

# Selective pre-processing of imbalanced data for improving classification performance

Jerzy Stefanowski<sup>1</sup> and Szymon Wilk<sup>1,2</sup>

<sup>1</sup> Institute of Computing Science, Poznań University of Technology,  
ul. Piotrowo 2, 60-965 Poznań, Poland  
`jerzy.stefanowski@cs.put.poznan.pl`, `szymon.wilk@cs.put.poznan.pl`

<sup>2</sup> Telfer School of Management, University of Ottawa,  
55 Laurier Ave East, K1N 6N5 Ottawa, Canada  
`wilk@telfer.uottawa.ca`

**Abstract.** The paper discusses problems of constructing classifiers from imbalanced data. Re-sampling approaches that change the original class distribution are often used to improve performance of classifiers for the minority class. We describe a new approach to selective pre-processing of imbalanced data which combines local over-sampling of the minority class with filtering difficult examples from majority classes. In experiments focused on rule-based and tree-based classifiers we compare our approach with two other well known preprocessing methods – NCR and SMOTE. The results show that NCR is too strongly biased towards the minority class and leads to deteriorated specificity and overall accuracy, while SMOTE and our approach do not demonstrate such behavior. Analysis of a degree to which the original class distribution has been modified also reveals that our approach does not introduce so extensive changes as SMOTE.

## 1 Introduction

The problem of discovering classification knowledge from imbalanced data received much research interest in recent years [2, 4, 10]. A data set is considered to be *imbalanced* if one of the classes (further called a *minority class*) contains much smaller number of examples than the remaining classes (majority classes). The minority class is usually of primary interest from the perspective of considered problems. Then, the imbalanced distribution of classes constitutes a difficulty for standard learning algorithms because they are biased towards the majority classes. As a result examples from the majority classes are classified correctly by created classifiers, whereas examples from the minority class tend to be misclassified. In this context, overall classification accuracy is not appropriate performance measure because it is biased towards the majority classes. Thus, classifiers are evaluated by measures derived from a binary confusion matrix, like sensitivity and specificity (sensitivity is defined as the ratio of correctly recognized examples from the minority class, while specificity is the ratio of correctly recognized examples from the majority classes). Moreover the Receiver Operating Characteristics curve and the area under this curve are often used [2].

Resampling methods that modify the original class distributions in pre-processing are the most popular approaches to deal with imbalanced data. In particular, methods SMOTE, NCR or their combinations, were experimentally shown to work well [1, 3]. However, some of their properties can be considered as shortcomings. Focused undersampling methods, like NCR [6] or one-side-sampling [5], may remove too many examples from the majority classes. As a result, improved sensitivity is associated with deteriorated specificity. Random introduction of synthetic examples by SMOTE [3] may be questionable or difficult to justify in some domains, where it is important to preserve a link between original data and a constructed classifier (e.g., to justify suggested decisions). Moreover, SMOTE may blindly "over-generalize" the minority area without checking positions of nearest examples from the majority class and lead to overlapping between classes. Finally, the number of synthetic samples generated by SMOTE has to be globally parameterized, thus reducing the flexibility of approach.

Although when constructing classifiers from imbalanced data attention is usually focused on improving sensitivity, in many problems it is still important to preserve satisfactory specificity (i.e., to sufficiently recognize the majority classes), thus keeping overall accuracy of a classifier at an acceptable level. We claim that in general there is a kind of trade-off between sensitivity and specificity, but too large drop of specificity may not be accepted. Moreover, we hypothesize that it is worth to develop more flexible, adaptive approaches based on analyzing local neighborhood of "difficult" examples rather than using global approaches with fixed parameters. Following the above motivations we introduce our own approach to selective pre-processing of imbalanced data. It combines filtering of these examples of the majority classes, which may result in a classifier misclassifying some examples from the minority class, with local over-sampling of examples from the minority class that are located in "difficult regions" (i.e., surrounded by examples from the majority classes). On the one hand, such filtering is less greedy than the one employed by NCR, and on the other hand, the over-sampling is more specific than this used by SMOTE.

The main aim of this paper is to experimentally evaluate usefulness of our approach combined with two different learning algorithms. Specifically we use C4.5 [13] for inducing decision trees and MODLEM [7] for decision rules. We compare our technique to SMOTE and NCR – two re-sampling methods that are closely related to our proposal. The second aim of these experiments is to study how much all compared methods change the class distribution (numbers of examples in the minority and majority classes).

## 2 Related Works on Focused Re-sampling

Here we discuss only focused re-sampling methods, as they are most related to our approach and further experiments – for more extensive review see [2, 10].

In [5] one-side-sampling was used to under-sample the majority classes in a focused way. Noisy and borderline (i.e., lying on a border between decision

classes) examples from the majority class are identified using Tomek links and are deleted. Another approach to focused removal of examples from the majority class is the neighborhood cleaning rule (NCR) introduced in [6]. NCR uses the edited nearest neighbor rule (ENNR) to find and delete these examples from the majority classes, whose class labels differ from class labels of their three nearest neighbors. Experiments demonstrated that both above approaches provided better sensitivity than simple random over-sampling. According to [6] NCR performs better than one-side sampling and processes noisy examples more carefully.

The Synthetic Minority Over-sampling Technique (SMOTE) selectively over-samples the minority class by creating new synthetic (artificial) examples [3]. It considers each example from the minority class; finds its  $k$ -nearest neighbors from the minority class; randomly selects  $j$  of these neighbors and randomly introduces new artificial examples along the lines joining it with the  $j$  nearest neighbors. SMOTE can generate artificial examples with quantitative and qualitative attributes and a number of nearest neighbors depends on how extensive over-sampling is required. Experiments showed that a combination of SMOTE and under-sampling yielded the best AUC [3]. This was also confirmed in a comprehensive study [1], where various re-sampling methods were evaluated with different imbalanced data sets. SMOTE was also used in combination with ensemble classifiers as SMOTEBoost [2]. There are also other proposals to focused over-sampling, e.g. Japkowicz used local over-sampling of sub-clusters inside the minority class.

Our past research was concerned with the use of rough set theory to detect inconsistent examples that were in turn removed or relabeled [8]. This technique was combined with rule induction and experimentally evaluated. Then, in our last paper [9] we preliminary sketched the idea of the selective pre-processing based on ENNR which forms a basis of the approach presented in the next section.

### 3 An Algorithm for Selective Pre-processing

Our approach uses “internal characteristic” of examples to drive their pre-processing. We distinguish between two types of examples – *noisy* and *safe*. Safe examples should be correctly classified by a constructed classifier, while noisy ones are likely to be misclassified and require special processing. We discover the type of an example by applying the nearest neighbor rule (NNR) with the heterogeneous value distance metric (HVDM) [11]. An example is *safe* if it is correctly classified by its  $k$  nearest neighbors, otherwise it is *noisy*.

The approach is presented below in details in pseudo-code. We use  $C$  for denoting the minority class and  $O$  for one majority class (i.e. for simplicity we group all majority classes into one). We also use flags *safe* or *noisy* to indicate appropriate types of examples. Moreover, we introduce  $classify\_knn(x, k)$  for a result of classifying  $x$  using its  $k$  nearest neighbors and  $knn(x, k, c, f)$  for a set of these examples among  $k$  nearest neighbors of  $x$  that belong to class  $c$  and are

flagged as  $f$ . Following the suggestion from [6] we set  $k$  to 3. Finally, we assume  $|\cdot|$  returns the number of items in a set.

```

1: for each  $x \in O \cup C$  do
2:   if  $classify\_knn(x, 3)$  is correct then
3:     flag  $x$  as safe
4:   else
5:     flag  $x$  as noisy
6:    $D \leftarrow$  all  $y \in O$  and flagged as noisy
7:   if weak amplification then
8:     for each  $x \in C$  and flagged as noisy do
9:       amplify  $x$  by creating its  $|knn(x, 3, O, safe)|$  copies
10:  else if weak amplification and relabeling then
11:    for each  $x \in C$  and flagged as noisy do
12:      amplify  $x$  by creating its  $|knn(x, 3, O, safe)|$  copies
13:    for each  $x \in C$  and flagged as noisy do
14:      for each  $y \in knn(x, 3, O, noisy)$  do
15:        relabel  $y$  by changing its class from  $O$  to  $C$ 
16:        remove  $y$  from  $D$ 
17:    else {strong amplification}
18:      for each  $x \in C$  and flagged as safe do
19:        amplify  $x$  by creating its  $|knn(x, 3, O, safe)|$  copies
20:      for each  $x \in C$  and flagged as noisy do
21:        if  $classify\_knn(x, 5)$  is correct then
22:          amplify  $x$  by creating its  $|knn(x, 3, O, safe)|$  copies
23:        else
24:          amplify  $x$  by creating its  $|knn(x, 5, O, safe)|$  copies
25:      remove all  $y \in D$ 

```

Our approach consists of two phases. In the first phase (lines 1–5) we identify type of each example by applying NNR and flagging it accordingly. Then, in the second phase (lines 6–25) we process examples according to their flags. As we want to preserve all examples from  $C$ , we assume only examples from  $O$  may be removed (lines 6 and 25, where we applied ENNR). On the other hand, unlike previously described methods, we want to modify  $O$  more carefully, therefore, we preserve all safe examples from this class (NCR removes some of them if they are too close to *noisy* examples from  $C$ ). We propose three different techniques for the second phase: *weak amplification*, *weak amplification and relabeling*, and *strong amplification*. They all involve modification of the minority class, however, the degree and scope of changes varies between techniques.

Weak amplification (lines 7–9) is the simplest technique. It focuses on noisy examples from  $C$  and amplifies them by adding as many of their copies as there are safe examples from  $O$  in their 3-nearest neighborhoods. Thus, the amplification is limited to “difficult” examples from  $C$ , surrounded by safe members of  $O$  (if there are no such safe neighbors, then an example is not amplified; However its noisy neighbors from  $O$  will be removed - line 16). This increases the “weight”

of such difficult examples and enables learning algorithms to capture them as patterns, while they could be discarded as noise otherwise.

The second technique – weak amplification and relabeling (lines 10–16) – results from our previous positive experience with changing class labels of selected examples from  $O$  [8]. It is also focused on noisy examples from  $C$  and extends the first technique with an additional relabeling step. In the first step (lines 11–12) noisy examples from  $C$  surrounded by safe examples from  $O$  are weakly amplified. In the next step (lines 13–16) noisy examples from  $O$  located in the 3-nearest neighborhoods of noisy examples from  $C$  are relabeled by changing their class assignment is from  $O$  to  $C$  (reabeled examples are no longer removed – line 16). Thus, we expand the “cover” around selected noisy examples from  $C$ , what further increases their chance of being captured by learned classifiers. Such increasing of density is similar to the technique employed by SMOTE, however, instead of introducing new artificial examples, we use relabeled examples from the other class.

Strong amplification (lines 17–24) is the most sophisticated technique. It focuses on all examples from  $C$  – safe and noisy. First, it processes safe examples from  $C$  and amplifies them by adding as many copies as there are safe examples from  $O$  in their 3-nearest neighborhoods (lines 17–18). Then, it switches to noisy examples from  $C$  (lines 19–23). Each such example it is reclassified using an extended neighborhood, i.e., 5 nearest neighbors. If an example is reclassified correctly, it is amplified according to its regular neighborhood (i.e., by adding as many of its copies as there are safe examples from  $O$  in its 3-nearest neighborhood), as it should be sufficient to form a “strong” classification pattern. However, if an example is reclassified incorrectly, its amplification is stronger and the number of copies is equal to the number of safe examples from  $O$  in the 5-nearest neighborhood. Such more aggressive intervention is caused by the limited number of examples from  $C$  in the considered extended neighborhood and it is necessary to strengthen a classification pattern.

## 4 Experiments

Our approach for selective pre-processing was experimentally compared to NCR and SMOTE. We combined all tested approaches with two learning algorithms – C4.5 [13] for inducing decision trees and MODLEM [7] for decision rules. We focused on these two algorithms because they are both sensitive to class imbalance. Moreover, MODLEM was used in our previous research on improving sensitivity of classifiers [8, 9]. Both algorithms were run in their unpruned versions to get more precise description of the minority class. To obtain baseline results, we also run them without any prior pre-processing of data. For NCR and our approach the nearest neighborhood was calculated with  $k = 3$ . Moreover, to find the best over-sampling degree for SMOTE, we tested it different values from 100% to 600% [3] and selected the best one. We implemented MODLEM and all tested preprocessing approaches in WEKA [12]. We also used an implementation of C4.5 (called J48) available in this environment.

The experiments were carried out on 9 data sets listed in Table 1. They were either downloaded from the UCI repository or provided by our medical partners (acl). We selected data sets that were characterized by varying degrees of imbalance and that were used in related works (e.g. in [6]). Several data sets originally included more than two classes, however, to simplify calculations we decided to collapse all majority classes into one.

**Table 1.** Characteristics of evaluated data sets ( $N$  – number of examples,  $N_A$  – number of attributes,  $C$  – minority class,  $N_C$  – number of examples in the minority class,  $N_O$  – number of examples in the majority class,  $R_C = N_C/N$  – ratio of examples in the minority class)

Data set	$N$	$N_A$	$C$	$N_C$	$N_O$	$R_C$
Acl	140	6	with knee injury	40	100	0.29
Breast cancer	286	9	recurrence-events	85	201	0.30
Bupa	345	6	sick	145	200	0.42
Cleveland	303	13	positive	35	268	0.12
Ecoli	336	7	imU	35	301	0.10
Haberman	306	3	died	81	225	0.26
Hepatitis	155	19	die	32	123	0.21
New-thyroid	215	5	hyper	35	180	0.16
Pima	768	8	positive	268	500	0.35

During experiments we evaluated sensitivity, specificity and overall accuracy – see Tables 2, 3 and 4 respectively. All these measures were estimated in the 10-fold cross validation repeated 5 times. Evaluation of these three measures, as well as the trade-off between sensitivity and specificity, was the primary goal of our experiments. We also examined AUC reported by all tested approaches and both learning algorithms, however, due to limited space, we do not include these results. We also observed a degree of changes in class distributions introduced by all approaches – see Table 5. It was evaluated in a single pass of pre-processing. In all tables with results we use *base* for denoting the baseline approach (without any pre-processing), *weak* for weak amplification, *relabel* for weak amplification and relabeling, and *strong* for strong amplification.

In order to compare performance of evaluated approaches on all data sets we used the Wilcoxon Signed Ranks Test (confidence  $\alpha = 0.05$ ). Considering sensitivity (Table 2), all other approaches significantly outperformed the baseline (only for new-thyroid the baseline with C4.5 performed best). NCR led to the highest increase of sensitivity and the largest gain was observed for bupa, breast cancer, ecoli and haberman. It was observed for both learning algorithms, although relative improvements of sensitivity were better for MODLEM. For C4.5, the baseline results were better for almost all data sets, thus an opportunity for improvement was limited. Two variants of our approach combined with MODLEM – weak amplification and relabeling and strong amplification – were second best after NCR (there were differences between them on specific data sets). For

**Table 2.** Sensitivity

Data set	MODLEM						C4.5					
	Base	SMOTE	NCR	Weak	Relabel	Strong	Base	SMOTE	NCR	Weak	Relabel	Strong
Acl	0.805	0.850	0.900	0.830	0.835	0.825	0.855	0.840	0.920	0.835	0.835	0.850
Breast can.	0.319	0.468	0.638	0.437	0.554	0.539	0.387	0.463	0.648	0.500	0.576	0.531
Bupa	0.520	0.737	0.873	0.799	0.838	0.805	0.491	0.662	0.755	0.710	0.720	0.700
Cleveland	0.085	0.245	0.343	0.233	0.245	0.235	0.237	0.260	0.398	0.343	0.395	0.302
Ecoli	0.400	0.632	0.683	0.605	0.643	0.637	0.580	0.730	0.758	0.688	0.687	0.690
Haberman	0.240	0.301	0.626	0.404	0.468	0.483	0.410	0.572	0.608	0.657	0.694	0.660
Hepatitis	0.383	0.382	0.455	0.385	0.438	0.437	0.432	0.537	0.622	0.513	0.580	0.475
New-thyr.	0.812	0.917	0.842	0.860	0.877	0.865	0.922	0.898	0.873	0.897	0.897	0.913
Pima	0.485	0.640	0.793	0.685	0.738	0.738	0.601	0.739	0.768	0.718	0.751	0.715

a few data sets (cleveland, ecoli and new-thyroid) SMOTE was comparable to these two variants. The last variant of our approach – weak amplification – resulted in the worst sensitivity. The tested approaches demonstrated a similar performance when combined with C4.5.

**Table 3.** Specificity

Data set	MODLEM						C4.5					
	Base	SMOTE	NCR	Weak	Relabel	Strong	Base	SMOTE	NCR	Weak	Relabel	Strong
Acl	0.942	0.914	0.890	0.934	0.922	0.930	0.940	0.922	0.898	0.924	0.908	0.918
Breast can.	0.804	0.657	0.523	0.710	0.621	0.606	0.767	0.676	0.525	0.630	0.609	0.614
Bupa	0.820	0.568	0.308	0.453	0.473	0.459	0.775	0.611	0.415	0.524	0.459	0.532
Cleveland	0.957	0.887	0.884	0.934	0.919	0.927	0.899	0.870	0.849	0.877	0.864	0.887
Ecoli	0.969	0.951	0.924	0.958	0.953	0.962	0.959	0.921	0.920	0.931	0.916	0.941
Haberman	0.816	0.782	0.658	0.746	0.720	0.713	0.805	0.747	0.698	0.597	0.565	0.591
Hepatitis	0.933	0.927	0.894	0.918	0.907	0.908	0.873	0.851	0.823	0.822	0.807	0.803
New-thyr.	0.987	0.986	0.984	0.990	0.990	0.984	0.973	0.984	0.974	0.971	0.972	0.976
Pima	0.856	0.778	0.658	0.774	0.720	0.698	0.814	0.716	0.656	0.681	0.667	0.687

In case of specificity (Table 3), the baseline for both learning algorithms was significantly better than all other approaches and specificity attained by NCR was significantly lowest. For MODLEM the largest deterioration occurred for bupa (decrease of 0.512), breast cancer (0.282), pima (0.2) and haberman (0.152), while for C4.5 the loss of specificity was slightly smaller. Differences between the remaining approaches and the baseline were not so large. Our approach with weak amplification combined with MODLEM was able to preserve satisfactory specificity for most of data sets. SMOTE with MODLEM behaved similarly on some data sets (acl, ecoli, haberman, hepatitis and pima). On the other hand, SMOTE with C4.5 was slightly better than our approach. Similar observations hold for overall accuracy (Table 4) – the baseline was usually the best, then there were SMOTE and our approach. In particular, the variant with weak amplification combined with MODLEM managed to maintain high (i.e., the best or second best) accuracy for 5 data sets (acl, breast cancer, cleveland, haberman and pima). SMOTE with C4.5 demonstrated similar behavior on 6

**Table 4.** Overall accuracy [in %]

Data set	MODLEM						C4.5					
	Base	SMOTE	NCR	Weak	Relabel	Strong	Base	SMOTE	NCR	Weak	Relabel	Strong
Acl	90.3	89.6	89.3	90.4	89.7	90.0	91.6	89.9	90.4	89.9	88.7	89.9
Breast can.	66.0	60.0	55.6	62.9	60.1	58.6	65.4	61.2	56.1	59.1	59.9	58.9
Bupa	69.4	63.9	54.5	59.8	57.9	60.4	65.6	63.2	55.7	60.2	56.8	60.2
Cleveland	85.6	81.3	82.1	85.3	84.0	84.6	82.3	79.9	79.7	81.5	81.0	81.9
Ecoli	91.0	91.8	90.0	92.2	92.1	92.8	91.9	90.1	90.4	90.6	89.2	91.5
Haberman	66.3	65.4	64.9	65.5	65.2	65.1	70.1	70.0	67.4	61.3	59.9	60.9
Hepatitis	81.9	81.5	80.4	81.0	81.0	81.2	78.5	78.9	78.2	75.9	76.2	73.7
New-thyr.	95.8	97.4	96.2	96.9	97.1	96.5	96.5	97.0	95.8	95.9	96.0	96.6
Pima	72.7	73.0	70.6	74.3	72.7	71.2	74.0	72.4	69.5	69.4	69.6	69.7

data sets (breast cancer, bupa, haberman, hepatitis, new-thyroid and pima). Finally, overall accuracy achieved by NCR was the worst.

**Table 5.** Changes in class distribution ( $N_C$  – number of examples in the minority class,  $N_O$  – number of examples in the majority class  $N_O$ ,  $N_R$  – number of relabeled examples,  $N_A$  – number of amplified examples)

Data set	SMOTE		NCR		Weak		Relabel				Strong	
	$N_C$	$N_O$	$N_C$	$N_O$	$N_C$	$N_O$	$N_C$	$N_O$	$N_R$	$N_A$	$N_C$	$N_O$
Acl	120	100	40	83	57	98	59	98	2	17	67	98
Breast cancer	255	201	85	101	173	167	197	167	24	88	253	167
Bupa	290	200	145	81	236	145	271	145	35	91	309	145
Cleveland	245	268	35	198	102	255	110	255	8	67	147	255
Ecoli	210	301	35	266	58	288	69	288	11	23	77	288
Haberman	162	225	81	121	162	182	193	182	31	81	223	182
Hepatitis	64	123	32	90	61	113	68	113	7	29	88	113
New-thyroid	175	180	35	174	40	179	40	179	0	5	47	179
Pima	536	500	268	280	430	409	493	409	63	162	573	409

Analysis of changes in class distributions (Table 5), showed that NCR removed the highest number of examples from the majority class, in particular for breast cancer, bupa, haberman and pima it was about 50% of the majority class. None of the other approaches was such “greedy”. On the other hand, SMOTE increased the cardinality of the minority class on average by 250% by introducing new random artificial examples. For cleveland and ecoli it led to the highest increase of cardinality of the minority class (by 600% and 500% respectively). Our approach was in the middle, only the variant with strong amplification increased cardinality of the minority class for 4 data sets (bupa, breast, haberman and pima) to a level similar to SMOTE. For our approach with relabeling, a number of weakly amplified examples was usually higher than a number of relabeled examples. It may signal that for many data sets difficult noisy examples of the minority class (located inside the majority class) occurred more frequently

than noisy examples on the borderline. This was somehow confirmed by introducing many additional examples by the variant with strong amplification as a result of considering a wider neighborhood for amplification. Also when analyzing changes in class distribution ratio, we noticed that usually larger changes led to better classification performance (e.g., for cleveland, breast cancer and haberman).

## 5 Conclusions

The two main contributions of this paper are the presentation and experimental evaluation of the new selective approach to pre-processing imbalanced data. The approach combines local over-sampling of difficult examples from the minority class with removing or relabeling noisy examples from the majority classes. Its construction results from our motivation to preserve satisfactory recognition of majority classes while improving sensitivity of the minority class. The experiments showed that NCR was not able to satisfy such requirements as it tended to remove too many examples from the majority classes. As a result high increase of sensitivity of induced classifiers was at a cost of significantly decreased specificity and consequently deteriorated overall accuracy. Our approach and SMOTE did not demonstrate such behavior and both kept specificity and overall accuracy at an acceptable level.

The analysis of a changes in class distribution revealed that SMOTE introduced much more extensive changes than our technique, what might have resulted in swapping the minority and majority classes. Our approach tended to be more limited without sacrificing the performance gain. Unlike SMOTE, it did not introduce any artificial examples, but replicated existing ones what may be more acceptable in some applications.

According to experimental results, classification performance of our approach is slightly better or comparable to SMOTE depending on a learning algorithm. Moreover, it does not require tuning the global parameterized degree of over-sampling, but in a more flexible way identifies difficult regions in the minority class and modifies only these examples, which could be misclassified. Thus, we claim our approach is a viable alternative to SMOTE.

Selection of a particular variant in our approach depends on the accepted trade-off between sensitivity and specificity. On the one hand, if a classifier learned from pre-processed data has to provide higher sensitivity, then the variant with relabeling is the best choice (in our experiments is offered the highest sensitivity among the three for most data sets). On the other hand, specificity is more important, then the simplest variant with weak amplification is suggested (it performed best on specificity during experimental evaluation). Finally, if balance between sensitivity and specificity and good overall accuracy are requested, then the variant with strong amplification is preferred. It acceptable acceptable specificity (in experiments it was usually the second best among the three variants) with good sensitivity, especially when combined with C4.5. The latter combination also leads to very good overall accuracy (in experiments it resulted

in the highest accuracy among the three variants for most data sets). Moreover we can remark that relabeling of examples may not be accepted in some specific domains - in such cases users will decide between amplification variants.

At the end we would like to mention possible performance issues that may arise when pre-processing large data sets. All tested approaches, including ours, rely on NNR or its variants (e.g., ENNR), thus the process of identifying and searching local neighborhoods may be computationally expensive. This problem could be diminished by applying advanced indexing structures, e.g., ball trees [14] to improve effectiveness of calculations.

## References

1. 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), (2004) 20–29
2. Chawla, N.: Data mining for imbalanced datasets: An overview. In: Maimon O., Rokach L. (eds.) *The Data Mining and Knowledge Discovery Handbook*, Springer 2005, 853-867.
3. Chawla, N., Bowyer, K., Hall, L., Kegelmeyer, W.: SMOTE: Synthetic Minority Over-sampling Technique. *J. of Artificial Intelligence Research*, **16** (2002) 341–378.
4. Japkowicz N., Stephen, S., The Class Imbalance Problem: A Systematic Study. *Intelligent Data Analysis*, 6 (5), (2002) 429-450.
5. Kubat, M., Matwin, S.: Addressing the curse of imbalanced training sets: one-side selection. In: *Proc. of the 14th Int. Conf. on Machine Learning*, (1997) 179-186.
6. Laurikkala, J.: Improving identification of difficult small classes by balancing class distribution. *Tech. Report A-2001-2*, University of Tampere (2001).
7. Stefanowski, J.: The rough set based rule induction technique for classification problems. In: *Proc. of the 6th European Conference on Intelligent Techniques and Soft Computing EUFIT'98*, Aachen, (1998) 109-113.
8. Stefanowski, J., Wilk, S.: Rough sets for handling imbalanced data: combining filtering and rule-based classifiers. *Fundamenta Informaticae*, **72**, (2006) 379-391.
9. 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, (2007) 54-65.
10. Weiss, G.M.: Mining with rarity: a unifying framework. *ACM SIGKDD Explorations Newsletter*, **6**(1), (2004) 7–19.
11. Wilson, D.R., Martinez, T.: Reduction techniques for instance-based learning algorithms. *Machine Learning Journal*, **38** (2000) 257–286.
12. Witten, I.H., Frank, E.: *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann (2005).
13. Quinlan, J. R.: *C4.5: Programs for Machine Learning*. Morgan Kaufmann (1993).
14. Moore, A.W.: The anchors hierarchy: using the triangle inequality to survive high-dimensional data. In: *Proc. of the 16th Conference on Uncertainty in Artificial Intelligence*, Stanford, (2000) 397-405.