J Intell Inf Syst
DOI 10.1007/s10844-015-0368-1

Types of minority class examples and their influence on learning classifiers from imbalanced data

Krystyna Napierala[1] · Jerzy Stefanowski[1]

Received: 9 January 2015 / Accepted: 20 May 2015

Abstract Many real-world applications reveal difficulties in learning classifiers from imbalanced data. Although several methods for improving classifiers have been introduced, the identification of conditions for the efficient use of the particular method is still an open research problem. It is also worth to study the nature of imbalanced data, characteristics of the minority class distribution and their influence on classification performance. However, current studies on imbalanced data difficulty factors have been mainly done with artificial datasets and their conclusions are not easily applicable to the real-world problems, also because the methods for their identification are not sufficiently developed. In our paper, we capture difficulties of class distribution in real datasets by considering four types of minority class examples: safe, borderline, rare and outliers. First, we confirm their occurrence in real data by exploring multidimensional visualizations of selected datasets. Then, we introduce a method for an identification of these types of examples, which is based on analyzing a class distribution in a local neighbourhood of the considered example. Two ways of modeling this neighbourhood are presented: with k-nearest examples and with kernel functions. Experiments with artificial datasets show that these methods are able to re-discover simulated types of examples. Next contributions of this paper include carrying out a comprehensive experimental study with 26 real world imbalanced datasets, where (1) we identify new data characteristics basing on the analysis of types of minority examples; (2) we demonstrate that considering the results of this analysis allow to differentiate classification performance of popular classifiers and pre-processing methods and to evaluate their areas of competence.

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

Krystyna Napierala
krystyna.napierala@cs.put.poznan.pl

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

Finally, we highlight directions of exploiting the results of our analysis for developing new algorithms for learning classifiers and pre-processing methods.

Keywords Class-imbalanced data · Learning classifiers · Data difficulty factors · Local analysis · k-nearest neighbourhood

1 Introduction

In many real life problems classifiers are faced with imbalanced data, which means that one of the target classes contains a much smaller number of instances than the other classes. For example, in detection of fraudulent telephone calls or credit card transactions the number of legitimate transactions is much higher than the number of fraudulent ones. A similar situation occurs in many medical problems, where the number of patients requiring special attention (e.g., therapy or treatment) is much smaller than the number of patients who do not need it. Class imbalances have been also observed in many other application problems such as detection of oil spills in satellite images, analysing financial risk, predicting technical equipment failures, managing network intrusion, text categorization and information filtering; for some reviews see, e.g. (He and Garcia 2009; He and Ma 2013). In all those problems the correct recognition of the minority class is of key importance. However, class imbalance is an obstacle for learning classifiers as they are biased toward the majority classes and tend to missclassify minority class examples.

The class imbalance problem has been receiving a growing research interest in the recent decade and several methods have been introduced; for their reviews see, e.g., (He and Garcia 2009; He and Ma 2013; Chawla 2005). They are usually divided into *data level pre-processing methods* (that rely on transforming the original data to change the distribution of classes, e.g. by re-sampling) and *methods modifying the algorithms*. Although several specialized methods already exist, the identification of conditions for the efficient use of a particular method is still an open research problem. In our opinion, it is related to more fundamental issues of better understanding the nature of the imbalanced data, key properties of its underlying distribution and their consequences.

Note that in most of experimental evaluations, where existing or newly introduced methods are compared (as e.g (Batista et al. 2004; Garcia et al. 2007; Van Hulse et al. 2007)), datasets are categorized with respect to the global ratio between imbalanced classes or to the size of the minority class only. Authors usually do not consider other more complex characteristics of data distributions. Nevertheless, it seems that these two factors do not sufficiently explain differences in classification performance of the compared methods. For instance, for some datasets even with a high imbalance ratio, the minority class can be sufficiently recognized by many standard classifiers.

Some researchers have already shown that the global *imbalanced ratio* between classes is not a problem itself and it may not be the main source of difficulties. The degradation of classification performance is linked to other factors related to data distribution, such as decomposition of the minority class into many rare sub-concepts playing a role of small disjuncts (Japkowicz 2001, 2003; Ting 1994; Weiss and Hirsh 2000), the effect of too strong overlapping between the classes (Prati et al. 2004b; Garcia et al. 2007), or a presence of too many minority examples inside the majority class regions (Napierala et al. 2010). 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. 2013; Napierala et al. 2010).

The role of the above mentioned factors has been usually examined with special artificial datasets, where the data distribution is given a priori and the impact of each factor can be precisely controlled. Although these studies give an insight into the important aspects of data distribution, their conclusions might not be easy to apply in the real-world settings, as it is not evident how to estimate the occurrence of these data factors in the real-world datasets. Up to now, only few methods have been proposed for the real data (see their review in Section 2). However, using them is often a non-trivial task and does not lead to clear results. For instance, clustering algorithms, applied to look for small disjuncts (Jo and Japkowicz 2004), are difficult to parametrize. Other proposals, as e.g. (Saez et al. 2015; Denil and Trappenberg 2011), are based on combinations of complex methods and strongly rely on a particular classifier – thus their applicability for other learning algorithms is limited. Moreover, these method address a single data factor and do not deal with the co-occurrence of other factors.

We claim that current studies on data difficulty factors and methods for their identification are still not sufficiently developed. In this paper, following our earlier studies (Napierala and Stefanowski 2012b), we hypothesize that most of the data difficulty factors can be approximated by analysing the *local characteristics* of the minority examples. Depending on it, and also inspired by earlier considerations of some pre-processing methods (Kubat and Matwin 1997; Laurikkala 2001), we will distinguish different types of examples creating the minority class distribution: *safe*, *bordeline*, *rare* examples and *outliers*. The last three types correspond to *unsafe* examples, which are more difficult to learn. Unlike many related studies, which are focused on examining a single data factor only, we hypothesize that in real world data a mixture of these four types of examples occurs.

In Section 4 we will show how to verify their occurrence in real world data by adapting two visualization methods (multidimensional scaling (Cox and Cox 1994) and t-SNE (van der Maaten and Hinton 2008)), which project multidimensional data into a two-dimensional space. However, these methods allow us to analyze the datasets visually.

Therefore, the next aim of our study is to propose a new method for an automatic identification of these four types of examples in the real-world datasets. Unlike the previous proposals, we want to consider a simple, intuitive method, which is more universal, as it does not depend on a particular classifier. Thus, we propose a method based on analysing the mutual positions of learning examples from different classes in the *local neighbourhood* of minority examples. Depending on the number of majority class examples in this neighbourhood of the minority example, we will evaluate how safe or unsafe this example is. Such a neighbourhood could be modeled in different ways. We will present two approaches based either on k-nearest neighbours or kernels built around minority examples.

Although our approach to model neighbourhoods is based on using known methods, we claim that such an approach has not been considered yet. Furthermore, it can be applied to several crucial aims for learning classifiers from imbalanced data:

1. To analyse internal characteristics of real-world datasets, often exploited in the related experimental studies, to show their differences, which have not been discussed yet.
2. To carry out more advanced experimental studies with popular classifiers as well as pre-processing methods. Considering observations from the analysis of data characteristics could help in better indicating differences in their prediction performance and to establish their areas of competence.
3. To construct new, specialized algorithms for improving classifiers learned from imbalanced data, which will take into account the local characteristics of the dataset.

In this paper we will carry out a comprehensive experimental study, applying our method to 26 real imbalanced datasets often considered in the related studies. With respect to the above aims we will analyse the differences in proportions of the types of minority examples in these datasets. Then, we will try to relate these data characteristics to the performance of basic classifiers and pre-processing methods, to show which types are the most difficult and which methods are sensitive to the particular types of examples. Our experiments should significantly extend the most related studies (Batista et al. 2004; Van Hulse et al. 2007; Batista et al. 2012; Lopez et al. 2013) as we take a different research perspective and perform a more detailed analysis of difficulties in the distribution of the minority class. In particular, it should focus the reader attention on the role of analysing the local data characteristics, which has not be sufficiently studied yet.

Furthermore, we hope that our study will inspire a future development of more advanced methods for studying data difficulty factors and conducting experimental studies in a more insightful way. Finally, although it is not the main aim of this paper, we will also briefly discuss the consequences of our study for developing new generalizations of ensembles and pre-procesing methods.

To sum up, the main contributions of this paper will be:

- focus researchers' attention on the local characteristics of the diversified distribution of the minority class and its influence on learning from imbalanced datasets,
- consider four types of minority examples: *safe*, *border*, *rare* and *outlier*,
- propose a method for identifying these types of examples in the real-world data,
- analyse the characteristics of the real-world datasets commonly used in the works concerning class imbalance,
- carry out a comprehensive study on real-world datasets, relating the data characteristics to the performance of classifiers and pre-processing methods.

The paper is organized as follows. The next section summarizes related works on data difficulty factors. Motivations for considering types of minority examples are discussed in Section 3 and supported by visualizations of several datasets in Section 4. In Section 5, the neighborhood-based method for identification of types of examples is introduced and validated in the experiments with simulated data distributions. An experimental analysis with real datasets in carried out in Section 6. It is followed by a comprehensive study of basic classifiers and pre-processing methods in Section 7, where we attempt to relate their performance to the discovered characteristics of the diversified minority class distributions. Section 8 discusses possible options of using the local information, in particular for ensembles. The final section draws conclusions and highlights future research directions of applying the proposed method for imbalanced data.

2 Related works

In this section we focus on the most related studies concerning the properties of imbalanced data and their consequences for learning classifiers or pre-processing methods. For a more comprehensive review of various methods proposed to deal with class imbalance, the reader is referred to He and Garcia (2009), He and Ma (2013), Weiss (2004), and Chawla (2005).

2.1 Studying data factors with artificial datasets

It has been shown that when the dataset is imbalanced, standard classifiers encounter difficulties while recognizing the minority class. Nevertheless, the discussion of reasons on the data level still goes on. Some researchers analysed the relationship between the *imbalance ratio* (defined as the ratio of the minority class examples to the total number of examples in the data) and the classification performance, showing that its high values deteriorate the evaluation measures (see, e.g., (Grzymala-Busse et al. 2004; Japkowicz and Shah 2011; Weiss and Provost 2003)). However, it has also been observed that in some problems characterized by strong imbalance (e.g., Sick, or New Thyroid datasets from the UCI repository), standard classifiers are capable to be sufficiently accurate. This shows that the class imbalance ratio is not the only factor that impedes learning and more systematic studies have been undertaken to examine which properties of the distribution of examples in the attribute space are other critical factors.

Japkowicz et al. carried out a large number of experiments with simulated data (generated over numerical attributes) studying the relationship between the fragmentation of the class, the size of the training set and the class imbalance ratio (Japkowicz 2003; Japkowicz and Stephen 2002; Jo and Japkowicz 2004). By manipulating with degrees of these data factors, their influence on the recognition of minority classes was analysed. Their results show an important role of sparsity of the minority class when it is decomposed into very small sub-groups. It is linked to the problem of *small disjuncts* (known from symbolic learning (Holte et al. 1989)). However, the detection of such small disjuncts in the real world data is not easy. Another comparative study with artificial data and various sampling methods gave recommendation to using a special cluster-based technique which takes into account a fragmentation of classes (Jo and Japkowicz 2004).

Other researchers studied the role of *overlapping* between the minority and majority classes. In (Prati et al. 2004b), the authors used the artificial, numerical datasets where minority and majority classes formed two spherical clusters, and considered the C4.5 classifier with respect to the AUC measure. By changing the imbalance ratio and the distance between the clusters, they noticed that increasing class overlapping was more influential than increasing class imbalance, leading to stronger deterioration.

A similar experiment, but concerning six classifiers compared with more evaluation measures, was carried out in (Garcia et al. 2007). For two-dimensional artificial, numerical datasets the degrees of overlapping and imbalance ratio were systematically changed. It was shown that increasing overlapping between the classes degraded the recognition of the minority class more than changing the imbalance ratio. However, it affected various classifiers in a different degree. In case of a very high overlapping, nearest neighbour classifier performed the best, while support vector machine (SVM) was the worst classifier. In the additional experiment, the authors noticed that the local imbalanced ratio inside the overlapping area is more influential than the global one. The same experimental setup was then used to analyse in more detail the kNN classifier, with k changing from 1 to 15 (Garcia et al. 2008). It showed that when the overlapping increased, more local classifiers (with smaller k) performed better on the minority class.

In Stefanowski (2013) the effect of overlapping was studied together with other factors such as decomposition of the minority class into smaller sub-concepts. The experiments were carried out on artificial two-dimensional datasets with more complicated nonlinear borders and the results showed that the combination of class decomposition with overlapping makes learning very difficult.

Finally, two other teams studied the influence of noisy examples (Anyfantis et al. 2007; Khoshgoftaar and Van Hulse 2009). They also used special data where class noise was introduced to real-world datasets by randomly re-labelling the learning examples. The experimental results showed that all compared classifiers were sensitive to noise and it affected stronger the minority class. However, some of them, as Naive Bayes and nearest neighbour, were often more robust than other, more complex classifiers.

2.2 Studies with real datasets

Only few similar systematic studies concern real-world datasets. These are comparative studies of several algorithms, where data are differentiated with respect to the imbalance ratio and the data size only

Van Hulse et al. carried out a comprehensive study with 35 real-world datasets, 11 classifiers and 7 pre-processing methods in (Van Hulse et al. 2007). They grouped their datasets into 4 categories with respect to the imbalance ratio and compared the learning algorithms within these categories. According to the authors, random undersampling worked better than other approaches for data with the most severe imbalance ratio ($< 5\%$). Unlike other studies, they claimed that simpler random re-sampling often performed better than more sophisticated informed re-sampling methods. Having many experimental configurations, they also noted that algorithms respond differently to various pre-processing methods and it depends on the evaluation measures (e.g., G-mean or F-measure showed higher improvements than AUC).

Yet another study concerning impact of the imbalance ratio was carried out in Batista et al. (2012), where 7 learning algorithms were compared over 20 real-world datasets with the AUC measure. Averaging the results for all the datasets, the authors observed that the loss of performance started to be significant when the minority class represented 10% of the data or less. SVM was less affected by changing the class imbalance ratio than other classifiers for all except the most imbalanced distributions. Then, they analysed the performance of two pre-processing methods, random oversampling and SMOTE, and concluded that the pre-processing methods usually could not improve the performance by more than 30%.

In Batista et al. (2004) Batista et al. developed another wider systematic experimental study with 15 real-world UCI datasets and 10 different pre-processing methods – all used with the C4.5 decision trees. The oversampling methods provided better AUC than the undersampling ones. Considering data factors, the authors took into account the data size and claimed that the pre-processing SMOTE method (Chawla et al. 2002) combined with informed undersampling (ENN or Tomek links), led to the best results for smaller data with few minority examples while simple random oversampling was competitive to other methods for datasets containing a relatively high number of the minority examples.

2.3 Limitations of identification methods

To summarize the works presented in previous sections, the studies with real-world datasets concentrate mostly on simple data factors, such as the data size and the imbalance ratio, as they can be directly measured in the data. The authors usually do not consider several data factors occurring together and do not study the influence of more complex factors. Studies on such factors are usually done only with artificial data, in which the distribution of the data is known a-priori and can be precisely controlled. These studies have clearly demonstrated that data factors, such as overlapping or noise, have a crucial impact on the performance of classifiers and pre-processing methods. However, their conclusions might not be easy to directly apply in the real-world settings, as the authors do not sufficiently discuss how to identify the occurrence of these data factors in the real-world datasets.

The identification methods are not numerous and they are usually complex, rely on a specific classifier or tuning their parameters is not easy. For instance, consider the decomposition of the minority class into rare sub-concepts. Up to now researchers have mainly applied clustering to look for sub-concepts in the data. Usually, k-means algorithm is used either to the minority class only or independently for each class (Jo and Japkowicz 2004). The main open question is how to identify the appropriate number of expected clusters, as in most cases the underlying class distribution is unknown and rather complex. Attempts to approximate overlapping are also quite rare and rely on a chosen classifier. For instance, in (Denil and Trappenberg 2011) (paper concerning the effects of overlapping and imbalance on the SVM classifier), the degree of overlapping in real-world datasets is estimated by measuring the number of support vectors (examples) which can be removed without deteriorating the classification accuracy. Methods for identifying unsafe or noisy examples are also quite complex and not intuitive to parametrize. For instance, there are two specialized approaches based on classifier ensembles that analyse the distribution of component classifier predictions (Khoshgoftaar and Van Hulse 2009; Saez et al. 2015). The most often misclassified examples are treated as possible noise and iteratively removed from the learning data until a certain value of accuracy is achieved. These methods depend on a number of parameters and a choice of a particular type of base classifier. What is more, removing the examples can be debatable especially for the minority class.

To sum up, there is a limited number of methods for the identification of data difficulty factors in real-world data sets and they focus on single factors only. Developing new methods, able to identify several data factors, not relying on a particular classifier and simpler to parametrize, would have a positive impact on practical aspects of learning from imbalanced data. If, for example, the results presented in Garcia et al. (2007) suggest that the performance of some classifiers is seriously downgraded when there is a very high overlapping of classes, then it would be interesting to know if such strong overlapping often occurs in real-world applications and to resign from applying a particular learning algorithm to a dataset.

Finally, by analysing a representative collection of real-world imbalanced datasets, we could observe what are the most common data distribution patterns. Such knowledge would help to point out the most promising directions for the development of new methods dedicated for class imbalance. Today state-of-the-art methods, either on data level or on algorithmic level, concentrate mostly on compensating the effects of the global imbalance ratio, not taking into account other, possibly more influential, data difficulty factors.

3 Distinguishing types of examples

In our study we claim that most of the above mentioned data difficulty factors can be linked with different types of examples forming the minority class distribution.

We remark that the diverse role of examples has already been noticed in a few pre-processing methods (Kubat and Matwin 1997; Laurikkala 2001; Han et al. 2005; Stefanowski and Wilk 2008). Following (Kubat and Matwin 1997; Laurikkala 2001), the most common distinction is between safe and unsafe examples. *Safe* examples, located in the homogenous regions populated by the examples from one class only, should be easier to learn by a classifier, while unsafe ones are considered to be more difficult and more likely to be misclassified.

In Kubat and Matwin (1997), unsafe examples are further discriminated between *borderline* and *noisy* examples. The term *borderline* has been assigned to examples located in the regions around decision boundary between classes. In our further considerations it is referred to two kinds of examples. Firstly, these are examples located inside overlapping regions of minority and majority classes. Secondly, these are also examples placed close to the complex, decision boundary between the classes, which could be misclassified by their neighbours from the opposite class located on the other side of the boundary.

Noisy examples (either referring to class or attribute errors) deteriorate the performance of standard classifiers and they are also particularly harmful for the minority class (Anyfantis et al. 2007; Khoshgoftaar and Van Hulse 2009). In the standard approaches, when the data is balanced, they could be specially handled inside the learning algorithm or removed during the pre-processing (Brodley and Friedl 1999). However, in imbalanced data more caution is advised. While single majority examples located inside the minority class may increase its fragmentation and cause additional difficulties in learning (so they could be processed as above), the situation is different for the single minority examples. Here, it is necessary to distinguish *outliers* from possible noise as such distant minority examples might often be a result of the insufficiently covered example space. As the minority class can be underrepresented in the data, we claim, similarly to Kubat and Matwin (1997), that these examples can often be outliers, representing a rare but valid subconcept of which no other representatives could be collected for training. Therefore, they should not be removed or re-labeled. We can also refer to studies on medical problems (Gamberger et al. 1999), where the authors showed that the results of noise identification filters were often identified by experts as valid outliers.

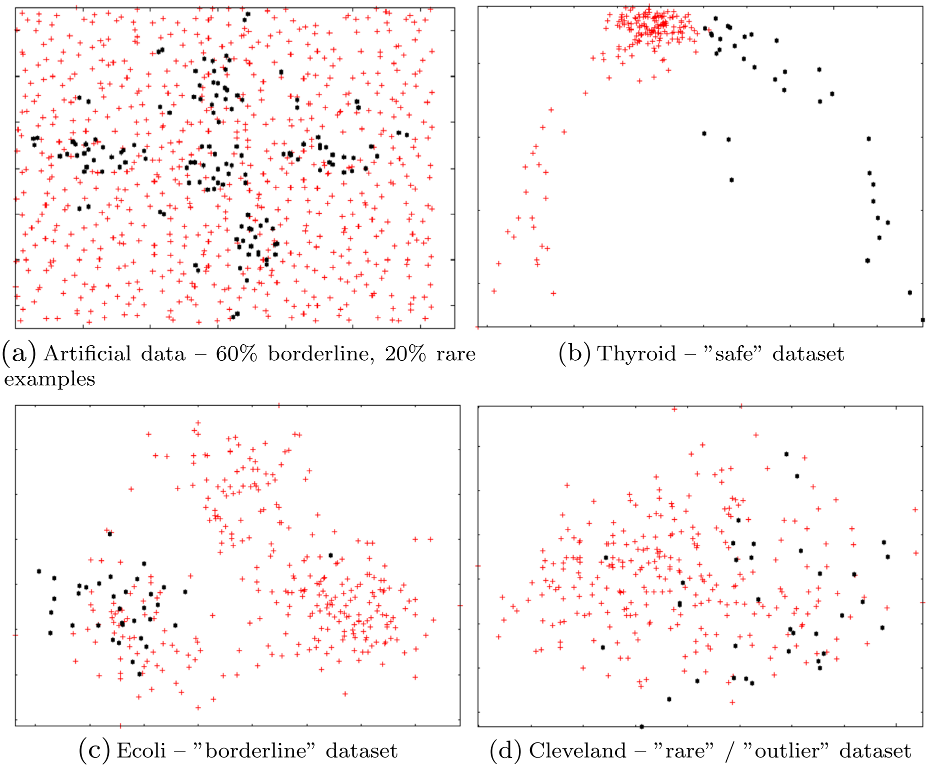
Finally, we want to distinguish yet another type of, so-called, *rare examples*. These are isolated pairs or triples of minority class examples, located in the majority class region, which are distant from the decision boundary so they are not borderline examples, and at the same time are not singular examples, so they are not exactly outliers. The role of these examples has been preliminary studied by us in the experiments with special artificial datasets (Napierala et al. 2010; Stefanowski 2013), where they strongly influenced the performance of classifiers.

To sum up, in this paper we will propose to relate the considered properties of imbalanced data distributions to four types of minority examples: safe, borderline, rare and outlying examples.

4 Visualization of imbalanced datasets

To verify whether these four types of minority examples can be observed in real-world datasets, we will use the visualisation methods, which project multi-dimensional data points into the low-dimensional space such that the structural properties of the data are preserved. Firstly, we choose the Multidimensional Scaling (MDS) as it is one of the most popular method for projections into new created dimensions. It performs a linear mapping of dimensions with the aim of preserving the pairwise distances between data points in the original high dimensional data space into the projected low dimensional space (Cox and Cox 1994). Moreover, we decided to confirm our analysis by another projection method based on a different principle, which uses a non-linear mapping. Among such non-linear methods, t-SNE method (*t-Distributed Stochastic Neighbour Embedding*) is one of the most recent dimensionality reduction methods, which does not concentrate on preserving *all* the pairwise distances, but it puts more emphasis on preserving *local* distances to keep similar examples together, rather than on preserving the exact distances between dissimilar examples (van der Maaten and Hinton 2008). According to the experiments in van der Maaten and Hinton (2008), t-SNE is capable of capturing much of the local structure of the high-dimensional data very well, while also revealing the global structure such as the presence of clusters. As considered UCI datasets have both numeric and nominal attributes, we calculate distances between the examples using the HVDM metric (Wilson et al. 1997) – its justification is given in the next section.

Due to space limits, we present the visualisations after the MDS projection of three imbalanced datasets from the UCI repository, often used in the experimental studies concerning class imbalance: thyroid, ecoli and cleveland (Fig. 1b, c and d). For these datasets, the percentage of preserved variance was high enough to analyse the projected data. Looking at Fig. 1b, c and d, one can notice that the three datasets are of different nature. In thyroid dataset (Fig. 1b), the classes are clearly separated (even linearly), so most of the minority examples represent safe examples. In ecoli dataset (Fig. 1c) the classes seriously overlap. The consistent region belonging solely to the minority class (on the very left) is rather small. Most examples lie in a mixed region between the classes. Finally, the cleveland dataset (Fig. 1) has an even more complex distribution, as the minority class is very


(a) Artificial data – 60% borderline, 20% rare examples

(b) Thyroid – "safe" dataset

(c) Ecoli – "borderline" dataset

(d) Cleveland – "rare" / "outlier" dataset

Fig. 1 MDS visualisation of selected imbalanced datasets

scattered – the examples form very small groups of few examples and there are some singular observations, surrounded by the opposite class. This dataset consists mostly of rare examples and outliers.

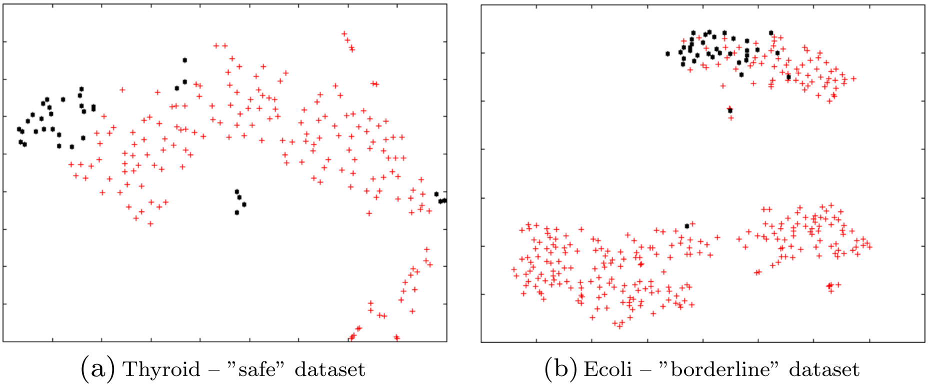
Figure 2 presents the results for thyroid and ecoli datasets after the t-SNE projection (run with the default parameters). The cleveland dataset is not used in this comparison as the implentation of t-SNE method does not handle nominal attributes. It can be observed that the dimensions to which the datasets were projected are different, e.g. for ecoli, the t-SNE visualisation is rotated. Also, the mutual positions of examples differ – the three clusters in the t-SNE projection of ecoli dataset are better separated than in the MDS method (which is consistent with the assumptions of t-SNE), and in the thyroid dataset the minority class forms several clusters instead of one. However, for both datasets the principle observations of distribution characteristics remain the same: in the thyroid dataset the classes can be easily separated, while in the ecoli dataset, in one of the clusters the examples from both classes strongly overlap.

5 Identifying types of examples

5.1 Motivations

In previous sections we distinguished four types of examples. The visualisation methods can help to inspect the distribution of examples in some real-world datasets and could confirm the occurrence of these types of minority examples. However, the applicability of these methods is limited. First, they are not applicable to very large datasets, as visualisations of thousands of points would be difficult to read. Secondly, the projection may need more than two dimensions. For instance, we were unable to visualize the imbalanced dataset hepatitis, as MDS with two dimensions preserved only 25% of variance in the dataset. Therefore, there is still a need for more flexible methods, which can identify types of examples in a quantitative way.

We would like to propose a universal method, which evaluates the occurrence of all four types of examples defined in Section 3 at once. We want to keep the method simple


(a) Thyroid – "safe" dataset

(b) Ecoli – "borderline" dataset

Fig. 2 T-SNE visualisation of selected imbalanced datasets

and intuitive to the user. Furthermore, we prefer the method, which is not directly related to a particular classifier, but relies on the natural distributions in the data. As described in Section 2, no such method has been proposed yet in the literature. These expectations and the way of defining types of examples in Section 3 lead our interests to analysing the mutual positions of the learning examples in the attribute space. This could allow us to assess the type of example by studying class labels of the other examples in its *local neighborhood*.

We think that analysing a ”local” distribution of examples may be better suited for this task than ”global” approaches, especially when the minority class is considered, as this class is often decomposed into smaller subconcepts with difficult, nonlinear borders between the classes (see results of visualizations in Section 4). What is more important, similar approaches to such ”local” analysis have already been used in some pre-processing methods dedicated for class imbalance (such as OSS (Kubat and Matwin 1997), NCR (Laurikkala 2001), SMOTE (Chawla et al. 2002) or SPIDER (Stefanowski and Wilk 2008)). Furthermore, local neighborhoods are also a basis for some density cluster algorithms, as e.g. DBSCAN (Ester et al. 1996), that are well suited for detecting difficult concept shapes as well as for discriminating them from noise or outliers.

The neighborhood of the example could be modeled in different ways. In this paper we will propose two different approaches based on k-nearest neighbour and on kernel functions.

The analysis of class labels of examples in the former approach concerns a fixed number of nearest examples (without taking into account their distances to the “seed examples”) while in the other approach all examples within a given radius are taken into account together with their distances. In further considerations and experiments we will use the first approach. It results both from the previous assumption of keeping the identification method simple and universal and inspirations from related experimental studies (as e.g., (Batista et al. 2004)). However, in Section 5.3 we will also present an approach based on kernel functions and then evaluate it experimentally in Section 6.4.

5.2 Modeling k-neighbourhood

To identify the type of example, we analyse the class labels of their k-nearest neighbours. Note that a proper choice of the value k and the distance function is needed for constructing this type of neighborhood. We will come back to this problem after introducing the general idea of our approach.

5.2.1 Labelling types of examples

For simplicity let us consider the neighbourhood of a fixed size $k = 5$ and the HVDM metric (*Heterogenous Value Difference Metric*) (Wilson et al. 1997) to calculate the distance between the examples (we will then justify these choices for further experiments). With $k = 5$, the proportion of neighbours from the same class against neighbours from the opposite class can range from 5:0 (all neighbours are from the same class as the analysed example) to 0:5 (all neighbours belong to the opposite class). Depending on this proportion, we propose to assign the labels to the minority examples, representing the four distinguished types, in the following way:

- 5:0 or 4:1 – an example is labelled as a safe example (further denoted as S).
- 3:2 or 2:3 – a borderline example (denoted as B). The examples with the proportion 3:2 are correctly classified by its neighbours, so they might still be safe. However, the number of neighbours from both classes is approximately the same, so we assume that this example could be located too close to the decision boundary between the classes.
- 1:4 – labelled as a rare example (denoted as R), only if its neighbour from the same class has the proportion of neighbours either 0:5 or 1:4 (additionally, in case of 1:4, it must point to the analysed example). Otherwise there are some other examples from the same class in the proximity (although not in the immediate surrounding of $k = 5$), which suggests that it could be rather a borderline example B.
- 0:5 – an example is labelled as an outlier and denoted as O.

This kind of labeling examples based on analysing proportions of class labels can be extended for higher values of k. For instance, types of examples with $k = 7$ will be following: proportions of the neighbours 7:0 or 6:1 or 5:2 – a safe example; 4:3 or 3:4 – a borderline example; 2:5 or 1:6 – a rare example; 0:7 – an outlier. For higher values of k we propose to adapt thresholds defined for the kernel approach (see Section 5.3).

5.2.2 Neighbourhood parametrization

An important issue is choosing an appropriate k value. In general, different values may be considered. Values smaller than 5, e.g. $k = 1$ and $k = 3$, may poorly distinguish the nature of examples, especially if we want to assign them to four types. Too high values, on the other hand, would be inconsistent with our assumption of the locality of the method (see the discussion in Section 5.1 why the locality is important for analysing complex minority class distributions in imbalanced data). To verify more precisely whether the parameter k could strongly influence the results of labelling minority examples, we will carry out an additional sensitivity analysis in Section 6. Revealing earlier these results, they will show that proportions of identified types of examples are quite stable while changing k values. Yet another issue concerns possible tuning of the size of the neighbourhood for each dataset individually. For instance, in case of constructing the standard kNN classifier for balanced data it could be done with respect to the data cardinality, see e.g. (Goldstein 1972). However, we think that for imbalanced datasets, complex data factors are more influential than the minority class cardinality only, so the locality of the neighbourhood may be important regardless of the dataset size. For these reasons, we have decided to stay with $k = 5$. It has also been inspired by earlier experiments with the related pre-processing methods for class imbalance (see e.g. typical options for running SMOTE (Chawla et al. 2002)).

As the distance measure we have decided to choose the HVDM metric (Wilson et al. 1997), as it provides more appropriate handling of a mixture of numerical and qualitative attributes. It belongs to special *heterogeneous distance functions* which aggregate normalized distance functions for numerical attributes with 0-1 functions for qualitative ones (Wilson et al. 1997). Instead of a simple value matching (used, e.g., in HOEM or Gower (Wilson et al. 1997)), HVDM makes use of the class information to compute attribute value conditional probabilities by means of the Stanfil and Valtz value difference metric for nominal attributes (Stanfill and Waltz 1986). For numeric attributes, it uses a normalized Euclidean distance.

We will not experimentally test other functions as according to the literature review, HVDM performs better than many other heterogeneous functions – see, e.g., the comparative study (Lumijarvi et al. 2004; McCane and Albert 2008).

5.3 Kernel based local neighbourhood

Notice that an alternative approach to fixing the number of neighbours is to fix the local area around the example and to estimate the number of neighbours and their class labels within it. It gives rise to kernel approaches (Bishop 2006).

In such an approach, a kernel function is used to determine which neighbours should be taken into account. Moreover, due to the form of the function, different weights (probabilities) could be assigned to the neighbours, based on their distance from the analysed minority example x. In preliminary experiments (Napierala 2013) we have considered several functions and have decided to apply the commonly used Epanechnikov function (see its definition in (Bishop 2006) and its illustration in Fig. 3), which gives more weight to the neighbours closer to the example x. Let us mention here that we have also experimentally tested other functions, such as Gaussian, triangular or uniform functions, but they did not influence the results too much.

We propose to set the width of the Epanechnikov function (which determines the maximum distance up to which the examples are treated as neighbours) for each dataset separately. It is equal to the average distance to the 5^{th} neighbour of each minority example in the dataset, to keep the average number of analysed neighbours comparable to the one used in our k neighbourhood method.

Given the definition of the kernel function we estimate a weighted sum of all minority neighbours, where weights depend on the distance from the analysed example. Comparing it to the weighted sum for the majority class neighbours we can estimate the probability that the analysed example x could belong to the minority class $p(C_{min}|x)$. To assess the type of a minority example, we need to discretize the range of this value into four subintervals. We

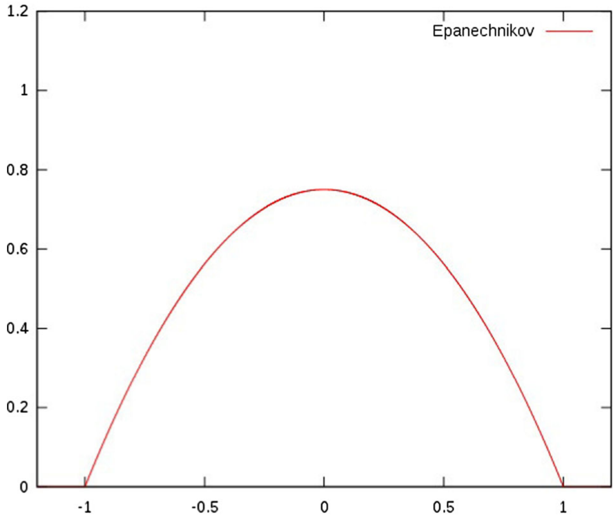


Fig. 3 Epanechnikov kernel function

propose to use the thresholds which could lead to the analogous labelling as in the previous method based on the proportions in k nearest neighbours.

Note that for $k = 5$ the proportion of neighbours 3:2 (in case of which we treat an example as borderline) is equivalent to the distribution estimation $p(C_{min}|x) = \frac{3}{5} = 0.6$, while proportion 4:1 (safe example) is equivalent to the distribution estimation 0.8. Interpolating between these values, we can say that our method labels the example as safe if its distribution is greater than 0.7. Following this schema, we propose to use the following rules: if $1 \geq p(C_{min}|x) > 0.7$ then label x as safe; if $0.7 \geq p(C_{min}|x) > 0.3$ then label x as borderline; if $0.3 \geq p(C_{min}|x) > 0.1$ then label x as rare; if $0.1 \geq p(C_{min}|x) > 0$ then label x as outlier.

These rules are also applicable for establishing the proportions of examples in our standard k-nearest-neighbour method given in Section 5.2.1, for numbers of neighbours higher than k=5.

5.4 Validation of the method with artificial data

The presented k neighbourhood method is based on a simple analysis of a fixed number of k neighbours. To check whether the assigned labels can precisely reflect the known distribution of examples, we verify it with the artificial datasets.

Inspired by a good experience with such data in (Napierala et al. 2010; Stefanowski 2013), we generated a number of such datasets, containing 800 examples described by 2 numerical attributes. The minority class forms elliptical subconcepts, surrounded by uniformly distributed majority class examples. The datasets are characterized by various imbalance ratios (from 1:5 to 1:9) and a different number of the minority class sub-concepts (from 1 to 5). In these datasets we changed the percentage of safe, borderline, rare and outlying minority examples. Table 1 presents the description of several analysed datasets and the labelling results.

The first three datasets are disturbed in the same way (60% of borderline examples and 20% of rare examples), but differ in the number of sub-concepts. One of them (with 5 sub-concepts) is plotted in Fig. 1a. Proportions of the identified labels show that our labelling method can correctly reconstruct the percentage of safe, borderline and rare examples, regardless of the number of sub-concepts. The other three datasets contain 10% of outliers and differ according to the imbalance ratio. Here, the labels also correctly reflect the percentage of outliers regardless of the changing imbalance ratio. However, although the classes in these datasets are not overlapped, a considerable number of examples is labelled

Table 1 Labelling of artificial datasets

Dataset Description					Identified Label			
Imbalance Ratio	Sub- concepts	Borderline [%]	Rare [%]	Outlier [%]	Safe [%]	Borderline [%]	Rare [%]	Outlier [%]
1:5	1	60	20	0	17.04	60.74	21.48	0.74
1:5	3	60	20	0	18.52	57.78	23.70	0.00
1:5	5	60	20	0	17.78	64.44	17.78	0.00
1:5	5	0	0	10	64.44	25.93	0.00	9.63
1:7	5	0	0	10	54.00	36.00	0.00	10.00
1:9	5	0	0	10	52.00	36.00	2.00	10.00

as borderline. Let us recall that by these examples we understand also the examples lying close to the borderline between the classes, even when there is no overlapping. The examples close to the border between the classes can contain in their neighbourhood some examples from the opposite class, so our labelling method will also assign them to the "borderline" category.

6 Experimental study – analysing real-world datasets

6.1 Experimental setup

The main aim of this part of experiments is to analyse the distribution of minority class examples in real world imbalanced datasets. Our hypothesis is that depending on this distribution, imbalanced datasets can constitute a different degree of difficulty.

First, we will check whether the studied datasets could be grouped into different categories according to the type of examples which prevail in the minority class. Then, in the next section we study the performance of popular classifiers as well as of pre-processing methods in each group of datasets separately.

The characteristics of 26 analysed imbalanced datasets is presented in Table 2. We have chosen the datasets which have been often studied in the most related experimental studies. They represent different sizes, imbalance ratios, domains and have both continuous and nominal attributes. Most of them come from the UCI repository.[1] Four datasets are retrospective medical datasets which were used in the earlier works of Stefanowski et al. on class imbalance.[2] In datasets with more than one majority class, they are aggregated into one class to have only binary problems, which is also typically done in the literature.

For the second part of experiments (Section 7), the classification performance will be evaluated by measures appropriate for class imbalance, i.e. *Sensitivity* (TPR – true positive rate, Recall or accuracy of the minority class), G-mean and F-measure. The G-mean aggregates Sensitivity of the minority class with its Specificity (for a binary classification specificity is also an accuracy of the majority class) and evaluates the balance between them. F-measure aggregates Precision with Recall - Sensitivity of the minority class. For their definition and justification see, e.g., (He and Garcia 2009; He and Ma 2013; Japkowicz and Shah 2011). We do not use the AUC (an area under ROC curve) measure, because most of the classifiers compared in our study give deterministic predictions while AUC reflects much better the performance of probabilistic or scoring classifiers (Japkowicz and Shah 2011). To estimate the selected measures, we use a stratified 10-fold cross validation repeated 5 times to reduce the variance of results.

To examine the importance of differences between the methods, we will apply a nonparametric ranked Friedman test ((Demsar 2006; Japkowicz and Shah 2011)), which globally compares the performance of several methods on multiple data sets with a null hypothesis saying that all methods perform equally. We also carry out a post-hoc analysis (a Nemenyi test (Japkowicz and Shah 2011)) of differences between the average ranks of classifiers. In both tests we will use a confidence level $\alpha = 0.05$. In some cases, where it could

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

[2] We are grateful to prof. W. Michalowski and the MET Research Group from the University of Ottawa for abdominal-pain and scrotal-pain datasets; and to prof. K. Slowinski from Poznan University of Medical Science for hsv and acl datasets.

Table 2 Characteristics of the datasets

Dataset	No of examples	Imbalance ratio [%]	No of attributes (numeric)	Minority class name
breast-w	699	34.47	9(9)	malignant
abdominal-pain	723	27.94	13 (0)	positive
acl	140	28.57	6 (4)	1
new-thyroid	215	16.28	5 (5)	hyper
vehicle	846	23.52	18 (18)	van
nursery	12960	2.53	8(0)	very-recom
satimage	4435	9.35	36(36)	4
car	1728	3.99	6 (0)	good
scrotal-pain	201	29.35	13 (0)	positive
credit-g	1000	30	20 (7)	bad
ecoli	336	10.42	7 (7)	imU
hepatitis	155	20.65	19 (6)	die
ionosphere	351	35.89	34 (34)	bad
haberman	306	26.47	3 (3)	died
cmc	1473	22.61	9 (2)	l-term
breast-cancer	286	29.72	9 (0)	rec-events
cleveland	303	11.55	13 (6)	positive
glass	214	7.94	9 (9)	v-float
hsv	122	11.48	11 (9)	4.0
abalone	4177	8.02	8 (7)	0-4 16-29
postoperative	90	26.66	8 (0)	S
seismic-bumps	2584	6.57	18(14)	1
solar-flare	1066	4.03	12 (0)	F
transfusion	748	23.8	4 (4)	yes
yeast	1484	3.44	8 (8)	ME2
balance-scale	625	7.84	4(4)	B

be interesting to compare more precisely the differences in performance of a given pair of classifiers, we will additionally refer to the Wilcoxon paired test. Unlike the Friedman test, which is based on rankings of many classifiers, the Wilcoxon test focuses on values of differences in performance of two classifiers (Demsar 2006; Japkowicz and Shah 2011).

6.2 Analysing types of minority examples

In this experiment we used k-neighbourhood method (with $k = 5$) to identify (label) types of the minority class examples in all the datasets. The results are presented in Table 3.[3] To facilitate the analysis, we have sorted the datasets from the “easiest” to the ”most difficult” (in terms of presence of unsafe examples in the data distribution).

[3] We abbreviate names of example types with their first capital letter.

Table 3 Labelling of datasets with respect to minority class examples and k-neighbourhood

Dataset	S [%]	B [%]	R [%]	O [%]
breast-w	91.29	7.88	0.00	0.83
abdominal-pain	59.90	22.28	8.90	7.92
acl	67.50	30.00	0.00	2.50
new-thyroid	68.57	31.43	0.00	0.00
vehicle	74.37	24.62	0.00	1.01
nursery	82.00	17.00	1.00	0.00
satimage	47.47	39.76	4.58	8.19
car	47.83	39.13	8.70	4.35
scrotal-pain	38.98	45.76	10.17	5.08
ionosphere	44.44	30.95	11.90	12.70
credit-g	9.33	63.67	10.33	16.67
ecoli	28.57	54.29	2.86	14.29
hepatitis	15.63	62.50	6.25	15.63
haberman	4.94	61.73	18.52	14.81
breast-cancer	24.71	25.88	32.94	16.47
cmc	17.72	44.44	18.32	19.52
cleveland	0.00	31.43	17.14	51.43
glass	0.00	35.29	35.29	29.41
hsv	0.00	0.00	28.57	71.43
abalone	8.36	20.60	20.60	50.45
postoperative	0.00	41.67	29.17	29.17
seismic-bumps	3.52	29.41	16.47	50.58
solar-flare	0.00	48.84	11.63	39.53
transfusion	18.54	47.19	11.24	23.03
yeast	5.88	47.06	7.84	39.22
balance-scale	0.00	0.00	8.16	91.84

The first observation is that most of the datasets contain minority examples of all four types. Moreover, a majority of datasets contains rather a small number of safe examples – only in the top six datasets (from breast-w to nursery) safe minority examples prevail and there are almost no rare or outlying examples. Some datasets, on the other hand, do not contain any safe examples – such as cleveland, glass, hsv, solar-flare or balance scale.

Datasets from satimage to ionosphere consist of safe and borderline minority class examples in quite comparable proportions and they do not have many rare or outlying examples. One could suspect that in these datasets a complicated border between the classes or some overlapping occurs.

Then, we can distinguish a group of datasets where borderline examples dominate in the distribution of the minority class – these are datsets from credit-g to haberman. A high number of borderline examples may suggest that there is a strong overlapping between the classes in these datasets.

Several datasets contain many rare minority class examples. Although they are not that numerous as borderline or safe examples, they can constitute even 20–30 % of the minority class. Datasets from haberman to postoperative have about 20 % or more of rare examples. Other datasets contain less than 10 % of these examples.

Finally, many datasets contain a relatively high number of outlier examples – datasets from cmc to balance-scale contain more than 20 % of these examples. Sometimes the outlying examples constitute more than a half of the whole minority class (see cleveland, hsv, balance-scale). This observation confirms the discussion in Section 3, in which we claimed that outlier minority examples cannot be treated entirely as noise. Finally, it is interesting to observe that for many datasets rare and outlying examples appear together.

Note that the results of this analysis are consistent with the observations of the MDS visualisations. The three datasets visualised in Fig. 1 b,c and d also show that new-thyroid contains mostly safe examples, ecoli has a lot of borderline examples, while cleveland constitutes mostly of rare and outlying examples.

We do not show such k neighbourhood analysis for the majority classes. However, we can summarize it by a contrary observation – nearly all datasets contain mainly safe majority examples (often over 90 %, e.g. yeast – 98.5 %, ecoli – 91.7 %) and sometimes a limited number of borderline examples (e.g. balance-scale – 84.5 % safe and 15.6 % borderline examples). What is even more important, nearly all datasets do not contain any majority outliers and at most 2 % of rare examples.

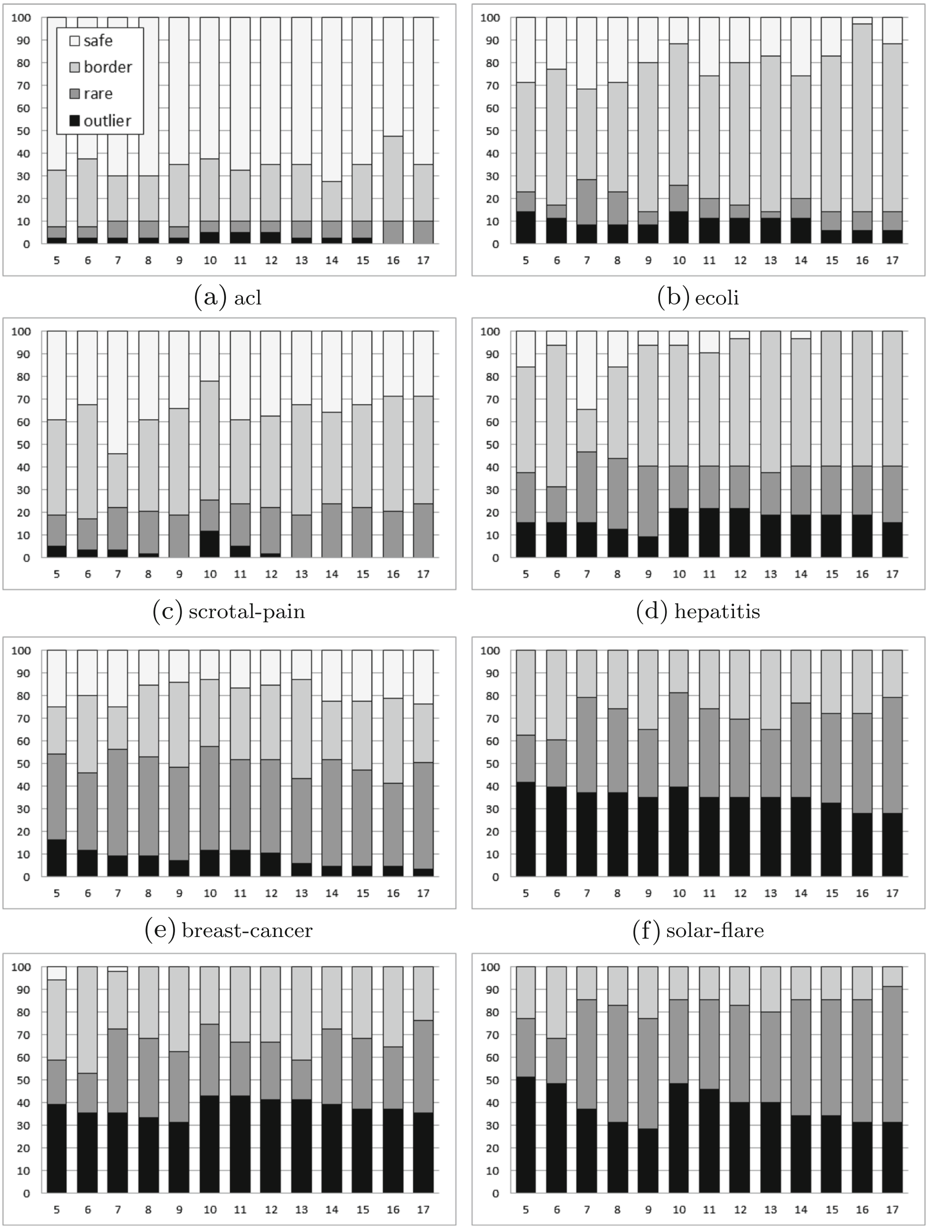
6.3 Studying influence of other neighbourhoods

Although the results of our proposed labelling method concur with the MDS and t-SNE visualisations, we would like to verify in more detail if the results presented in Table 3 are not related to the used identification method and its parametrization rather than to the distribution of the analysed datasets.

First, we will study the influence of increasing k value on the results of labelling types of the minority class examples. In order to assign the examples to the four types for higher k values, new thresholds have to be established. We have defined them in the same way as for the kernel method in Section 5.3.

We have analysed again the same 26 datasets. Due to space limits, we show only graphical summaries for selected datasets, see Fig. 4a b, c, d, e ,f, g, and h. Each graph presents structured bars corresponding to different values of the neighbourhood size. We consider $k = 5, \ldots, 17$. Each bar is stratified into 4 parts representing the percentage of the given type of example. By estimating how the share of each type of example varies between the bars, we can analyse the sensitivity of our method to the size of the neighbourhood.

We do not expect that the precise numbers (percentages of examples types) will be exactly the same for all considered k values. It is more interesting to check whether the distribution of the minority example types varies a lot depending on k. Looking at Fig. 4a, b, c, d, e, f, g, and h, one can notice that for most of the datasets, the differences in percentage of examples assigned to a given type (visible in parts of bars) are rather small (around 5-10%). For a few datasets we can sometimes observe a bit higher differences, as in hepatitis dataset with $k = 7$ (changes between safe and borderline categories – see Fig. 4d) or cleveland dataset (changes between rare and outlier categories – see Fig. 4h). Despite these shifts between particular categories (occurring only in few datasets among all studied ones),



(a) acl

(b) ecoli

(c) scrotal-pain

(d) hepatitis

(e) breast-cancer

(f) solar-flare

(g) yeast

(h) cleveland

Fig. 4 Distribution of types of examples depending on the size of neighbourhood k. X-axis – k. Y-axis – percentage of types of minority examples. Legend on chart 4a applies to all other charts

we observe that the general trends of type distributions are quite stable for all the datasets. Therefore, we can conclude that the results of our labelling method do not depend too much on the k parametrization.

6.4 Using kernel analysis

One could also ask whether analysing the local neighbourhood based on fixed *k* does not influence negatively the results, as the datasets might have different densities in different regions. Therefore, we compare the results of this method with an alternative kernel approach, described in Section 5.3, where different numbers of neighbours could appear inside the given width of the kernal function. Recall also that in this approach widths of kernel functions are also tuned individually for each dataset.

We remark that using this method, we have observed that some examples do not have any neighbours closer than *width*. In this case we assume that we do not have enough information about these examples and do not take them into account in our analysis. In practice, such examples constituted no more than few percents of the dataset. Only in the ionosphere dataset, there were more such examples, so the results for this dataset should be treated with caution. The labelling of examples based on the kernel approach is presented in Table 4.

Comparing the results in Tables 3 and 4, we can observe that using the kernel method does not change the results more than by 5–10 % for most of the datasets. Only in

Table 4 Labelling of datasets – the kernel density method

Dataset	S [%]	B [%]	R [%]	O [%]
breast-w	90.9	5.8	0.0	3.4
abdominal-pain	62.0	21.9	5.3	10.7
acl	72.2	22.2	0.0	5.6
new-thyroid	62.5	37.5	0.0	0.0
vehicle	77.4	18.9	0.0	3.7
nursery	93.3	6.7	0.0	0.0
satimage	56.7	30.2	2.3	10.7
car	47.8	43.5	8.7	0.0
scrotal-pain	24.4	53.3	11.1	11.1
ionosphere	12.9	62.9	1.4	22.9
credit-g	13.9	63.3	6.4	16.3
ecoli	25.8	61.3	3.2	9.7
hepatitis	13.6	63.6	9.1	13.6
haberman	15.1	56.2	16.4	12.3
breast-cancer	18.8	46.3	33.8	1.3
cmc	17.2	44.3	10.4	28.2
cleveland	6.7	30.0	13.3	50.0
glass	6.7	40.0	26.7	26.7
hsv	0.0	0.0	16.7	83.3
abalone	7.8	23.7	11.4	57.1
postoperative	0.0	65.2	30.4	4.3
seismic-bumps	6.4	20.0	8.6	65.0
solar-flare	7.1	45.2	7.1	40.5
transfusion	15.1	57.8	9.6	17.5
yeast	15.2	37.0	2.2	45.7
balance-scale	0.0	0.0	0.0	100.0

three datasets the differences are more visible. In postoperative dataset, 24 % of examples changed its label from borderline to outlier. However, it should be remembered that it is a very small dataset, and this difference refers in fact to only 5 minority examples. Then in breast-cancer dataset, the kernel approach labelled more examples as outliers and less as borderline. Finally, there are also differences for the ionosphere dataset (there are shifts between safe and borderline examples and between rare and outlier examples). Let us recall that in this dataset nearly 40 % of examples remained unlabelled by the kernel method which might have influenced the results. Finally, let us mention that we have also tested other kernel functions, such as Gaussian, triangular or uniform functions; we have also tested other kernel widths (calculated as the average distances to the 3^{rd}, 7^{th} and 9^{th} neighbour), but it did not influence too much the results.

To sum up, these experimental results and the sensitivity analysis presented in the previous section we noted that the general categories of considered datasets are the same. Both approaches, although based on different principles, have discovered the similar proportions of the type categories characterizing the minority class distributions.

We can also say that the simpler method based on analysing the neighbourhood of fixed size 5 is sufficient to analyse the distribution of a dataset and we will use its results to study the performance of learning algorithms in the following sections.

7 Experimental study of learning abilities with respect to types of examples

7.1 Comparing base classifiers

Having shown that the analysed imbalanced datasets differ in their distribution of minority examples, the aim of the next experiment is to verify whether the identified categories of data constitute a different degree of difficulty for the learning algorithms, and whether different classifiers reveal different sensitivity to the particular types of examples.

In this experiment we focus on studying basic classifiers only, as for complex classifiers, e.g. ensembles, there are more elements which could be influenced by data characteristics (see e.g. (Blaszczynski et al. 2013; Galar et al. 2012)). We have decided to compare six learning algorithms, which have been often considered in related works and which represent different learning strategies. They are: tree learning by C4.5, rule induction with PART and RIPPER algorithms, k-nearest neighbour (kNN), neural network based on radial functions (RBF) and support vector machine (SVM).[4] C4.5 (J48) and PART are run without pruning as it should be better for recognizing the minority class (Prati et al. 2004a; Lopez et al. 2013). In RIPPER (JRIP) the default parameters are used. kNN is used with $k = 1$ and $k = 3$, following the suggestion in (Garcia et al. 2008) that for difficult imbalanced datasets more local classifiers (with smaller k) perform better on the minority class. Standard values of parameters for RBF and SVM have failed to recognize the minority class. For RBF we have scanned several configurations trying to get the best values of sensitivity measures on all 26 datasets. As a result, we changed a number of clusters to 5 and minimum standard deviation to 0.1. Similar scanning of parameters has been done for the SVM classifier –

[4] The WEKA implementations are used. We use J48 version of C4.5, SMO version of SVM and JRIP version of RIPPER.

which leads to choosing the RBF kernel, complexity $C = 50$ and gamma parameter 1.0. We will come back to the issue of SVM parametrisation at the end of this section.

7.1.1 Analysing classifiers with general evaluation measures

In the first step of experiments, we compare the performance of classifiers on all the 26 imbalanced datasets. Again we can present the details of selected experiments only due to page limits. Table 5 presents the Sensitivity measure. The datasets are sorted in the same way as in Table 3 – from the simple to the more complex distributions. Relating the results to the labelling of the datasets presented in Table 3, we can observe how with the increasing difficulty of the dataset distribution, the performance of all classifiers decreases. For datasets where safe minority examples prevail (breast-w – nursery), all classifiers learn the minority class quite well – they recognize 70-90% of the minority examples. Only on nursery dataset, kNN classifier works worse. In datasets with more borderline examples (satimage – haberman), the classifiers usually recognize 40-60% of the minority class. When many rare and/or outlying example are observed (haberman – balance-scale), the sensitivity measure

Table 5 Sensitivity [%] of compared classifiers

Dataset	1NN	3NN	J48	JRIP	PART	RBF	SVM
breast-w	93.5	96.3	90.1	95.9	94.7	95.4	90.8
abdominal-pain	76.4	78.5	69.8	72.5	72.6	75.0	71.8
acl	72.0	78.5	85.5	84.5	80.0	84.0	82.5
new-thyroid	96.3	90.2	92.2	86.7	93.3	99.5	89.8
vehicle	89.1	87.9	87.0	89.0	88.3	88.0	95.2
nursery	41.1	41.1	89.7	68.0	99.4	76.2	97.2
satimage	71.4	67.2	52.8	50.3	55.6	37.8	62.1
car	3.1	3.1	77.7	47.0	90.0	49.6	88.2
scrotal-pain	58.4	58.7	55.3	53.4	63.4	62.5	65.9
ionosphere	69.4	65.5	82.7	84.4	84.0	94.2	89.0
credit-g	50.3	39.9	46.5	37.6	47.7	43.6	52.2
ecoli	52.2	50.8	58.0	59.7	42.0	54.7	58.5
hepatitis	44.0	37.0	43.2	31.2	45.7	60.7	51.5
haberman	30.1	26.9	41.0	34.0	33.4	18.3	1.3
breast-cancer	40.4	27.6	38.7	32.4	41.1	40.8	45.3
cmc	37.6	33.8	39.2	30.0	37.7	12.1	5.2
cleveland	20.3	12.5	23.7	6.3	25.2	9.5	9.0
glass	30.0	16.0	30.0	7.0	34.0	25.0	0.0
hsv	0.0	0.0	0.0	0.0	2.0	1.0	0.0
abalone	20.5	16.5	30.4	29.7	18.8	12.3	0.2
postoperative	4.3	0.0	4.7	0.0	10.3	13.7	7.0
seismic-bumps	16.5	9.3	9.3	2.4	9.1	0.0	0.9
solar-flare	9.1	8.2	20.9	3.7	18.7	10.2	15.7
transfusion	31.9	34.3	41.3	39.7	42.9	32.9	2.2
yeast	38.1	26.2	30.9	36.7	26.7	15.1	0.0
balance-scale	0.0	0.4	0.0	0.0	0.0	0.8	0.0

Classifier	Avg. rank
PART	5.0
J48	4.3
RBF	4.2
1NN	4.1
SVM	3.8
JRIP	3.5
3NN	3.2

(a) All datasets

Classifier	Avg. rank
SVM	4.7
PART	4.5
RBF	4.4
1NN	3.9
JRIP	3.9
J48	3.6
3NN	3.0

(b) Safe and border datasets

Classifier	Avg. rank
PART	5.5
J48	5.3
1NN	4.3
RBF	3.7
3NN	3.3
JRIP	3.3
SVM	2.5

(c) Rare and outlying datasets

Fig. 5 Rankings of classifiers depending on the dataset characteristics (based on Sensitivity)

varies between 0% and 40%. Finally, for the datasets with a lot of outlying examples (e.g. cleveland, hsv, balance-scale), it is impossible to recognize more than 30 % of the minority examples (for some data even no examples are correctly classified). As for the majority class, we observed that all the classifiers could recognize this class in a similar degree, reaching 80–100% on the accuracy of this class for all the datasets.

The second observation is that different classifiers reveal different sensitivity to the particular category of data (with respect to types of minority examples). To examine the importance of such differences we apply the Friedman test (see Section 6.1). It uses ranks of all classifiers on each of the datasets. We applied the interpretation ”the higher rank, the better classifier”. First, we compare how the classifiers perform on average on all datasets, not taking into account the types of examples. The null hypothesis, saying that all the classifiers perform equally, is rejected ($p < 0.035$). The ranking of classifiers according to their average ranks is presented in Fig. 5a. The best classifiers are PART and J48, while SVM, JRIP and 3NN perform the worst. The critical difference (CD) according to the Nemenyi test is 1.76 – so we cannot say that differences between the best performing classifiers are significant, however the first best classifiers are better than the last ones.

In our opinion, averaging over the datasets of different characteristics (with respect to minority class distribution) might hide the interesting characteristics of the learning methods. Our aim is rather to analyse the influence of types of examples on the performance of classifiers. To carry out such analysis, we have decided to divide the collection of datasets into two groups. In the first group we place the datasets with many safe and borderline examples, and only a small number of outlier or rare examples – these are datasets from breast-w to haberman. In the second group we put the datasets where many rare and/or outlying examples were observed – these are datasets from haberman to balance-scale.[5] If we consider only the datasets from the first group, the ranking of best performing classifiers changes (although the differences between average ranks are smaller than before) – see ranking in Fig. 5b. SVM becomes the best classifier.

In the second group of datasets, according to Friedman test we can also reject the null hypothesis ($p < 0.001$) and CD=1.92. For these datasets, we observe an opposite behaviour (Fig. 5c). PART and J48 perform the best. 1NN also becomes one of the best classifiers. RBF becomes much worse and SVM fails at these datasets. It is also reflected by the results of the Wilcoxon test, which we use as a supplementary test to analyse more precisely the

[5] Let us note that we initially wanted to divide the datasets into four groups, one for each type of minority examples. However, the number of datasets in each group would be too small to be able to draw meaningful conclusions.

differences between the selected pairs of classifiers. The results of this test showed that, e.g., PART or J48 were better than RBF with $p \leq 0.005$. If we look solely on the most difficult datasets with a lot of outlying examples (e.g. cleveland, abalone, glass, yeast) in Table 5, J48, PART or 1NN can classify a few examples while SVM usually cannot recognize the minority class at all. The results on G-mean and F-measure have demonstrated quite similar trends. On safe and borderline datasets, on both measures SVM and RBF are the best classifiers, while on rare and outlier datasets PART and J48 dominate other methods.

The results of RBF and SVM classifiers need an additional comment. It is important to remember that these two classifiers are particularly sensitive to the configuration of parameters. As we did not want to favor any of the classifiers, in our experimental setup they were used with one configuration for all the datasets (similarly to other compared classifiers), which was on average the best on the whole collection of datasets. A similar approach was taken, e.g., in the experimental setup in (Van Hulse et al. 2007). In total, more than one hundred combinations of parameters were tested. However, tuning the parameters for each group of datasets separately (e.g. selecting another configuration for datasets with mostly safe and borderline examples and another one for those with a lot of rare and outlying examples) or even for each dataset separately, might improve the performance of these classifiers, especially on the rare and outlying examples. We have done a preliminary study in this topic. When we performed tuning for a whole group of rare and outlying datasets, we could not identify a single configuration which would improve all the datasets. When we optimized the parameters for each dataset separately, for some datasets with rare and outlying examples the results of SVM and RBF could be improved. However, there were only a few configurations among a hundred of combinations tested which yielded an improvement. Therefore, although the results of some rare and outlying datasets presented in Table 5 could be higher if a more fine-grained tuning was performed, the general conclusion remains the same: the SVM and RBF classifiers are very sensitive to the rare and outlying minority examples.

7.1.2 Classifier differences with respect to types of testing examples

Note that examples of different types often occur together inside the same dataset. So, one can ask a question which types of examples actually contribute to the ”global sensitivity” and which types are the most difficult. For instance, yeast dataset was included in the second group of datasets as it contains a lot of outlying examples (O) and almost no safe examples, but still many of its examples are of borderline type (B); see Table 3. As a result, ”global sensitivity” in yeast may rely more on the recognition of B than of O examples. As our labelling method enables to identify a type of each testing example, we record the ”local accuracy” for each type of testing examples separately. In Table 6 we present the results for two classifiers – PART and 1NN – which represent different learning strategies and achieve a high ”global sensitivity”, especially for more difficult datasets. When analysing these results, one should keep in mind that the minority class is small, and partitioning the testing examples with respect to their types makes the representation of each type even more sparse. Thus, we do not present the results for too small datasets (less than 300 examples) to ensure that there are enough representatives in each category. Also, as some datasets do not have any examples of a particular type or their number is too small (e.g. cleveland, vehicle) – in Table 6 we left the corresponding cells empty.

The results in Table 6 confirm the previous observations. Safe examples are easy to recognize for both classifiers (except for car dataset where 1NN cannot recognize the

Table 6 Local accuracies for labelled testing examples [%]

	1NN				PART			
Dataset	S	B	R	O	S	B	R	O
abalone	81.4	45.3	22.5	2.1	74.3	27.8	12.4	10.4
abdominal-pain	99.5	74.5	5.7	0.0	91.7	69.5	17.1	8.8
balance-scale			0.0	0.0			0.0	0.0
breast-w	97.3	58.7			97.5	72.0		
car	5.5	1.5	0.0	0.0	91.5	91.1	100	46.7
cleveland		35.0	8.9	20.0		45.0	22.2	16.7
cmc	81.0	35.9	26.3	18.2	63.1	37.5	34.9	19.1
credit-g	72.9	53.9	42.1	39.2	65.7	53.3	40.5	32.0
ecoli	100.0	49.4		0.0	84.0	32.9		12.0
haberman	100.0	43.5	18.1	0.0	90.0	48.2	20.6	5.0
ionosphere	96.4	69.7	49.5	0.0	98.6	92.1	67.6	37.5
nursery	48.7	7.5			99.6	98.6		
satimage	96.1	66.1	36.3	0.0	76.2	46.4	22.9	19.4
seismic-bumps	83.3	26.3	20.4	6.0	26.7	22.5	8.3	3.3
solar-flare		16.3	14.0	0.0		27.5	32.0	2.4
transfusion	80.6	41.9	12.4	0.0	86.1	64.5	21.2	1.6
vehicle	97.6	71.8			93.1	74.1		
yeast	100.0	62.2	52.0	0.0	73.3	50.0	20.0	2.0

minority class at all and seismic-bumps dataset which is more difficult for PART). Borderline examples are more difficult, but still a large number of them can be correctly learnt. Rare examples are usually recognized within the range of 10–30%. Outlier examples are extremely difficult for these and also all other classifiers. Only for cmc, credit-g and cleveland datasets both classifiers can recognize some of them, while for other datasets these examples are mostly neglected.

7.2 Comparing pre-processing methods

The pre-processing methods, which transform the distribution of examples between the classes, are popular approaches for improving standard classifiers. They are based on different principles, and in general they either clean (undersample) the majority class or oversample the minority class. For their review, see e.g. (Batista et al. 2004; He and Garcia 2009; He and Ma 2013). Some conclusions of the most related experimental comparative studies, as e.g. (Batista et al. 2004; Stefanowski and Wilk 2008) indicate that there is no single best pre-processing method. For instance, informed over-sampling, like SMOTE, has been reported to perform better for some dataset, while under-sampling is superior for others. However, there are no clear explanations of these differences. Therefore, we have decided to carry out the next experiment, where we want to compare the performance of the popular pre-processing methods, depending on the type of minority examples in the dataset.

We consider 4 different pre-processing methods: Random Oversampling, SMOTE (Chawla et al. 2002), NCR (Laurikkala 2001) and SPIDER (Stefanowski and Wilk 2008).

Random oversampling aims to balance the class distribution by random replication of minority examples. SMOTE is a best known representative of informed oversampling. Its main idea is to generate new synthetic minority examples by interpolating between k nearest neighbours of the minority example. We use it with $k = 5$ and the oversampling ratio aimed to balance the distribution between the classes (the configuration suggested, e.g., in (Van Hulse et al. 2007) and many other studies). NCR (Neighbour Cleaning Rule) (Laurikkala 2001), on the other hand, represents an undersampling technique. It applies the Edited Nearest Neighbour Rule to identify and remove noisy and borderline examples from the majority classes. SPIDER (Stefanowski and Wilk 2008) is a hybrid approach which combines oversampling of the minority class with undersampling of the majority class in the difficult regions, while leaving safe regions for both classes unchanged. We run it with $k = 5$ to preserve consistency with SMOTE. All the methods use the HVDM distance measure.

We compare the pre-processing methods for four best previously identified classifiers – PART, J48, RBF and 1NN classifier. We resigned from the SVM classifier as it was the worst on rare and outlier datasets and it was difficult to parametrize for all datasets. We also did not analyse the JRIP and 3NN classifiers. Due to the page limits detailed results will be presented only for PART, while other classifiers will be discussed in the main text if their results differ from the PART's results.

7.2.1 Studying general evaluation measures

First, we compare the pre-processing methods on all 26 datasets, not taking into account their differentiation with respect to data difficulty category. The Sensitivity on each dataset is presented in Table 7.[6] The null hypothesis in the Friedman test is rejected. The order of pre-processing methods is given in Table 6a . The critical difference CD in post-hoc test is 1.19. When all the datasets are concerned, a cleaning method (NCR) is followed by SPIDER method, which is followed by oversampling SMOTE, however the differences are statistically insignificant. On the other hand, all informed methods perform significantly better than Random Oversampling (RO) and all pre-processing methods are better than no pre-processing.

If we split the datasets into two groups, as it has been done in the experiments described in the previous section, the order of methods is slightly changed (and in both groups the null hypothesis in the Friedman test is rejected with p close to 0.0001). The critical difference (CD = 1.63) again allows us to differentiate only between informed methods and random oversampling with no processing, however, we can still observe some trends. For datasets where safe and borderline examples prevail (see ranks in Fig. 6b), the NCR cleaning method performs the best. SPIDER, which also performs some cleaning, is the second in this order of averaged ranks.

For datasets with a lot of rare and outlier examples, on the other hand, the ranking (Fig. 6c) shows that the oversampling method SMOTE and hybrid method SPIDER are at the first positions of the ranking and they practically equal to each other (CD = 1.69). Moreover, Random Oversampling becomes closer in the ranking to informed methods on these data. When we look only on the datasets with a lot of outliers in Table 7 (e.g., hsv, abalone, yeast), the advantage of SMOTE is even more visible.

[6] Random oversampling is denoted as RO, SPIDER – SP, SMOTE – SM, no pre-processing – None.

Table 7 Sensitivity for PART and pre-processing [%]

Dataset	None	RO	NCR	SM	SP
breast-w	94.2	92.5	96.3	96.3	96.3
abdominal-pain	72.6	76.0	86.1	73.8	85.0
acl	80.0	83.5	91.0	86.5	87.5
new-thyroid	93.3	90.2	86.3	94.0	91.0
vehicle	88.3	90.6	92.6	92.4	91.4
nursery	99.4	96.7	100.0	99.1	100.0
satimage	55.1	58.8	69.7	64.1	63.6
car	90.0	75.6	92.6	88.3	91.2
scrotal-pain	63.4	66.6	74.9	69.7	72.1
ionosphere	84.0	84.0	86.8	88.9	85.0
credit-g	47.7	47.5	69.7	53.3	60.6
ecoli	42.0	55.0	71.2	74.0	72.8
hepatitis	45.7	57.3	63.3	54.8	56.7
haberman	33.4	55.3	59.7	68.3	70.3
breast-cancer	41.1	43.7	67.9	44.3	55.9
cmc	37.7	48.5	59.8	49.7	55.9
cleveland	25.2	16.7	42.2	28.8	24.5
glass	34.0	33.0	56.0	46.0	43.0
hsv	2.0	9.0	7.0	15.0	10.0
abalone	18.8	38.2	31.1	52.8	50.2
postoperative	10.3	21.7	42.3	17.0	36.0
seismic-bumps	10.6	13.5	18.2	24.7	21.8
solar-flare	18.7	35.6	45.5	33.9	46.1
transfusion	42.9	59.1	50.3	72.4	70.3
yeast	26.7	33.3	30.3	47.9	37.2
balance-scale	0.0	69.5	0.0	22.0	12.0

Considering F-measure and G-mean, the results are similar, with NCR becoming more comparable or better than SMOTE in case of datasets with rare and outlying examples, which might suggest that SMOTE's increased performance on the minority class comes at a too high cost of the majority class recognition. This is consistent with some studies based on artificial datasets (Maciejewski and Stefanowski 2011; Batista et al. 2004).

Rankings of pre-processing methods used with J48 are the same. For 1NN, SMOTE is often a better classifier. For RBF we observed a different behaviour of Random Oversampling, which seems to work better than for the former classifiers. With respect to Sensitivity, it is particularly good for the datasets with rare and outlying examples.

7.2.2 Performance of pre-processing methods with respect to the types of testing examples

Again, to get a more precise insight into classification results, we analyse the local accuracies for each type of testing examples separately. Tables 8, 10 present the values of accuracies for pre-processing with PART. As already mentioned in Section 7.1, some

Pre-process	Avg. rank
NCR	3.9
SPIDER	3.8
SMOTE	3.6
RO	2.2
None	1.5

(a) All datasets

Pre-process	Avg. rank
NCR	4.3
SPIDER	3.8
SMOTE	3.4
RO	1.9
None	1.6

(b) Safe and border datasets

Pre-process	Avg. rank
SMOTE	3.9
SPIDER	3.8
NCR	3.4
RO	2.5
None	1.3

(c) Rare and outlying datasets

Fig. 6 Rankings of pre-processing methods used with PART, depending on the dataset characteristics (based on sensitivity)

datasets contain too few examples of a given type to provide reliable results. Therefore, in Tables 8, 10 we present only the datasets which have a sufficient number of examples of a given type.

When comparing the results of all methods to the use of PART without pre-processing (None column in Tables 8, 10), PART without pre-processing can recognize about 40-60% of the borderline examples (except for ionosphere and car, where PART performs better), 10-20% of rare examples (Table 9) and usually not more than 10% of outlying examples (Table 10). The pre-processing methods can increase the results on each type of minority examples by 10-30%.

If one considers borderline testing examples in the datasets given in Table 8, Friedman test rejects the null hypothesis and gives the ranking where NCR method is the first. Average ranks for PART are presented in Fig. 7a. However, one should be more careful with interpteting them as according to post-hoc test their differences are not significant. J48 and 1NN produce the similar order. For the RBF classifier and borderline testing examples, Random Oversampling is ordered before SMOTE.

The results for PART on rare testing examples are given in Table 9. Depending on the classifier, SMOTE or SPIDER are the first methods in the ranking. Ranking for PART is presented in Fig.7b – SPIDER is on the first position ex aequo with NCR. Again RBF works a bit better with Random Oversampling.

For outlier testing examples (in the datasets presented in Table 10), SMOTE, followed by SPIDER, perform the best for all the classifiers. Ranking for PART is given in Fig. 7c. J48 and 1NN produce the same order with slightly different values. For RBF, Oversampling becomes the second best method after SMOTE.

Table 8 PART and pre-processing: local accuracies recorded on borderline testing examples [%]

Dataset	None	RO	NCR	SM	SP
ionosphere	92.1	92.1	95.2	93.3	92.7
car	91.1	69.6	92.6	89.6	86.7
scrotal-pain	64.0	69.6	74.4	68.8	77.6
satimage	46.4	46.2	57.8	59.1	56.7
credit-g	53.3	54.1	76.9	58.8	67.9
ecoli	32.9	60.0	78.8	90.6	80.0
hepatitis	65.7	80.0	82.9	80.0	80.0
haberman	48.2	69.4	73.5	85.3	86.5

Table 9 PART and pre-processing: local accuracies recorded on rare testing examples [%]

Dataset	None	RO	NCR	SM	SP
haberman	20.6	49.0	48.4	62.6	64.5
cmc	34.9	40.4	56.1	41.4	45.1
breast-cancer	26.7	28.7	59.3	35.3	44.7
cleveland	22.2	22.2	33.3	22.2	22.2
glass	25.0	25.0	45.0	37.5	35.0
hsv	0.0	30.0	0.0	20.0	20.0
abalone	12.4	37.1	26.5	52.1	48.8
postoperative	8.0	18.0	42.0	6.0	32.0
seismic-bumps	8.3	15.7	22.2	24.3	24.3

8 Using neighbourhood analysis for developing new algorithms

Previous experimental sections have shown the usefulness of the presented identification method to point out differences in classification performance of learning algorithms and pre-processing methods.

In this section we additionally briefly show yet another perspective of exploiting our proposal. We claim that the results of the analysis of neighbourhood could be also useful for developing new learning algorithms specialized for class imbalance.

First, recall that partly similar ideas of exploiting the class distribution among nearest neighbours of the minority examples have been already incorporated in two generalizations of the informed pre-processing method SMOTE (Maciejewski and Stefanowski 2011). Some researchers have noticed that minority examples could be oversampled in a different way depending on their role. In Borderline SMOTE (Han et al. 2005) authors identify minority examples incorrectly re-classified by its neighbours and use only these examples as seeds for oversampling. Maciejewski and Stefanowski have recently introduced Local Neighbourhood extension of SMOTE (Maciejewski and Stefanowski 2011) which uses the local analysis of neighbours of the seed minority example and its reference neighbour to

Table 10 PART and pre-processing: local accuracies recorded on outlier testing examples [%]

Dataset	None	RO	NCR	SM	SP
cmc	19.1	24.0	28.0	25.5	30.2
cleveland	16.7	11.1	37.8	21.1	10.0
glass	28.0	16.0	48.0	52.0	32.0
hsv	4.0	4.0	12.0	16.0	8.0
abalone	10.4	27.7	16.6	41.5	39.1
postoperative	5.7	5.7	28.6	22.9	14.3
seismic-bumps	3.3	4.4	7.7	10.2	10.2
solar-flare	2.4	16.5	12.9	12.9	27.1
transfusion	1.6	22.9	4.9	45.3	49.4
yeast	2.0	7.0	9.0	26.0	13.0
balance-scale	0.0	57.8	0.0	27.6	18.7

Pre-process	Avg. rank
NCR	4.3
SPIDER	3.6
SMOTE	3.6
RO	1.9
None	1.6

(a) Border testing examples

Pre-process	Avg. rank
SPIDER	3.8
NCR	3.7
SMOTE	3.2
RO	2.8
None	1.4

(b) Rare testing examples

Pre-process	Avg. rank
SMOTE	4.0
SPIDER	3.7
NCR	3.4
RO	2.4
None	1.4

(c) Outlying testing examples

Fig. 7 Rankings of pre-processing methods used with PART, depending on the dataset characteristics (based on accuracies on testing examples of a given type)

restrict the range of the line between them where new examples are generated. Experimental results are encouraging.

These methods are just first proposals and we believe that other, better pre-processing techniques could be still developed using the inspirations coming from our study. Several problems for informed pre-processing methods (also for SPIDER or NCR) remain still open. For instance, identified types of examples could be used to select the minority examples for over-sampling, e.g. one could over-sample more unsafe examples, as according to our experiments these examples are more difficult to learn. Another application could concern tuning the oversampling ratio in SMOTE or SPIDER. Currently one global value is used for all examples. In our opinion, it would be more profitable to tune this ratio dynamically depending on the local data characteristics and on the varying density of examples. Evaluating types of examples could be exploited in such dynamic approaches.

It would be also promising to integrate the information about the local neighbourhood inside the learning phase of the algorithms. For instance, in our recent paper on a new rule induction algorithm BRACID (Napierala and Stefanowski 2012a) we modify the generalization of rule candidates depending on the type of minority examples. Incorporating it has improved classification predictions.

Finally, we hypothesize that our results could be particularly useful for constructing the ensembles for imbalanced data. Most of current proposals are generalizations of known techniques as bagging, boosting or random forests. Many of them employ pre-processing methods before learning component classifiers. In particular, it concerns the generalizations of bagging (see their review in Galar et al. (2012)). However, in case of balancing bootstraps (see e.g. the Roughly Balanced Bagging (Hido and Kashima 2008)), all examples are treated as equally important and sampled with the same probabilities. We think that drawing minority examples chould depend on the difficulty of the example, e.g. using the neighbourhood analysis as discussed in our paper. The first research on such a new type of bagging ensemble, called Nearest Neighborhood Bagging, which modifies sampling minority examples has been just started in Blaszczynski and Stefanowski (2015). The results of its experimental evaluation are promising. We hope that more advanced proposals could be still developed.

9 Conclusions

Our paper intends to increase a research interest on data difficulty factors that may deteriorate classifiers learnt from imbalanced data. Although most of the current research on class imbalance concerns the development of new algorithms, we claim that it is still worth to study the nature of imbalanced data, characteristics of the minority class distribution and its influence on classification performance. The main objectives of our study have been:

- to show that characteristics of the minority class distribution can be captured by considering different types of examples creating the minority class,
- to distinguish safe and unsafe examples – borderline, rare and outliers,
- to propose the method for their identification in real-world data.

Considering these four types of examples is a novel proposal concerning data difficulty factors. In particular, we pay more attention to rare examples and outliers. Other contribution concerns focusing the interest on the local characteristics of the minority class distributions (Napierala and Stefanowski 2012b). Following it, we have introduced an identification method based on the analysis of class labels of examples belonging to the local neighbourhood of learning examples. We have shown two ways of constructing such neighbourhood: either with *k-nearest* examples or with kernel functions. In both cases, finding the number of majority class examples in the neighbourhood of the given minority example allows us to identify a type of this example.

We have experimentally shown that both these methods lead to similar results. Moreover, the experimental sensitivity analysis has shown that changing the main parameter k has not influenced too much the results. In the first part of experiments, the results of applying this method have been successfully validated on the artificial datasets. Moreover, the results on the real-world datasets have been confirmed (where it was possible) by the visualisation methods, based on the MDS and t-SNE projections of the multidimensional dataset into two dimensions.

Although the proposed method is simple and adapts known methodological elements, such a formulation has not been considered yet in the literature on imbalanced data. What is more important, its further experimental evaluation on a large collection of real datasets with several learning algorithms, led us to many interesting observations which have not been discussed yet in the earlier related works. Below we summarize the most interesting conclusions:

- Imbalanced datasets usually contain all types of minority examples, but in different proportions. Most of the datasets do not contain too many safe examples but they are rather characterized by unsafe distributions of the minority class. On the other hand, majority classes contain mainly safe examples.
- An interesting finding is that outlier examples can constitute an important part of the minority class – in some datasets they even prevail in the minority class. We think that one should be cautious in treating them as noise and applying noise-handling methods such as relabelling or removing these examples from the learning set. In general, distinguishing between noise and outliers in the minority class is an important, but challenging issue, which requires future research. Secondly, rare examples also occur in many datasets.
- The global imbalance ratio and the size of the data are not as influential as the above example types. Using simply the imbalance ratio to differentiate the datasets as done in (Van Hulse et al. 2007; Batista et al. 2012), or size of data as in Batista et al. (2004), is not sufficient to explain the differences in the classification performance according to our experiments. For instance, datasets with a high imbalance ratio, e.g. car or nursery (2-3%), are easier to learn than transfusion (23%). Similarly, large datasets are often more difficult than the small ones – compare, e.g., the results of abalone (over 4000 examples) with acl (less than 150 examples). Analysing the types of examples

provides information more relevant to the classification performance. Our observations are consistent with some earlier works on artificial datasets. In Garcia et al. (2007), Jo and Japkowicz (2004), and Napierala et al. (2010) it has also been shown that the imbalance ratio is not the main source of difficulty. However, these earlier works did not consider so many data factors and did not attempt to analyse real-world datasets.

- Collecting the information about local characteristics of the minority class and distinguishing between safe, borderline, rare and outlier examples is useful to differentiate the performance of basic classifiers. In general, our experiments show that safe datasets are easy to learn for all the classifiers. Datasets with a lot of borderline examples are more difficult, however the RBF and SVM classifiers work quite well on these datasets. Rare and especially outlier examples are extremely difficult to recognize. PART, J48 and sometimes 1NN may classify some of them but at a very low level. SVM and RBF are very sensitive to these types of data.
- Similarly, taking into account the information on types of minority examples is useful to analyse differences between popular pre-processing methods. We have observed that they can improve the recognition of the minority class examples by 10-30%. NCR (representative of informed undersampling) is better for safe and borderline examples, while SMOTE and SPIDER (informed oversampling and hybrid approach) are more accurate on rare examples. SMOTE seems to be slightly more effective for recognizing outliers. All the informed sampling methods are significantly better than simple Random Oversampling for *k*NN, tree- and rule-based classifiers. Random Oversampling sometimes performs better when RBF is used instead of the above classifiers.
- Our results confirm some of the results of the related works conducted on artificial datasets. For instance, similarly to Garcia et al. (2008) we have observed that for datasets with more difficult distributions (i.e. with many rare and outlier examples), a more local *k*NN (1NN) performs better compared to 3NN. Our results confirm also the hypothesis from Khoshgoftaar and Van Hulse (2009), in which 1NN performed better than SVM on difficult distributions (here with a lot of outliers). Concerning the results of pre-processing methods, we have also observed that they usually improve the recognition of the minority class by no more that 30%, which is consistent with the results presented in (Batista et al. 2012). Finally, the observation that informed re-sampling is better than simple random re-sampling follows the conclusions from Batista et al. (2004) and Chawla et al. (2002). Van Hulse et al. in (Van Hulse et al. 2007) gave a bit contradictory recommendations in favour of simple random oversampling. However, their analysis is mainly based on averaging over datasets having quite different characteristics.

To sum up, according to our best knowledge such an experimental study of analysing so many data difficulty factors occurring together in the real-world datasets has not been carried out before. Moreover, we hope that our analysis of the common data distribution patterns carried out on many imbalanced datasets can contribute to the further development of new learning algorithms as well as pre-processing methods dedicated for class imbalance. Although it is not the main aim of our paper, we have briefly highlighted such research directions in the previous section.

Acknowledgments The authors' research was partially funded by the Polish National Science Center under Grant No. DEC-2013/11/B/ST6/00963.

References

Anyfantis, D., Karagiannopoulos, M., Kotsiantis, S.B., & Pintelas, P.E. (2007). Robustness of learning techniques in handling class noise in imbalanced datasets. In *Proc. of AIAI 07* (pp. 21–28).

Batista, G., Prati, R.C., & Monard, M.C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter*, *6*(1), 20–29.

Batista, G., Silva, D., & Prati, R. (2012). An experimental design to evaluate class imbalance treatment methods. In *Proc. of ICMLA'12* (Vol. 2, pp. 95–101). IEEE.

Bishop, C.h.M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics).* New York: Springer.

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

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

Brodley, C.E., & Friedl, M.A. (1999). Identifying mislabeled training data. *Journal of Artificial Intelligence Research*, *11*, 131–167.

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

Chawla, N.V., Bowyer, K.W., Hall, L.O., & Kegelmeyer, W.P. (2002). SMOTE: synthetic minority oversampling technique. *J. Artif. Int. Res. (JAIR)*, *16*, 321–357.

Cox, T., & Cox, M. (1994). *Multidimensional Scaling.* Chapman and Hall.

Demsar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, *7*, 1–30.

Denil, M., & Trappenberg, T.P. (2011). A characterization of the combined effects of overlap and imbalance on the SVM classifier. *CoRR*, 1–24.

Ester, M., Kriegel, H., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases. In *Proc. Int. Conf. KDD'96* (pp. 226–231).

Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., & Herrera, F. (2012). A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches. *IEEE Transactions on Systems, Man, and Cybernetics Part C*, *42*(4), 463–484.

Gamberger, D., Boskovic, R., Lavrac, N., & Groselj, C. (1999). Experiments with noise filtering in a medical domain. In *Proc. of 16th ICML,* (pp. 143–151). Morgan Kaufmann.

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

Garcia, V., Sanchez, J., & Mollineda, R. (2007). An empirical study of the behavior of classifiers on imbalanced and overlapped data sets. In *Proceedings of the 12th Iberoamerican Conf. on Progress in Pattern Recognition, Image Analysis and Applications of LNCS*, (Vol. 4756 pp. 397–406).

Goldstein, M. (1972). K_n-nearest neighbour classification. *IEEE Trancs. on Inform. Theory*, 627–630.

Grzymala-Busse, J.W., Stefanowski, J., & Wilk, Sz. (2004). A comparison of two approaches to data mining from imbalanced data. In *Proceedings of the KES 2004–8th Int. Conf. on Knowledge-based Intelligent Information Engineering Systems of LNCS* (Vol. 3213, pp. 757–763). Springer.

Han, H., Wang, W., & Mao, B. (2005). Borderline-SMOTE, A new over-sampling method in imbalanced data sets learning. In *Proc. of ICIC of LNCS,* (Vol. 3644, pp. 878–887). Springer.

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

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

Hido, S., & Kashima, H. (2008). Roughly balanced bagging for imbalanced data. In *Proc. of 8th SIAM Int. Conf. Data Mining* (pp. 143–152).

Holte, R.C., Acker, L.E., & Porter, B.W. (1989). Concept learning and the problem of small disjuncts. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence* (pp. 813–818).

Japkowicz, N. (2001). Concept-learning in the presence of between-class and within-class imbalances. In *Proceedings of the Canadian Conference on AI 2001* (pp. 67–77).

Japkowicz, N. (2003). Class imbalance: Are we focusing on the right issue. In *Proc. of 2nd Workshop on Learning from Imbalanced Data Sets (ICML)* (pp. 17–23).

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

Japkowicz, N., & Stephen, S. (2002). The class imbalance problem: A systematic study. *Intelligent Data Analysis*, *6*(5), 429–450.

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

Khoshgoftaar, T.M., & Van Hulse, J. (2009). Knowledge discovery from imbalanced and noisy data. *Data & Knowledge Engineering*, *68*, 1513–1542.

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

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

Lopez, V., Fernandez, A., Garcia, S., Palade, V., & Herrera, F. (2013). Empirical results and current trends on using data intrinsic characteristics. *Information Sciences*, *250*, 113–141.

Lumijarvi, J., Laurikkala, J., & Juhola, M. (2004). A comparison of different heterogeneous proximity functions and Euclideandistance. *Stud Health Technol Inform*, *107*(Pt 2), 1362–6.

Maciejewski, T., & Stefanowski, J. (2011). Local neighbourhood extension of SMOTE for mining imbalanced data. In *Proc. of the IEEE Symposium on Computational Intelligence and Data Mining,* (pp. 104–111). IEEE Press.

McCane, B., & Albert, M. (2008). Distance functions for categorical and mixed variables. *Pattern Recogn Lett.*, *29*, 986–993.

Napierala, K. (2013). *Improving rule classifiers for imbalanced data.* Ph.D dissertation: Poznan University of Technology.

Napierala, K., & Stefanowski, J. (2012). BRACID: a comprehensive approach to learning rules from imbalanced data. *Journal of Intelligent Information Systems*, *39*(2), 335–373.

Napierala, K., & Stefanowski, J. (2012). Identification of different types of minority class examples in imbalanced data. In *Proc. of HAIS, volume 7209 of Springer LNCS* (pp. 139–150).

Napierala, K., Stefanowski, J., & Wilk, Sz. (2010). Learning from imbalanced data in presence of noisy and borderline examples. In *Proc. of 7th Int, Conf. Rough Sets and Current Trends in Computing, volume 6086 of Springer LNAI* (pp. 158–167).

Prati, R.C., Batista, G., & Monard, M.C. (2004). Class imbalances versus class overlapping: An analysis of a learning system behavior. In *Proc. of MICAI'04* (pp. 312–321).

Prati, R.C., Batista, G., & Monard, M.C. (2004). Learning with class skews and small disjuncts. In *Proc. of SBIA'04* (pp. 296–306).

Saez, J., Luengo, M., Stefanowski, J., & Herrera, F. (2015). SMOTE-IPF: Addressing the noisy and borderline examples problem in imbalanced classification by a re-sampling method with filtering. *Information Sciences*, *291*, 184–203.

Stanfill, C., & Waltz, D. (1986). Toward memory-based reasoning. *Commun. ACM*, *12*, 1213–1228.

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, of Smart Innovation, Systems and Technologies* (Vol. 13, pp. 277–306). Berlin Heidelberg: Springer.

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

Ting, K.M. (1994). The problem of small disjuncts: its remedy in decision trees. In *Proceeding of the 10th Canadian Conference on Artificial Intelligence* (pp. 91–97).

van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*, 2579–2605.

Van Hulse, J., Khoshgoftaar, T.M., & Napolitano, A. (2007). Experimental perspectives on learning from imbalanced data. In *Proc. of the 24th Int. Conf. on ML (ICML)* (pp. 17–23).

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

Weiss, G.M., & Provost, F. (2003). Learning when training data are costly: the effect of class distribution on tree induction. *Journal of Artificial Intelligence Research*, *19*, 315–354.

Weiss, G.M., & Hirsh, H. (2000). A quantitative study of small disjuncts. In *Proc. the 17th National Conference on Artificial Intelligence – AAAI00* (pp. 665–670).

Wilson, D.R., Artif, T., & Martinez, R. (1997). Improved heterogeneous distance functions. *J. Artificial Intell. Res. (JAIR)*, *6*, 1–34.