Dealing with Data Difficulty Factors while Learning from Imbalanced Data

Jerzy Stefanowski

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

Abstract. Learning from imbalanced data is still one of challenging tasks in machine learning and data mining. We discuss the following data difficulty factors which deteriorate classification performance: decomposition of the minority class into rare sub-concepts, overlapping of classes and distinguishing different types of examples. New experimental studies showing the influence of these factors on classifiers are presented. The paper also includes critical discussions of methods for their identification in real world data. Finally, open research issues are stated [1].

1 Introduction

Data mining and machine learning have shown tremendous progress in last decades and have become ones of the main sub-fields of computer sciences. The supervised learning of object classification is one of the most common tasks considered both in theory and practice. Discovered classification knowledge is often used as a classifier to predict class labels for unclassified, new instances. This task has been intensively studied and a large number of approaches, based on different principles, have been already introduced; for some reviews the reader can consult, e.g. [4, 49].

Nevertheless many real world problems still reveal difficulties for learning accurate classifiers and require new solutions. One of these challenges is *learning from imbalanced data*, where at least one of the target classes contains a much smaller number of examples than the other classes. This class is usually referred to as the *minority class*, while the remaining classes are denoted as *majority ones*. For instance, in medical problems the number of patients requiring special attention is much smaller than the number of patients who do not need it. Class imbalances have been also observed in many other application domains such as fraud detection in telephone calls or credit cards transactions, bank risk analysis, technical diagnostics, network intrusion detection, image recognition, detecting specific astronomical objects in sky surveys, text categorization, information filtering; for some reviews see, e.g., [10, 29, 30, 18, 72].

[1] Will be published as a chapter in S. Matwin and J. Mielniczuk (eds.), Challenges in Computational Statistics and Data Mining, Springer Studies in Computational Intelligence vol. 605, 2016, pp. 333–363

In all those problems, the correct recognition of the minority class is of key importance. For instance, in a medical diagnosis a failure in recognizing an illness and not assigning a proper treatment to a patient is much more dangerous than misdiagnosing a healthy person, whose diagnosis may be verified by additional examinations. Although focusing attention on a critical class and considering misclassification errors are similar to cost sensitive learning [16], dealing with imbalanced classes is not the same, as the costs of misclassifications are rather unknown in advance [50]. Even if they could be somehow approximated, they may be different for particular instances of the given class.

The standard classifiers do not work sufficiently well with imbalanced data [29, 30, 41, 74]. They mainly concentrate on larger classes and often fail to classify sufficiently accurately minority class examples. For instance, [45] describes an information retrieval system, where the minority class contained only 0.2% of all examples. Although all considered classifiers achieved the overall accuracy close to 100%, they were useless because they failed to deliver requested documents from this class. While this degradation of classification performance has been known earlier from applications, improving classifiers for imbalanced data has received a growing research interest in the last decade and a number of specialized methods have been proposed, for their review see, e.g., [10, 18, 30, 29].

Although several specialized methods exist, the identification of conditions for their efficient use is still an open research problem. It is also related to more fundamental issues of better understanding the nature of the imbalance data and key properties of its underlying distribution which makes this problem too difficult to be handled.

Note that many authors introducing their new method usually carry out its experimental evaluation over some data sets and show its better performance than some reference methods. However, these evaluations are usually quite limited and authors do not ask the above mentioned questions on data characteristics. In more comprehensive comparative studies, as [2, 69], data sets are categorized with respect to the global ratio between imbalanced classes or the size of the minority class only. Nevertheless, it seems that these numbers do not sufficiently explain differences between classification performance of the compared methods. For instance, for some data sets even with a high imbalance ratio, the minority class can be sufficiently recognized by many standard classifiers.

Some researchers claim that the global imbalance ratio is not a problem itself and it may not be the main source of difficulties for learning classifiers. Following related works [34, 36, 60, 23, 47] and earlier studies of Stefanowski et al. [65, 53, 54, 52] we claim that other, as we call them, *data difficulty factors*, referring to characteristics of minority class distributions, are also influential. They include:

- decomposition of the minority class into many rare sub-concepts - also known as small disjuncts [32, 34, 35, 67, 73],
- an effect of too strong overlapping between the classes,
- a presence of too many minority examples inside the majority class region.

When these data difficulty factors occur *together* with class imbalance, they may seriously hinder the recognition of the minority class, see e.g. a study [47, 64].

Moreover, in earlier paper of Stefanowski at al. we have proposed to capture some of these data difficulty factors by considering the local characteristics of learning examples from the minority class [54, 53].

We claim that the studies on data difficulty factors are still not sufficiently developed and even well known among machine learning or data mining communities. Furthermore, most of these studies have been carried out on special synthetic data with assumed distributions of the minority class, while good identification of these factors in the real data sets is not easy and it poses still open questions and requires new solutions.

The aim of this paper is to survey the main current research on the above mentioned data difficulty factors including our own new experimental results. We will present consequences of these data factors on the classification performance. Then, we critically discuss current methods for their identification and put open questions on the directions of their future developments. Finally, we will claim that the proper analyzing of these data factors could be the basis for developing new specialized algorithms for imbalanced data.

The paper is organized as follows. Section 2 summarizes related pre-processing methods and evaluation issues. Difficulties caused by a fragmentation of the minority class into rare sub-concept are described in section 3. It is followed by a discussion of class overlapping in section 4. Then, the novel view of types of minority examples, the method for their identification in real world data sets, its experimental evaluation are presented. The final section concludes the paper.

2 Pre-processing Methods for Class Imbalances

Methods addressing class imbalances are usually categorized into two groups:

- **Data level methods** – these are classifier-independent methods that rely on transforming the original data distribution of classes into the better one for learning classifiers, e.g., by re-sampling or focused filtering some examples.
- **Algorithmic level methods** – they involve modifications of the algorithm.

In this paper we do not intend to provide a comprehensive review of all proposed methods and rather will briefly present the selected data level methods only as they will be considered in further experiments. The comprehensive reviews can be found in, e.g., [10, 18, 29, 30, 52, 66, 72].

The methods on the algorithmic level include the following sub-categories: adaptations to cost-sensitive learning, changing of internal algorithm bias (either in search strategies, evaluation criteria or classification strategies), generalizations of ensembles or one-class learning. On the other hand, methods on data level modify imbalanced data to provide the class distribution more appropriated for learning classifiers. Many of these proposed methods offer a more balanced distribution of classes. In general, changing the class distribution towards a more balanced one improves the performance for most data sets and classifiers [29]. We describe the best well known pre-processing methods below.

2.1 Random Re-sampling and Informed Pre-processing

The most popular re-sampling methods are random *over-sampling* which replicates examples from the minority class, and random *under-sampling* which randomly eliminates examples from the majority classes until a required degree of balance between class cardinalities is reached. However, several authors showed the simple random re-sampling methods were not sufficiently good at improving recognition of imbalanced classes. Random under-sampling may potentially remove some important examples and simple over-sampling may also lead to overfitting [11, 42]. The recent research focuses on particular examples, taking into account information about their distribution in the attribute space [29].

Kubat and Matwin claim in [42] that characteristics of mutual positions of examples is a source of difficulty for learning from imbalanced data, see also their more applied study [43]. They introduced *one-side-sampling* method (OSS), which filters the majority classes in a focused way [42]. It is based on distinguishing different types of learning examples: safe examples, borderline (located near the decision boundary) and noisy examples. They propose to use Tomek links (two nearest examples having different labels) to identify and delete the borderline and noisy examples from majority classes.

Then, the *Nearest Cleaning Rule* (NCR) method is introduced in [44] and it is based on the focused removal of examples from the majority class. It applies the edited nearest neighbour rule (ENNR) to the majority classes [75]. ENNR first looks for a specific number of *nearest neighbours* ([44] recommends using 3 neighbours) of the "seed" example, re-classifies it with them and then removes these majority examples, which cause the wrong re-classification. Experiments have shown that NCR outperforms OSS [44].

The best well know informative sampling method is the Synthetic Minority Over-sampling Technique (SMOTE) [11]. It is also based on the k nearest neighbourhood, however it exploits it to selectively over-sample the minority class by creating *new synthetic examples*. It considers each minority class example as a "seed" and finds its k-nearest neighbours also from the minority class. Then, according to the user-defined *over-sampling* ratio – o_r, SMOTE randomly selects o_r of these k neighbours and randomly introduces new examples along the lines connecting the seed example with these selected neighbours. It generate artificial examples for both qualitative and quantitative attribute.

Some of the assumptions behind SMOTE could still be questioned. First, using the same over-sampling ratio to all minority examples may be doubtful for some data. Several researchers claim that unsafe examples are more liable to be misclassified, while safe examples located inside the class regions are easier to be learned and do not require such a strong over-sampling. What is more important, SMOTE may over-generalize the minority class as it blindly over-samples the attribute space without considering the distribution of neighbours from the majority class. To overcome such limitations several generalizations of SMOTE have been recently introduced; for reviews see [62, 48]. They usually follow one of the two directions: (1) an integration of standard SMOTE with an extra post-processing step or (2) a modification of an internal sampling strategy.

The first solution is to integrate SMOTE with a post-processing phase including filtering the most harmful examples. For instance, using ENNR after SMOTE performs quite well with tree classifiers [2] and rules [52]. Yet a more elaborated approach is presented in [62], where an additional bagging ensemble is used to identify the most misclassified examples and iteratively remove them if it improves evaluation measures. The other group of more "internal" extensions includes two general solutions. The first generalizations over-sample some types of minority examples only. For instance, in *Borderline-SMOTE* only the borderline examples could be seeds for over-sampling [27]. The other generalizations attempt to modify localizations for introducing the new synthetic examples. In *Safe Level SMOTE* and *LN-SMOTE* the distribution of local sub-areas around the seed example and its selected neighbour are analysed and the new example is generated closer to a safer one [48].

Hybrid methods combine of over-sampling with cleaning difficult examples. Besides a simple integration of SMOTE with either ENN or Tomek links [68] other more complex methods offer sophisticated internal combinations of different operations, e.g. by using evolutionary algorithms to optimize some parameters, as the balancing ratio, combinations of over-sampling vs. under-sampling amount, see e.g. [71, 21].

SPIDER is another hybrid method that selectively filters out harmful examples from the majority class and amplifies the difficult minority examples [65]. In the first stage it applies ENNR to distinguish between safe and unsafe examples (depending how k neighbours reclassify the given "seed" example). For the majority class - outliers or the neighbours which misclassify the seed minority example are either removed or relabeled. The remaining unsafe minority examples are additionally replicated depending on the number of majority neighbours.

Note that in all the above mentioned methods k nearest neighbourhood is often calculated with the HVDM metric (*Heterogeneous Value Difference Metric*) [75]. Recall that it aggregates normalized distances for both continuous and qualitative attributes, however it uses the Stanfil and Valtz value difference metric for qualitative attributes.

Many generalizations of ensembles are based on integrating re-sampling to modify contents of training samples in bagging or boosting. For instance, SMOTEBoost is an integration of SMOTE with classical AdaBoost to focus successive classifiers on the minority class [10]. Another representative is IIvotes, where SPIDER is added to Breiman's importance sampling of bootstraps [5]. Other extensions of bagging re-balance the class distribution inside each bootstrap sample into fully balanced ones, by either simple random over-sampling of the minority examples, or by under-sampling the majority class - for their review and experimental comparison see [6, 19].

2.2 Evaluation Issues

Imbalanced data constitutes a challenge not only when constructing a classifier, but also when evaluating its performance. Indeed, an overall classification accuracy is not the best criterion characterizing performance of a classifier as it is

biased toward the majority classes. A good recognition of the minority is more preferred, thus a classifier should be characterized rather by other specialized measures, e.g. by its *sensitivity* and *specificity* for the minority class.

Both these and other similar measures are defined with the confusion matrix for two class only, where typically the class label of the minority class is called positive and the class label of the majority class is negative [29, 37]. Even if data contains more majority classes the classifier performance on these classes are usually aggregated into one negative class.

The *sensitivity* (also called a True-Positive Rate or *Recall* of the minority class) is defined as the ratio of correctly recognized examples from the minority class while the *specificity* is the ratio of correctly excluded examples from the majority classes (in a case of binary classification the specificity of the minority class is the recall of the majority class). More attention is usually given to sensitivity than to specificity [24]. However, in general there is trade-off between these two measures, i.e., improving the sensitivity too much may lead to deterioration of specificity at the same time - see experimental results in [25]. Thus, some measures summarizing both points of view are considered. One of them is *G-mean* [42], calculated as a geometric mean of sensitivity and specificity. Its key idea is to maximise the recognition of each of minority and majority classes while keeping these accuracies balanced. An important, useful property of the G-mean is that it is independent of the distribution of examples between classes. An alternative criterion aggregating precision and recall for the minority class is *F - measure*; for a deeper discussion of its properties see e.g. [29]. Other less frequently used measures are nicely reviewed in [38].

Several authors also use the *ROC* (*Receiver Operating Characteristics*) *curve* analysis in case of scoring classifiers. A ROC curve is a graphical plot of a true positive rate (sensitivity) as a function of false positive rate (1 – specificity) along different threshold values characterizing the performance of the studied classifier [37]. The quality of the classifier performance is reflected by the area under a ROC curve (so called AUC measure) [10, 37, 38]. Although AUC is a very popular tool, some researchers have discussed some limitations, e.g. in the case of highly skewed data sets it could lead to an overoptimistic estimation of the algorithm's performance [28]. Thus, other proposals include Precision Recall Curves or other special cost curves (see their review in [29, 13]).

3 Nature of Imbalanced Data

A data set is considered imbalanced when it is characterized by an unequal distribution between classes. N.Japkowicz refers it to the *between-class imbalance* [33]. It is evaluated by the *global class imbalance ratio* IR. Assume that the data set D contains n learning examples assigned to two classes: the minority class MK with N_{min} representatives and the majority class WK having N_{maj} examples. Depending on the literature sources, IR is usually expressed as either N_{maj}/N_{min} or the percentage of N_{min} in the total number of examples n.

There is no unique opinion about the threshold for the degree of such imbalance between the class cardinalities to establish data to be imbalanced. Some researchers have studied the data sets where one class was several times smaller than other classes, while others have considered more severe imbalance ratios as, e.g., with $IR = 10/1$, $100/1$ or even greater. Without showing a precise threshold value for this ratio, we repeat after [72] that the problem is associated with lack of data (absolute rarity), i.e. the number of examples in the rare (minority) class is too small to recognize properly the regularities in the data.

Although this description implies binary (two-class) problems, data with many majority classes are often aggregated into one global majority class - which is a case considered in this paper. However, note that some authors also consider multi-class data sets, where imbalances may exist between various classes.

The imbalance of a learning data set can be either *intrinsic* (in the sense that it is a direct result of the nature of the data space) or *extrinsic* (caused by reasons external to the data space). Extrinsic imbalances can be caused by too high costs of acquiring the examples from the minority class, e.g. due to economic or privacy reasons [72] or comes from technical time or storage factors. For instance, He et al. give in [29] examples of learning from continuous balanced data stream where due to technical sporadic interruptions in transmissions of some sub-blocks inside the analyzed stream would become an extrinsic imbalanced data set.

Gary Weiss also discusses problems of data rarity and distinguishes between *relative imbalance* and *absolute rarity*. In the former case, the data set contains too small minority class. However, if it is possible to collect / sample more examples and to increase the total size of data while keeping the same global imbalance ratio, it may happen that the absolute cardinality of the minority class will not be rare anymore and it may be easier to be learned [72].

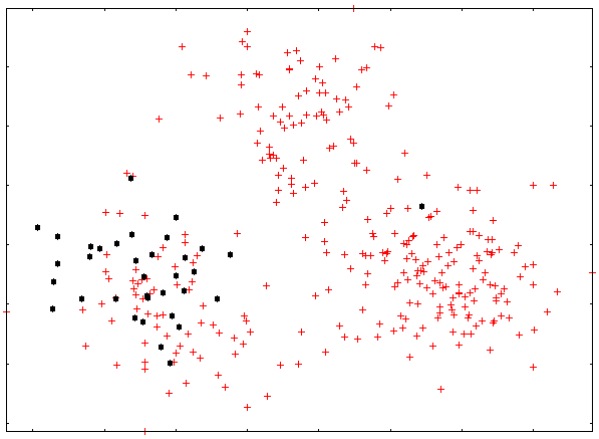

Fig. 1. MDS visualization of class distribution in ecoli imbalanced data.

On the other hand, some studies have shown that for even highly imbalanced data the minority class can be sufficiently accurately learned by all standard algorithms [2]. Examples of such popular UCI benchmark data sets are `new-thyroid` or `vehicle` - see their experimental analysis in [52]. Indeed one can image binary class distributions which could be linearly separated with not so much disturbance from even high imbalances assuming that the minority class does not represent an absolute rarity. In case of a clear separation the minority class boundary could be easily approximated by many algorithms.

Distributions of real world imbalance data usually are not similar to the above examples. For instance, Napierala in her attempts to visualize imbalanced data with either multi-dimensional scaling or non-linear projections [52] to low dimensional (2 or 3 variables) has showed such distributions as presented in Figure 1. One can notice that in `ecoli` data both classes are not separated, instead they seriously overlap. The consistent region belonging solely to the minority class is rather very small – most examples lie in a mixed region between the classes. Another observation is presence of small sub-groups of the minority class, having sometimes few instances only.

Furthermore, well known comprehensive experimental studies where many specialized approaches over large collections of imbalanced data show that simply discussing the global imbalance ratio does not sufficiently explain differences of classification performance of these approaches [36, 23, 47, 54, 64].

All these results lead us to conclude that the global imbalance ratio is not the only, and possibly not the main, data factor that hinders deterioration of learning classifiers. As some researchers claims one should rather consider data set *complexity* which should be more influential. *Data complexity* can be understood as the difficult properties distribution of examples from both classes in the attribute space. It is not particularly surprising that it shows a crucial impact on learning, as one could expect that data complexity should affect learning also in balanced domains. However, when data complexity occurs *together* with the class imbalance data difficulty factors, the deterioration of classification performance is amplified and it affects mostly (or sometimes only) the minority class.

The term "data complexity" can comprise different data distribution patterns. Up to now, the researchers have distinguished several *data difficulty factors* which hinder learning in imbalanced domains, such as: decomposing the minority class into rare sub-concepts, overlapping, and presence of outliers, rare instances or noise. We will discuss their role in the next sections.

4 Decomposition of the Minority Class

4.1 Rare Sub-concepts and Small Disjuncts

Nearly all research on data difficulty factors were carried out by experimental studies with synthetic data. The most well known and inspiring studies are research of Nathalie Japkowicz and her co-operators. They focused on *within-class imbalance*, i.e. target concepts (classes) were decomposed into several sub-concepts [33, 36]. To check how much increasing the level of such a decomposition

could influence the classification performance, they carried our many experiments with specially generated data. They controlled three parameters: the size of the training set, the imbalance ratio, and so called *degree of concept complexity* (understood as a decomposition of the class into a number of sub-concepts). Two classes were considered - the minority vs. the majority class. In their first experiments each data set was generated over a one-dimension interval. Depending on the concept complexity, the input interval was divided into a number of sub-intervals of the same size (up to five), each associated with a different class label. Following similar assumptions, in further studies they generated additional data sets in five-dimensional space, where an occurrence of classes was modeled by separate clusters.

Decision tree (C4.5) and multi layered perceptron neural networks (MLP) were learned from these data sets. The results of their experimental evaluation showed that imbalance ratio did not cause the degradation of classifiers' performance as much as increasing the degree of complexity (the number of sub-intervals). The worst classification results were obtained for the highest decomposition of classes (5 sub-intervals), in particular if they contained too small number of examples. On the other hand, in much larger data, where sub-clusters were represented by a reasonable number of examples, the imbalance ratio alone did not decrease the classification performance as much [36].

According to Japkowicz [33], if such imbalanced sub-concepts contain quite a small number of minority class examples, then the deterioration of classification performance is associated with the problem of so called *small disjuncts* – which was originally introduced by Holte et al. in standard (balanced) learning of symbolic classifiers [32]. Briefly speaking, a classifier learns a concept by generating disjunct forms (e.g. rules of tree) to describe it. Small disjuncts are these parts of the learned classifier which cover a too small number of examples [32, 67, 72]. It has been observed in the empirical studies that small disjuncts contribute to the classification error more than larger disjuncts. In case of fragmented concepts (in particular in the minority class) the presence of small disjunct arises [29]. The impact of small disjuncts was also further studied by other researchers, see e.g. [59, 73]. In particular, additional experiments with applying other classifiers on the artificial data constructed in the similar way as [34] showed that decision trees were the most sensitive to the small disjuncts, then the next was multi layered perceptron, and support vector machines were the less sensitive to them.

Stefanowski studied in [64] more complicated decision boundaries in two dimensional, numerical data. First data sets, called `sub-clus`, contained rectangles defining the minority class distributions. All these sub-rectangles are surrounded by the uniformly distributed examples from the majority class. Figure 2 represents the next shape, called a `clover`, a more difficult, non-linear setting, where the minority class resembles a flower with elliptic petals (here 3 sub-concepts - petals). The examples of majority class were uniformly distributed in all the free parts. Similarly to earlier Japkowicz et al. research [36] , the size of data was changed (from 200 to 1200 examples) and the imbalance ratio changed from fully balanced IR=1 till more highly imbalanced IR=9. The minority class was

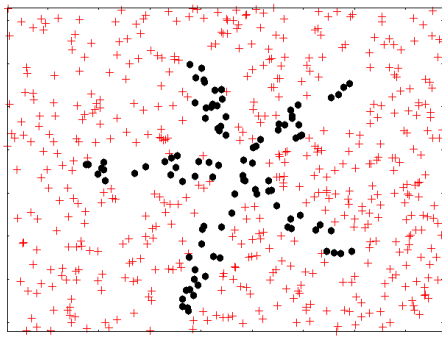

Fig. 2. Clover data set

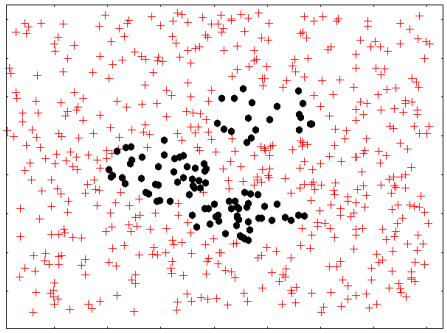

Fig. 3. Paw data set

also stepwise decomposed from 2 to 6 sub-parts. Finally, other non-linear shapes of the minority class sub-concepts were presented in `paw` data, see Figure 3.

Three algorithms: k–nearest neighbor (k-NN), decision tree (J4.8)– and rule (JRIP)–based classifiers were considered. Representative results of the sensitivity measure are shown in Table 1 for k-NN classifier and in Table 2 for decision trees. One can notice that while changing the size of the data - larger number 600 and 400 did not influence so much as 200 ones. The highest decrease of evaluation measures (also for G-mean) was observed for increasing the number of sub-regions of the minority class combined with decreasing the size of a data set - for all sizes of data it degraded the performance of a classifier much more than increasing the imbalanced ratio. The tree and rule classifiers showed the similar performance. The degradation of performance was larger if the decision boundary became non-linear even for larger data set. It is illustrated in Table 2 by results for tree classifier and `clover` data. The stepwise growth of the number of sub-regions (from 2 to 6) in `clover` shape decreases much more the sensitivity measure than stepwise increase of the class imbalance ratio (from 3 to 9).

Table 1. Sensitivity of k-NN classifier with respect to decomposing the minority class into sub-concepts and changing other parameters of sub-class data.

Number of	IR=5			IR=9		
sub-clusters	600	400	200	600	400	200
2	0.82	0.8	0.78	0.78	0.76	0.45
3	0.78	0.72	0.70	0.66	0.74	0.25
4	0.75	0.70	0.68	0.64	0.50	0.15
5	0.73	0.68	0.42	0.58	0.45	0.11
6	0.64	0.62	0.36	0.42	0.32	0.10

Table 2. Sensitivity of a tree classifier with respect to decomposing the minority class into sub-concepts and changing imbalance IR. Data size – 600 and 400 examples.

Number of sub-clusters vs. IR	600				400			
	3	5	7	9	3	5	7	9
2	0.92	0.92	0.83	0.80	0.94	0.85	0.82	0.80
3	0.90	0.85	0.80	0.78	0.84	0.78	0.72	0.70
4	0.85	0.80	0.78	0.74	0.82	0.75	0.68	0.60
5	0.75	0.35	0.24	0.06	0.14	0.10	0	0
6	0.22	0.10	0	0	0.06	0	0	0

4.2 Dealing with Small Disjuncts

As a consequence of this research special approaches to handle the problem of small disjuncts were proposed in [34, 36]. They are based on specialized over-sampling of the minority class, sometimes also the majority class, with respect to inflate small disjuncts. The most well known proposal is *cluster-based over-sampling* [36]. Its idea is to consider not only the between class imbalance but also the within-class imbalance (imbalance between discovered sub-clusters of each class) and to over-sample the data set by improving these two types of imbalances simultaneously. More precisely, the approach workflow is the following:

1. Apply a clustering algorithm to examples from each class separately. In this way, one discovers C_{min} clusters in N_{min} examples from the minority class MK and C_{maj} clusters in N_{maj} examples from the majority class WK.
2. Inside the majority class all the clusters C_{maj}, except the largest one, are randomly oversampled so as to get exactly the same number of examples as inside the largest cluster. In this way the current size of the majority class increases from N_{maj} to Max_{maj}.
3. In the minority class, each cluster is randomly over-sampled until it will contain Max_{maj}/C_{min} examples, where C_{min} is the number of clusters inside the minority class.

As the over-sampled data set will finally contain the same number of examples and all sub-clusters will also be of the same size, the authors claim that no between-class and no within-class imbalances remain inside the transformed data. They successfully applied this approach to several artificial data as well as to 2 real world problems of letter recognition [35] and text classification [57]. In these applications they applied k-means clustering algorithm, although they did not give precise hints how to tune an appropriate k value.

Similarly Borowski [8] considered this pre-processing in text categorization of two larger collection of documents. The first collection was Reuters 21578 and its subset, called MadApte[2], where 9603 documents constituted a training

[2] Reuters data is at http://www.daviddlewis.com/resources/testcollections/reuters21578/

set (the minority class – `corn` – contained 181 examples) while 3299 ones were used a testing set. The other collection was OHSUMED containing text summaries restricted to sub-parts from 23 cardiovascular diseases[3]. The training set contained 10 000 summaries (the minority class – `C01 disease` – has 423 documents) while the testing sets was build on 10 000 summaries. In both collections NLP techniques were applied to extract around 5000 terms in a vector space representation. Then features were selected to around a few hundred by using chi-square and entropy gain filters. Table 3 and 4 summarize the main classification results of using different pre-processing methods with the following classifiers: Naive Bayes (abbreviated as NB), k-nearest neighbour (k-NN), logistic regression (Reg-Log) support vector machines (SVM). For cluster over-sampling we tested 6 values $k = 4, \ldots, 10$ - the best values were 6 and 7 depending on data. SMOTE was applied with 5 neighbours and testing over-sampling ratios o_r between 100% and 1000% (with a step 100) - the best ratio was 400. Note that the cluster over-sampling improved both G-mean and F-measure. However, these improvements were not as high as those achieved by using SMOTE.

Table 3. Applying cluster over-sampling and SMOTE to Reuters data.

	Classifiers				Evaluation
Method	NB	k-NN	Reg-Log	SVM	measure
cluster-	0.42	0.41	0.49	0.45	F
oversample	0.77	0.71	0.77	0.69	G-mean
SMOTE	0.38	0.46	0.47	0.46	F
	0.88	0.83	0.90	0.91	G-mean
no pre-	0.0	0.34	0.18	0.4	F
processing	0.0	0.56	0.33	0.59	G-mean

A quite similar conclusion was reached by another study of Napierala et al. [53] with synthetic data sets – `subclass`, `clover` and `paw` – which were affected by different amounts of disturbance (increasing amount of overlapping and rare examples – this type of examples is further defined in Section 6.1). The representative results are presented in Table 5 where base denotes using a classifier without any pre-processing, RO is a simple random over-sampling, CO – cluster over-sampling, NCR – nearest cleaning rule, and the last column refers to SMOTE. While analyzing these results one can notice that cluster over-sampling is competitive with other methods for data sets containing the minority class without any perturbations. Then, the more complex, overlapped and affected shapes of the minority class sub-parts, the better are other pre-processing methods as SMOTE and SPIDER.

[3] OSHSUMED available at http://ir/ohsu.edu/ohsumed/ohsumed.html

Table 4. Applying cluster over-sampling and SMOTE to Oshumed data.

	Classifiers				Evaluation
Method	NB	k-NN	Reg-Log	SVM	measure
cluster-	0.46	0.40	0.48	0.43	F
oversample	0.72	0.64	0.71	0.68	G-mean
SMOTE	0.34	0.41	0.47	0.49	F
	0.81	0.77	0.83	0.82	G-mean
no pre-	0.13	0.38	0.34	0.46	F
processing	0.27	0.61	0.51	0.65	G-mean

Table 5. G-mean for synthetic data sets with varying degrees of the disturbance ratio.

Data set	Pre-processing method					
	Base	RO	CO	NCR	SPIDER	SMOTE
subclus-0	0.937	0.937	0.948	0.925	0.929	0.938
subclus-30	0.733	0.724	0.724	0.702	0.715	0.712
subclus-50	0.559	0.565	0.602	0.664	0.621	0.704
subclus-70	0.407	0.442	0.469	0.596	0.578	0.541
clover-0	0.739	0.742	0.761	0.778	0.791	0.738
clover-30	0.636	0.637	0.651	0.722	0.676	0.665
clover-50	0.506	0.554	0.549	0.696	0.607	0.601
clover-70	0.418	0.465	0.489	0.658	0.568	0.571
paw-0	0.904	0.913	0.918	0.918	0.902	0.968
paw-30	0.763	0.776	0.771	0.785	0.778	0.833
paw-50	0.657	0.686	0.686	0.752	0.712	0.786
paw-70	0.508	0.582	0.569	0.718	0.651	0.718

Yet another approach to deal with the above-mentioned within class decomposition was presented in [26]. Gumkowski and Stefanowski proposed to use a two phase approach including: (1) clustering and (2) constructing a hierarchical classifiers. More precisely,

1. Use a clustering algorithm to identify sub-concepts of the minority class.
2. Construct Voronoi diagram sub-regions around centroids of the identified minority class clusters; Assign also majority class examples to these sub-regions following the distance to the nearest centroid of the minority class cluster.
3. Learn separate classifiers from learning examples (from both classes) located in each sub-region.
4. Built the arbiter for the set of classifiers - i.e. for a new instance, find to which Voronoi region it belongs and use its classifier to make a final decision.

This approach was implemented in WEKA with X-means clustering algorithm and J4.8 decision trees and its resulting classifier will be further abbreviated as *HC*. X-means is a kind of wrapper around running k-means with different k. The resulting clustering is chosen with respect to optimizing BIC criterion [51].

Table 6. G-mean performance of the hierarchical classifiers with cluster analysis (HC) against a standard decision tree (J4.8)

Data	Classifier	Sensitivity	F	G-mean
paw-0	treeJ48	0.855	0.968	0.713
	HC	0.940	0.975	0.844
paw-separ	treeJ48	0.98	0.925	0.739
	HC	0.961	0.946	0.864
paw-overlap	treeJ48	0.0	0.0	0.0
	HC	0.741	0.81	0.614
paw-outliers	treeJ48	0.0	0.0	0.0
	HC	0.86	0.89	0.729

In Table 6 we show results of using this approach with J4.8 trees over several versions of the synthetic data set `paw`. We used it as it models three different sub-concepts inside the minority class (see its idea in Figure 3). The first data, called `paw-0` is just a version illustrated in this figure without any disturbance. In this construction two sub-concepts are quite close to each other, so may mislead the clustering algorithm (X-means has a tendency to propose 2 clusters instead of three clusters). Therefore, we constructed a version with more separated clusters (moving clusters away) – this is called `paw-separ`. Then, we additionally disturbed minority class shapes by introducing overlapping (`paw-overlap`) and moving more minority examples inside the majority class as outliers.

In case of these synthetic data sets `paw`, where sub-parts are relatively well separated, this algorithm can divide the space into three sub-areas and the hierarchical classifier *HC* improves slightly the sensitivity and other measures comparing to using a single, standard tree. The improvements are a bit higher for `paw-0`, with more difficult separation. For more disturbed data `paws` with overlapping and outliers the standard trees deteriorates its performance while the *HC* classifier maintains its good performance - although values of evaluation measures are smaller than in cleaner shapes. However, we can conclude that in all cases the proposed approach improves evaluation measures.

4.3 Open Issues

Although the idea of identifying and dealing with small disjuncts sounds quite appealing, several open issues remain critical if one needs to analyse real-world data sets. Note that most of the above discussed studies have been carried out with special synthetic data while for real ones the underlying distribution of the

minority class is unknown and it is not easy to approximate (or even to guess) the possible number and structure of sub-concepts.

Up to now most researchers have used clustering algorithms to find these sub-concepts. Other, quite rare studies concern analyzing classification or association rules, mainly their evaluation measures as coverage or similar ones, to distinguish between small and large disjuncts.

For clustering nearly all authors applied k-means algorithm. The main problem is to tune k number of searched clusters. However, other problems include dealing with non-spherical, complex shapes of clusters, an influence of overlapping, noise or outliers. It is also not obvious which clustering evaluation criteria should be considered as the most common ones were proposed for standard unsupervised framework [49]. Here, one deals with at least partly supervised and imbalanced case where one has to distinguish between minority and majority examples inside clusters. Even if clustering algorithms are applied separately to each class the algorithm may improperly agglomerate smaller sub-parts of the minority into too large ones (see experiences with paw data in [26]).

Tuning clustering algorithm parameters in the wrapper framework is also non-trivial. First, it refers to choosing an appropriate quality criterion. Some authors propose to consider tuning clustering together with learning the final classifier and evaluate the integrated approach with special imbalance measures (as e.g. G-mean, AUC). To avoid possible over-fitting it is necessary to use an extra validation set or an internal cross validation inside the training set. This was a case in experimental studies as [8, 61]. However, one should take into account that the data set may be highly imbalanced and it may be difficult, or even impossible, to select a sufficient number of minority examples inside learning, validation and test parts. Perhaps new solutions of partly informed bootstrap sampling could be developed. One should also remember that scanning too many k may be too time consuming or even not realistic.

Nevertheless, k-means may not be the best algorithm to be used in this context, in particular for more complex distributions of the minority class which we will discuss in further sections. Besides non-linear and complex decision shapes of clusters, over-lapping many minority examples could be either singletons like outliers or rare cases (a kind of pairs or triples). Additional experiments with real data sets showed that approaches such as clustering or building hierarchical classifiers are not effective for such difficult data sets [26, 53]. Moving toward density based clustering algorithms is one of the solutions. They can distinguish between core instances (creating clusters) and noisy ones (referring to outliers or rare cases). However tuning parameters of DBSCAN or OPTICS is also not an easy task even in a classical unsupervised version [17]. The current heuristics do not take into account a distinction between minority and majority examples but treat them in the same unsupervised way. Some recent proposals of working with DBSCAN try to look for new heuristics [58]. However, we think that it is necessary to develop a new family of *semi-supervised density algorithms* which take into account labels of examples while constructing neighbour clusters. Finally as

imbalanced data sets usually contain minority class outliers new approaches for their detection are still necessary.

5 Overlapping Between Minority and Majority Classes

Researchers also study different difficulty factors characterizing imbalanced data. An *overlapping* between minority and majority classes is one of them. Although many previous studies in classical machine learning have shown that overlapping of different classes deteriorates the total classification accuracy, its influences on the minority class is thoroughly examined. As the minority class is under-represented in the data set, it will be more likely under-represented also in the overlapping region. As a result, the algorithms may shift the decision boundary too close to the minority class, in the worst case treating the whole overlapping area as belonging to the majority class definition.

5.1 Experimental Studies

Prati et al. started more systematic studies on the role of overlapping [60]. They generated artificial data sets where the minority and the majority class were represented by two clusters in five dimensional space (examples where generated around centroids following the Gaussian distribution). Two parameters were investigated: the imbalance ratio, and the distance between centroids – so classes could be moved from clear separation to high overlapping. For the C4.5 classifier they showed that increasing the overlapping ratio was more responsible for decreasing AUC results than decreasing cardinality of the minority class.

Then, an influence of increasing overlapping was more precisely examined in [23]. Garcia et al. generated two-dimensional data sets with two classes separated by a line orthogonal to one of the axis. They assumed a fixed size of data and changed the overlapping amount for a given imbalance ratio and vice versa. Results of experiments with 6 different classifiers showed that increasing overlapping degraded their performance more (with respect to minority class) than changing the imbalance ratio. Moreover, in the other experiments they fixed the amount of overlapping and changed the distribution of the minority examples by increasing their number in the overlapping area. Again the results confirmed that increasing the local imbalance ratio and the size of the overlapping area were more influential than changing the overall imbalance ratio. However, these factors influenced performance of particular classifiers in a different way. For instance k – nearest neighbor classifier was the most sensitive to changes in the local imbalance region. Naive Bayes, MLP and J4.8 worked better in the dense overlapping region. These conclusions have been later verified in additional experiments (focusing on performance of k-NN and other evaluation measures), see [22]. One of their conclusions was that when overlapping regions increased, the more local classifiers - like k-NN with smaller values of k - performed better with recognition of the minority class.

The other study in [14] focused on the effects of overlapping and class imbalance on support vector machines (SVM). The authors showed that when the overlap level was high, it was unlikely that collecting more training data would produce a more accurate classifier. They also observed that the performance of SVM decreased gradually with the increasing imbalance ratio and overlapping, and that there was a sudden drop when the imbalance ratio equaled to 20% and the overlapping level exceeded 60%, regardless of the training set size.

Prati et al. have recently come back to studying the overlapping in class imbalance [3]. Comparing to their previous work [60] they investigated the usefulness of five different re-sampling methods on the same difficult artificial data sets: popular random-over sampling, random under-sampling, Nearest Cleaning Rule (NCR) [44], SMOTE and SMOTE + ENN [11]. Their main conclusion was that appropriate balancing of training data usually led to a performance improvement of C4.5 classifiers for highly imbalanced data sets with highly overlapped classes. However, the improvements depend on the particular pre-processing method and the overlapping degree. For the highest degree of overlapping it was not clear which method was the best (NCR worked there quite well). Results for other overlapping degrees showed that over-sampling methods in general, and SMOTE-based methods in particular, were more effective than under-sampling. Then, the data cleaning step used in the SMOTE + ENN seemed to be especially suitable in situations having a higher degree of overlapping.

Finally, we come back to our studies [39, 64] where the effect of overlapping was studied together with other factors such as decomposition of the minority class into smaller sub-concepts and more complicated non-linear borders. The k-NN, rules (MODLEM [63]) and J4.8 decision tree classifier were applied to a collection of specially generated artificial data sets sub-class, clover (described in the previous section). Table 7 shows influence of stepwise increase of the amount overlapping on the tree classifier. The degree of overlapping is measured as a percentage of the size of the minority class. It was observed that stepwise increase of overlapping more strongly decrease the sensitivity. For instance, let us analyse the first column (%) - the sensitivity changes from 0.96 to 0.94. While for any of the number of sub-clusters the sensitivity decreases in range of nearly 0.2 (see, e.g. 4 sub-clusters, the sensitivity decreases from 0.96 to 0.78). The similar tendency can be observed for rule and k-NN classifiers.

The influence of overlapping on specialized pre-processing was also studied in [53]. The tree and rule classifiers (J4.8 and MODLEM) were integrated with standard random over-sampling, cluster over-sampling, nearest cleaning rule and SPIDER. All these methods were applied to artificial data sets as `sub-clus`, `clover` and also more complicated versions of `paw` data. The results clearly showed that all methods of pre-processing improved the sensitivity of both classifiers. However, simpler random over-sampling and cluster over-sampling performed comparably on all non-disturbed data sets. While on more difficult sets (disturbance over 30%) both NCR and SPIDER methods were clearly better than there over-sampling methods.

Table 7. Influence of overlapping on the sensitivity of the tree classifier learned from subclass data. Overlapping is expressed by % of borderline examples from the minority class. Total number of examples – 800.

Number of	IR = 5			IR = 9		
sub-clusters	0%	10%	20%	0%	10%	20%
3	0.96	0.91	0.85	0.94	0.9	0.75
4	0.96	0.89	0.78	0.94	0.87	0.74
5	0.96	0.87	0.76	0.90	0.81	0.66
6	0.94	0.84	0.74	0.88	0.68	0.38

5.2 Detecting Overlapping in Real World Data Sets

Note that the data difficulty factors, as overlapping, were examined using mostly artificial data sets [64, 60, 23], in which the data distribution was given a priori and the degree of each factor could be precisely controlled by augmenting or diminishing the degree of overlapping [60, 23] as well as the number and cardinality of small disjuncts [35, 36]. Moreover, the data sets were usually two-dimensional.

The difficult issue is to analyse data factors in real-world imbalanced data sets where the *natural* underlying distribution of minority examples is unknown and has to discovered or at least approximated. Although some researchers published wide comprehensive experimental studies with several popular UCI benchmark data - see e.g. [2, 19, 40, 69], nearly all of them are just comparative experiments of using different standard classifiers [47, 59, 69], ensembles [6, 19] or pre-processing methods [2]. The authors have mainly tried to identify general differences of studied algorithms, quite often without a deeper analysis of hidden data characteristics, or referred to averaged values of these data factors which were easier to be calculated as the total number of minority examples or the global imbalance ratio.

There is not so much research on direct evaluation of overlapping in the real world data sets. For example, in [14] (concerning the effects of overlapping and imbalance on the SVM classifier), the authors propose to estimate the degree of overlapping in real-world data sets by measuring a number of support vectors which can be removed from the classifier without deteriorating the classification accuracy. In the next chapter we will present a simpler and intuitive method based on analyzing local characteristics of minority examples.

6 Types of Minority Examples with Respect to their Local Characteristics

6.1 Motivations

The first paper discussing different types of minority examples is [42] where Kubat and Matwin have distinguished between safe, borderline and noisy examples.

Borderline examples are located in the area surrounding class boundaries, where the minority and majority classes overlap. However, they are not only located in the overlapping zone (discussed in the previous section) as they could also be difficult examples located in specific sub-areas near highly non-linear decision boundaries. *Safe examples* are placed in relatively homogeneous areas with respect to the class label. So, they are inside typical clear parts of target concepts, which located further from the decision boundary between classes. By *noisy examples* they understand individuals from one class occurring in safe areas of the other class. However, authors applied this term to majority class individuals inside the minority class and proposed to remove them from the training set [42].

Here we share these authors' opinions saying that as the minority class is often under-represented in the data, one should be careful with the similar treatment of the singletons from the minority class and rather not recognizing them as noise. Moreover, it is worth to stress that the typical understanding of noisy examples in machine learning corresponds to a kind of data imperfections or errors (see e.g. [20, 62, 70]) which come from either random errors in descriptions of examples or from an inappropriate description language. Researchers usually distinguish between class noise (errors with assigning a correct class label to a given example) or attribute noise (erroneous attribute values which could lead to wrong decisions, in particular, if such an example is located too close to decision boundaries) [9]. The typical approaches to deal with such noisy examples include: (1) identification of suspicious examples and eliminating or correcting them (e.g., by using edited approaches for k-nearest neighbour classifiers) or (2) omitting them during the learning phase to solve overfitting phenomena (e.g., by pruning in decision trees or rules). These approaches may improve the total classification accuracy in the standard learning perspective, see e.g. [9, 20].

However, the role of noisy examples in imbalanced data has not been deeply studied yet. Some authors randomly injected changes of class labels or attribute values to noise free data [1, 70, 62]. In such a way in [1] and [70] effectiveness of standard techniques for handling class noise was evaluated. These two independent experimental results showed that all learning algorithms were sensitive to noise in the minority examples, however some of them, such as Naive Bayes and k - nearest neighbor classifiers, were often more robust than more complex learners, such as support vector machines or Random Forests. In more recent our studies [62], the authors introduced both class noise and attribute noise, by either changing the class label or the attribute values, respectively. The comparison concerned the SMOTE pre-processing method and its several extensions. It showed that SMOTE was sensitive to the noisy data and its extensions which also clean noise introduced by SMOTE, were necessary. In particular, the new proposed specialized noise filter added as post-processing to SMOTE, called SMOTE-IPF, can deal with some of these noisy examples [62].

Napierala and Stefanowski in their papers [53–55, 52] claimed that one should be very careful with directly transferring standard methods for dealing with noise to difficult minority class examples, as it may lead to removal or relabel too high number of minority examples, or to prune too many elements of classifiers mainly

for the minority class. This claim is also consistent with research of Koshgoftar et al. [70] which also stated that in the class imbalance setting, using standard approaches for handling noise "can be catastrophic". The study in [9] also showed that when there is an abundance of data, it is better to detect properly "bad data" at the expense of throwing away "good data", while in case when the data are rare, more conservative filters are better.

What is even more important – the noisy examples are often misclassified with singletons playing a role of *outliers*. Note that the outlier is just an untypical example not coming from erroneous measurements. As the minority class can be under-represented in the data, the minority class singletons located in the majority class areas can be outliers, representing a rare but valid sub-concept of which no other representatives could be collected for training. A quite similar opinion was expressed e.g. in [42], where the authors suggested that minority examples should not be removed as they are too rare to be wasted, even considering the danger that some of them are noisy. In [76], which concerns the detection of noise in balanced data sets, the authors suggest to be cautious when performing automatic noise correction, as it may lead to ignoring outliers which is "questionable, especially when the users are very serious with their data". In our opinion, the minority class examples conform to this case.

We claim that the minority and majority distant examples should be treated in a different way. Majority examples located inside the minority class regions are more likely to be a true noise and they could be candidates for removing or relabeling. In general, noisy majority examples are undesired as they can cause additional fragmentation of the minority class and can increase the difficulties in learning its definition. On the other hand, minority examples considered as outliers should be rather kept in the learning set and properly treated by next pre-processing methods or specialized algorithms for imbalanced data.

Moreover, it is worth to distinguish yet another type of so-called *rare examples*. These are pairs or triples of minority class examples, located inside the majority class region, which are distant from the decision boundary so they are not borderline examples, and at the same time are not pure singletons. The role of these examples has been preliminarily studied in the experiments with special artificial data sets [53, 64]. It has been shown that rare examples significantly degraded the performance of classifiers. Additionally, various preprocessing methods performed differently on such rare examples. Finally, works on graphical visualizations of real-world imbalanced data sets [54] have shown existence of such types of examples. The reader can also analyse Figure 1 where the minority class contains mainly unsafe examples: many borderline, pairs or triples of rare small "islands" and many outliers.

Napierala and Stefanowski in their earlier research [53, 54] claimed that many of considered data difficulty factors could be linked to the distinguishing the following types of examples forming the minority class distribution:

- safe examples
- borderline examples
- rare examples

– outliers

They also claimed that distinguishing these types of examples can be useful to focus attention on difficulties of the minority class distributions, to support interpretations of differences in the performance of classifiers or specialized methods applied to imbalanced data as well as to develop new specialized algorithms. In the next subsection we will briefly discuss some of these issues.

6.2 Identification of Example Types

Distinguishing four types of examples refers to most of previously discussed data difficulty factors. If the minority class distribution will contain mainly unsafe examples, it could indicate that the minority class does not constitute a whole concept but is affected by different disturbances. Although one cannot directly discover sub-concepts, it is possible to indirectly show possible decomposition. A larger number of borderline examples will directly approximate overlapping difficulty factors. Furthermore, rare examples and outliers also express data difficulty discussed in the previous sub-section. Finally, it should be stressed that authors of related works focus rather on studying single data factors and usually do not consider several data factors occurring together. What is even more important to notice, they usually carried out their experiments with artificially generated data, where given perturbations were introduced to assumed data distribution, and rarely attempt to transfer such studies to real world methods.

Therefore, while considering our distinguishing of four types of examples, the research open issue is – how does one can automatically and possibly simply identify these example type in real world data sets (with unknown underlying class distributions).

Note that the visualisation projection methods – discussed in [52] – could confirm the occurrence of different types of examples in some real-world data sets but they cannot be directly applied in the real-world settings. First of all, they cannot be used for very large data sets, as the visualisation of thousands of points would be difficult to read. Secondly, the projection to two dimensions may not always be feasible, as the data set may be intrinsically characterized by more dimensions.

Furthermore, as we attempt to stress in earlier sections, it is practically easy to directly measure only the simplest data characteristics as the global imbalanced ratio, data size, etc. while other more influential data factors are rather difficult to precisely estimate in real world, not trivial data sets. Some of already proposed methods may rather very roughly indicate the presence of the given data factors. For instance, in [14] (concerning the effects of overlapping and imbalance on the support vector machine classifier), the authors proposed to approximate the possible amount of overlapping in real-world data sets by measuring a number of support vectors which can be removed from the classifier without deteriorating the classification accuracy. Other methods for identification of unsafe or noisy examples are based on an extensive using cross-validated ensembles, bagging and boosting. However, their parameters are not easy to tune.

Moreover, not all instances misclassified by ensembles may be noisy examples as some of them could be rather difficult but valid examples.

Therefore, Napierala and Stefanowski have looked for new simple techniques which should more directly identify the difficult types of example distributions in imbalanced data. Moreover, they could be more intuitive for user with respect to principles and rules of their parametrization.

The proposed method origins from the hypotheses [54] on role of the mutual positions of the learning examples in the attribute space and the idea of assessing the type of example by analyzing class labels of the other examples in its *local neighbourhood*. By a term local we understand that one should focus on the processing characteristics of the nearest examples due to the possible sparse decomposition of the minority class into rather rare sub-concepts with non-linear decision boundaries. Considering a larger size of the neighbourhood may not reflect the underlying distribution of the minority class.

In general, such a neighbourhood of the minority class example could be modeled in different ways. In further considerations we will use an analysis of the class labels among *k-nearest neighbours* [54, 52]. An alternative approach to model the local neighbourhood with *kernel functions* has been recently presented in [52] – however, its experimental evaluation has given similar conclusions as to data characteristics.

Constructing the k – neighbourhood involves decisions on choosing the value of k and the *distance function*. In our previous considerations we have followed results of analyzing different distance metrics [46] and chose the HVDM metric (*Heterogeneous Value Difference Metric*) [75]. Its main advantage for mixed attributes is that it aggregates normalized distances for qualitative and quantitative attributes. In particular, comparing to other metrics HVDM provides more appropriate handling of qualitative attributes as instead of simple value matching, as it makes use of the class information to compute attribute value conditional probabilities by using a Stanfil and Valtz value difference metric for nominal attributes [75]. Tuning k value should be done more carefully. In general, different values may be considered depending on the data set characteristic. Values smaller than 5, e.g. $k = 3$, may poorly distinguish the nature of examples, especially if one wants to assign them to four types. Too high values, on the other hand, would be inconsistent with the assumption of the locality of the method and not useful while dealing with complex, non-linear and fragmented distributions of the minority class. In this paper we do not solve the problem of an automatic tuning this value with respect to complexity of the minority class distribution and its difficulty factors, leaving it for future research.

Experiments from [52] over many UCI data sets have showed that choosing $k = 5$, 7, 9 and 11 values has led to quite similar categorizations of data with respect to proportions of the minority class types. Below we will show assigning types minority class for the smallest k values.

Depending on the number of examples from the majority class in the local neighbourhood of the given minority class example, we can evaluate whether this example could be safe or unsafe (difficult) to be learned. If all, or nearly

all, its neighbours belong the minority class, this example is treated as the safe example, otherwise it is one of unsafe types. For instance, in case of $k = 5$ the type of example x is defined as follows:

- if 5 or 4 of its neighbours belong to the same class as x, it is treated as a safe example;
- if the numbers of neighbours from both classes are similar (proportions 3:2 or 2:3) – it is a borderline example;
- if it has only one neighbour with the same label (1:4) it is a rare example;
- if all neighbours come from the opposite class (0:5) – it is an outlier.

Similar interpretations can be extended for larger values of k. For instance, in case of $k = 7$ and the neighbourhood distribution 7:0 or 6:1 or 5:2 – a safe example; 4:3 or 3:4 – a borderline example; again the number of neighbours from both classes are approximately the same; 2:5 or 1:6 – a rare example; and 0:7 – an outlier. Such an interpretation can be extended for larger neighbourhoods and even tuning bandwidth in kernels – see such an analysis in [52].

The analysis of this neighbourhood has been applied in experiments with UCI imbalanced real-world data sets [54, 52]. The results of labeling types of minority class examples are presented in Table 8. Note that many data sets contain rather a small number of safe minority examples. The exceptions are three data sets composed of almost only safe examples: `flags, breast-w, car`. On the other hand, there are data sets such as `cleveland, balance-scale` or `solar-flare`, which do not contain any safe examples. We carried out a similar neighbourhood analysis for the majority classes and made a contrary observation – nearly all data sets contain mainly safe majority examples (e.g. `yeast`: 98.5%, `ecoli`: 91.7%) and sometimes a limited number of borderline examples (e.g. `balance-scale`: 84.5% safe and 15.6% borderline examples). What is even more important, nearly all data sets do not contain any majority outliers and at most 2% of rare examples. These results show that outliers and rare examples can constitute an important part of the minority class – there are some data sets where they even prevail in the minority class. Therefore, one should be cautious with considering all of them as noise and applying noise-handling methods such as relabeling or removing these examples from the learning set.

6.3 Influence of Example Types on Classification Performance

The results of labeling the minority class examples can also be used to categorize data sets. depending on the dominating type of examples from the minority class. Only in `abdominal-pain, acl, new-thyroid and vehicle` data sets, safe minority examples prevail. Therefore, we can treat these 4 data sets as representatives of *safe* data sets. In the next category the borderline examples dominate in the distribution of the minority class. As could be observed in Table 8, even in data sets with clean borders a considerable amount of examples (up to 36%) can be labeled as borderline ones. So, the percentage of borderline examples must be even higher to represent some overlapping between classes.

Table 8. Labeling minority examples expressed as a percentage of each type of examples occurring in this class.

Data set	Safe	Border	Rare	Outlier
abdominal_pain	61.39	23.76	6.93	7.92
balance-scale	0.00	0.00	8.16	91.84
breast-cancer	21.18	38.82	27.06	12.94
breast-w	91.29	7.88	0.00	0.83
bupa	20.69	76.55	0.00	2.76
car	47.83	47.83	0.00	4.35
cleveland	0.00	45.71	8.57	45.71
cmc	13.81	53.15	14.41	18.62
credit-g	15.67	61.33	12.33	10.67
ecoli	28.57	54.29	2.86	14.29
flags	100.00	0.00	0.00	0.00
haberman	4.94	61.73	18.52	14.81
hepatitis	18.75	62.50	6.25	12.50
hsv	0.00	0.00	28.57	71.43
ionosphere	44.44	30.95	11.90	12.70
new-thyroid	68.57	31.43	0.00	0.00
pima	29.85	56.34	5.22	8.58
postoperative	0.00	41.67	29.17	29.17
scrotal_pain	50.85	33.90	10.17	5.08
solar-flareF	2.33	41.86	16.28	39.53
transfusion	18.54	47.19	11.24	23.03
vehicle	74.37	24.62	0.00	1.01
yeast-ME2	5.88	47.06	7.84	39.22

We could treat a data set as a *borderline* data set if it contains more than 50% of borderline examples - for instance these are `credit-g, ecoli, haberman, hepatitis`. Additional data sets - as `car` and `scrotal-pain` – are located somewhere between safe and borderline categories. As the amount of safe examples is too low, they are mostly inside the borderline category. Then, several data sets contain many rare examples. Although they are not as numerous as borderline examples, they constitute even 20-30% of the minority class. The rare category includes `haberman` (also assigned to borderline category), `cmc, breast-cancer, cleveland, glass, hsv` and `abalone` data sets, which have at least 20% of rare examples. Other data sets contain less than 10% of these examples. Finally, some data sets contain a relatively high number of outlier examples – sometimes more than a half of the whole minority class. We can assign the data set to outlier category if more than 20% of examples are labeled as outliers.

In previous studies [54, 52] we compared different learning algorithms and shown that distinguishing these data characteristics is co-related with differentiating differences in the performance of classifiers. First, for the safe data nearly all compared single classifiers (SVM, RBF, k-NN, J4.8 decision trees or PART rules) perform quite well with respect to sensitivity, F-measure or G-

mean. The larger differentiation of classifiers occurs for more unsafe data sets. For instance, SVM and RBF classifiers work much better for safe category, while rare or outlier data strongly deteriorate their classification performance. Rare and especially outlier examples are extremely difficult to recognize. PART, J48 and sometimes 1NN may classify them but at a very low level. On the other hand, SVM and RBF fail to classify minority examples in these data sets.

Similar analysis has been carried out for the most representative pre-processing approaches, showing that the competence area of each method depends on the data difficulty level, based on the types of minority class examples [56]. Again in the case of safe data there are no significant differences between the compared methods - even random over-sampling works quite accurate. However, for borderline data sets Nearest Cleaning Rules performs best. On the other hand, SMOTE [11] and SPIDER [65], which can add new examples to the data, have proved to be more suitable for rare and outlier data sets.

For more details on the competence of each studied single classifier and pre-processing methods see [52]. Moreover, our results often confirm the results of the related works conducted on artificial data sets, see [2, 22, 53].

Finally, yet another analysis for different generalizations of bagging ensembles specialized for class imbalances, have been carried out in our recent papers [6, 7]. For safe data sets nearly all bagging extensions for imbalanced data achieve similar high performance. The strong differences between classifiers occur for the most difficult data distributions with a limited number of safe minority examples. Furthermore, the best improvements of all evaluation measures for Roughly Balanced Bagging and Nearest Balanced Bagging are observed for the most unsafe data sets with many rare examples and outliers [7].

7 Final Remarks and Open Research Challenges

This paper concerns problems of learning classifiers from imbalanced data. Although many specialized methods have been introduced, it is still a challenging problem. We claim that besides developing new algorithms for improving classifiers, it is more interesting to ask more general research questions on the nature of the class imbalance problem, properties of an underlying distribution of the minority class in data, and its influence on performance of various classifiers and pre-processing methods.

The main aim of this study is to discuss the data difficulty factors which correspond to sources of difficulties in recognizing the minority class. Following the literature survey and own studies we have focused our attention on the following factors:

- decomposition of the minority class into rare sub-concepts,
- overlapping of classes and borderline examples,
- distinguishing different types of the minority examples.

For each difficulty factor we have discussed its influence of classification performance and details of its practical identification in real data sets. The main

lesson from various experiments is that these factors are more influential than the global imbalance ratio or the absolute size of the minority class which have been more often considered in the related literature up to now.

Our experiments with synthetics data have clearly showed that increasing data complexity (understood as decomposition of the minority class into many sub-parts) decreased evaluation measures more than changing the imbalance ratio or the absolute size of the class. We have also showed that combining the minority class decomposition with non-linear decision boundaries and overlapping makes the learning task extremely difficult. However, as it has been discussed and showed on several illustrative examples, identification of sub-clusters (corresponding to small disjuncts) in real world data, e.g. by clustering algorithms, is still an open research challenge. In particular, it is not obvious how to tune algorithm parameters (e.g. a number of expected clusters in k-mean) and to deal with complex shapes or outliers. We think that developing a new kind of a semi-supervised density based algorithm (where it is necessary to deal with presence of minority vs. majority examples inside clusters) could be a promising research direction. Similar limitations are manifested by current methods for identification of overlapping minority and majority class distributions.

The other novel contributions are distinguishing different types of minority examples and proposing a new method for their identification in real world data sets. This identification method is based on analyzing class distribution inside the local k-neighbourhood of the minority examples. It can also approximate many discussed data difficulty factors, except discovering small disjuncts. Its experimental evaluation has led us to several novel observations with respect to earlier studies on imbalanced data. First, analyzing types of examples in many UCI imbalanced data sets has showed that safe examples are uncommon in most of the imbalanced data. They rather contain all types of examples, but in different proportions. Depending on the dominating type of identified minority examples, the considered data sets could be categorized as: safe, border, rare or outlier. Borderline examples appear in most of the data sets and often constitute more than a half of the minority class. We could also observe that rare and outlier examples are not only extremely difficult for most of the learning methods, but they are often quite numerous in the imbalanced data sets.

Our other comparative experiments have showed that the classifier performance could be related to the above mentioned categories of data. First, for the safe data nearly all compared single classifiers perform quite well. The larger differentiation occurs for more unsafe data set. For instance, support vector machines and RBF neural networks work much better for safe data category, while rare or outlier data strongly deteriorate their classification performance. On the other hand, unpruned decision trees and k-NN classifiers work better for more unsafe data sets. Similar analysis has been carried out for the most representative pre-processing approaches, showing that the competence area of each method also depends on the data difficulty level; For more details see [52]. The other experiments for different generalizations of bagging ensembles for class imbalances, have been carried out in the recent paper [6].

We also claim that the appropriate treatment of these factors, in particular types of minority example, within new proposals of either pre-processing or classifiers, should lead to improving their classification performance. Although it is not inside the scope of this paper, we mention that such research has already been undertaken and resulted in proposing: informed pre-processing method LN-SMOTE [48], rule induction algorithm BRACID [55] and nearest neighbour generalization of bagging, called NBBag [7].

On the other hand, several topics still remain open issues for future research. Besides already mentioned semi-supervised clustering for detecting small disjuncts, one could look for a more flexible method of tuning k in the local neighborhood method for identification of types of examples with respect to the given data set; studying differences between outliers and real noise; detecting singleton examples in empty spaces (which is an absolute rarity different to the situation of single examples surrounded by k-neighbours from opposite classes), developing a new method for dealing with such examples, re-considering k-neighbourhood methods in highly dimensional spaces, studying different over-sampling with respect to identified different characteristics of sub-areas of data. Finally, it is worth to consider mutli-class imbalanced problems, where at least two smaller classes are particularly interesting to experts and they prefer to improve their recognition separately and do not allow to aggregate them together. Although some authors have already attempted to decompose this problem into one-against all or pairwise coupling classifiers, we think it would be more beneficial to look for another framework with unequal costs of misclassifications between classes.

Acknowledgment. The research was funded by the the Polish National Science Center, grant no. DEC-2013/11/B/ST6/00963. Close co-operation with Krystyna Napierala in research on types of examples is also acknowledged.

References

1. Anyfantis, D., Karagiannopoulos, M., Kotsiantis, S., Pintelas, P.: Robustness of learning techniques in handling class noise in imbalanced datasets. In Proc. of the IFIP Conf. AIAI 2007, 21–28 (2007).
2. Batista, G.,Prati, R., Monard, M.: A study of the behavior of several methods for balancing machine learning training data. ACM SIGKDD Explorations Newsletter, 6(1), 20–29 (2004).
3. Batista, G., Prati, R., Monard, M.: Balancing strategies and class overlapping. In Proc. IDA 2005, Springer LNCS vol. 3646 , 24–35 (2005).
4. Bishop, Ch .: Pattern Recognition and Machine Learning (Information Science and Statistics), Springer-Verlag New York, Inc., (2006).
5. Błaszczyński J., Deckert M., Stefanowski J., Wilk Sz.: Integrating selective preprocessing of imbalanced data with Ivotes ensemble. In Proc. of 7th Int. Conf. RSCTC 2010, Springer, LNAI vol. 6086, 148-157 (2010).
6. Błaszczyński, J., Stefanowski, J., Idkowiak L.: Extending bagging for imbalanced data. In Proc. of the 8th CORES 2013, Springer Series on Advances in Intelligent Systems and Computing 226, 269–278 (2013).
7. Błaszczyński, J., Stefanowski, J.: Neighbourhood sampling in bagging for imbalanced data. Neurocomputing, 150 (Part B) , 529-542 (2015).

8. Borowski, J.: Constructing data representations and classification of imbalanced text documents. Master Thesis, Poznan University of Technology (supervised by Stefanowski J.) (2014).
9. Brodley, C. E., Friedl, M. A: Identifying mislabeled training data. Journal of Artificial Intelligence Research, vol. 11, 131–167 (1999).
10. Chawla, N.: Data mining for imbalanced datasets: An overview. In Maimon O., Rokach L. (eds.): The Data Mining and Knowledge Discovery Handbook, Springer, 853–867 (2005).
11. Chawla, N., Bowyer, K., Hall, L., Kegelmeyer, W.: SMOTE: Synthetic Minority Over-sampling Technique. J. of Artificial Intelligence Research, 16, 341-378 (2002).
12. Cost, S., Salzberg, S.: A Weighted Nearest Neighbor Algorithm for Learning with Symbolic Features. Machine Learning Journal, vol. 10 (1), 1213–1228 (1993).
13. Davis, J., Goadrich, M.: The Relationship between Precision- Recall and ROC Curves, In Proc. Int. Conf. on Machine Learning ICML 2006, 233–240 (2006).
14. Denil, M., Trappenberg, T.: A characterization of the combined effects of overlap and imbalance on the SVM classifier. In Proc. CoRR Conf., 1–10, (2011).
15. Drummond C., Holte R.: Cost curves: An improved method for visualizing classifier performance. Machine Learning, vol. 65(1), 95-130 (2006).
16. Elklan C.: The foundations of cost-sensitive learning. In Proc. Int. Joint Conf. on Artificial Intelligence IJCAI-01, 63–66 (2001).
17. Ester, M., Kriegel, H., Sander, J., Xu, X.: A density-based algorithm for discovering clusters in large spatial databases. In Proc. Int. Conf. KDD'96, 226–231, (1996).
18. Fernandez, A., Garcia, S., Herrera, F.: Addressing the Classification with Imbalanced Data: Open Problems and New Challenges on Class Distribution. In Proc. HAIS Conf. (part. 1), 1-10, (2011).
19. Galar, M., Fernandez, A., Barrenechea, E., Bustince, H. Herrera, F.: A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches. IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, vol. 99, 1–22 (2011).
20. Gamberger, D., Boskovic, R., Lavrac, N., Groselj, C.: Experiments With Noise Filtering in a Medical Domain. In Proc. of the 16th International Conference on Machine Learning ICML'99, 143–151 (1999)
21. Garcia, S., Herrera, F.: Evolutionary undersampling for classification with imbalanced datasets: Proposals and taxonomy. Evolutionary Computation 17 (3), 275-306 (2009).
22. Garcia, V., Mollineda, R., Sanchez, J.S.: On the k-nn performance in a challenging scenario of imbalance and overlapping. Pattern Analysis and Applications, 11(3-4):269–280, 2008.
23. Garcia, V., Sanchez, J.S., Mollineda, R.A.: An empirical study of the behaviour of classifiers on imbalanced and overlapped data sets. In Proc. of Progress in Pattern Recognition, Image Analysis and Applications 2007, Springer, LNCS, vol. 4756, 397–406 (2007).
24. Grzymala-Busse, J.W., Goodwin, L.K., Grzymala-Busse, W., Zheng, X.: An approach to imbalanced data sets based on changing rule strength. In Proc. of Learning from Imbalanced Data Sets, AAAI Workshop at the 17th Conference on AI, 69–74, (2000).
25. Grzymala-Busse, J.W., Stefanowski, J., Wilk, S.: A comparison of two approaches to data mining from imbalanced data. Journal of Intelligent Manufacturing, 16 (6), 565-574 (2005).
26. Gumkowski, M.: Using cluster analysis to classification of imbalanced data. Master Thesis, Poznan University of Technology (supervised by Stefanowski J.) (2014).

27. Han H., Wang W., Mao B.: Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning. In Proc. ICIC, Springer LNCS vol. 3644, 878-887 (2005).
28. Hand, D.: Measuring classifier performance. A coherent alternative to the area under the ROC curve. Machine Learning, vol. 42, 203-231, (2009).
29. He, H., Garcia, E.: Learning from imbalanced data. IEEE Transactions on Data and Knowledge Engineering, vol. 21 (9), 1263–1284 (2009).
30. He, H., Yungian, Ma (eds): Imbalanced Learning. Foundations, Algorithms and Applications. IEEE - Wiley, (2013).
31. Hido, S., Kashima, H.: Roughly balanced bagging for imbalance data. Statistical Analysis and Data Mining, 2 (5-6), 412–426 (2009).
32. Holte, C., Acker, L.E., Porter, B.W.: Concept Learning and the Problem of Small Disjuncts. In Proc. of the 11th IJCAI Conference, 813–818 (1989).
33. Japkowicz, N.: Concept-Learning in the Presence of Between-Class and Within-Class Imbalances. In Proc. Canadian Conference on AI 2001: 67-77 (2001).
34. Japkowicz, N.: Class imbalance: Are we focusing on the right issue? In Proc. II Workshop on Learning from Imbalanced Data Sets, ICML Conference, 17–23, (2003).
35. Japkowicz, N., Stephen, S.: Class imbalance problem: a systematic study. Intelligent Data Analysis Journal, vol. 6 (5), 429-450 (2002).
36. Jo, T., Japkowicz, N.: Class Imbalances versus small disjuncts. ACM SIGKDD Explorations Newsletter, 6 (1), 40–49 (2004).
37. Japkowicz, N., Shah, Mohak: Evaluating Learning Algorithms: A Classification Perspective. Cambridge University Press (2011).
38. Japkowicz, N.: Assessment metrics for imbalanced learning. In He, H., Yungian, Ma (eds): Imbalanced Learning. Foundations, Algorithms and Applications. IEEE - Wiley, 187-206, (2013).
39. Kaluzny, K: Analysis of class decomposition in imbalanced data. Master Thesis (supervised by J.Stefanowski), Poznan University of Technology, (2009).
40. Khoshgoftaar, T., Van Hulse, J., Napolitano, A.: Comparing boosting and bagging techniques with noisy and imbalanced data. IEEE Transactions on Systems, Man, and Cybernetics–Part A, 41 (3), 552–568 (2011).
41. Krawczyk B., Wozniak M., Schaefer G.: Cost-sensitive decision tree ensembles for effective imbalanced classification. Applied Soft Computing 14, 544-562 (2014).
42. Kubat, M., Matwin, S.: Addressing the curse of imbalanced training sets: one-side selection. In Proc. of the 14th Int. Conf. on Machine Learning ICML-97, 179-186 (1997).
43. Kubat, M., Holte, R. and Matwin, S.: Machine Learning for the Detection of Oil Spills in Radar Images. Machine Learning Journal, vol. 30, 195-215 (1998)
44. Laurikkala, J.: Improving identification of difficult small classes by balancing class distribution. Tech. Report A-2001-2, University of Tampere, (2001).
45. Lewis, D., Catlett, J.: Heterogeneous uncertainty sampling for supervised learning. In Proc. of 11th Int. Conf. on Machine Learning, 148–156 (1994).
46. Lumijarvi, J., Laurikkala, J., Juhola, M.: A comparison of different heterogeneous proximity functions and Euclidean distance. Stud Health Technol. Inform., 107 (Part 2), 1362–1366 (2004).
47. Lopez, V., Fernandez, A., Garcia, S., Palade, V., Herrera, F.: An Insight into Classification with Imbalanced Data: Empirical Results and Current Trends on Using Data Intrinsic Characteristics. Information Sciences 257, 113-141,(2014).

48. Maciejewski, T., Stefanowski, J.: Local neighbourhood extension of SMOTE for mining imbalanced data. In Proc. IEEE Symp. on Computational Intelligence and Data Mining, 104–111 (2011).
49. Maimon O., Rokach L. (eds.): The Data Mining and Knowledge Discovery Handbook, Springer, 2005.
50. Maloof M.: Learning When Data Sets are Imbalanced and When Costs are Unequal and Unknown. In Proc. II Workshop on Learning from Imbalanced Data Sets, ICML Conference, (2003).
51. Moore, A., Pelleg, D.: X-means: extending k-means with efficient estimation of the numbers of clusters, In Proc. 17th ICML, 727-734, (2000).
52. Napierala, K.: Improving rule classifiers for imbalanced data. Ph.D. Thesis. Poznan University of Technology, (2013).
53. Napierala, K., Stefanowski, J., Wilk, Sz.: Learning from imbalanced data in presence of noisy and borderline Examples. In Proc. of 7th Int. Conf. RSCTC 2010, Springer, LNAI vol. 6086, 158-167 (2010).
54. Napierala, K., Stefanowski, J.: The influence of minority class distribution on learning from imbalance data. In Proc. 7th Conf. HAIS 2012, LNAI vol. 7209, Springer, 139-150 (2012).
55. Napierala, K., Stefanowski, J.: BRACID: a comprehensive approach to learning rules from imbalanced data. Journal of Intelligent Information Systems, vol. 39 (2), 335-373 (2012).
56. Napierala, K., Stefanowski, J., Trzcielinska, M.: Local characteristics of minority examples in pre-processing of imbalanced data. In Proc. ISMIS 2014: 123-132, (2014).
57. Nickerson, A., Japkowicz, N., Milios, E.: Using unsupervised learning to guide resampling in imbalanced data sets. In Proc. of the 8th Int. Workshop on Artificial Intelligence and Statistics, 261–265 (2001).
58. Niemann, U., Spiliopoulou, Volzke, H., Kuhn J.P.: Subpopulation discovery in epidemiological data with subspace clustering. Foundations of Computing and Decision Sciences, vol. 39 (4), (2014).
59. Prati, R., Gustavo, E., Batista, G., Monard, M.: Learning with class skews and small disjuncts. In Proc. SBIA 2004, Springer LNAI vol. 3171, 296–306 (2004).
60. Prati, R., Batista, G., Monard, M.: Class imbalance versus class overlapping: an analysis of a learning system behavior. In Proc. 3rd Mexican Int. Conf. on Artificial Intelligence, 312–321 (2004).
61. Parinaz, S., Victor H., Matwin S.: Learning from Imbalanced Data Using Ensemble Methods and Cluster-based Undersampling. In Proc. NFMCP 2014 Workshop at ECML-PKDD 2014, Nancy (2014).
62. Saez, J.A., Luengo, J., Stefanowski, J., Herrera, F.: Addressing the noisy and borderline examples problem in classification with imbalanced datasets via a class noise filtering method-based re-sampling technique. Information Sciences, 291, 184-203 (2015).
63. Stefanowski, J.: On combined classifiers, rule induction and rough sets. Transactions on Rough Sets, 6, 329–350 (2007).
64. Stefanowski, J.: Overlapping, rare examples and class decomposition in learning classifiers from imbalanced data. In S.Ramanna, L.C. Jain, and R.J. Howlett (eds), Emerging Paradigms in Machine Learning, 277-306 (2013).
65. Stefanowski, J., Wilk, Sz.: Selective pre-processing of imbalanced data for improving classification performance. In Proc. of the 10th Int. Conf. DaWaK 2008. LNCS vol. 5182. Springer, 283–292 (2008).

66. Stefanowski J., Wilk Sz.: Extending rule-based classifiers to improve recognition of imbalanced classes. Z.W. Ras, A. Dardzinska (eds): Advances in Data Management, Studies in Computational Intelligence, Springer Verlag, 223, 131–154, (2009).
67. Ting K.: The problem of small disjuncts. Its remedy in decision trees. In Proc. of the 10th Canadian Conf. on AI, 91-97, (1997).
68. Tomek, I.: Two Modifications of CNN. IEEE Transactions on Systems, Man and Communications 6, 769–772 (1976).
69. Van Hulse, J., Khoshgoftarr, T., Napolitano, A.: Experimental perspectives on learning from imbalanced data. In Proc. of ICML 2007, 935–942 (2007).
70. Van Hulse, J., Khoshgoftarr, T.: Knowledge discovery from imbalanced and noisy data. Data and Knowledge Engineering, vol. 68 1513–1542 (2009)
71. Verbiest, N., Ramentol, E., Cornelis, C., Herrera F.: Improving SMOTE with fuzzy rough prototype selection to detect noise in imbalanced classification data. In Proc. Int. Conf. IBERAMIA 2012, 169-178 (2012).
72. Weiss, G.M.: Mining with rarity: a unifying framework. ACM SIGKDD Explorations Newsletter, vol. 6 (1), 7-19 (2004).
73. Weiss, G.M., Hirsh, H.: A quantitative study of small disjuncts. In Proc. the 17th National Conference on Artificial Intelligence – AAAI00, 665-670, (2000).
74. Weiss, G.M., Provost, F.: Learning when training data are costly: the efect of class distribution on tree induction. Journal of Artificial Intelligence Research 19, 315–354 (2003).
75. Wilson, D.R., Martinez, T.R.: Improved heterogeneous distance functions. Journal of Artificial Intelligence Research, 6, 1-34 (1997).
76. Zhu, X., Wu, X., Yang, Y.: Error detection and impact-sensitive instance ranking in noisy data sets. In Proc. 19 Nat.Conf. on AI, AAAI'04 (2014).