) = y] \tag{1}$$

Considering its applicability to imbalanced data, note that bootstrap sampling is performed on all data elements, regardless their class labels (majority or minority). Therefore, the imbalanced class distribution will be hold in each bootstrap and the ensemble will fail to sufficiently classify the minority class.

Most of the current proposals overcome this drawback by applying pre-processing techniques to each bootstrap sample, which change the balance between classes – usually leading to the same, or similar, cardinality of the minority and majority classes in each sample.

For instance, the oversampling methods typically replicate the minority class data (either by random sampling or by generating synthetic examples) to balance classes in bootstraps. In this way, the number of minority examples is increased (e.g., by a random replication), while the majority class is not reduced. This idea was realized in many ways as authors considered integrations with different oversampling techniques. *OverBagging* is the simplest version which applies a simple random oversampling to transform each training bootstrap sample S_i. The number of N_i^{maj} minority class examples is sampled with replacement to exactly balance the cardinality of the minority and the majority class in each bootstrap sample. Another approach is used in *SMOTEBagging* to increase the diversity of component classifiers (Wang and Yao 2009). First, SMOTE is used instead of the random oversampling of the minority class. Then, SMOTE resampling rate (α) is stepwise changed in each iteration from smaller to higher values. Quite a similar way of varying ratio α to construct bootstrap samples is also used in "from underbagging to overbagging" ensemble mentioned in Wang and Yao (2009).

In *under-bagging* the number of the majority class examples in each bootstrap is reduced to the cardinality of the minority class (N_{min}) in the original training set. In the simplest proposals, as *Exactly Balanced Bagging* (Chang 2003), the entire minority class is just copied to the bootstrap and then combined with the randomly chosen subset of the majority class to exactly balance the cardinality between classes.

Other variations of under-bagging have been also proposed. For instance, the method proposed in Chan and Stolfo (1998) partitions the majority class into a set of non-overlapping subsets, with each subset having approximately N_{min} examples. Then, each of these

majority subsets and all examples from the minority class forms a bootstrap sample 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 (Liu and Zhu 2013). The other option is to construct *Balanced Random Forests* as an extension of classical Random Forests (Chen et al. 2004). 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 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). An interesting extension of Random Forests for massive and imbalanced data, which is paralelly implemented in MapReduce and Hadoop frameworks, has been studied in Rio et al. (2014).

Yet another approach has been considered in *Neighbourhood Balanced Bagging*, where sampling probabilities of examples to the bootstraps are modified according to the class distribution in their neighbourhood (Błaszczyński and Stefanowski 2015). It shifts the sampling toward the examples located in the most difficult sub-regions of the minority class (identified with the safe level of examples (Napierala and Stefanowski 2012)).

This chapter will not discuss all such extensions, as we focus on Roughly Balanced Bagging which according to many experimental studies is the most accurate at imbalanced datasets.

2.4 Roughly balanced bagging

While such under-bagging strategies seem to be intuitive and work efficiently in some studies, Hido et al. (2009) have claimed that they do not truly reflect the philosophy of bagging and could be still improved. In the original bagging ensemble, the class distribution of each sampled subset varies according to the binomial distribution while in the aforementioned under-bagging strategies each subset has the same class ratio as the desired balanced distribution.

Hido et al. have introduced the *Roughly Balanced Bagging* (RBBag) (Hido and Kashima 2009), where the numbers of instances for both classes are determined in a different way by equalizing the sampling probability for each class. The number of minority examples (N_i^{min}) in each bootstrap S_i is set to the size of the minority class N_{min} in the original data D. On the contrary, the number of majority examples is decided probabilistically according to the negative binomial distribution. This distribution[1] is a probability distribution of the number of m failures given the number of n successes in the sequence of Bernoulli trials. It is defined by the following probability mass function:

$$p(m|n) = \binom{m+n-1}{n} p^n q^m \tag{2}$$

where p is the probability of success and $q = 1 - p$ is the probability of failure. For the imbalanced data, these authors set both probabilities p and q to 0.5. After fixing the number of minority examples to N_{min} and setting the probability of success equal to 0.5, they used this distribution to find the number of the majority examples for each bootstrap. Note that the size of the majority examples (N_i^{maj}) varies over the bootstraps in the ensemble, however its average value is N_{min} (Hido and Kashima 2009).

[1] The negative binomial distribution with an integer parameter n is also called Pascal distribution.

The other elements of constructing RBBag ensemble are the same as in the earlier under-bagging extensions, i.e. component classifiers are induced by the same learning algorithm from each i-th bootstrap sample and their predictions form the final decision with the equal weight majority voting - although (Hido and Kashima 2009) promotes using probability outputs of component classifiers. Algorithm 2 presents the pseudocode of the RBBag algorithm.

Algorithm 2 Roughly Balanced Bagging

Input: $D = D_{min} \cup D_{maj}$: original training set of examples of size N, k: number of bootstrap samples, LA: learning algorithm;
Output: C^* bagging ensemble with k component classifiers
Learning phase:

1: **for** $i = 1 \rightarrow k$ **do**
2: $N_i^{min} \leftarrow |D_{min}|$
3: $N_i^{maj} \leftarrow$ following negative binomial distribution with $n = N_i^{min}$ and $p = q = 0.5$
4: $S_i^{min} \leftarrow N_i^{min}$-element sample drawn with replacement from D_{min}
5: $S_i^{maj} \leftarrow N_i^{maj}$-element sample drawn with replacement from D_{maj}
6: $C_i \leftarrow LA(S_i^{min} \cup S_i^{maj})$
7: **end for**

Prediction phase:

$$C^*(x) = \arg\max_y \sum_{i=1}^{k} p_{C_i}(y|x)$$

Hido et al. compared RBBag with several classifiers showing that it was better on G-mean and AUC measures (Hido and Kashima 2009). Another study (Khoshgoftaar et al. 2011) demonstrated that under-bagging ensembles, including RBBag, significantly outperformed best extensions of boosting and the difference was even more significant when data were noisier. Then, the comparative experiments from Błaszczyński et al. (2013) showed that under-bagging ensembles, as Exactly and Roughly Balanced bagging, were significantly better than several main oversampling extensions of bagging (either using random oversampling or SMOTE) with respect to all evaluated measures. Roughly Balanced Bagging was also slightly better than Exactly Balanced one. This is why we have chosen RBBag as the best performing specialized ensemble for this paper. Recall the introductory motivation that there have not been so many attempts to experimentally examine properties of this ensemble or to more theoretically explain why and when it should outperform other methods.

2.5 Feature ensembles and class imbalance

Most of the current research on imbalanced data concerns problems with a relatively small or medium number of attributes. The proposed methods often do not work sufficiently well with a higher number of attributes. For instance, popular informed pre-processing methods, such as SMOTE (Chawla et al. 2002), NCR (Laurikkala 2001) or SPIDER (Stefanowski and Wilk 2008), intensively exploit calculations of distances between learning examples. As they use typical metrics – usually being variants of Euclidean or HVDM distances (Wilson

and Martinez 1997), they meet difficulties for problems with a higher number of attributes. The dimensionality curse also concerns the methods which modify algorithms. On the other hand, several practical classification problems in domains, such as text categorization, image analysis, medical data analysis or genetics, are characterized by many attributes.

Recall that this problem is also a challenge for standard classifier learning as it increases risks of overfitting as well as spurious findings. However, considering it with class-imbalance presents an additional source of difficulties for prediction, as it biases classification towards majority class for most classifiers (see, e.g. experimental analyses from Blagus and Lusa (2010)). The attribute (feature) selection is often applied in standard balanced classification to enhance predictive performance. Although these selection methods have been extensively studied, see surveys as Tang et al. (2014), many popular filtering methods are too biased toward majority class. Thus, some new class imbalance techniques have been recently introduced, see e.g. FAST (Chen and Wasikowski 2008), or others surveyed in Pant and Srivastava (2015).

This dimensionality challenge can be solved in another way in ensemble classifiers. Note that the motivation for feature subset selection is slightly different as it is often used as an additional mechanism for introducing the *diversity of component classifiers*. According to it, the learning sets for creating the ensemble could be obtained by using different subsets of attributes for each of them. Improving global accuracy and diversity of the ensemble is also known under the name *ensemble feature selection*; see Kuncheva (2014) for a review of these approaches. One of the most well-known approaches is *Random Subspace Method* (RSM) (Ho 1998), where in each iteration of constructing the ensemble a subset of all available attributes is randomly drawn and a component classifier is built using only this subset. This method has been further generalized where firstly learning examples are bootstrap sampled, like in the standard bagging, then the random attribute selection is done in each of these bootstraps, see Latinne et al. (2000). Recall that Breiman combined bootstrap sampling with random selection of attributes in nodes of trees inside the Random Forest ensemble. Nowadays, Random Forest seems to be often used in many highly dimensional practical problems. In particular, see review of various its modifications for bio-medical problems (Draminski et al. 2016).

However, there are not so many proposals of new feature selection ensembles specialized for imbalanced datasets with many attributes, see e.g. a review in Lin and Chen (2013). Typically, only *Balanced Random Forest* (Chen et al. 2004) is considered. Its adaptation mainly includes undersampling of the majority class to exactly balance class cardinalities in each bootstrap. Then, attributes are randomly selected in each tree node following ideas from the original Random Forest. Recent experiments of Liu and Zhu (2013) demonstrated that Balanced Random Forests makes it competitive to other good generalizations of ensemble - however these experiments do not concern datasets with too many attributes. On the other hand, Roughly Balanced Bagging, which is not dependent on a particular tree induction, has not been considered and generalized yet to this challenging data characteristics. Therefore, it motivates our research in the next sections of this paper.

2.6 Multiple imbalanced classes

A binary classification task is mostly studied in the case of imbalanced data. This formulation is justified by focus an interest in the most important class and its real-life semantics, like in medical diagnosis (distinguishing sick vs. healthy patients), spam detection (e.g.,

valid activity vs. malicious one), image recognition (target object vs. background), etc. Even if the dataset includes more majority classes, then they are aggregated into one global majority class as in most applications it is necessary to improve the recognition of the minority class – for yet stronger its justification see He and Ma (2013), Krawczyk (2016), Sun et al. (2009), Weiss (2004). A straightforward discrimination between the minority class and the majority one have led to a development of several methods that take into account only this relation, such as random re-sampling. It is also the basis for balanced bootstrap samples in the current generalizations of bagging, such as Roughly Balanced Bagging.

On the other hand, in some situations it may be reasonable to distinguish more classes with low cardinalities. Consider for instance, technical diagnostics where the experts distinguish an intermediate status of the working machine besides considering bad (damaged) and good technical status. Similar situations may occur in some medical problems. Usually, this intermediate class contains fewer examples than the majority class. As it is also quite difficult to recognize examples of this class, there are needs for improving its classification. However, the classifier will fail to do it when this class is aggregated with either minority or majority ones.

Considering multiple minority classes makes the learning task more difficult as relations between particular classes become more complex (Wang and Yao 2012). Internal data distributions or decision boundaries will be different than in the case when some classes are aggregated. Techniques developed for binary imbalanced problems are usually not directly applicable to multi-class problems. Quite often they lose performance on one class while trying to gain it on another. Therefore, more specialized techniques have been recently proposed; Their review is available in Seaz et al. (2016).

In this paper we are more interested in ensemble based methods for multi-class imbalanced data. Nearly all of them adapt solutions already introduced in non-imbalanced ensembles to deal with many classes, see the review of ECOC, pairwise-coupling and other methods in Kuncheva (2014). Practically, the analysis of Wang and Yao (2012) is the only exception.

The decomposition of the multi-class imbalanced dataset to a set of binary problems is a dominating strategy so far, see a review in (Krawczyk 2016). Researchers consider either one-vs-all classes or one-vs-one decompositions. These techniques pre-process the binary decomposed datasets (usually by means of balancing classes with various known re-sampling methods), apply identical learning algorithms to learn a set of component classifiers, then aggregate their predictions using known combination rules. For a representative family of such decomposition-based ensembles and their experimental evaluation refer to Fernandez et al. (2013). Although these decomposition approaches are quite easy to implement and some experimental results seem to be promising, this problem still requires new solutions.

In particular, we share the opinions expressed in the position paper (Krawczyk 2016). According to it while decomposing the multiple imbalanced classes, pairwise relations between two classes only may be a too strong over-simplification and they do not reflect more complex relations between several of classes, as one class influences several neighboring classes at the same time. The researchers may risk a loss of somehow balanced improved performance on all minority classes and reject a more global view of data nature with respect to all classes.

In this paper we follow such a critical perspective and will consider new generalizations of Roughly Balanced Bagging that will deal with changing bootstraps with respect to recognizing all multiple classes.

3 Studying properties of Roughly Balanced Bagging

3.1 Experimental setup

In this section we carry our comprehensive experiments, where we study the following basic properties of constructing Roughly Balanced Bagging (RBBag):

1. Using different learning algorithms to build component classifiers.
2. The influence of the number of component classifiers on final classification performance.
3. The role of diversity of component classifier's predictions.
4. The influence of data difficulty factors on confidence of predictions.

According to our best knowledge, these characteristics of Roughly Balanced Bagging have not been studied in the literature yet.

We choose 24 UCI datasets which have been used in the most related experimental studies (Błaszczyński and Stefanowski 2015; Khoshgoftaar et al. 2011; Napierala and Stefanowski 2012, 2016). They represent different imbalance ratios, different numbers of attributes and they come from different domains. Due to the fact that the original version of RBBag is not able to handle multi-class classification problems, we first consider binary class versions of these data as it is done in the earlier related studies (Galar et al. 2011; Hido and Kashima 2009; Khoshgoftaar et al. 2011). A generalization of RBBag for multiple classes will be further studied in Section 5.

Moreover, these datasets represent different difficulty factors referring to distributions of the minority class which is the additional issue to study in our experimental analysis. Recall that different difficulty factors could be considered: a fragmentation of the minority class into small disjuncts, overlapping of decision boundaries, presence of rare cases, outliers, noise (Stefanowski 2016a). Here we follow the methodology from Napieraha and Stefanowski (2012, 2016), where most of these data difficulty factors can be modeled by distinguishing the following types of examples: *safe examples* (located in the homogeneous regions populated by examples from one class only); *borderline* (placed close to the decision boundary between classes); *rare examples* (isolated groups of few examples located deeper inside the opposite class), or *outliers*.

Following the method introduced in Napierala and Stefanowski (2012) our approach to identify these types of examples in data is based on analyzing class label distribution inside the neighbourhood of the minority class example. Such an analysis has been implemented either with k-nearest neighbours or kernels – according to experiments from Napierala and Stefanowski (2016) both approaches provided comparative labeling for studied datasets. In this study, we decided to use the method where the type of example can be identified by analysing class labels of the k-nearest neighbours of this example. For instance, if $k = 5$, the type of the example is assigned in the following way (Napierala and Stefanowski 2012; 2016): 5:0 or 4:1 – an example is labeled as a safe example; 3:2 or 2:3 – a borderline example; 1:4 – labeled as a rare example; 0:5 – example is labeled as an outlier. This rule can be generalized for higher k values, however, results of recent experiments (Napierala and Stefanowski 2016) show that they lead to a similar categorization of considered datasets. Therefore, in the following study we stay with $k = 5$.

Basing on this method for an identification of example types, we were able to distinguish between easier (safe) distributions and more difficult ones, including borderline, rare or outlier examples. In Table 1 we present characteristics of the chosen datasets with respect

Table 1 Datasets characteristics

Dataset	# examples	# attrib.	IR	Difficulty type
breast-w	699	9	1.90	safe
vehicle	846	18	3.25	safe
new-thyroid	215	5	5.14	safe
abdominal-pain	723	13	2.58	safe
acl	140	6	2.50	safe
scrotal-pain	201	13	2.41	safe/borderline
car	1728	6	24.04	safe/borderline
ionosphere	351	34	1.79	safe/borderline
pima	768	8	1.87	borderline
bupa	345	6	1.38	borderline
hepatitis	155	19	3.84	borderline
credit-g	1000	20	2.33	borderline
haberman	306	4	2.78	borderline
ecoli	336	7	8.60	borderline/rare
cmc	1473	9	3.42	borderline/rare
transfusion	748	4	3.20	rare
yeast	1484	8	28.10	rare
solar-flareF	1066	12	23.79	rare
postoperative	90	8	2.75	rare
cleveland	303	13	7.66	rare
hsv	122	11	7.71	rare
breast-cancer	286	9	2.36	outlier
abalone	4177	8	11.47	outlier
balance-scale	625	4	11.76	outlier

to these properties, the number of attributes, number of examples and the global imbalance ratio (IR).

The performance of ensembles is evaluated using measures developed for binary imbalanced problems. They are defined on the basis of the binary confusion matrix. We have chosen the following measures:

1. *Sensitivity* of the minority class (the local accuracy of the minority class),
2. *Specificity* of the minority class (the local accuracy of majority classes),
3. Their aggregation to the *geometric mean* (G-mean).

For their definitions see, e.g., Japkowicz and Shah (2011). We have chosen these point measures instead of AUC, as the most of considered learning algorithms employed in RBBag produce deterministic outputs. These measures are estimated with the stratified 10-fold cross-validation repeated several times to reduce the variance. All experiments were performed in the WEKA framework in which we extended the previous implementation of RBBag done by L. Idkowiak for Błaszczyński et al. (2013).

3.2 Choosing algorithms to learn component classifiers

The related works show that Roughly Balanced Bagging, as well as other undersampling extensions of bagging, are usually constructed with decision trees. In this study, we check whether classification performance of this ensemble may depend on using other learning algorithms. Besides J4.8 unpruned tree we considered linear classifiers such as logistic regression and Support Vector Machines, decision rule algorithms – Ripper and PART as well as probabilistic algorithms: Naive Bayes and BayesNet classifiers. We also checked the performance of different tree algorithms: Naive Bayes tree and REPTree. All these algorithms are available in the WEKA framework. The RBBag ensemble was constructed with different numbers (30, 50 and 70) of component classifiers. For comparison, we also added the standard bagging to our experiment.

In Table 2 we report the average G-mean values and average ranks (the smaller, the better) from Friedman test for this measure. However, quite similar rankings were obtained for other measures. For all considered evaluation measures we were unable to reject the null hypothesis on equal performance of all versions of RBBag (e.g. for G-mean $p = 0.5045$). On the contrary to the standard bagging for which we observe significant differences between algorithms and we were able to reject the null hypothesis with rather small p-values (e.g. for G-mean $p = 0.0003$)

All these results did not show significant differences of using any of these algorithms inside RBBag and show its robustness to the change of the classifier. It is important to notice that for each single algorithm RBBag was significantly better than its standard bagging equivalent (according to the paired Wilcoxon test).

3.3 The influence of the number of component classifiers

Related works showed that RBBag was constructed with rather a high number of component classifiers. Hido and Kashima (2009) tested it with 100 C4.5 trees. In the study (Khoshgoftaar et al. 2011) authors applied a dozen of components. Then, 30, 50 or 70 trees were considered in Błaszczyński et al. (2013). Thus, we have decided to examine more systemically other (also smaller) sizes of this ensemble and its influence on classification

Table 2 Average values of G-mean and average ranks in Friedman test

	Average rank		Average G-mean	
Component classifier	Bagging	RBBag	Bagging	RBBag
Bayes Net	2.92	5.00	0.5854	0.7081
J4.8 unpruned tree	4.64	5.08	0.5475	0.7373
Logistic regression	5.60	5.16	0.4893	0.7234
Naive Bayes	3.60	5.44	0.5772	0.6935
Naive Bayes Tree	5.04	4.08	0.5135	0.7254
PART	4.44	5.20	0.5441	0.7314
REPTree	4.88	5.84	0.5426	0.7106
Ripper	4.12	4.60	0.4935	0.7412
SVM	5.20	4.56	0.4464	0.7275

performance. We stayed with learning components with J4.8 unpruned trees, and for each dataset we constructed a series of Roughly Balanced Bagging ensembles increasing its size one by one - so the number of component classifiers changed from 2 trees up to 100 ones. We present the changes of G-mean values for all datasets in Fig. 1.

For almost all considered datasets increasing the number of component classifiers improves the evaluation measures up to the certain size of the ensemble. Then, values of measures stay at a stable level or slightly vary around a certain level. Note that the RBBag ensemble achieves good performance for a relatively small number of component classifiers. For most datasets, the stable highest value of G-mean is observed approximately between 10 and 15 trees. In case of the sensitivity or F-measure we noticed similar tendencies.

Changing the number of components gives a slightly different effect on `hsv` and `postoperative` datasets. Observe that augmenting the number of components on these datasets causes a slight decrease of sensitivity and G-mean values, instead of increasing them. Then, values start to fluctuate around certain levels. However, both datasets are the smallest ones as well as the distributions of the minority class are the most sparse and the most difficult ones (Napierala and Stefanowski 2016).

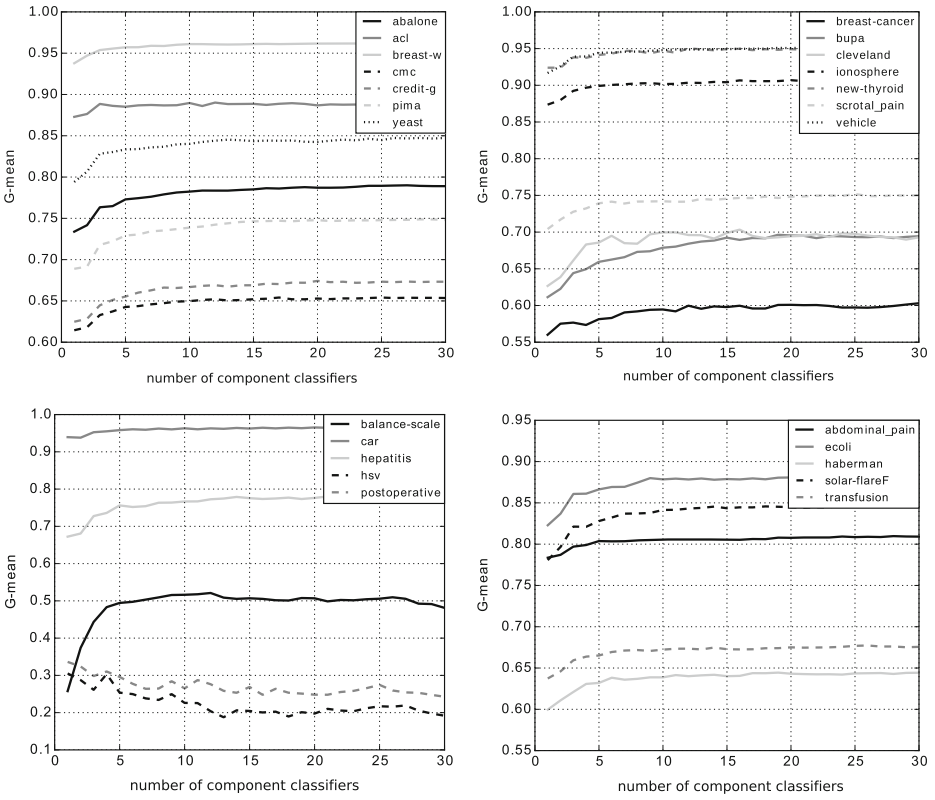

Fig. 1 G-mean vs. a number of component classifiers in RBBag

Moreover, we decided to examine confidence of the final decision of RBBag. We refer to a *margin* of the ensemble prediction. For standard ensembles, it is defined as a difference between the number of votes of components for the most often predicted class label and the number of votes for the second predicted label. Here, we modified it as the relative ratio:

$$marg = \frac{n_{cor} - n_{incor}}{n_{cptclas}} \tag{3}$$

where n_{cor} is the number of votes for the correct class, n_{incor} is the number of votes for the incorrect class and $n_{cptclas}$ is the number of component classifiers in the ensemble. The higher absolute value of $marg$ is interpreted as high confidence while values closer to 0 indicate uncertainty in making a final decision for a classified instance. It is worth noticing that the margin value close to -1 means the highly confident but incorrect decision.

In Fig. 2 we present a representative trend of changes of the relative margin with the size of RBBag for `ecoli` and `cmc` data. For many other datasets the trend line of the margin also stabilizes after a certain size (Note the resolution of the margin scale is more detailed than G-mean, so margin values achieve a satisfactory level also quite fast). We can conclude that good performance of Roughly Balanced Bagging comes from rather a small number of component classifiers.

3.4 Diversity of component classifiers

The final accuracy of ensembles may be also related to their diversity - which is usually understood as the degree to which component classifiers make different decisions on one problem (in particular, if they do not make the same wrong decisions). Although such an intuition behind constructing diverse component classifiers is present in many solutions, research concerns the total accuracy perspective (Kuncheva 2014). It is still not clear how diversity of components affects ensemble classification performance, especially for minority classes. The only work on ensembles dedicated for imbalance data (Wang and Yao 2009) does not provide a clear conclusion. Its authors empirically studied diversity of specialized oversampling ensembles and noticed that larger diversity improved recognition of the minority class, but at the cost of deteriorating the majority classes. However, nobody has analysed diversity of RBBag.

A popular group of measures to evaluate ensemble diversity are pairwise diversity measures (Kuncheva 2014; Tang et al. 2006). They are designed to compare the differences in predictions of two classifiers, which are treated as oracle outputs, i.e. it is only known

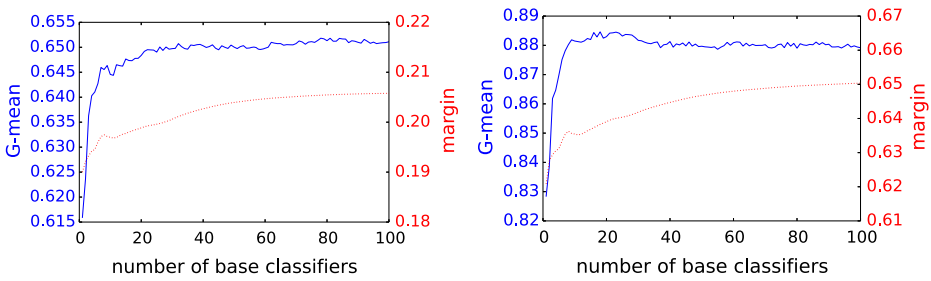

Fig. 2 G-mean and margin vs. a number of component classifiers in RBBag for `cmc` (*left*) and `ecoli` (*right*) datasets

whether the classifier prediction for a given object is correct or wrong[2] (Kuncheva 2014). The measures are defined for a pair of classifiers on the 2×2 oracle matrix containing the number of examples for which both classifiers makes correct decision (n_{11}), the number of misclassified examples by one of the classifiers (n_{10}, n_{01}) and the number of examples which were incorrectly classified by both algorithms (n_{00}). To evaluate the diversity of the whole ensemble, one calculates these measures for each pair of component classifiers. Then, the globally averaged value for an ensemble is averaged over all pairs of classifiers.

In this study we use two popular pairwise diversity measures, one of them is the *disagreement measure* defined by following equation:

$$D = \frac{n_{10} + n_{01}}{n_{11} + n_{10} + n_{01} + n_{00}} \tag{4}$$

The larger its value is, the more diverse classifiers are, but the maximal value of this measure depends on the accuracy of classifiers (Kuncheva 2014). For this reason the *Q-statistics* will be also used as another diversity measure:

$$Q = \frac{n_{11}n_{00} - n_{10}n_{01}}{n_{11}n_{00} + n_{10}n_{01}} \tag{5}$$

It has a constant range of possible values from -1 to 1 which also makes it easier to interpret. It is worth to notice that this measures simply take into account the number of examples classified by a pair of algorithms regardless its class label. This causes that the diversity of classifiers when predicting the majority classes has big influence on the final value of diversity measures. For this reason besides the global diversity measures for predictions in both classes (D, Q) we also calculated these measures for the minority class only (denoted with min in the subscript). The results for RBBag with 30 component J4.8 trees are presented in Table 3. For some pairs of classifiers we were unable to calculate Q-statistics for minority examples only due to the zero denominator – if we were unable to calculate this metric for more than 25 % of components pairs then we present the result with a star symbol.

Notice that values of disagreement measures are relatively low. For nearly all datasets they are between 0.1 and 0.3. This small diversity concerns both class predictions (D) and minority class (D_{min}), although D_{min} is usually lower than D. We also checked that changing the number of component classifiers in RBBag did not influence values of the disagreement measure. The conclusions from the results of Q-statistic are consistent with those for the disagreement measure.

To sum up, this high accuracy of RBBag may not be directly related to its higher diversity. We have also analysed predictions of particular pairs of classifiers and noticed that they quite often make the same correct decisions.

3.5 Influence of the type of examples

Experiments carried out in the related works indicates the superiority of RBBag over other specialized ensembles for imbalance data. However, it is unclear whether good predictive results of RBBag come from its special abilities to deal with particular data difficulty factors or just from handling very efficiently the global class imbalance. In this section, we conduct

[2]This interpretation of the classifier outputs disregards the precise information on which class label has been assigned to the classified object. Nevertheless, for binary imbalanced classes it is sufficient. For multiple classes other generalizations of these pairwise diversity measures could be considered, see e.g. recent proposals such as Mikami et al. (2015).

Table 3 Diversity measures, calculated for examples from both classes (D, Q) and from the minority class only (D_{min}, Q_{min}), for Roughly Balanced Bagging

Dataset	D	D_{min}	Q	Q_{min}
abalone	0,2072	0,1642	0,8001	0,7939
abdominal-pain	0,1564	0,1310	0,8444	0,8479
acl	0,0756	0,0843	0,8551	0,2533
balance-scale	0,4884	0,4963	−0, 0112	* 0,0080
breast-cancer	0,3154	0,2876	0,5576	0,5265
breast-w	0,0361	0,0476	0,9523	−0, 1040
bupa	0,2371	0,2759	0,5192	0,6298
car	0,0951	0,0201	0,8707	*−0, 7852
cleveland	0,2807	0,2470	0,6680	*−0, 4505
cmc	0,2798	0,2481	0,7259	0,7364
credit-g	0,2648	0,2279	0,6167	0,6558
ecoli	0,1197	0,1040	0,9144	0,6538
haberman	0,2667	0,2482	0,8337	0,9417
hepatitis	0,2476	0,2127	0,6343	0,0865
hsv	0,4529	0,3384	0,3311	* 0,1667
ionosphere	0,0733	0,0909	0,5370	* 0,2471
new-thyroid	0,0525	0,0297	0,6198	* 0,1036
pima	0,2100	0,1949	0,8136	0,8143
postoperative	0,4011	0,3837	0,3424	0,0889
scrotal-pain	0,1871	0,1670	0,7888	0,7553
solar-flareF	0,1062	0,0999	0,9318	0,8047
transfusion	0,1931	0,1897	0,8625	0,9536
vehicle	0,0592	0,0509	0,9222	*−0, 3853
yeast	0,1335	0,0885	0,9127	*−0, 3308

experiments which should provide us to get an insight into the work of RBBag on different types of examples, following the methodology from Napierala and Stefanowski (2012) which was also described earlier in this paper.

Similarly to Napieraha and Stefanowski (2012, 2016) we observed that the most of the datasets considered in this paper contain rather a small number of safe examples from the minority class. The exceptions are two datasets composed of many safe examples: `new-thyroid` and `car`. Many datasets such as `cleveland`, `balance-scale` or `solar-flare` do not contain any safe examples but many outliers and rare cases. The similar analysis of the majority class shows that the datasets contain mostly safe types of majority examples. Recalling recent experiments from Błaszczyński and Stefanowski (2015) we add that differences between the performance of various generalizations of bagging are smaller for datasets where safe minority examples dominate inside the distribution. On the other hand, RBBag stronger outperforms other generalizations if datasets contain many unsafe minority examples.

In the current experiments, we identified a type of the testing example and recorded whether it was correctly classified or not. Additionally, we refer types of examples in both (minority and majority) classes to the relative margins of the RBBag predictions (these are presented as histograms of numbers of testing examples with a given value of the margins).

In Figs. 3 and 4 we present a representative results of RBBag and the standard bagging for `cleveland` dataset. Histograms for other datasets present similar observations.

Notice that RBBag quite well recognizes the borderline examples from the minority class. Rare minority examples are more difficult, however, on average RBBag can still recognize many of them. It classifies them much better than the standard bagging. Outliers are the most difficult, but RBBag classifies correctly some of them and again this is the main difference to the standard bagging and other its oversampling extensions evaluated in Błaszczyński and Stefanowski (2015). The similar tendency is observed for other unsafe datasets which are not visualized due to page limits. If the dataset contains some safe minority examples, nearly all of them are correctly classified with high margins.

On the other hand, for the majority class, one can notice that RBBag correctly classifies most of safe examples while facing some difficulties with borderline ones. It also holds for other non-visualized datasets (where the margin's median for borderline majority examples is always worse than the median for borderline minority examples). The majority class does not contain any rare or outlying examples for nearly all considered datasets. For few exceptions as `pima`, `breast-cancer` or `cmc`, these rare majority examples are misclassified with the high negative margin.

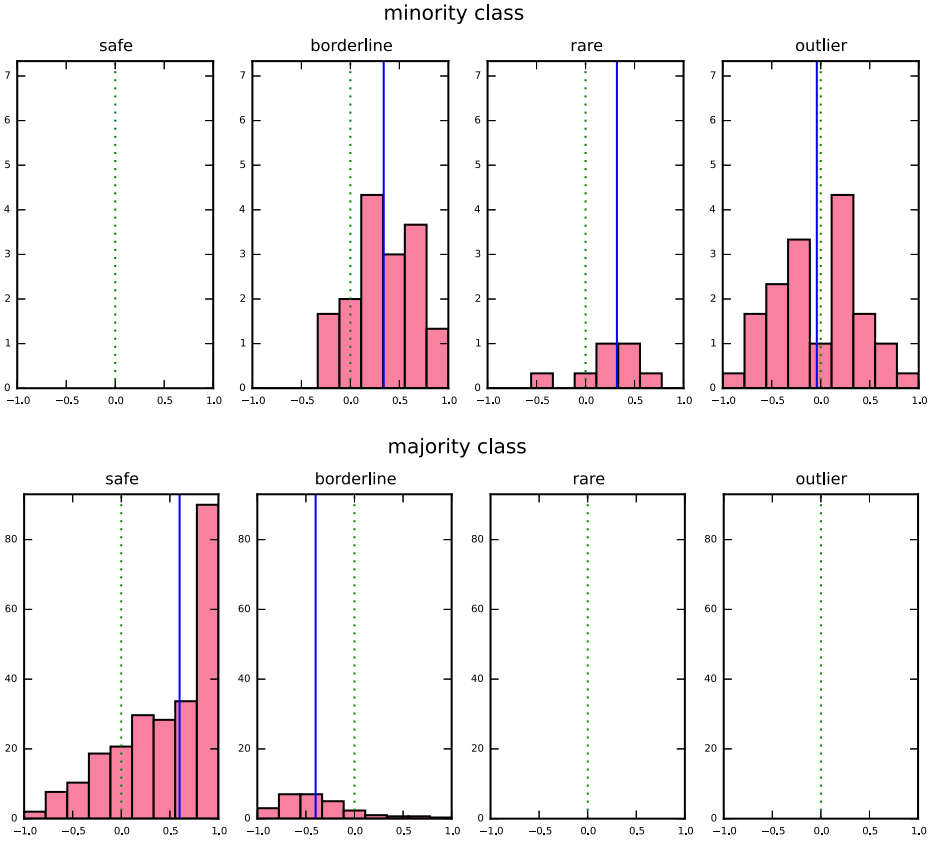

Fig. 3 Histogram of RBBag margins for `cleveland` dataset with respect to a class and a type of an example. Blue vertical line shows the value of the margin's median

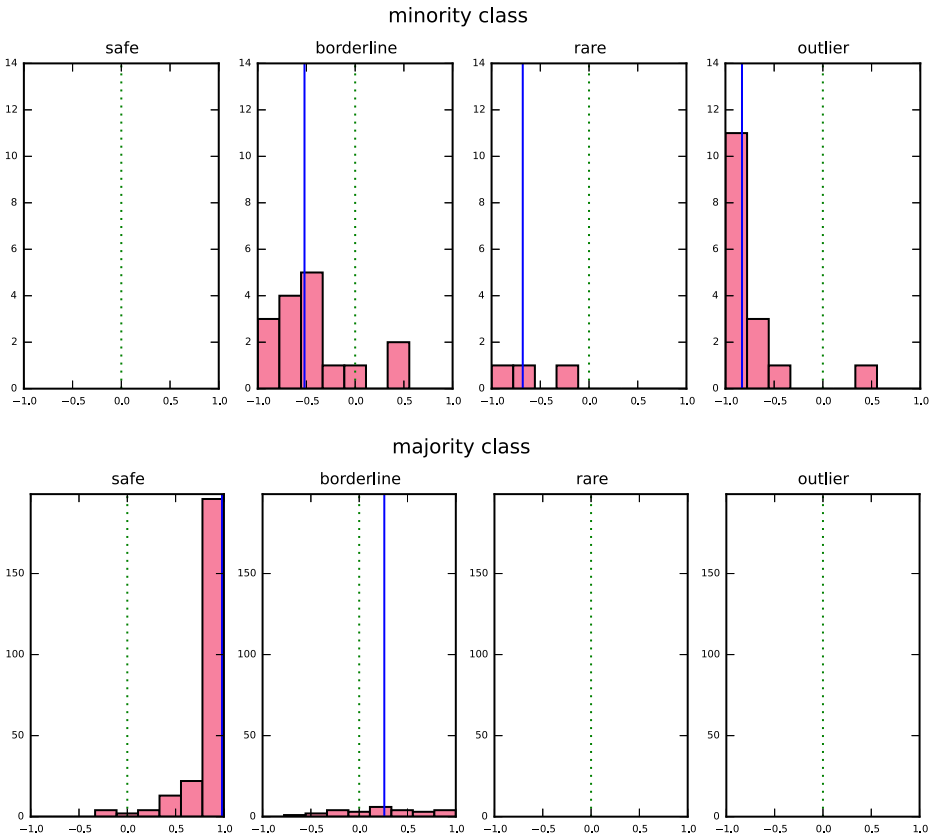

Fig. 4 Histogram of standard bagging margins for `cleveland` dataset with respect to a class and a type of an example. Blue vertical line shows the value of the margin's median

In conclusion, we can hypothesize that Roughly Balanced Bagging improves recognition of unsafe minority examples, but at the cost of worse dealing with unsafe majority examples. However, as the number of unsafe examples is relatively small in the majority class, the final performance of RBBag is improved.

4 Generalization of Roughly Balanced Bagging with selection of attributes

Although the previous experiments have shown very good predictive abilities of Roughly Balanced Bagging, we have decided to study its applicability to some other complex classification problems and directions of potential improvements. Firstly, we will consider dealing the higher number of attributes. A higher number of attributes in datasets is common in text categorization, image recognition, and bio-informatics and it requires additional processing of attributes. For instance, earlier studies with medical images, such as Jelonek and Stefanowski (1997), have shown that some classes are well discriminating with smaller subsets of attributes.

Recall the discussion presented in Section 2.5 which concludes that the most of the current research on class-imbalanced problems covers the relatively small number of attributes. Yet another motivation comes from noticing a rather low diversity of component classifiers in the typical undersampling ensemble, in particular, see our results presented in Section 3.4.

In our proposals, we will follow the line of modifying the construction of bootstrap samples of bagging, which is inspired by earlier research on applying random attribute selection while constructing standard ensembles, see e.g. a review available in Kuncheva (2014). More precisely we refer to Ho's proposal of *Random Subspace* method (RSM) (Ho 1998), and Breiman's Random Forest ensemble. Besides simplicity, while dealing with the higher number of attributes, this random selection of attributes increases the diversity of component classifiers.

While recent experiments of Liu and Zhu (2013) with Balanced Random Forest demonstrated its usefulness for class imbalance, here we are more interested in adapting Random Subspace into the context of Roughly Balanced Bagging as it is a classifier independent strategy. To the best of our knowledge, it has not been considered in this ensemble yet. In the only related work (Hoens and Chawla 2010) authors successfully applied this method to SMOTE based oversampling bagging.

Our proposal named RBBag+RSM[3] extends Roughly Balanced Bagging with the random attribute selection in the following way. After sampling examples to each bootstrap a subset of F attributes is randomly drawn from the set of all attributes (where each attribute has the same probability to be selected). One should decide about the number of the drawn attributes. The inspiration of Random Subspace Method may result in taking 50 % of the number of attributes in the original dataset. On the other hand, if we want to deal with both higher number of attributes and improving diversity we will follow Breiman's rule from Random Forest which takes a smaller number of attributes. Subsequently, we train component classifier with such bootstrap sample with selected attributes only, e.g. $\lceil\sqrt{F}\rceil$, $\lceil\log_2 F + 1\rceil$. As these component classifiers may be more diversified than in the standard RBBag, we expect that the number of components of the ensemble should be higher. Their predictions are aggregated into the final decision of the ensemble in the same way as in the standard version.

Below we experimentally evaluate it. Since it is an approach designed to deal with a higher number of attributes, we have focused our experiments only on these datasets from earlier phases of experiments, which contain more than 11 attributes. As this condition holds for 9 datasets only, we added 4 new, high-dimensional imbalanced datasets from UCI repository (Lichman 2013). Finally, in this experiment we examine 13 following datasets: `abdominal-pain` (13 attributes), `cleveland` (13), `credit-g` (20), `dermatology` (35), `hepatitis` (19), `ionosphere` (34), `satimage` (37), `scrotal-pain` (13), `segment` (20), `seismic-bumps` (19), `solar-flare` (12), `vehicle` (18) and `vowel` (14).

We tested it with J4.8 decision tree (without pruning) and SVM as base classifiers. Following the literature review, we considered setting f parameter to $\lceil\sqrt{F}\rceil$, $\lceil\log_2 F + 1\rceil$ and $\lceil{}^1/_2 F\rceil$, where F is the total number of attributes in the dataset. Due to space limit we present detailed results only for J4.8 decision trees and $f = \lceil\sqrt{F}\rceil$, since this parameter setting gives, on average, the highest increments.

Another issue concerns the size of RBBag+RSM. Although bagging can be constructed with a small number of components (our experiments may recommend RBBag approx. 15),

[3]RSM is an abbreviation from Random Subspace Method.

their number in case of attribute selection should be higher as the randomization of attributes increases the variance of bootstrap samples. This is why we will compare RBBag against the RBBag+RSM ensemble with more components: 30, 50, 70 and 90.

The values of G-mean and sensitivity are presented in Tables 4 and 5, respectively. One can notice increases on the values of both measures, in particular for RBBag+RSM with more trees. For instance, the increase on sensitivity (`abdominal-pain`, `hepatitis` – above 6 %) and G-mean (`abdominal-pain`, `hepatitis`, `scrotal-pain`, `seismic-bumps` – above 3 %). We performed the paired Wilcoxon test to compare RBBag+RSM against RBBag. With the confidence $\alpha = 0.05$, RBBag+RSM is better on G-mean for 50 ($p = 0.021$), 70 ($p = 0.013$) and 90 ($p = 0.006$) trees and nearly for 30 trees ($p = 0.054$). Similar results we obtained for the sensitivity measure.

In additional experiments, we also observed that RBBag+RSM needs more components than RBBag, e.g. for 15 trees there were no significant differences in values of G-mean ($p = 0.11$). It confirms our expectations and earlier literature opinions saying that while introducing random attribute selection one should use more components than in the standard bagging. However, as we do not want to increase computational costs too much comparing to the basic version of RBBag, so we prefer to stay with the considered sizes of the ensemble.

We also analysed the results of specificity to see whether good recognition of the minority class is not achieved at a high cost of majority class accuracy. Surprisingly, for most datasets, this measure has actually increased while the highest decrease does not exceed 2 %.

Similar results were obtained for RBBag+RSM with SVM as a component classifier. Since SVMs are more robust, the improvements on G-mean and sensitivity measure were smaller, but still significant on many datasets. For example on `scrotal-pain` dataset we observe above 3 % improvement on G-mean and over 5 % increase of sensitivity.

Additionally, we calculated the disagreement measure for all examples (D) and also the minority class (D_{min}). The values presented in Table 6 are calculated for 30 trees. For

Table 4 G-mean for Roughly Balanced Bagging (RBBag) and its modification by a random attribute selection (RBBag+RSM)

	RBBag				RBBag+RSM			
Dataset	30	50	70	90	30	50	70	90
abdominal-pain	0.8077	0.8072	0.8062	0.8050	0.8336	0.8411	0.8358	0.8363
cleveland	0.7161	0.7247	0.7208	0.7086	0.6938	0.7197	0.7410	0.7347
credit-g	0.6735	0.6755	0.6792	0.6704	0.6930	0.6923	0.7007	0.7036
dermatology	0.9868	0.9864	0.9873	0.9925	0.9986	1.0000	1.0000	1.0000
hepatitis	0.7663	0.7947	0.7920	0.7841	0.8131	0.8113	0.8029	0.8083
ionosphere	0.9063	0.9079	0.9098	0.9098	0.9068	0.9104	0.9152	0.9142
satimage	0.8727	0.8734	0.8752	0.8800	0.8677	0.8678	0.8698	0.8701
scrotal-pain	0.7484	0.7414	0.7455	0.7452	0.7869	0.7846	0.7884	0.7831
segment	0.9892	0.9895	0.9896	0.9890	0.9945	0.9955	0.9953	0.9951
seismic-bumps	0.6824	0.6945	0.6937	0.6914	0.7103	0.7153	0.7124	0.7123
solar-flare	0.8499	0.8511	0.8529	0.8548	0.8351	0.8437	0.8458	0.8500
vehicle	0.9525	0.9548	0.9552	0.9546	0.9590	0.9588	0.9599	0.9596
vowel	0.9623	0.9604	0.9606	0.9616	0.9751	0.9766	0.9789	0.9805

Table 5 Sensitivity for Roughly Balanced Bagging (RBBag) and its modification by random attribute selection (RBBag+RSM)

	RBBag				RBBag+RSM			
Dataset	30	50	70	90	30	50	70	90
abdominal-pain	0.7955	0.7975	0.7925	0.7811	0.8523	0.8623	0.8563	0.8560
cleveland	0.7067	0.7175	0.7100	0.6883	0.6800	0.7117	0.7567	0.7325
credit-g	0.6610	0.6637	0.6657	0.6243	0.6493	0.6407	0.6540	0.6323
dermatology	0.9900	0.9950	0.9950	1.0000	1.0000	1.0000	1.0000	1.0000
hepatitis	0.7500	0.7917	0.7950	0.7717	0.8200	0.8267	0.8267	0.8217
ionosphere	0.8553	0.8561	0.8593	0.8529	0.8660	0.8737	0.8796	0.8715
satimage	0.8690	0.8726	0.8753	0.8816	0.8738	0.8720	0.8777	0.8777
scrotal-pain	0.7400	0.7330	0.7360	0.7240	0.7467	0.7560	0.7453	0.7277
segment	0.9863	0.9875	0.9875	0.9857	0.9918	0.9933	0.9930	0.9930
seismic-bumps	0.6312	0.6547	0.6529	0.6488	0.6624	0.6629	0.6612	0.6653
solar-flare	0.8690	0.8705	0.8730	0.8785	0.8450	0.8670	0.8670	0.8760
vehicle	0.9688	0.9703	0.9724	0.9679	0.9990	0.9990	0.9990	0.9995
vowel	0.9667	0.9667	0.9667	0.9667	0.9911	0.9911	0.9900	0.9889

the reader's convenience, we present results together with the difference of disagreement between RBBag+RSM and original RBBag.

One can notice that the selection of attributes resulted in an increase of disagreement on almost all datasets (except `seismic-bumps`). Interestingly, despite a decline of the disagreement measure on this dataset we observed improvements on both G-mean and

Table 6 Disagreement measures, calculated for examples from both classes (D) and from the minority class only (D_{min}), for Roughly Balanced Bagging (RBBag) and its modification by random attribute selection (RBBag+RSM)

	RBBag		RBBag+RSM		Difference	
Dataset	D	D_{min}	D	D_{min}	D	D_{min}
abdominal-pain	0.1564	0.1310	0.2995	0.2580	0.1431	0.1269
cleveland	0.2807	0.2470	0.3506	0.3050	0.0700	0.0581
credit-g	0.2648	0.2279	0.4075	0.3951	0.1427	0.1672
dermatology	0.0211	0.0162	0.1815	0.1384	0.1604	0.1222
hepatitis	0.2476	0.2127	0.3156	0.2915	0.0680	0.0788
ionosphere	0.0733	0.0909	0.1158	0.1650	0.0424	0.0741
satimage	0.1549	0.1160	0.1782	0.1448	0.0233	0.0288
scrotal-pain	0.1871	0.1670	0.3522	0.3139	0.1651	0.1469
segment	0.0168	0.0106	0.0659	0.0293	0.0491	0.0187
seismic-bumps	0.2891	0.2373	0.2470	0.2383	−0.0421	0.0010
solar-flare	0.1062	0.0999	0.2362	0.2395	0.1300	0.1396
vehicle	0.0592	0.0509	0.1461	0.0972	0.0869	0.0463
vowel	0.0461	0.0251	0.2126	0.0825	0.1665	0.0574

sensitivity. The further analysis of diversity shows that the highest increase occurs for safe majority and borderline minority examples.

The analysis of histograms of decision margins for RBBag+RSM (done in the same way as for RBBag in Section 3.5) shows that it increases the margin on rare and outlier minority examples. Moreover, more safe and borderline minority instances are classified correctly, although the average of the margin slightly decreases. Due to increased diversity, fewer examples are classified with a maximum decision margin. Sometimes it also decreases the decision margin too much, ending with misclassification of some examples. Then, RBBag+RSM decreases the margin of safe majority examples, but this does not adversely affect the final prediction.

5 Multi-class extensions of RBB bagging for imbalanced data

5.1 Multi-class Roughly Balanced Bagging

In this section, we introduce a new extension of Roughly Balanced Bagging for multi-class imbalanced data. In opposition to the related works it does not decompose the multi-class problem into many binary problems, but it learns all the classes at once. This property of our extension is obtained by considering simultaneously all classes during the construction of bootstrap samples.

Recall that in original bagging, a bootstrap sample is constructed by sampling examples one by one from the uniform joint distribution $p(x, c)$ where x is an example of a class c. This can be simply decomposed into the conditional probability of selecting an example given the class c and the probability of selecting example from a particular class (i.e. $p(x, c) = p(c)p(x|c)$). Thus, the bootstrap construction can be realized by repetitively sampling examples in the following way: first select (with a proper probability) the class from which the example will be taken and then choose an example from the selected class (Hido and Kashima 2009).

Note that the number of each class examples in the bootstrap samples of original bagging varies according to the multinomial distribution – which in the case of binary classification is simply the binomial distribution. The authors of RBBag (which works for binary classification only) exploit it by fixing the number of selected minority class examples to N_{min} and using the negative binomial distribution to estimate a proper number of majority examples. Then, these numbers of examples are sampled from each class with replacement (where each example has the same probability of being selected); see Section 2.4.

Hido et al. (2009) claims that this approach better reflects the philosophy of bagging than other specialized bagging approaches for imbalance data. However, in contrast to the original bagging which constructs bootstrap samples of a constant size, Roughly Balanced Bagging creates bootstraps of different sizes. In our approach, we propose yet another sampling schema, which is also coherent with the ideas of bagging. However, it creates samples of equal size – just like in the original Breiman's proposal (Breiman 1996). Nevertheless, our main objective it to adapt it to a multi-class scenario.

Therefore, in Multi-class Roughly Balanced Bagging (further abbreviated as MRBBag) the main modification concerns a construction of bootstrap samples, which is realized in the following way. First, the number of examples to be selected from each class has to be estimated for each bootstrap. Following the aforementioned probabilistic interpretation

we will estimate it from the multinomial distribution, which is defined by the following probability mass function:

$$p(n_1, n_2, ..., n_c) = \frac{n!}{n_1!n_2!\cdots n_c!} p_1^{n_1} p_2^{n_2} \cdots p_c^{n_c}$$

where $p_1, p_2, ..., p_c$ and $n = \sum_{i=1}^{c} n_i$ are the parameters of the distribution.

In the original bagging the $p_1, p_2, ..., p_c$ values should be selected proportionally to the presence of classes in the training set. In our algorithm we want to handle the class imbalance problem by obtaining roughly balanced bootstrap samples also with respect to class probabilities, so we fix values $p_1, p_2, ..., p_c$ to the same constant value equal to $\frac{1}{c}$, such that $\sum_{i=1}^{c} p_i = 1$. Then, just like in the Roughly Balanced Bagging we sample a proper number of examples from each class.

The value of n which determines the sample size is a parameter of our algorithm. In this study, we set $n = N$ which is the size of training set. In MRBBag, which creates roughly balanced samples this will bring the oversampling of minority classes – this version we will be later denoted as oMRBBag. Alternatively, following earlier observations from binary imbalance classification where undersampling strategies gave better results (Błaszczyński and Stefanowski 2015), we also consider setting n to the size of the smallest minority class in the original training dataset (i.e., $n = \min_{i \in \{1,2,...,c\}} N_i$). We refer to this version of the algorithm later as uMRBBag.

Algorithm 3 Multi-class Roughly Balanced Bagging

Input: $D = \cup_{j=1}^{c} D_j$: original training set with c classes, N: size of each bootstrap sample, k: number of bootstrap samples, LA: learning algorithm;
Output: C^* bagging ensemble with k component classifiers
Learning phase:

1: **for** $i = 1 \rightarrow k$ **do**
2: $S_i = \emptyset$
3: $[n_1, n_2, ..., n_c] \leftarrow$ following multinomial distribution with $n = N$ and $q_i = 1/c$ for $i = 1, 2, ..., c$
4: **for** $j = 1 \rightarrow c$ **do**
5: $S_{i,j} \leftarrow n_j$-element bootstrap sample drawn with replacement from D_j
6: $S_i \leftarrow S_i \cup S_{i,j}$
7: **end for**
8: $C_i \leftarrow LA(S_i)$
9: **end for**

Prediction phase:

$$C^*(x) = \arg\max_y \sum_{i=1}^{k} p_{C_i}(y|x)$$

5.2 Evaluation of Multi-class Roughly Balanced Bagging on artificial data

Firstly, we investigate experimentally the performance of MRBBag on artificially generated data that are affected by different difficulty factors. We apply a special generator (Wojciechowski and Wilk 2014). The datasets have three classes: two minority classes and one majority class. The examples of both minority classes are generated randomly inside

predefined spheres and the majority class examples are randomly distributed in an area surrounding them. We consider three configurations of these spheres positions:

– no overlap between classes– the centers of spheres are far away each other and the minority classes are separated by majority class examples,
– small overlap – the borders of spheres which contains examples of minority classes are touching each other,
– overlap – the spheres have a larger common part in the attribute space.

The majority examples are randomly generated in a cube which contains both spheres. However, they are added in a way which ensures that the minority classes will have predefined a number of safe and borderline examples. Later, additional minority examples are placed among the majority ones to create the predefined number of rare and outlier examples.

We generated datasets with the following configurations:

– 50 % of safe examples, 35 % of borderline, 10 % of rare and 5 % of outliers (50-35-10-5);
– 50 % of safe, 30 % of borderline, 10 % of rare and 10 % of outliers (50-30-10-10);
– 70 % of safe and 30 % of borderline examples (70-30-0-0).

These three configurations are combined with three different setups of spheres positions which finally results in nine artificial datasets. All datasets are generated with three attributes and 900 examples – 100 of them belong to the first minority class, 200 belong to the second one and 600 belong to the majority class. We have also generated data with more attributes, however as their analysis gave similar results as three dimensions, we skip their presentation due to page limits. Figure 5 visualizes two of three-dimensional artificial datasets.

In our experiment, we decided to compare classification performance of proposed multi-class oMRBBag and uMRBBag generalizations against original bagging and a single J4.8 decision tree as baseline multi-class classifiers. Moreover, as the original RBBag cannot deal directly with multi-class datasets, we have used its slight modification (denoted as RBBag*). There were two issues to resolve, in order to enable RBBag to handle multiple classes. First of all, the choice of a base learner algorithm has been limited to multi-class classifiers only. Secondly, the process of bootstrap samples construction was slightly modified. Normally, while constructing bootstrap samples, RBBag divides the training set into

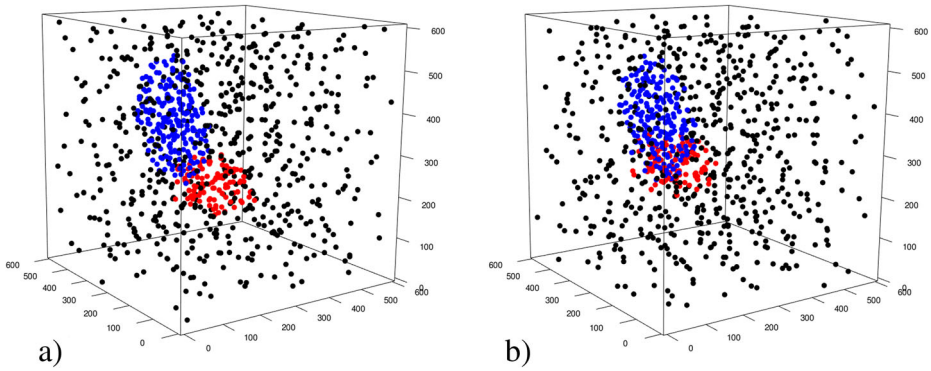

Fig. 5 Visualization of two artificial datasets with different levels of overlapping: "small overlap" (**a**) and "overlap" (**b**). Minority classes in both datasets contains 70 % of safe and 30 % of borderline examples

a set of minority examples D_{min} and a set of majority examples D_{maj}. Later, these sets are utilized to establish the parameter of the reverse binomial distribution ($n = N_i^{min} = |D_{min}|$) and to perform sampling of minority and majority examples separately. Unfortunately, partitioning of the training set in this way is not possible in the multi-class scenario since we have multiple minority classes. In our modification, we simply substituted the size of minority class (N_i^{min} in step 2 of Algorithm 2) with the number of all minority examples (a sum of all minority classes sizes). Later, instead of taking a sample from the set of minority class examples, the stratified sampling of all minority classes is used (step 4). By doing this simple adjustments RBBag* is able to handle classification problems with multiple classes.

As the evaluation of classification performance concerns now the non-binary case, we calculated sensitivity of each class and aggregated it into the geometric mean. All ensembles were learned with the same algorithm J4.8 learning decision unpruned trees. Always the same number of 30 component classifiers was induced. The values of the generalized G-mean are presented in Table 7.

Note that the undersampling proposal – uMRBBag – is the best performing ensemble. The next ones are oMRBBag and RBBag*. In all cases, they are better than the standard bagging and a single J4.8 tree classifier. As it could be expected the single tree is always the worst classifier. The interesting observation concerns a simpler modification RBBag*. It is comparable to our extensions only for one dataset which is the simplest and the easiest one – it contained only safe and borderline minority examples (70-30-0-0) with no overlapping. However, it is sometimes better than oversampling strategy in MRBBag. Note that oMRBBag outperforms RBBag* for two datasets only. On contrary, uMRBBag is almost always better than RBBag* and when it is not, the difference is not significant. The differences between uMRBBag and RBBag* are also higher on more difficult datasets. Especially, on the dataset with the highest number of difficult examples (50-30-10-10) and with class overlapping our modification improved G-mean above 3 %.

5.3 Evaluation of Multi-class Roughly Balanced Bagging on real data sets

The results on artificial datasets have led us to carry out more experiments on real datasets. We selected several multi-class imbalanced datasets from UCI repository. Some of them

Table 7 G-mean and average ranks (the lower, the better) for multi-class artificial data

Difficulty types	Overlapping	Bagging	uMRBBag	oMRBBag	RBBag*	J4.8
50-30-10-10	no overlap	0,7863	0,8111	0,7980	0,8104	0,7853
50-30-10-10	small overlap	0,7735	0,7945	0,7989	0,7962	0,7666
50-30-10-10	overlap	0,6822	0,7293	0,6839	0,6955	0,6576
50-35-10-5	no overlap	0,8195	0,8523	0,8522	0,8444	0,8101
50-35-10-5	small overlap	0,8169	0,8438	0,8428	0,8345	0,8010
50-35-10-5	overlap	0,6686	0,7449	0,7193	0,7280	0,6168
70-30-0-0	no overlap	0,9401	0,9496	0,9525	0,9566	0,9340
70-30-0-0	small overlap	0,9154	0,9378	0,9402	0,9386	0,9045
70-30-0-0	overlap	0,8245	0,8707	0,8335	0,8439	0,8022
average rank		4,0000	1,6666	2,2222	2,1111	5,0000

Table 8 G-mean and average ranks (the lower, the better) for multi-class real datasets

Dataset	Bagging	uMRBBag	oMRBBag	RBBag*	J4.8
car	0,8603	0,9016	0,9516	0,8680	0,7890
cleveland	0,0000	0,0128	0,0034	0,0000	0,0037
cleveland-sm	0,0833	0,1746	0,1238	0,1910	0,0755
dermatology	0,9542	0,9668	0,9658	0,9512	0,9436
dermatology-3	0,9494	0,9602	0,9569	0,9569	0,9275
ecoli	0,6534	0,7800	0,7460	0,7108	0,6095
ecoli-3	0,6872	0,8474	0,8004	0,8272	0,6613
glass	0,2819	0,4169	0,4424	0,4229	0,2591
glass-3	0,1386	0,6191	0,5236	0,5076	0,2885
new-thyroid	0,8937	0,9224	0,9215	0,9276	0,8778
thyroid	0,9425	0,9232	0,9455	0,9383	0,9420
vehicle	0,7162	0,7141	0,7181	0,7243	0,6825
yeast	0,0000	0,0336	0,0348	0,0000	0,0058
yeast-sm	0,0000	0,1307	0,0699	0,0109	0,0058
yeast-3	0,5602	0,8296	0,7104	0,8150	0,5874
average rank	4,0000	1,7142	2,0000	2,73331	4,5454

contained classes which have an extremely small number of examples. Note that some of these datasets contain extremely rare classes – containing few examples only. Thus, we removed classes which have less than 5 examples from the datasets and created simpler versions of these datasets. We denote them by adding to the original dataset name a suffix `-sm`. Additionally, we created a few three-class datasets by choosing two minority classes and treating all other examples as majority class (suffix `-3`). Proceeding in that way we prepared 15 datasets. The classification results of the same classifiers as considered in the previous sub-section are give in Table 8.

The best performing algorithm was uMRBBag. For two datasets the improvement of G-mean measure was above 11 % while comparing to RBBag*. Then, oMRBBag was better than RBBag* on more than a half of datasets. Again, a single tree classifier and original bagging had the worst performance on G-mean. We performed the Friedman test and we were able to reject the null hypothesis about lack of significant differences between classifiers with a very small p-value ($p < 0.0001$). The critical difference between ranks according to the post-hoc Nemenyi test $CD = 1.575$, which supports significant differences between a single classifier, original bagging and all modifications of Roughly Balanced Bagging. The differences between modifications of RBBag were not statistically significant.

6 Conclusions

Our study has covered two aims: (1) to experimentally study properties of Roughly Balanced Bagging (RBBag), and (2) to extend it for dealing with additional complexities of data referring either to a higher number of attributes or multiple class imbalances.

The experimental study of the properties of the RBBag ensemble has led us to the following main observations:

- It can be constructed with a relatively small number of component classifiers (approx. 15 ones).
- The choice of the considered algorithms for learning component classifiers does not influence the final performance of RBBag.
- Component classifiers in RBBag are characterized by quite low diversity according to the disagreement measure and Q-statistics.
- Studying the local recognition of types of classified examples shows that RBBag improves classification of unsafe minority examples. Its power for dealing with borderline, rare and outlying examples distinguishes it from other ensembles.

Comparing quite low diversity of Roughly Balanced Bagging to earlier results (Błaszczyński et al. 2013) we argue that RBBag is less diversified than over-bagging or SMOTE-based bagging (Wang and Yao 2009). On the other hand, RBBag is more accurate than these more diversified ensembles. We have also checked that its components are quite accurate and pairs of classifiers often make the same correct decisions. It may open another research on studying the trade-off between accuracy and diversity of ensembles for imbalanced data.

In spite of a good performance of Roughly Balanced Bagging, we have also asked questions about its further extensions. Our paper presents two types of methodological contributions.

Firstly, we have proposed to integrate a random selection of attributes into this ensemble. In experiments, we have shown that this proposal improves G-mean and sensitivity measures for more dimensional complex datasets. We have also observed that: (1) it increases the diversity of component classifiers and (2) using the higher number of components improves classification, differently than in the original RBBag.

Secondly, we have also introduced a generalization of Roughly Balanced Bagging for multiple imbalanced classes, which exploits the multinomial distribution to estimate cardinalities of class examples in bootstrap samples. The experiments with synthetic and real datasets have clearly demonstrated that the undersampling version of our proposed Multi-class RBBag improves G-mean and it is better than the oversampling variant and simpler multi-class classifiers.

Other lines of further research could also concern modifications of bootstrap sampling with information about types of minority examples and directing sampling more toward the unsafe examples. Recall that experiments from Section 3.5 have shown that RBBag also improves recognition of unsafe minority examples while worsening classification of borderline majority examples. Therefore, looking for other modifications of sampling which capture the trade-off between choosing examples from both classes could be still undertaken. Furthermore, a decomposition of classes into sub-concepts (Jo and Japkowicz 2004) could be considered. In Parinaz et al. (2015) authors applied k-means clustering to stratify sampling majority examples inside their modifications of standard bagging. Looking for another semi-supervised clustering to better handle complex boundaries of data distributions could be yet another direction for future research. Finally, considering both multiple classes and attribute selection could be also a line of future research.

Acknowledgments The research was supported by NCN grant DEC-2013/11/B/ST6/00963.

References

Anyfantis, D., Karagiannopoulos, M., Kotsiantis, S., & Pintelas, P. (2008). Creating ensembles of classifiers by distributing an imbalance data set to reach balance in each resulting training set. In *Proceedings of the IEEE DHMS Conference*.

Blagus, R., & Lusa, L. (2010). Class prediction for high- dimensional class-imbalanced data. *BMC Bioinformatics*, *11*, 523.

Błaszczyński, J., & Stefanowski, J. (2015). Neighbourhood sampling in bagging for imbalanced data. *Neurocomputing*, *150 A*, 184–203.

Błaszczyński, J., Deckert, M., Stefanowski, J., & Wilk, Sz. (2010). Integrating Selective Pre-processing of Imbalanced Data with Ivotes Ensemble. In *Proceedings of 7th International Conference RSCTC 2010, Springer, LNAI vol. 6086*, (pp. 148–157).

Błaszczyński, J., Stefanowski, J., & Idkowiak, L. (2013). Extending bagging for imbalanced data. In *Proceedings of the 8th CORES 2013, Springer Series on Advances in Intelligent Systems and Computing*, (Vol. 226 pp. 269–278).

Branco, P., Torgo, L., & Ribeiro, R. (2016). A survey of predictive modeling under imbalanced distributions. *ACM Computing Surveys (CSUR)*, *49*(2), 31. CoRR, arXiv:1505.01658.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123–140.

Chang, E.Y. (2003). Statistical learning for effective visual information retrieval. In *Proceedings of the ICIP 2003*, (Vol. 3 pp. 609–612).

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

Chawla, N. (2005). Data mining for imbalanced datasets: An overview. Chapter in Maimon O., Rokach L. (eds.): *The Data Mining and Knowledge Discovery Handbook*, (pp. 853–867): Springer.

Chawla, N., Bowyer, K., Hall, L., & Kegelmeyer, W. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 341–378.

Chen, X., & Wasikowski, M. (2008). FAST: A ROC–based feature selection metric for small samples and imbalanced data classification problems. In *Proceedings of the 14th ACM SIGKDD*, (pp. 124–133).

Chen, C., Liaw, A., & Breiman, L. (2004). *Using random forest to learn imbalanced data*. Berkeley: Technical Report, University of California.

Dal Pozzolo, A., Caelen, O., & Bontempi, G. (2015). When is undersampling effective in unbalanced classification tasks?. In *Proceedings of the ECML PKDD 2015. LNCS vol. 9284* (pp. 200–215): Springer.

Draminski, M., Dabrowski, M., Diamanti, K., Koronacki, J., & Komorowski, J. (2016). Discovering networks of interdependent features in high-dimensional problems. In Japkowicz, N., & Stefanowski, J. (Eds.) *Big Data Analysis: New Algorithms for a New Society* (pp. 285–304): Springer.

Fernandez, A., Garcia, S., & Herrera, F. (2011). Addressing the Classification with Imbalanced Data: Open Problems and New Challenges on Class Distribution. In *Proceedings HAIS Conference (part. 1)* (pp. 110).

Fernandez, A., Lopez, V., Galar, M., Jesus, M., & Herrera, F. (2013). Analysis the classification of imbalanced data sets with multiple classes, binarization techniques and ad-hoc approaches. *Knowledge Based Systems*, *42*, 97–110.

Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., & Herrera, F. (2011). 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.

Garcia, V., Sanchez, J.S., & Mollineda, R.A. (2007). 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, Springer, LNCS 4756, 397–406*.

He, H., & Garcia, E. (2009). Learning from imbalanced data. *IEEE Transactions on Data and Knowledge Engineering*, *21*(9), 1263–1284.

He, H., & Ma, Y. (eds.) (2013). Imbalanced Learning Foundations, Algorithms and Applications, IEEE - Wiley.

Hido, S., & Kashima, H. (2009). Roughly balanced bagging for imbalance data. *In Proceedings of the SIAM International Conference on Data Mining, 143-152 (2008) - an extended version in Statistical Analysis and Data Mining*, *2*(5-6), 412–426.

Ho, T. (1998). The random subspace method for constructing decision forests. *Pattern Analysis and Machine Intelligence*, *20*(8), 832–844.

Hoens, T., & Chawla, N. (2010). Generating diverse ensembles to counter the problem of class imbalance. In *Proceedings of PAKDD 2010* (pp. 488–499).

Japkowicz, N. (2003). Class imbalance: Are we focusing on the right issue?. In *Proceedings II Workshop on Learning from Imbalanced Data Sets, ICML Conference, 17–23*.

Japkowicz, N., & Shah, M. (2011). Evaluating Learning Algorithms: A Classification Perspective, Cambridge University Press.

Jo, T., & Japkowicz, N. (2004). Class Imbalances versus small disjuncts. *ACM SIGKDD Explorations Newsletter*, *6*(1), 40–49.

Jelonek, J., & Stefanowski, J. (1997). Feature subset selection for classification of histological images. *Artificial Intelligence in Medicine*, *9*, 227–239.

Khoshgoftaar, T., Van Hulse, J., & Napolitano, A. (2011). Comparing boosting and bagging techniques with noisy and imbalanced data. *IEEE Transactions on Systems, Man, and Cybernetics–Part A*, *41*(3), 552–568.

Krawczyk, B. (2016). Learning from imbalanced data: open challenges and future directions. *Progress Artificial Intelligence*, *5*(4), 221–232.

Kubat, M., & Matwin, S. (1997). Addresing the curse of imbalanced training sets: one-side selection. In *Proceedings of the 14th International Conference on Machine Learning ICML-97* (pp. 179–186).

Kuncheva, L. (2014). Combining pattern classifiers. Methods and Algorithms, Wiley, 2.

Lango, M., & Stefanowski, J. (2015). The usefulness of roughly balanced bagging for complex and high-dimensional imbalanced data. In *Proceedings of International ECML PKDD Workshop on New Frontiers in Mining Complex Patterns NFmCP 2015, Springer LNAI 9607* (pp. 93–107).

Latinne, P., Debeir, O., & Decaestecker, C.h. (2000). Different ways of weakening decision trees and their impact on classification accuracy of decision tree combination. In *Proceedings of the 1st International Workshop of Multiple Classifier Systems, Springer Verlag LNCS 1857*.

Laurikkala, J. (2001). Improving identification of difficult small classes by balancing class distribution. Techical Report A-2001-2, University of Tampere.

Lichman, M. (2013). UCI machine learning repository. University of California School of Information and Computer Science.

Lin, W., & Chen, J. (2013). Class–imbalanced classifiers for high-dimensional data. *Briefings in Bioinformatics*, *14*(1), 13–26.

Liu, A., & Zhu, Z. h. (2013). Ensemble methods for class imbalance learning. In He, H., & Ma, Y. (Eds.), *Imbalanced Learning. Foundations, Algorithms and Applications* (pp. 61–82): Wiley.

Lopez, V., Fernandez, A., Garcia, S., Palade, V., & Herrera, F. (2014). An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Information Sciences*, *257*, 113–141.

Mikami, A., Kudo, M., & Nakamura, A. (2015). Diversity measures and margin criteria in multiclass majority vote ensemble. In *Proceedings of the 12th International Workshop of Multiple Classifier Systems, MCS 2015* (pp. 27–37): Springer.

Napierala, K., & Stefanowski, J. (2012). The influence of minority class distribution on learning from imbalance data. In *Proceedings 7th Conference HAIS 2012, LNAI vol. 7209* (pp. 139–150): Springer.

Napierala, K., & Stefanowski, J. (2016). Types of minority class examples and their influence on learning classifiers from imbalanced data. *Journal of Intelligent Information Systems*, *46*(3), 563–597.

Napierala, K., Stefanowski, J., & Wilk, Sz. (2010). Learning from imbalanced data in presence of noisy and borderline Examples. In *Proceedings of 7th International Conference RSCTC 2010, Springer, LNAI*, (Vol. 6086 pp. 158–167).

Pant, H., & Srivastava, R. (2015). A survey on feature selection methods for imbalanced datasets. *International Journal of Computer Engineering and Applications, 9*(2).

Parinaz, S., Victor, H., & Matwin, S. (2015). Learning from imbalanced data using ensemble methods and cluster-based undersampling. In *Post-Proceedings 3rd Workshop New Frontiers of Mining Complex Patterns at ECML-PKDD 2014, Nancy, LNAI vol. 8983* (pp. 69–86): Springer.

Pio, G., Malerba, D., D'Eila, D., & Ceci, M. (2014). Integrating microRNA target predictions for the discovery of gene regulatory networks: a semi-supervised ensemble learning approach. *BMC Bioinformatics*, *15*(Suppl. 1), S4.

Rio, S., Lopez, V., Bemlez, J., & Herrera, F. (2014). On the use of MapReduce for imbalanced big data using Random Forests. *Information Sciences*, *285*, 112–130.

Seaz, J., Krawczyk, B., & Wozniak, M. (2016). Analyzing the oversampling of different classes and types in multi-class imbalanced data. *Pattern Recognition*, *57*, 164–178.

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

Stefanowski, J. (2016a). Dealing with data difficulty factors while learning from imbalanced data. In Mielniczuk, J., & Matwin, S. (Eds.), *Challenges in Computational Statistics and Data Mining* (pp. 333–363): Springer.

Stefanowski, J. (2016b). On properties of under-sampling bagging and its extensions for imbalanced data. In *Proceedings of the Conf. on Computer Recognition Systems, CORES 2015* (pp. 407–417): Springer.

Stefanowski, J., & Wilk, Sz. (2008). Selective pre-processing of imbalanced data for improving classification performance. In *Proceedings of the 10th International Conference DaWaK 2008. LNCS vol. 5182. Springer* (pp. 283–292).

Sun, Y., Wong, A., & Kamel, M. (2009). Classification of imbalanced data: a review. *International Journal Pattern Recognition Artificial Intelligence*, *23*(4), 687–719.

Tang, E., Suganthan, P., & Yao, X. (2006). An analysis of diversity measures. *Machine Learning*, *65*(1), 247–271.

Tang, J., Alelyani, S., & Liu, H. (2014). Feature selection for classification: A review. *Data Classification: Algorithms and Applications, 37*.

Van Hulse, J., Khoshgoftarr, T., & Napolitano, A. (2007). Experimental perspectives on learning from imbalanced data. In *Proceedings of ICML*, (Vol. 2007 pp. 935–942).

Wallace, B., Small, K., Brodley, C., & Trikalinos, T. (2011). Class Imbalance, Redux. In *Proceedings 11th IEEE International Conference on Data Mining*, (pp. 754–763).

Wang, S., & Yao, X. (2009). Diversity analysis on imbalanced data sets by using ensemble models. In *Proceedings IEEE Symposium Computer Intelligence Data Mining*, (pp. 324–331).

Wang, S., & Yao, X. (2012). Mutliclass imbalance problems: analysis and potential solutions. *IEEE Transaction System, Man Cybernetics Part B*, *42*(4), 1119–1130.

Weiss, G.M. (2004). Mining with rarity: a unifying framework. *ACM SIGKDD Explorations Newsletter*, *6*(1), 7–19.

Wilson, D., & Martinez, T. (1997). Improved heterogeneous distance functions. *Journal of Artificial Intelligence Research*, *6*, 1–34.

Wojciechowski, S., & Wilk, Sz. (2014). The generator of synthetic multi-dimensional data. Poznan University of Technology Report RB-16/14.