Chapter 11
Overlapping, Rare Examples and Class Decomposition in Learning Classifiers from Imbalanced Data

Jerzy Stefanowski

Abstract. This paper deals with inducing classifiers from imbalanced data, where one class (a minority class) is under-represented in comparison to the remaining classes (majority classes). The minority class is usually of primary interest and it is required to recognize its members as accurately as possible. Class imbalance constitutes a difficulty for most algorithms learning classifiers as they are biased toward the majority classes. The first part of this study is devoted to discussing main properties of data that cause this difficulty. Following the review of earlier, related research several types of artificial, imbalanced data sets affected by critical factors have been generated. The decision trees and rule based classifiers have been generated from these data sets. Results of first experiments show that too small number of examples from the minority class is not the main source of difficulties. These results confirm the initial hypothesis saying the degradation of classification performance is more related to the minority class decomposition into small sub-parts. Another critical factor concerns presence of a relatively large number of borderline examples from the minority class in the overlapping region between classes, in particular for non-linear decision boundaries. The novel observation is showing the impact of rare examples from the minority class located inside the majority class. The experiments make visible that stepwise increasing the number of borderline and rare examples in the minority class has larger influence on the considered classifiers than increasing the decomposition of this class. The second part of this paper is devoted to studying an improvement of classifiers by pre-processing of such data with re-sampling methods. Next experiments examine the influence of the identified critical data factors on performance of 4 different pre-processing re-sampling methods: two versions of random over-sampling, focused under-sampling NCR and the hybrid method SPIDER. Results show that if data is sufficiently disturbed by borderline and rare examples SPIDER and partly NCR work better than over-sampling.

Jerzy Stefanowski
Institute of Computing Science, Poznań University of Technology, ul. Piotrowo 2, 60–965 Poznań, Poland
e-mail: jerzy.stefanowski@cs.put.poznan.pl

S. Ramanna, L.C. Jain, & R.J. Howlett (Eds.): Emerging Paradigms in ML, SIST 13, pp. 277–306.
springerlink.com

11.1 Introduction

Supervised learning of classifiers from examples is one of the main tasks in machine learning and data mining. Many approaches based on different principles have been introduced in last decades, for reviews, see e.g. [34, 41]. However, their usefulness for obtaining high predictive accuracy in real life data depends on different factors, including also difficulties of the learning problem and its data characteristics. *Class imbalance* is one of the sources of these difficulties.

A data set is considered to be imbalanced if one of target classes contains much smaller number of examples than the other classes. The under-represented class is called the *minority class*, while the remaining classes are referred to as *majority classes*.

Many real life problems are characterized by a highly imbalanced distribution of examples in classes. Typical examples are rare medical diagnosis [26], recognition oil spills in satellite images [36], detecting specific astronomical objects in sky surveys [45] or technical diagnostics of equipment failures. Moreover, in fraud detection, either in card transactions [17] or in telephone calls [5] the number of legitimate transactions is much higher than the number of fraudulent ones. Similar situations occur either in direct marketing where the response rate class is usually very small in most marketing campaigns [39] or information filtering where some important categories contain few messages only [38]. Other practical problems are also discussed in [11, 18, 19, 62].

If imbalance in the class distribution is extensive, i.e. some classes are *strongly under-represented*, then the typical learning methods do not work properly. An even class distribution is often assumed (also non explicitly) and the classifiers are "somehow biased" to focus searching on the more frequent classes while "missing" examples from the minority class. As a result constructed classifiers are also biased toward recognition of the majority classes and they usually have difficulties (or even are unable) to classify correctly new objects from the minority class. In [38] authors described an information retrieval system, where the minority class (being of a primary importance) contained only 0.2% of all examples. Although the classifiers achieved the overall accuracy close to 100%, they were useless because they failed to deliver requested documents from this class. Similar degradation of classifier's performance for the minority class was also reported for other imbalanced problems, see e.g. [9, 26, 29, 35, 43, 62].

Learning from imbalanced data is considered by some researchers as one of the most challenging topics in machine learning and data mining [65]. It has received growing research interest in the last decade and several specialized methods have already been proposed, see [11, 12, 18, 62] for a review. These methods are usually categorized in two groups:

- The first group includes classifier-independent methods that rely on transforming the original data to change the distribution of classes, e.g., by re-sampling.
- The other group involves modifications of either a learning phase of the algorithm, classification strategies, construction of specialized ensembles or adaptation of cost sensitive learning.

This paper concerns the first group as these methods are more universal and they can be used in a pre-processing stage before applying many learning algorithms. While the other group includes many quite specialized methods based on different principles. For instance, many authors changed search strategies, evaluation criteria or parameters in the internal optimization of the algorithm, see e.g [23, 27, 31, 61, 62, 63]. A survey of special changes in ensembles is given in [21], while adaptations to cost sensitive learning are reviewed in [18].

Before focusing our interest on some of pre-processing methods, we want to ask a more general question about the nature of class imbalance problem and to study the key properties of data distribution which make learning classifiers so difficult. A small number of examples in the minority class is not the only source of difficulties for classifiers. Recent works also suggest that there are other factors that contribute to difficulties. The most well known studies with artificial data are the works of Japkowicz [29, 30], who showed that simple class imbalance ratio was not the main difficulty. The degradation of performance was also related to other factors, mainly to decomposition of the minority class into many sub-clusters with very few examples. The rare sub-concepts correspond to, so called, *small disjuncts*, which lead to classification errors more often than examples from larger parts of the class [20]. Other researchers also explored the effect of *overlapping* between imbalanced classes – more recent experiments on artificial data with different degrees of overlapping also showed that overlapping was more important than the overall imbalance ratio [24, 47].

However, the authors of the above mentioned papers considered these factors independently each other. It could be worth to investigate them occurring together in the data also in presence of other factors. In earlier studies Stefanowski and his cooperators have noticed that many imbalanced data (e.g. coming from UCI repository [2] and used in many papers on new approaches to class imbalance) contain also minority class examples located inside the majority class [40, 42]. They could treated as *outliers* (in particular, if they are single examples surrounded by many examples from majority classes) or *rare cases* (if they are not single ones). They should not be considered as noise as they are too rare and too precious for the minority class. According to the best knowledge this kind of rare examples has not been examined in studies with imbalanced data. Furthermore, it could be interesting to consider the role of changing decision boundary between classes from linear to non-linear shapes. Let us remind that rather simpler shapes were previously studied [24, 29].

To sum up, studying the role of these factors in class imbalance is still an open research problem. Therefore, the main aim of this study is to experimentally examine which of these factors are more critical for the performance of the classifier. Carrying out such experiments requires preparing a new collection of artificial data sets which are affected by the above mentioned factors. Proposing such data sets is another sub-aim of this paper.

Then, assuming that performance of classifiers could be deteriorated by these data factors one could examine competence of pre-processing methods to deal with particular factors.

In this paper we are particularly interested in *focused* (also called *informed*) *re-sampling* methods, which modify the class distribution taking into account local characteristics of examples. Representative such methods are SMOTE for selective over-sampling of the minority class [9], one side sampling [35] and NCR for removing examples from the majority classes [37] or hybrid SPIDER method [54].

Therefore, an experimental comparison of chosen focused re-sampling methods and simpler random replication of the minority class methods applied to previously generated data sets and establishing competence of these methods for dealing with particular data factors are the next aims of this paper.

The following paper contains many new experimental results (in particular in studying the role of data factors). However, its also summarizes some results coming from co-operation with other colleagues, in particular concerning SPIDER methods and already published in [42, 55].

The paper is organized as follows. Section 2 describes main evaluation measures used for imbalanced data. The review of related research with data factors in imbalanced data is given in Section 3. Then, the generation of artificial data sets is presented in Section 4. The next section contains results of experiments study of the influence of data critical factors on the tree, rule based and k-NN classifiers. In Section 6 the most related focused pre-processing methods, including SPIDER, are briefly presented. Their comparative experimental evaluation is summarized in Section 7. The paper concludes with a discussion in Section 8.

11.2 Evaluation Measures for Learning Classifiers from Imbalanced Data

Imbalanced data constitutes a problem not only when inducing a classifier, but also when evaluating its performance. The overall classification accuracy is not the only and the best criterion characterizing performance of a classifier in case of class imbalance [62].

As the overall classification accuracy is biased towards the majority classes, in most of the studies on imbalanced data, measures defined for two-class classification are considered, where typically the class label of the minority class is called positive and the class label of the majority class is negative [18]. Even if data contains more majority classes the classifier performance on these classes could be aggregated into one negative class. Therefore, the performance of the classifiers is presented in a confusion matrix as in Table 1.

Table 1 Confusion matrix for performance evaluation

	Predicted Positive	**Predicted Negative**
True Positive	TP	FN
True Negative	FP	TN

From the confusion matrix, apart from other more elaborated measures (see e.g. reviews as [18]), one can construct simple metrics concerning recognition of the positive (minority) and negative (majority) classes:

$$TruePositiveRate = TP/(TP+FN)$$

$$TrueNegativeRate = TN/(TN+FP)$$

$$FalsePositiveRate = FP/(TN+FP)$$

$$Precision = TP/(TP+FP)$$

True Positive Rate is called *Sensitivity* (also Recall) while True Negative Rate is referred to *Specificity*. As the improvement of recognizing the minority class is associated with changes of recognizing other majority classes, aggregated measures are considered to characterize the performance of the classifiers. First of all, several authors use the *ROC (Receiver Operating Characteristics) curve* analysis [16]. A ROC curve is a graphical plot of a true positive rate (sensitivity) as a function of a false positive rate (1 − specificity) along different threshold values characterizing overall performance of a studied classifier. The quality of the classifier performance is reflected by the area under a ROC curve (so called AUC measure) [11, 62]. AUC varies between 0 and 1. Its larger values indicate better classifier performance. Although AUC is a very popular tool, some researchers have showed that it has some limitations as in the case of highly skewed data sets it could lead to an overoptimistic estimation of the algorithm's performance. Thus, other proposals include Precision Recall Curves [15] or other special cost curves (see their review in [18]).

One can also use simpler measures to characterize classifiers, in particular if they have a purely deterministic prediction (see discussions on applicability of ROC analysis in [61]). Kubat and Matwin [35] proposed to use the geometric mean of sensitivity an specificity defined as a:

$$G\text{-}Mean = \sqrt{Sensitivity \cdot Specificity},$$

This measure relates to a single point on the ROC curve and it 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 [24]. An alternative criterion aggregating precision and recall is F measure; for discussion of its properties see e.g. [18].

11.3 Earlier Studies with Data Factors in Class Imbalance

In this section we discuss earlier, related works on studying data properties which influence learning classifiers from imbalanced data sets.

First, one can notice that the majority of researchers proposing new approaches to handle class imbalance validated them in experiments conducted mainly on

real-life data sets (usually imbalanced data coming from the UCI Machine Learning Repository [2]). Moreover, other comparative studies of various basic classifiers or pre-processing methods are carried out in a similar way on such data sets, see e.g. such comprehensive comparative studies [3, 59]. However, working with artificial data sets, where it is possible to change their characteristics in the controlled way, is more valuable if someone attempts to study more precisely the impact of a chosen factor characterizing distributions of examples.

Several experimental studies have already showed that the performance of standard classifiers decreased in imbalanced data. However, some researchers hypothesized, that the *class imbalance ratio* (i.e. too low cardinality of the minority class referred to the total number of examples or to the majority class) is not necessarily the only, or main, problem causing this performance decrease and dealing only with it may be insufficient for improving classification results. In other words, besides this imbalanced ratio, data could be accompanied with other factors, which in turn cause the degradation of classification performance.

Japkowicz and her co-operators focused on *within-class imbalance*, i.e. target concepts (classes) were decomposed into sub-concepts [30]. To check the influence of increasing the level of decomposition she and her co-authors carried our many experiments with artificially generated data. Let us describe their construction just to have a reference point for our further solutions. Three parameters were controlled: the size of the training set, the imbalance ratio, and so called *degree of concept complexity* (understood as decomposition of the class into a number of subclasses). Two classes the minority vs. the majority class were considered only and each of data sets was generated in one-dimension interval. This input interval was divided into a number of subintervals of the same size (up to five), each associated with a different class label. The examples were uniformly distributed within subintervals. The degree of complexity corresponds to the number of alternating subintervals. For these assumption 27 data sets were generated with various combinations of the above mentioned three parameters. Following similar assumptions they also generated additional data sets in five-dimensional space, where an alternance of classes was modelled by separate clusters.

Then, C4.5 tree and multi layered perceptron (MLP) with back propagation algorithms were run over these data sets. Their results showed that imbalance ratio did not cause the degradation of classifiers' performance so much as increasing degree of complexity. The worst classification results were obtained for the highest decomposition of classes (e.g. into 5 parts) in particular existing with too small number of examples. Their main result is that "the true nature of the class imbalance problem (...) is only if the size of the small class is very small with respect to the concept complexity; i.e. it contains very small subclusters". On the other hand, this also means that in much larger data where subclusters could be represented by a reasonable number of examples, the imbalance ratio alone will not decrease so much the classification performance [30].

According to Japkowicz, if the such within-class imbalanced sub-concepts contain quite a small number of minority class examples it is associated with the problem of *small disjuncts* while building classifiers – which was originally introduced

by Holte in standard (balanced) learning of symbolic classifiers [20]. 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 [20, 62]. 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 [18].

As a practical consequence special approaches to handle the problem of small disjuncts of imbalanced concepts were proposed in [29, 30]. They are based on specialized over-sampling of minority class, i.e. a required number of examples are randomly replicated until balancing majority and minority class in the certain degree. By appropriate increasing the amount of the smaller class the classifiers learned from such modified data are less sensitive to original rare concepts. An example of using cluster analysis to identify the sub-concept and their random oversampling is described in Section 11.6. The impact of small disjuncts was also further studied by other researchers, see e.g. [28, 46]. In particular, additional experiments with applying other classifiers on the artificial data constructed in the above mention way showed that decision trees were the most sensitive to the small disjuncts, then the next was MLP and support vector machines were the less sensitive[1].

Recently some researchers have also focused on different factors characterizing data distributions. Prati et al studied the role of *overlapping* between minority and majority classes [47]. 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 Gaussian distribution). Two parameters were changed: the imbalance ratio, and the distance between centroids – so classes could be moved from clear separation to high overlapping. Using C4.5 classifier and AUC criterion they showed that increasing the overlapping ratio was more responsible for decreasing AUC results than decreasing cardinality of the minority class (for some data AUC decreased from 0.99 to 0.5). Another observation, consistent with intuition, was that for clearly separate and distant clusters the classification measures did not decrease even with high under-representation of the minority class.

Then, influence of increasing overlapping was more precisely examined in [24]. Garcia et al. generated two-dimensional data sets with two classes separated by a line orthogonal to one of the axis. Depending on the amount of overlapping examples of the majority class were uniformly generated inside the minority class part in a stepwise way moving from the decision boundary until covering completely minority class. Garcia et al. 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 more degraded their performance (with respect to TPR and TPN) 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. In this way they achieved balance between classes in this boundary, and then

[1] These results are also summarized in the report [8]

the minority class dominated the majority one. Again the results of experiments confirmed that increasing such a local imbalance ratio and the size of the overlapping area were more influential than changing the overall imbalance ratio. However, these factors influenced in a different way performance of particular classifier. For instance k nearest neighbor classifiers (K-NN) was the most sensitive to changes in the local imbalance region. Naive Bayesian, MLP and J4.8 were better working in the dense overlapping region. These conclusions have been later verified in additional experiments (more focusing on performance of K-NN and other evaluation measures), see [25].

Prati et al. have recently come back to studying the overlapping in class imbalance [4]. Comparing to their previous work [47] where they identified that performance degradation was not solely caused by class imbalances, but it was strongly related to the degree of class overlapping, in the new work, they decided to investigate the usefulness of five different re-sampling methods on the same difficult artificial data sets. The chosen methods were: popular random-over sampling, random under-sampling, Nearest Cleaning Rule (NCR) [37], SMOTE and SMOTE + ENN [9]. We will briefly describe them in further section 6. Now we can say that the main conclusion was that appropriate balancing training data usually led to a performance improvement of C4.5 classifiers for highly imbalanced data sets with highly overlapped classes. However, the improvements depends on the particular 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 showed that over-sampling methods in general, and Smote-based methods in particular, were more effective than under-sampling. Moreover, the Smote-based methods were able to achieve the best performance even for the most skewed distributions. Then, the data cleaning step used in the Smote + ENN seemed to be especially suitable in situations having a higher degree of overlapping. The quite good performance of SMOTE over-sampling integrated with data cleaning (as edited nearest rule ENN or Tomek Links [58]) was also confirmed in other experiments on many UCI real data sets [3].

Prati et al. have also noticed in their conclusions that ”it is also worthwhile to consider the generation of artificial data sets where the distribution of examples of the minority class is separated into several small clusters” [4].

Finally, quite a few researchers noticed that the other factor which could influence degradation of classifiers performance on imbalanced data could be *noisy examples* [1, 60]. Traditionally noise in supervised learning is understood as an (random) *error in labeling examples* (i.e. an example is assigned to wrong class) or erroneous values of attributes describing some examples [7]. These researchers wrote that existing methods for handling imbalanced data sets were studied under an assumption saying that the input data are noise-free or noise in the data sets is not significant. They claimed that real-world data are rarely perfect and can often suffer from corruptions that may impact decision of models created from these data. An investigation of both class and attribute noise in case of typical machine learning (i.e. balanced data) was conducted by many researchers which conclusion that the presence of noise can be harmful to a classifier, in particular when it is applied to previously unseen or testing

examples. In case of imbalanced data authors of [1, 60] proposed to identify noisy examples and remove them from the input data. However, their experiments were either conducted over UCI typical data sets or specific software data [60]. Moreover, they used special classification filters with identify so called *mislabeled examples* [7] by sophisticated ensembles. In these approaches the meaning of a mislabeled example has a broader sense as besides random errors it also includes outliers and other nontypical class representatives [22]. In this paper, a different view on such examples is presented.

11.4 Generation of New Artificial Data Sets

First, we will discuss assumptions for preparing new data sets in our experimental study.

Let us summarize that authors of related works indicated the role of following factors characterizing data distribution:

- decomposition of classes into sub-concepts,
- too low number of examples in sub-concepts (small disjuncts),
- high overlapping between classes.

One can notice that these authors studied the impact of single factors without considering other ones in the same experimental setup. Here, we want to consider all of them occurring together.

Besides the above mentioned factors we decide to consider two additional factors:

- different, more difficult shapes of the decision boundaries between classes
- and additional type of examples belonging to the minority class.

Explaining the second factor let us try to categorize types of minority class members. The examples located more deeply inside this class (even its sub-concepts) could be treated as *safe* ones. Such examples could be easier for correct recognition as they are surrounded by examples from the same class. Then, some other examples could be called *borderline* examples, if they are located inside the overlapping region. They are more unsafe or difficult to be learned as they occur closer to the decision boundary between classes with mixed distribution of majority class neighbors and possible noise could more influence changing the classification decision.

Moreover, it would be worthy to distinguish yet another type of so called *rare examples*. These examples are located in the majority class region and being distant enough to the decision boundary to be distinguished from borderline examples. In some of earlier experiments Stefanowski and his co-operators analysed local neighborhoods of minority class examples (see e.g. [40]). This analysis of local neighborhood was done with k-NN classifiers. An example was treated as safe, if its was correctly re-classified by its closest k neighbors. If the ratio neighbors from opposite classes was similar which could lead to mis-classification of the example it was treated as a borderline example. Then, if all its neighbours belong to opposite classes it was identified as an outlier. Moreover we noticed that some minority examples locally created pairs or sometimes triples also more distant from borderline

examples. Analysing in this way several UCI data sets we noticed that they could be difficult with respect to recognizing the minority class as they contained much less safe examples than others. Moreover, the number of outliers or rare pairs/triples was sometimes relatively high. For instance, in cleveland data the minority class contained 35 examples with 22 outliers or rare examples and 13 borderline ones. Similar situation ocured for a few data while some other data sets, e.g. haberman or ecoli, contained more bordeline examples than outlier ones (e.g. haberman has 51 and 20 respectively with 10 safe examples only). As the number of outliers or rare examples is quite high comparing to the size of the minority class we claim that they could be precious and cannot be just skipped while building classifiers. In our point of view they are not treated as simple noise but rather less typical rare cases of the scattered minority class. We will further call them *rare examples*[2].

To sum up, the following factors are chosen to be present in new artificial data sets considered in our experiments:

- decomposition of the minority class into sub-parts (subclusters, etc.),
- size of the overlapping region between classes (majority class examples are generated into the minority class inside the "borderline zone" of the given width and it is parametrized by the relative number of examples from the minority class that are located in this region),
- presence of rare examples (also parametrized by the relative number of examples from the minority class),
- linear vs. non-linear shape of decision boundaries,
- imbalance ratio (denoted as $i : j$, where i represents the minority class and j the majority one),
- total number of learning examples.

Similarly to other researchers we chose binary classification problems (the minority vs. the majority class) with examples randomly (either uniformly or not) distributed in the two-dimensional space (both attributes were real-valued). Following the literature on experiments with the factors influencing the performance of classifiers, we decided to prepare several artificial data sets in order to control these factors. We considered three different shapes of the minority class: *subclus*, *clover* and *paw*.

In *subclus*, examples from the minority class are located inside rectangles all surrounded uniformly by the majority class. The examples of the minority class are also uniformly distributed within its sub-region. This shape is a kind of two-dimensional generalization of data from related works on data decomposition and small disjuncts [29]. Fig. 1 shows such a shape with 3 subclusters; two zoom windows focus a reader attention on the exemplary borders.

Then, the next shape, called a *clover*, represents a more difficult, non-linear setting, where the minority class resembles a flower with elliptic petals. We decided to analyse two types of such clover shapes. In the first type the examples of majority class were also distributed inside the elliptic shapes fitted within the minority class

[2] In this sense it could be close to rare cases as discussed in the second section of G.Weiss study [62]. Although it is still class imbalance problem not a case of very rare data as sometimes considered in the context of one-class-learning

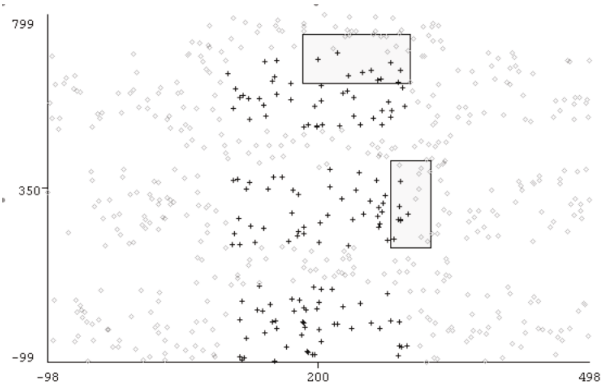

Fig. 1 Subclass data set – a minority class is decomposed into 3 parts (subclusters)

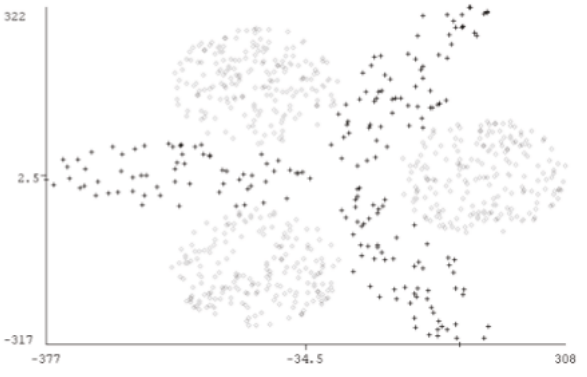

Fig. 2 Clover data set with 3 elements and sub-clusters of majority classes

”petals”. Figure 2 shows just such a *clover* with 3 petals. Then, we created another clover versions where examples from the majority class were uniformly distributed in all the free parts – see at Fig. 5 for *clover* with 5 petals (as this shape will be mainly used in further experiments – see Section 11.7 – examples are represented by points marked with different symbols).

Finally, in *paw* the minority class is decomposed into 3 elliptic sub-regions of varying cardinalities, where two subregions are located close to each other, and the remaining smaller sub-region is separated – two representatives of such shapes are showed in Fig. 3 and Fig. 4. In case of *02a* example the majority class are generated closer to the minority class regions and they are distributed more uniformly, while *02b* is more similar to *clovers* where the majority class is arranged in elliptical shapes. However, in both cases the majority class was also generated within some elliptical shapes. We also constructed another versions of *paw* where examples from the majority class are uniformly distributed in the allowed area - see an illustrative

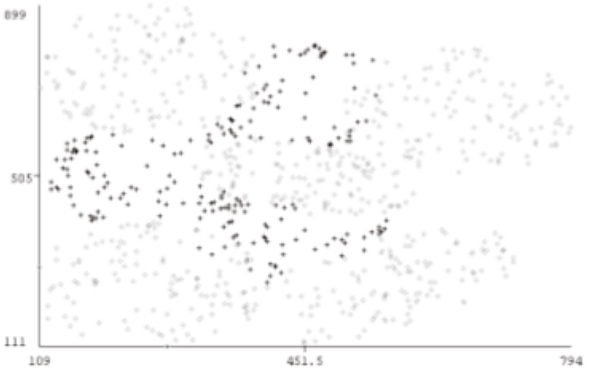

Fig. 3 Paw data set - version 02a

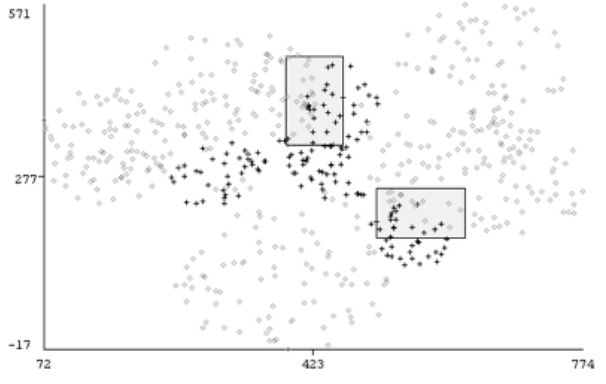

Fig. 4 Another paw data set - version 02b

example in Fig 6. We constructed such *paw* figure as it should better represents real-life data than the *clover*. Moreover, both *clover* and *paw* should be more difficult to learn than simple circles (or spheres) that were considered in some related works.

We generated a large collection of data sets with different numbers of examples (ranging from 200 to 1200) and imbalance ratios (from 1:3 to 1:9). Additionally, following Japkowicz's research on data complexity and splitting data shapes into interleaved sub-parts [29], we considered a series of the *subclus* and *clover* shapes with the number of sub-regions ranging from 1 to 5, and from 2 to 5 respectively.

Technically, this generator was implemented by Krzysztof Kaluzny in Java as a program compatible with WEKA platform; for more details see [32].

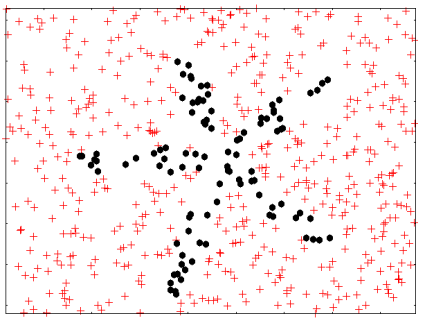

Fig. 5 Clover data set

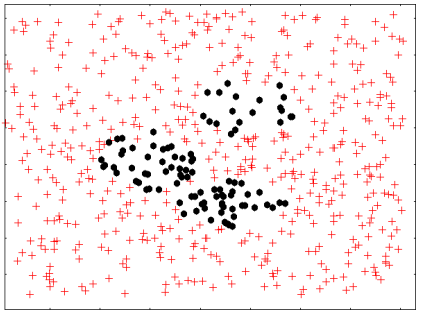

Fig. 6 Paw data set

11.5 Experimental Analysis of Influence of Critical Factors on Classifiers

In experiments three different classifiers were applied to the generated data sets (as presented in the previous section). *K*–nearest neighbod (K-NN), tree- and rule–based classifiers were chosen following the related experiments. K-NN was parametrized with $k = 3$ (we also tested 1 neighbour). While looking for these nearest neighbours, the distance is calculated with HVDM distance [64], i.e. aggregation of the Euclidean distance metric for numerical features and the Value Distance Metric [14] for the qualitative attributes. Decision trees were induced by Quinlan algorithm [44] - version J4.8 available in WEKA. Then, rules were generated by Ripper algorithm [13] (also JRip implementation from WEKA). Trees and rule were run without pruning to get more precise descriptions of the minority class. Evaluation measures were estimated by stratified 10 fold-cross validation. We focus mainly on the sensitivity (TPR) to study recognition of the minority class and AUC as a secondary criterion.

Due to the space limit, we are not able to present the complete results of all experiments but focus on the most interesting ones. For more details the reader is referred to the report describing much more experimental results [53].

In first experiments we studied the impact of the imbalance ratio combined with the size of data (i.e. total number of examples varied from 1200 to 200). Other critical factors were not considered, i.e. there was no overlapping or noisy examples. Let us only comment that our results were consistent with the observations reported in [30]. For the number of examples greater than 400 examples there was no significant influence of decreasing the imbalance ratio. Small decreases of sensitivity and partly AUC measures was observed only for very small cardinality of the data (smaller than 200) and a very high imbalanced ratio (1:11). It concerned all studied classifiers.

We also noticed that non-linear shapes of decision boundaries (as visible in *clover* or 02a, 02b data) were more difficult to recognize than linear rectangles *subclass* – see also Table 3.

Table 2 Influence of decomposing the minority class into sub-concepts on the sensitivity measure for K-NN classifier: Subclass data set. Two imbalance ratios and five different cardinalities of data sets are reported in columns.

Number of subclusters	1:5 600	1:5 400	1:5 200	1:9 600	1:9 400	1:9 200
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 3 Influence of non-linear decision boundaries on AUC measure. Size of data – 1200 examples and 3 different imbalance ratios.

Type of data	Trees 1:1	Trees 1:3	Trees 1:5	Rules 1:1	Rules 1:3	Rules 1:5	KNN 1:1	KNN 1:3	KNN 1:5
02a	0.95	0.85	0.68	0.94	0.91	0.89	0.94	0.92	0.9
02b	0.95	0.92	0.89	0.95	0.92	0.85	0.95	0.93	0.9
subclass3	0.98	0.97	0.96	0.96	0.94	0.92	0.97	0.96	0.94
subclass5	0.98	0.96	0.94	0.96	0.92	0.90	0.96	0.96	0.94
clover3	0.94	0.93	0.92	0.94	0.92	0.90	0.96	0.94	0.92
clover5	0.93	0.92	0.89	0.91	0.88	0.84	0.94	0.92	0.90

In the next phase of experiments we studied more precisely the impact of increasing the decomposition of the minority class into sub–parts. Generally speaking, the obtained results were also consistent with earlier related research, in particular works of Japkowicz – increasing the number of sub-regions of the minority class combined with decreasing the size of a data set degraded the performance of a classifier [28, 29, 30]. For illustration see Table 2, where for two imbalance ratios (1:5 and 1:9) and stepwise decrease of examples (from 600 to 200) we divide the rectangle of the minority class into appropriate subclusters (of the same size). One can notice that for smaller number of examples increasing the number of subclusters degraded the values of sensitivity much larger than changing the imbalance ratio. If the number of subclusters is higher than 4 and the number of examples is no greater than 400 examples, they could be seen as small sub-regions (e.g. for 1:5 the total number of 66 minority examples are divided into rare areas having less than 15 examples comparing to 333 majority examples – which could refer to the idea of small disjuncts [30]). Decreasing the total number of examples to 200 makes the problem definitely more difficult. The tree and rule classifiers also decreased their performance for these *subclass* shapes.

Table 4 Influence of decomposing the minority class on the sensitivity of tree classifier: Clover data set. Data size – 600 and 400 examples

Number of elements	600				400			
	1:3	1:5	1:7	1:9	1:3	1:5	1:7	1:9
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

Table 5 Influence of overlapping on the sensitivity of the tree classifier learned from subclass data. Overlapping is expressed by % of borderline examples in the minority class. Total number of examples – 800.

Number of subclusters	1:5			1:9		
	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

The degradation of performance was larger if the decision boundary became non-linear even for larger data set. Table 3 illustrates results for all classifiers applied to data sets characterized by different imbalanced ratio and smaller or higher decomposition. As the number of examples is rather highest (1200 examples), for nearly balanced data we did not notice the decrease. Non-linear and more complicated shapes (e.g. 02a, or increasing a number of parts in *clover5*) made the problem more difficult, in particular for tree or rule- classifiers, when data became more imbalanced (ratio 1:5). K-NN classifier is rather more local approach than global classifiers (trees or rules) and it works better with more complicated, non-linear classes. Values of AUC decrease in a more visible way if the number of examples is smaller than 600 ones.

Knowing that non-linear shapes were more difficult, we studied more precisely the impact of decomposition the minority class in presence of smaller number of examples. As it is more visible for the sensitivity measure we show a representative results for tree classifier, see Table 4. One can notice that stepwise increasing the number of sub-regions (from 2 to 6) in *clover* shape degrades much more the sensitivity measure than stepwise increasing the class imbalance ratio (from 1:3 to 1:9). Rule and K-NN classifiers showed similar behaviour - although in case of K-NN the degradation was not so radical.

Table 6 Influence of overlapping and rare examples of the minority class on the sensitivity of tree classifier: Subclass data set.

Number of	800				600				400			
subclusters	0%	10%	20%	30%	0%	10%	20%	30%	0%	10%	20%	30%
3	0.96	0.84	0.70	0.56	0.94	0.85	0.70	0.55	0.9	0.82	0.7	0.42
4	0.94	0.84	0.68	0.4	0.92	0.82	0.58	0.3	0.89	0.7	0.4	0.34
5	0.9	0.82	0.56	0.36	0.9	0.78	0.52	0.32	0.87	0.68	0.24	0.18
6	0.88	0.64	0.40	0.34	0.85	0.6	0.36	0.3	0.5	0.22	0.14	0.08

Table 7 Influence of rare and borderline examples: 02a data set.

Classifier	Sensitivity				AUC			
	0%	10%	20%	30%	0%	10%	20%	30%
Tree	0.45	0.38	0.17	0.04	0.82	0.8	0.64	0.5
Rules	0.82	0.70	0.65	0.58	0.92	0.85	0.82	0.78
KNN	0.84	0.72	0.70	0.62	0.95	0.92	0.9	0.87

The next phase of experiments concerned the influence of overlapping in the boundary between classes and the presence of rare minority examples located inside the majority class area. Starting from the overlapping factor, we established the width of the overlapping inside the area of the minority class and parametrized it by percentage of examples from the minority class which were located in this overlapping boundary. The majority examples are uniformly generated inside this boundary with their number was equal to the number of the minority class located there.

Table 5 shows influence of such overlapping on the tree classifier. Although comparing number of sub-regions to amount of overlapping may be not justified we can "roughly" say that stepwise increasing of overlapping could decrease more the sensitivity than stepwise increasing decomposition. For instance, let us analyse the first column (%) - the sensitivity changes from 0.96 to 0.94. While for any of the number of subclusters the sensitivity decreases in range of nearly 0.2 (see, e.g. 4 subclusters, the sensitivity decreases from 0.96 to 0.78). The similar tendency can be observed for rule and K-NN classifiers, also for smaller data sets and non-linear shapes (however, decreases of the sensitivity are even higher).

The next factor was the presence of rare examples from the minority class. We studied their impact together with overlapping of classes in these data sets. More precisely, if the parameter is set to $x\%$ it means that half of these examples are generated inside the overlapping and the rest are generated as rare examples. For instance, value 20% means that it is previous case of 10% overlapping extended by 10% minority examples treated as rare one. The appropriate new experiment referring to previous Table 5 is now presented in the next Table 6.

Table 8 Influence of rare and borderline examples: clover4 data set.

Classifier	Sensitivity				AUC			
	0%	10%	20%	30%	0%	10%	20%	30%
Tree	0.5	0.25	0.10	0.08	0.72	0.62	0.55	0.52
Rules	0.68	0.44	0.38	0.35	0.8	0.72	0.68	0.62
KNN	0.90	0.88	0.72	0.62	0.95	0.92	0.82	0.78

Although it could be questionable to compare directly the step of changing rare level with the step of increasing the class decomposition, we can say that stepwise increasing the level of rare and borderline examples decreases much more the evaluation measures than dividing a class into smaller parts (e.g. for 600 examples and no rare moving from 3 to 6 subclusters decreased the sensitivity around 0.1, while adding up to 30% noise or borderline examples introduced a change of the sensitivity over 0.5). Moreover, as it could be expected, adding rare made the problem more difficult (compare results from Table 5 to Table 6).

The similar results were obtained for other shapes and classifiers, see summaries presented in Tables 7 and 8.

Finally we checked all these parameters in case of the other versions of non-linear shapers (as illustrated in Fig. 6 and Fig. 5) where the examples from the majority class are surrounding more closely the minority class. These versions were more difficult to learn and values of evaluation measures were smaller than in the above presented tables.

To sum up, results of the experiments shows that besides decomposition of the minority class, the next important critical factors are:

- overlapping between classes (expressed by the number of borderline examples)
- rare examples from the minority class (in particular if they occur together with similar number of borderline examples).

We can also hypothesize that these factors could cause higher degradation of classification performance than decomposition itself - this is a new result comparing to previous related works. Moreover, presence of all these factors together in the data set causes larger classification deterioration than too low imbalance ratio – in particular for non-linear decision boundaries.

11.6 Improving Classifiers by Focused Re-sampling Methods

In the previous section we experimentally showed critical factors for degrading the performance of the selected tree, rule and K-NN classifiers. Pre-procesing methods that change distribution of examples in classes are one of the main types of specialize methods for improving classifiers in case of class imbalance [11, 18]. These methods, sometimes also called *re-sampling techniques*, are classifier-independent and consist in transforming an original data distribution to change the balance

between classes. Some them can also handle other properties of data distributions [3, 18].

Therefore, the next research problem of the following paper is to check the sensitivity of different re-sampling methods to overlapping, rare examples and class decomposition factors in the considered artificial data sets. We have chosen 4 methods some related to the proposal of the SPIDER method, introduced by Stefanowski and Wilk [54, 55]. More precisely they are *simple random over-sampling, cluster based over-sampling* and *nearest cleaning rule* NCR and SPIDER.

In the following review we briefly describe only these methods and SMOTE, which is also related to our proposal of SPIDER; for more extensive reviews see, e.g., [3, 11, 18, 62].

First of all, the simplest re-sampling techniques are random *over-sampling* which replicates examples from the minority class and random *under-sampling* which randomly eliminates examples from the majority classes until a required degree of balance between classes is reached (Many researchers attempted to obtain the same cardinality of the minority class as the majority one). However, several authors showed the random under-sampling or over-sampling 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 [9, 35]. Furthermore it is not easy to find an optimal ratio for balancing classes. Several authors have already shown that used "even" distribution (i.e. obtaining the same cardinality in classes) is not optimal when dealing with such rare classes. For instance, the reader can consult the comprehensive study with many data sets and classifiers showing that depending on combination of data and classifiers the ratios of modified majority vs. minority class cardinalities like 3:1 and 2:1 quite often outperformed the most popular ratio 1:1 [33]

Therefore, researchers proposed more elaborated methods that attempt at taking into account data characteristics and factors influencing nature of class imbalance.

Following critical observations on the role of small disjuncts Japkowicz proposed an advanced oversampling method (*cluster oversampling*) that takes into account not only *between-class imbalance* but also *within-class imbalance*, where classes are additionally decomposed into smaller sub-clusters [30]. First, random oversampling is applied to individual clusters of the majority classes so that all the subclusters are of the same size. Then, minority class clusters are processed in the same way until class distribution becomes balanced. This approach was successfully verified in experiments with decomposed classes [30, 43].

11.6.1 Informed Undersampling

Kubat and Matwin in their paper on *one-side selection* claim that characteristics of mutual positions of learning examples is a source of difficulty for learning classifiers from imbalanced data [35]; see also their more application study [36]. They focus attention on *noisy* majority class examples located inside the minority class and *borderline* examples. According to their approach, such examples are removed from

the majority classes, while the minority class is kept unchanged (these examples can be identified with so called Tomek links [58]). As result of such "focused" under-sampling ambiguous regions around the minority class are "cleaned". Moreover, some examples from "safer" regions of the minority class can be also discarded as they could be correctly classified by other learning examples.

Then, the *Nearest Cleaning Rule* (NCR) method was introduced in [37]. This method is based on the Edited Nearest Neighbor Rule (ENNR) and it removes these examples from the majority classes that are misclassified by its k nearest neighbors. Experimental results confirmed that both methods improved the sensitivity of the minority class comparing to simpler over- or under-sampling methods [35, 37].

11.6.2 Informed Oversampling Methods

A most well-known representative of informed over-sampling is SMOTE (Synthetic Minority Over-sampling Technique) introduced by Chawla et al. [9], which considers each example from the minority class and generates new synthetic examples along the lines between this example and some of its randomly selected k nearest neighbours (also belonging to the minority class). Experiments reported in [9] with C4.5 trees, Ripper rules and Naive Bayes classifiers showed that SMOTE improved recognition of the minority class. Moreover, its combination with under-sampling of the majority class was able to achieve better results than other under-sampling methods as ENNR alone – see, e.g., [3]. There are also other proposals of hybridization of the basic SMOTE with additional "filtering" step, see e.g. the use of some rough sets inspired solutions [48] or more sophisticated ensemble noise filtering [49].

Although SMOTE and NCR showed to be promising in experimental evaluation, they also demonstrated several shortcomings that became motivations for introducing SPIDER – we will further discuss them in the next section. We should note that recently some researchers have also tried to propose various generalizations of SMOTE following similar critical observations – see discussion in [18]. Two most interesting generalizations of SMOTE are Borderline SMOTE that takes into account the different nature of examples from the minority class, and Safe-Level SMOTE, where also the distribution of the majority class is considered while generating synthetic examples from the minority class. Both methods are described in [18, 40]. Yet another proposal is based on controlling distributions in local neighborhoods of the seed example and its nearest neighbors from both minority and majority classes [40].

11.6.3 SPIDER Method

The critical analysis of undesirable properties of well-known focused re-sampling methods, especially NCR and SMOTE, became a starting point for developing by Stefanowski and Wilk the SPIDER method [54]. NCR and in particular one-side-selection are too strongly biased toward cleaning overlapping regions between classes and interior areas of the majority class. However, both methods may *remove*

too many examples from the majority classes. Such greedy "cleaning" definitely leads to the increased sensitivity for the minority class but too extensive changes in the majority classes may deteriorate the ability of an induced classifier to recognize examples from these classes.

One of the main shortcoming of SMOTE is the *overgeneralization* problem. SMOTE blindly generalizes regions of the minority class without checking positions of the nearest examples from the majority classes. This strategy is particularly problematic in the case of skewed class distribution where the minority class is very sparse in comparison to the majority classes. In such a situation SMOTE may increase overlapping between classes by locating synthetic examples from the minority class among existing examples from the majority classes. Moreover, the number of synthetic examples generated by SMOTE has to be *globally parametrized*, thus reducing the flexibility of the approach. Results of experimental studies with simulated data sets [24] imply that an efficient method should be rather focused on local distributions of difficult examples than being controlled by a global parameter. Let us also notice that according to experimental results reported in the literature an appropriate value of this parameter strongly influences SMOTE's performance and its proper tuning requires a computationally costly procedure (iterative testing of various possible values). Finally, random introduction of synthetic examples by SMOTE may be difficult to justify in some domains where it is important to preserve a link between original data and a constructed classifier in order to explain suggested decisions.

The SPIDER method relies on the local characteristics of examples (i.e., characteristics of their local neighborhood) and distinguishing between different types of examples. Two types of examples are distinguished – *safe* and *not-safe*. Safe examples should be correctly classified by a constructed classifier, while not-safe ones are likely to be misclassified and require special processing. In SPIDER the type of an example is discovered by applying the *nearest neighbor rule* (NNR) with the heterogeneous value distance metric (HVDM) [64] – i.e., the distance is calculated with the Euclidean distance metric for numerical features and with the value distance metric [14] for qualitative features. According to NNR an example is *safe* if it is correctly classified by its k nearest neighbors, otherwise it is *not-safe*.

More precise categorization of an example is based on the analysis of its neighbors from the other classes (i.e., different than the class of a considered example). If an example is not-safe and its nearest examples belong to other classes, then this example is identified as *certain-not-safe* and we interpret it as a rare case (or an outlier) located deeply inside the other classes. Such example should be treated in a different way than a not-safe example with some neighbours from the same class – this one is rather located in / or closer to an overlapping region between classes. Unlike related methods that distinguish the type of examples in the minority class only, SPIDER identifies the nature of examples in all classes. SPIDER assumes two decision classes – the minority class c_{min} and the majority class c_{maj} – if an original data set contains several majority classes they are collapsed together.

The method consists of two main phases – *identification* and *pre-processing*. In the first phase the type of examples is identified according to the "local"

characteristics of their neighbors and 'flagged' accordingly. Firstly, examples from the majority class c_{maj} are processed. In particular, depending on the *relabel* option the method either removes or relabels certain-not-safe / noisy examples[3] from c_{maj} (i.e., changes their classification to c_{min}). Then, in the second phase, it identifies the characteristic of examples from c_{min} considering changes introduced in the first phase. Not-safe examples from this class require special processing – i.e they are amplified (by replicating them with different degree) according to the *ampl* option.

All options of SPIDER involve modification of the minority and majority classes, however, the degree and scope of changes varies between options. *Weak* amplification is the simplest and less greedy modification of the minority class. It focuses on not-safe examples from c_{min} and slightly over-samples them by adding as many of their copies as there are safe examples from c_{maj} in their neighborhoods. The second option – *relabel* – also changes these certain noisy examples from c_{maj} which could be interpreted as noisy outliers located more deeply inside the minority class. For the last option – *strong* – the degree of amplification of not-safe examples c_{min} could be higher depending on analysis of an extended neighborhood. Much more thorough description of the method is provided in [54, 55], including precise pseudocode of the algorithm.

11.7 Experiments with Focused Re-sampling Methods

In the next experiments we will study the impact of overlapping, rare examples and partly class decomposition on the performance of selected pre-processing methods for handling class imbalance, including our proposal of SPIDER. Here, we summarize some of the main results of the more comprehensive study on this topic recently carried out by Napierala, Stefanowski and Wilk [42].

According to the results of Stefanowski's earlier experiments (presented in Section 11.5) a group of data sets with 800 examples, the imbalance ratio of 1:7, and 5 sub-regions for the *subclus* and *clover* shapes is selected for experiments. We chose their more difficult versions where the majority class is uniformly distributed around the minority class shapes – see the Figures 5 and 6. Let us remind that all these data sets presented a significant challenge for a stand-alone classifier. Similar behaviour was observed for data sets with 600 examples, but due to space limit we did not describe these data sets in the paper. Due to specific aims of study [42], two symbolic rule and tree classifiers were applied. Trees were induced by C4.5 algorithm, while rules were induced by MODLEM algorithm (introduced by Stefanowski in [50]; see also for its description in [51, 52].

Firstly, the impact of disturbing the borders of sub-regions in the minority class was evaluated. It was simulated by increasing the ratio of borderline examples from the minority class subregions. This ratio (further called the *disturbance ratio*) was

[3] In case of the majority class we can consider possible noise example if a not-safe example is located deeply inside the region of the minority class (all its k-nearest neighbors belong to the opposite class) and it could be wrongly re-classified by its neighbors

changed from 0 to 70%. The width of the borderline overlapping areas was comparable to the width of the sub-regions (sub-parts in data shapes).

The constructed classifiers were combined with the following focused preprocessing methods:

- standard random oversampling (abbreviated as RO),
- Japkowicz's cluster oversampling (CO),
- nearest cleaning rule NCR,
- and SPIDER (SPID).

Cluster oversampling was limited to the minority class, and SPIDER method was used with the strong amplification as such combination performed best in our earlier studies [54, 55]. For baseline results (Base), both classifiers were ran without any pre-processing.

Table 9 G-mean for artificial data sets with varying degree of the disturbance ratio in the overlapping region.

Data set	Rules					Trees				
	Base	RO	CO	NCR	SPID	Base	RO	CO	NCR	SPID
subclus-0	0.9373	0.9376	0.9481	0.9252	0.9294	0.9738	0.9715	0.9715	0.9613	0.9716
subclus-30	0.7327	0.7241	0.7242	0.7016	0.7152	0.6524	0.7933	0.7847	0.7845	0.8144
subclus-50	0.5598	0.5648	0.6020	0.6664	0.6204	0.3518	0.7198	0.7113	0.7534	0.7747
subclus-70	0.4076	0.4424	0.4691	0.5957	0.5784	0.0000	0.7083	0.7374	0.6720	0.7838
clover-0	0.7392	0.7416	0.7607	0.7780	0.7908	0.6381	0.8697	0.8872	0.6367	0.6750
clover-30	0.6361	0.6366	0.6512	0.7221	0.6765	0.2566	0.7875	0.7652	0.6758	0.7686
clover-50	0.5066	0.5540	0.5491	0.6956	0.6013	0.1102	0.7453	0.7570	0.6184	0.7772
clover-70	0.4178	0.4658	0.4898	0.6583	0.5668	0.0211	0.7140	0.7027	0.6244	0.7665
paw-0	0.9041	0.9126	0.9182	0.9184	0.8918	0.6744	0.9318	0.9326	0.6599	0.7330
paw-30	0.7634	0.7762	0.7701	0.7852	0.7780	0.3286	0.8374	0.8334	0.8527	0.8337
paw-50	0.6587	0.6863	0.6865	0.7517	0.7120	0.3162	0.8013	0.7858	0.8200	0.8075
paw-70	0.5084	0.5818	0.5691	0.7182	0.6506	0.0152	0.7618	0.7472	0.7824	0.8204

We do not report all results from [42] but summarize the most representative ones. However, let us remark that with respect to recognition of the minority class alone, expressed by the sensitivity measure, results clearly showed that all methods of pre-processing improved the sensitivity of both classifiers in comparison to Base classifiers (in particular for more difficult decision boundaries and larger disturbance). Generally speaking, simpler over-sampling RO and CO performed comparably on all data sets, and on non-disturbed data sets they often over-performed focused methods NCR and SPIDER. On more difficult sets (disturbance = 50–70%) both methods NCR and SPIDER were significantly better than oversampling methods. Then, we took into account balance between sensitivity and specificity, so recognizing examples also from the other majority class. Results of the geometric mean (G-mean) are presented in Table 9.

These experiments also showed that the degradation in performance of a classifier is strongly affected by the number of borderline examples. If the overlapping area is large enough (in comparison to the area of the minority sub-clusters), and at least

30% of examples from the minority class are located in this area, then focused resampling methods (NCR, SPIDER) strongly outperform random and cluster oversampling with respect to sensitivity and G-mean. Moreover, the performance gain increases with the number of borderline examples. On the contrary, if the number of borderline examples is small, then oversampling methods sufficiently improve the recognition of the minority class and they are comparable to focused method with respect to G-mean.

Table 10 Sensitivity for artificial data sets with different types of testing examples

Data set	Rules					Trees				
	Base	RO	CO	NCR	SPID	Base	RO	CO	NCR	SPID
subcl-safe	0.58	0.58	0.62	0.78	0.64	0.32	0.84	0.86	0.98	1.00
subcl-B	0.84	0.84	0.84	0.86	0.84	0.00	0.82	0.84	0.36	0.92
subcl-C	0.12	0.1 0	0.16	0.24	0.26	0. 00	0.54	0.00	0.00	0.52
subcl-BC	0.48	0.47	0.50	0.55	0.55	0.00	0.68	0.42	0.18	0.72
clover-safe	0.30	0.38	0.44	0.70	0.60	0.02	0.96	0.92	0.04	0.98
clover-B	0.84	0.82	0.82	0.84	0.86	0.04	0.94	0.92	0.04	0.94
clover-C	0.14	0.08	0.14	0.24	0.36	0.00	0.30	0.02	0.00	0.40
clover-BC	0.49	0.45	0.48	0.54	0.61	0.02	0.62	0.47	0.02	0.67
paw-safe	0.84	0.92	0.84	0.84	0.80	0.42	0.90	0.96	0.74	1.00
paw-B	0.88	0.88	0.86	0.88	0.90	0.14	0.90	0.90	0.40	0.92
paw-C	0.16	0.14	0.12	0.26	0.16	0.04	0.20	0.00	0.00	0.34
paw-BC	0.52	0.51	0.49	0.57	0.53	0.09	0.55	0.45	0.00	0.63

In [42] we carried out additional experiments where we studied the impact of rare examples from the minority class, located outside the borderline area, on the performance of a classifier. To achieve this, we introduced new rare examples (single and pairs) and denoted them with C. Similarly to the first series of experiments we used data sets of three shapes (*subclus*, *clover* and *paw*), 800 examples and the imbalance ratio of 1:7. We also employed rule- and tree-based classifiers combined with the same pre-processing methods. However, we changed the 10-fold cross validation to the train-test verification in order to ensure that learning and testing sets had similar distributions of the C examples. In each training set 30% of the minority class examples were safe examples located inside sub-regions, 50% were located in the borderline/overlapping area (we denote them with B), and the remaining 20% constituted the C rare examples.

For each training set we prepared 4 testing sets containing the following types of examples from the minority class: only safe examples, only B examples, only C examples, and B and C examples combined together (BC). Results are presented in Table 10. They clearly show that for the "difficult" rarity (C or BC) SPIDER and in most cases NCR were superior to RO, CO and Base. SPIDER was also comparable to RO and CO in case of safe and (sometimes) B examples.

To sum up these experiments reveal the superiority of SPIDER and in most cases NCR in handling rare examples located inside the majority class (also accompanied with borderline ones). Such result has been in a way expected, as both methods were introduced to handle such situations. The experiments also demonstrated that even random oversampling is comparable to SPIDER and better than NCR in classifying safe examples from the minority class.

Besides the above mentioned experiments with artificial data, the focused re-sampling methods, including SPIDER, were also compared on some real data sets coming from UCI Machine Learning Repository [2]. As this kind of experiments is not within the main aim of this paper, and its page size is limited, we do not show these precise results but attempt at summarizing the general conclusions from two earlier papers ([54] - early results with rule based classifiers, and more extended comparison [55] including additional methods and classifiers). In these experiments SPIDER applied together with C4.5 and MODLEM algorithms was compared to competitive methods (NCR and SMOTE) and basic classifiers used without any pre-processing. The results of experiments showed that although NCR often led to the highest increase of sensitivity, at the same time it significantly deteriorated specificity and overall accuracy. SPIDER was the second best with respect to improving the sensitivity of the minority class (improvement was more visible for rule than trees), and slightly better than SMOTE or comparable to it. Moreover, it did not deteriorate the recognition of the majority classes as much as NCR. In case of SPIDER and possible pre-processing options, *weak* resulted in the best specificity and overall accuracy (often at the cost of sensitivity), *strong* resulted in a good balance between specificity and sensitivity – evaluated by G-mean – and *relabel* improved sensitivity in the similar range as *strong*, however at the cost of specificity.

Recently we showed another way of balancing recognition of the minority and majority classes (expressed by optimizing G-mean) which include using SPIDER inside the generalized framework of the adaptive ensembles called IIvotes [6].

Considering results of the experiments with UCI imbalanced data, we could also refer to the analysis of the "nature" of these data sets taking into account local neighbourhood characteristic for the focused re-sampling methods. In [40] we used $k-NN$ analysis of each minority class example and considered is as certain unsafe (outlier) or borderline (unsafe-possible) as defined in SPIDER. Results (for $k=3$ or partly 5) showed that all studied data sets are rather difficult with respect to classifier ability for recognizing the minority class. First of all, we noticed that for some data sets the number of outlier examples is quite high comparing to the size of the minority class. Other data sets contained also many borderline examples without too many safe regions of the minority class. Referring to the earlier comparison of oversampling methods we noticed that for such data sets SPIDER and NCR led to improvements of the sensitivity. On the other hand, for a two data sets (as e.g. new thyroid) with more safer examples, base classifiers without preprocessing or simpler oversampling worked sufficiently good. Let us remind that such a data as new thyroid is more imbalanced than other data sets. In our opinion this very simple analysis confirm our earlier observations on the role of critical factors from experiments with controlled artificial data sets.

11.8 Final Remarks

The problem of learning classifiers from imbalanced data has been considered. It is one of the most challenging topics in machine learning and data mining [65]. Moreover, it is still "open" from a theoretical point of view and it is very important in many application domains. On the other hand, one can notice by studying related literature that it has received growing research interest in the last decade and several specialized methods have already been proposed. Although some of them have been validated in experiments, it is still a need to ask more general research questions about the class imbalance, data characteristics and competence of some popular methods. This paper is an attempt to partly answer to these questions.

Firstly, the nature of this problem and sources of difficulties for achieving good recognition of the minority class are discussed. As it has been already noticed by other researchers, the small number of examples in the minority class is not the main source of difficulty [24, 28, 29, 47]. The degradation of classification performance is rather related to other critical factors as decomposition of the classes into smaller sub-parts including too few examples (so called small disjuncts [30]), overlapping between classes (existence of too many borderline examples in the minority class), presence of rare or noisy examples located farther from the decision boundary (deeper inside the distribution of the opposite class).

Following this literature review, in the first part of this paper we decided to carry out an experimental study on the impact of the critical factors on re-sampling methods dealing with imbalanced data. Unlike the related works we decided to consider them occurring together in the data. Moreover, we pay more attention to presence of borderline and rare examples. We also considered more complicated shapes of classes than in earlier works. This is why we introduced new types of artificial data sets for experimental evaluation. Generated artificial data include simpler rectangle shapes of the minority class (*subclass* following inspirations from earlier works) and more complicated non-linear boundaries as *clover*, *paw* and similar *02* shapes.

In the first phase of experiments we studied impact of critical factors in these artificial data sets on performance of the most popular rule-, tree- and K-NN classifiers. First of all, some results confirmed earlier results obtained for simpler artificial data on the importance of the minority class decomposition into smaller sub-concepts [30]. We also showed that for more nonlinear decision boundaries increasing decomposition of the class into small sub-parts decreased sensitivity or AUC measures. It was also observed that K-NN could classify the non-linear shapes better than trees - it could be explained by its local performance comparing to a more global way of constructing trees (see also more extensive discussion of specific properties of K-NN in [46]). On the other hand, both tree and rule classifiers worked better for rectangle shapes of the minority class. Moreover, rule classifier Ripper was slightly better than C4.5 tree. Its behaviour could be caused by a specific way of constructing a decision list in the final classifier, i.e. the algorithm induced rules only for the minority class (with a controlled pruning level) and they are ordered in a kind of an exception list [13]. If the new / testing example does not match any of these rules it is classified by a default rule to the majority class. As we ran Ripper without strong

pruning, it could be better suited to class imbalance in non-linear, complicated, decision boundaries (clover, paw, 02) than more global tree classifiers.

Then, experimental results clearly showed that the combination of class decomposition with overlapping makes learning very difficult (in particular for sensitivity measure and tree classifiers). Focussing attention on the rare examples from the minority class is an original contribution of this study – as according to the best knowledge they have not been studied yet in experimental studies. It was clearly visible that the presence of rare example significantly degraded performance of all classifiers. We could also say that stepwise increasing numbers of borderline and rare examples in the minority class decreased all evaluation measures more that increasing decomposition of this class into new sub-parts. This is also a new observation comparing to previous related research.

To sum up this part of experiments, we hope that our results expand the body of knowledge on the critical role of borderline and rare examples with respect to earlier results based on simpler artificial data sets and other factors [24, 28, 29, 47].

The next part of this paper concerns problems of handling these difficulties by the following re-sampling methods: random over-sampling, cluster over-sampling, informed under-sampling by NCR and SPIDER. Our experiments showed that the degradation in classification performance was strongly affected by the number of borderline examples. If the overlapping area was large enough (in comparison to the area of the minority sub-clusters), and at least 30% of examples from the minority class were located in this area (i.e., they are borderline examples), then focused resampling methods (as SPIDER and partly NCR) strongly outperformed random and cluster oversampling with respect to sensitivity and G-mean measures. Moreover, it seams that the performance gain increased with the number of borderline examples. The other experiments revealed the superiority of SPIDER and in some cases NCR in handling rare examples located inside the majority class (also accompanied with borderline ones).

We hope that the above mentioned comparative studies with artificial simulated data also extended by experiments on UCI data sets could give more insight into conditions of the usefulness of particular re-sampling methods to improve classifiers learned from imbalanced data.

Acknowledgements. The research has been partly supported by the Polish Ministry of Science and Higher Education, grant no. N N519 441939.

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 International Federation for Information Processing Comf. AIAI 2007, pp. 21–28 (2007)
2. Asuncion, A., Newman, D.J.: UCI Machine Learning Repository. School of Information and Computer Science. University of California, Irvine, CA (2007), `http://www.ics.uci.edu/~mlearn/MLRepository.html`

3. 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)
4. Batista, G.E.A.P.A., Prati, R.C., Monard, M.C.: Balancing Strategies and Class Overlapping. In: Famili, A.F., Kok, J.N., Peña, J.M., Siebes, A., Feelders, A. (eds.) IDA 2005. LNCS, vol. 3646, pp. 24–35. Springer, Heidelberg (2005)
5. Bay, S., Kumaraswamy, K., Anderle, M.G., Kumar, R., Steier, D.M.: Large scale detection of irregularities in accounting data. In: Proc. of the ICDM Conf., pp. 75–86 (2006)
6. Błaszczyński, J., Deckert, M., Stefanowski, J., Wilk, S.: Integrating Selective Preprocessing of Imbalanced Data with Ivotes Ensemble. In: Szczuka, M., Kryszkiewicz, M., Ramanna, S., Jensen, R., Hu, Q. (eds.) RSCTC 2010. LNCS, vol. 6086, pp. 148–157. Springer, Heidelberg (2010)
7. Brodley, C.E., Friedl, M.A.: Identifying Mislabeled Training Data. Journal of Artificial Intelligence Research 11, 131–167 (1999)
8. Casagrande, N.: The class imbalance problem: A systematic study. Research Report IFT 6390. Montreal University
9. Chawla, N., Bowyer, K., Hall, L., Kegelmeyer, W.: SMOTE: Synthetic Minority Oversampling Technique. J. of Artifical Intelligence Research 16, 341–378 (2002)
10. Chawla, N.V., Lazarevic, A., Hall, L.O., Bowyer, K.W.: SMOTEBoost: Improving Prediction of the Minority Class in Boosting. In: Lavrač, N., Gamberger, D., Todorovski, L., Blockeel, H. (eds.) PKDD 2003. LNCS (LNAI), vol. 2838, pp. 107–119. Springer, Heidelberg (2003)
11. Chawla, N.: Data mining for imbalanced datasets: An overview. In: Maimon, O., Rokach, L. (eds.) The Data Mining and Knowledge Discovery Handbook, pp. 853–867. Springer (2005)
12. Chawla, N., Japkowicz, N., Kolcz, A.: Editorial: Special Issue on Learning from Imbalanced Data Sets. ACM SIGKDD Explorations Newsletter 6(1), 1–6 (2004)
13. Cohen, W.: Fast effective rule induction. In: Proc. of the 12th Int. ICML Conf., pp. 115–123 (1995)
14. Cost, S., Salzberg, S.: A Weighted Nearest Neighbor Algorithm for Learning with Symbolic Features. Machine Learning Journal 10(1), 1213–1228 (1993)
15. Davis, J., Goadrich, M.: The Relationship between Precision- Recall and ROC Curves. In: Proc. Int. Conf. on Machine Learning, ICML 2006, pp. 233–240 (2006)
16. Fawcett, T.: ROC Graphs: Notes and Practical Considerations for Data Mining Researchers. Technical Report HPL-2003-4. HP Labs (2003)
17. Fawcett, T., Provost, F.: Adaptive Fraud Detection. Data Mining and Knowledge Discovery 1(3), 29–316 (1997)
18. He, H., Garcia, E.: Learning from imbalanced data. IEEE Transactions on Data and Knowledge Engineering 21(9), 1263–1284 (2009)
19. He, J.: Rare Category Analysis. Ph.D Thesis. Machine Learning Department. Carnegie Mellon University Pittsburgh (May 2010), CMU-ML-10-106 Report
20. Holte, C., Acker, L.E., Porter, B.W.: Concept Learning and the Problem of Small Disjuncts. In: Proc. of the 11th JCAI Conference, pp. 813–818 (1989)
21. Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., Herrera, F.: A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches. IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews 99, 1–22 (2011)
22. Gamberger, D., Boskovic, R., Lavrac, N., Groselj, C.: Experiments With Noise Filtering in a Medical Domain. In: Proceedings of the Sixteenth International Conference on Machine Learning, pp. 143–151 (1999)

23. Garcia, S., Fernandez, A., Herrera, F.: Enhancing the effectiveness and interpretability of decision tree and rule induction classifiers with evolutionary training set selection over imbalanced problems. Applied Soft Computing 9, 1304–1314 (2009)
24. García, V., Sánchez, J., Mollineda, R.A.: An Empirical Study of the Behavior of Classifiers on Imbalanced and Overlapped Data Sets. In: Rueda, L., Mery, D., Kittler, J. (eds.) CIARP 2007. LNCS, vol. 4756, pp. 397–406. Springer, Heidelberg (2007)
25. Garcia, V., Mollineda, R.A., Sanchez, J.S.: On the k-NN performance in a challenging scenario of imbalance and overlapping. Pattern Analysis and Applications 11, 269–280 (2008)
26. Grzymala-Busse, J.W., Goodwin, L.K., Zheng, X.: An approach to imbalanced data sets based on changing rule strength. In: AAAI Workshop at the 17th Conference on AI Learning from Imbalanced Data Sets, Austin, TX, pp. 69–74 (2000)
27. 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)
28. Japkowicz, N., Stephen, S.: Class imbalance problem: a systematic study. Intelligent Data Analysis Journal 6(5), 429–450 (2002)
29. Japkowicz, N.: Class imbalance: Are we focusing on the right issue? In: Proc. II Workshop on Learning from Imbalanced Data Sets, ICML Conference, pp. 17–23 (2003)
30. Jo, T., Japkowicz, N.: Class Imbalances versus small disjuncts. ACM SIGKDD Explorations Newsletter 6(1), 40–49 (2004)
31. Joshi, M.V., Agarwal, R.C., Kumar, V.: Mining needles in a haystack: classifying rare classes via two-phase rule induction. In: Proc. of SIGMOD KDD 2001 Conference on Management of Data (2001)
32. Kaluzny, K.: Analysis of class decomposition in imbalanced data. Master Thesis (supervised by J.Stefanowski). Faculty of Computer Science and Managment, Poznan University of Technology (2009)
33. Khoshgoftaar, T., Seiffert, C., Van Hulse, J., Napolitano, A., Folleco, A.: Learning with Limited Minority Class Data. In: Proc. of the 6th Int. Conference on Machine Learning and Applications, pp. 348–353 (2007)
34. Kononenko, I., Kukar, M.: Machine Learning and Data Mining. Horwood Pub. (2007)
35. Kubat, M., Matwin, S.: Addresing the curse of imbalanced training sets: one-side selection. In: Proc. of the 14th Int. Conf. on Machine Learning ICML 1997, pp. 179–186 (1997)
36. Kubat, M., Holte, R., Matwin, S.: Machine Learning for the Detection of Oil Spills in Radar Images. Machine Learning Journal 30, 195–215 (1998)
37. Laurikkala, J.: Improving identification of difficult small classes by balancing class distribution. Tech. Report A-2001-2, University of Tampere (2001); Another version was published in: Laurikkala, J.: Improving Identification of Difficult Small Classes by Balancing Class Distribution. In: Quaglini, S., Barahona, P., Andreassen, S. (eds.) AIME 2001. LNCS (LNAI), vol. 2101, pp. 63–66. Springer, Heidelberg (2001)
38. Lewis, D., Catlett, J.: Heterogenous uncertainty sampling for supervised learning. In: Proc. of 11th Int. Conf. on Machine Learning, pp. 148–156 (1994)
39. Ling, C., Li, C.: Data Mining for Direct Marketing Problems and Solutions. In: Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD 1998), pp. 73–79. AAAI Press, New York (1998)
40. Maciejewski, T., Stefanowski, J.: Local Neighbourhood Extension of SMOTE for Mining Imbalanced Data. In: Proceeding IEEE Symposium on Computational Intelligence in Data Mining, within Joint IEEE Series of Symposiums of Computational Intelligence, April 11-14, pp. 104–111. IEEE Press, Paris (2011)

41. Mitchell, T.: Machine learning. McGraw Hill (1997)
42. Napierała, K., Stefanowski, J., Wilk, S.: Learning from Imbalanced Data in Presence of Noisy and Borderline Examples. In: Szczuka, M., Kryszkiewicz, M., Ramanna, S., Jensen, R., Hu, Q. (eds.) RSCTC 2010. LNCS(LNAI), vol. 6086, pp. 158–167. Springer, Heidelberg (2010)
43. Nickerson, A., Japkowicz, N., Milios, E.: Using unspervised learning to guide resampling in imbalanced data sets. In: Proc. of the 8th Int. Workshop on Artificial Intelligence and Statistics, pp. 261–265 (2001)
44. Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kaufmann (1992)
45. Pelleg, D., Moore, A.W.: Active learning for anomaly and rare-category detection. In: Proc. of NIPS (2004)
46. Prati, R.C., Batista, G.E.A.P.A., Monard, M.C.: Learning with Class Skews and Small Disjuncts. In: Bazzan, A.L.C., Labidi, S. (eds.) SBIA 2004. LNCS (LNAI), vol. 3171, pp. 296–306. Springer, Heidelberg (2004)
47. 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, pp. 312–321 (2004)
48. Ramentol, E., Caballero, Y., Bello, R., Herrera, F.: SMOTE-RSB*: A Hybrid Preprocessing Approach based on Oversampling and Undersampling for High Imbalanced Data-Sets using SMOTE and Rough Sets Theory. Knowledge and Information Systems Journal (2011) (accepted)
49. Saez, J.A., Luengo, J., Stefanowski, J., Herrera, F.: Addressing the Noisy and Borderline Examples Problem in Classication with Imbalanced Datasets via a Class Noise Filtering Method-based Re-sampling Technique. Manuscript submitted to Pattern Recognition (2011)
50. Stefanowski, J.: The rough set based rule induction technique for classification problems. In: Proc. of the 6th European Conf. on Intelligent Techniques and Soft Computing EUFIT 1998, pp. 109–113 (1998)
51. Stefanowski, J.: Algorithms of rule induction for knowledge discovery. Habilitation Thesis published as Series Rozprawy no. 361. Poznan University of Technology Press (2001) (in Polish)
52. Stefanowski, J.: On Combined Classifiers, Rule Induction and Rough Sets. In: Peters, J.F., Skowron, A., Düntsch, I., Grzymała-Busse, J.W., Orłowska, E., Polkowski, L. (eds.) Transactions on Rough Sets VI. LNCS, vol. 4374, pp. 329–350. Springer, Heidelberg (2007)
53. Stefanowski, J.: An experimental analysis of impact class decomposition and overlapping on the performance of classifiers learned from imbalanced data. Research Report of Institute of Computing Science, Poznan University of Technology, RB- 010/06 (2010)
54. Stefanowski, J., Wilk, S.: Improving Rule Based Classifiers Induced by MODLEM by Selective Pre-processing of Imbalanced Data. In: Proc. of the RSKD Workshop at ECML/PKDD, Warsaw, pp. 54–65 (2007)
55. Stefanowski, J., Wilk, S.: Selective Pre-processing of Imbalanced Data for Improving Classification Performance. In: Song, I.-Y., Eder, J., Nguyen, T.M. (eds.) DaWaK 2008. LNCS, vol. 5182, pp. 283–292. Springer, Heidelberg (2008)
56. Stefanowski, J., Wilk, S.: Extending Rule-Based Classifiers to Improve Recognition of Imbalanced Classes. In: Ras, Z.W., Dardzinska, A. (eds.) Advances in Data Management. SCI, vol. 223, pp. 131–154. Springer, Heidelberg (2009)
57. Sun, A., Lim, E.P., Liu, Y.: On strategies for imbalanced text classication using SVM: A comparative study. Decision Support Systems 48(1), 191–201 (2009)

58. Tomek, I.: Two Modications of CNN. IEEE Transactions on Systems, Man and Communications 6, 769–772 (1976)
59. Van Hulse, J., Khoshgoftarr, T., Napolitano, A.: Experimental perspectives on learning from imbalanced data. In: Proceedings of ICML 2007, pp. 935–942 (2007)
60. Van Hulse, J., Khoshgoftarr, T.: Knowledge discovery from imbalanced and noisy data. Data and Knowledge Engineering 68, 1513–1542 (2009)
61. Wang, B., Japkowicz, N.: Boosting support vector machines for imbalanced data sets. Knowledge and Information Systems 25(1), 1–20 (2010)
62. Weiss, G.M.: Mining with rarity: a unifying framework. ACM SIGKDD Explorations Newsletter 6(1), 7–19 (2004)
63. 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)
64. Wilson, D.R., Martinez, T.: Reduction techniques for instance-based learning algorithms. Machine Learning Journal 38, 257–286 (2000)
65. Wu, J., Xiong, H., Wu, P., Chen, J.: Local decomposition for rare class analysis. In: Proc. of KDD 2007 Conf., pp. 814–823 (2007)
66. Yang, Q., Wu, X.: 10 challenging problems in data mining research. International Journal of Information Technology and Decision Making 5(4), 597–604 (2006)