Selective Pre-processing of Imbalanced Data for Improving Classification Performance

Jerzy Stefanowski[1] and Szymon Wilk[1,2]

[1] 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
[2] Telfer School of Management, University of Ottawa,
55 Laurier Ave East, K1N 6N5 Ottawa, Canada
wilk@telfer.uottawa.ca

Abstract. In this paper we discuss problems of constructing classifiers from imbalanced data. 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 the majority classes. In experiments focused on rule-based and tree-based classifiers we compare our approach with two other related pre-processing methods – NCR and SMOTE. The results show that NCR is too strongly biased toward the minority class and leads to deteriorated specificity and overall accuracy, while SMOTE and our approach do not demonstrate such behavior. Analysis of the 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

Discovering classification knowledge from imbalanced data received much research interest in recent years [2,4,11]. 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 in a given application. The imbalanced distribution of classes constitutes a difficulty for standard learning algorithms because they are biased toward 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. As an overall classification accuracy is not appropriate performance measure in this context, such classifiers are evaluated by measures derived from a binary confusion matrix, like *sensitivity* and *specificity*. Sensitivity is defined as the ratio of the number of correctly recognized examples from the minority class (also called positive examples) to the cardinality of this class. On the other hand, specificity corresponds to ability of classifying negative examples of the minority class, so it is defined the ratio of correctly recognized examples from all the majority classes. Receiver Operating Characteristics curve and the area under this curve are also often used to summarize performance of a classifier [2].

I.-Y. Song, J. Eder, and T.M. Nguyen (Eds.): DaWaK 2008, LNCS 5182, pp. 283–292, 2008.

Several methods have been proposed to improve classifiers learned from imbalanced data, for a review see [4,11]. Re-sampling methods that modify the original class distributions in pre-processing are the most popular approaches. In particular, methods such as SMOTE, NCR or their combinations, were experimentally shown to work well [1,3,10]. However, some of their properties can be considered as shortcomings. Focused under-sampling 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 the nearest examples from the majority classes and lead to overlapping between classes. Finally, the number of synthetic samples has to be globally parameterized, thus reducing the flexibility of this approach.

Our main research thesis is that focusing on improving sensitivity, which is typical for many approaches to class imbalance, cannot cause too high decrease of specificity at the same time. In many problems sufficiently accurate recognition of the majority classes and preserving the overall accuracy of a classifier at an acceptable level are still required. Moreover, we hypothesize that it is worth to develop more flexible approaches based on analyzing local neighborhood of "difficult" examples rather than using global approaches with fixed parameters. Following these motivations we introduce our own approach to selective preprocessing of imbalanced data. It combines filtering of these examples from the majority classes, which may result in 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).

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 for inducing decision trees and MODLEM [7] for decision rules. We compare our approach to SMOTE and NCR – two 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 (the 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 reviews see [2,11]. In [5] one-side-sampling is 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 classes are identified using Tomek links and deleted. Another approach to the focused removal of examples from the majority class is the neighborhood cleaning rule (NCR) introduced in [6]. It applies the edited nearest neighbor rule (ENNR) to the majority classes [12]. ENNR first uses the nearest neighbor rule (NNR) to classify examples using a specific number of nearest neighbors (NCR sets it to 3) and then removes incorrectly classified ones. 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 majority classes, randomly selects j of these neighbors and randomly introduces new artificial examples along the lines joining it with the j-selected 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 among tested techniques [3]. This was 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]. Finally, there are other proposals to focused over-sampling, e.g. Japkowicz used local over-sampling of sub-clusters inside the minority class [4]. Our past research was concerned with the use of rough set theory to detect inconsistent examples in order to remove or relabel them [8]. This technique was combined with rule induction algorithms and experimentally evaluated. Then, in our last paper [9] we preliminary sketched the idea of the selective pre-processing based on ENNR which forms the basis of the approach presented in the next section.

3 An Algorithm for Selective Pre-processing

Our approach uses the the "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 NNR with the heterogeneous value distance metric (HVDM) [12]. An example is *safe* if it is correctly classified by its k nearest neighbors, otherwise it is *noisy*. We pre-process examples according to their type, and handle noisy examples from the majority classes following the principles of ENNR.

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 the majority classes into one). We also use the flags *safe* or *noisy* to indicate appropriate types of examples. Moreover, we introduce two functions: $classify\_knn(x, k)$ and $knn(x, k, c, f)$. The first function classifies x using its k nearest neighbors and returns information whether the classification is correct or not. The second function identifies k nearest neighbors of x and returns a set of those that belong to class c and are flagged as f (the returned set may be empty if none of the k neighbors belongs to c or is flagged as f). 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: $\mathbf{D} \leftarrow$ 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:          $D \leftarrow D \setminus \{y\}$
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 the type of each example by applying NNR and flagging it accordingly. Following the suggestion from [6] we set the number of the nearest neighbors to 3. 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 apply the principles of 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.

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). This increases the "weight" of such difficult examples and enables learning algorithms to capture them, 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 (relabeled examples are no longer removed – they are excluded from removal in 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 in similar to the technique employed by SMOTE, however, instead of introducing new artificial examples, we use relabeled ones from O.

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 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 – Quinlan's C4.5 for inducing decision trees and MODLEM [7] for decision rules. We focused on these two algorithms because they are both sensitive to the imbalanced distribution of classes. Moreover, MODLEM was introduced by one of the authors and successfully applied to many tasks including 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 $k = 3$, as suggested in [6]. Moreover, to find the best over-sampling degree for SMOTE, we tested it with different values from 100% to 600% [3] and selected the best one in terms of obtained sensitivity and

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

specificity. We implemented MODLEM and all tested pre-processing approaches in WEKA. We also used an implementation of C4.5 available in this environment.

The experiments were carried out on 9 data sets listed in Table 1. They were either downloaded from 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 the majority classes into one.

During experiments we evaluated sensitivity, specificity and overall accuracy – see Tables 2, 3 and 4 respectively (for easier orientation the best result for each data set and classifier is marked with boldface and italics, and the second best with italics). All these measures were estimated in the 10-fold cross validation repeated 5 times. Such a selection of evaluation measures allowed us to observe the degree of trade-off between abilities to recognize the minority and majority classes for the tested pre-processing approaches, what was the primary goal of our experiments. According to it we wanted to examine precisely the decrease of specificity and accuracy at the same time, which is not directly visible in ROC analysis. This is also the reason why we did not report values of AUC. Following the secondary goal of experiments, we 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 a performance of pairs of approaches with regard to results on all data sets we used the Wilcoxon Signed Ranks Test (confidence $\alpha = 0.05$). Considering sensitivity (Table 2), the baseline with both learning algorithms was significantly outperformed by all other approaches (only for new-thyroid the baseline with C4.5 performed best). NCR led to the highest gain of sensitivity, especially for haberman (0.386), bupa (0.353) and breast cancer (0.319). NCR improved sensitivity of both learning algorithms, although relative improvements

Table 2. Sensitivity

	MODLEM						C4.5					
Data set	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

were better for MODLEM – for C4.5 the opportunity for improvement was limited (the baseline results were better than for MODLEM). Relabeling and strong amplification was the second best after NCR for 7 data sets (all except acl and haberman). Two other variants of our approach – weak and strong amplification – resulted in worse sensitivity than the former one, but still they were better than SMOTE on the majority of data sets. The tested approaches demonstrated a similar performance when combined with C4.5.

Table 3. Specificity

	MODLEM						C4.5					
Data set	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. Specificity attained by NCR was significantly the lowest comparing to other methods. NCR combined with MODLEM led to the lowest specificity for all data sets. In particular the highest decreases for occurred for bupa (0.512), breast cancer (0.282), pima (0.200) and haberman (0.152) – these are also the sets for which we noted large increase of sensitivity. Slightly smaller loss of specificity occurred for C4.5. Our approach with weak amplification combined with MODLEM was able to preserve satisfactory specificity for most of the data sets. SMOTE with MODLEM behaved similarly on selected 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

Table 4. Overall accuracy [in %]

	MODLEM						C4.5					
Data set	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

high (i.e., the best or second best) accuracy for 6 data sets (acl, breast cancer, cleveland, ecoli, haberman and pima). SMOTE with C4.5 demonstrated similar behavior also for 6 data sets (breast cancer, bupa, haberman, hepatitis, new-thyroid and pima). Finally, overall accuracy achieved by NCR was the worst.

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

	SMOTE		NCR		Weak		Relabel				Strong	
Data set	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 largest number of examples from the majority class, in particular for breast cancer, bupa, haberman and pima it was about 50% of this 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 the level similar to SMOTE. For our approach with relabeling, the number of weakly amplified examples was usually higher than the number of relabeled examples. It may signal that 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 the changes in the class distribution ratio, we noticed that usually larger changes led to better classification performance (e.g., for cleveland, breast cancer and haberman).

Finally, we would like to note that in our on-going experiments we also consider two additional pre-processing approaches (random under- and over-sampling) and one additional rule learning algorithm (Ripper), however, due to limited space we can only say that the Ripper's performance was between the performance of MODLEM and C4.5, only for cleveland it gave the highest observed sensitivity at the cost of specificity and overall accuracy. Under- and over-sampling demonstrated performance do not outperform the remaining pre-processing approaches.

5 Conclusions

The main research idea in our study focused on improving sensitivity of the minority class while preserving sufficiently accurate recognition of the majority classes. This was our main motivation to introduce the new selective approach for pre-processing imbalanced data and to carry out its experimental comparison with other related methods. The results of experiments clearly showed that although NCR led to the highest increase of sensitivity of induced classifiers, it was obtained at a cost of significantly decreased specificity and consequently deteriorated overall accuracy. Thus, NCR was not able to satisfy our requirements. Our approach and SMOTE did not demonstrate such behavior and both kept specificity and overall accuracy at an acceptable level.

The analysis of the changes in the class distribution showed that NCR tended to remove too many examples from the majority classes – although it could ”clean” borders of minority class, it might deteriorate recognition abilities of induced classifiers for the majority classes. Moreover, the results revealed that SMOTE introduced much more extensive changes than our approach, what might have also resulted in swapping the minority and majority classes. Finally, one can notice that our approach tended to introduce more limited changes in the class distribution without sacrificing the performance gain. Additionally, unlike SMOTE, it did not introduce any artificial examples, but replicated existing ones what may be more acceptable in some applications.

To sum up, according to the experimental results, the 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 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 a particular variant in our approach depends on the accepted trade-off between sensitivity and specificity. If a user prefers classifiers characterized by higher sensitivity, then the variant with relabeling is the best choice. When specificity is more important, then the simplest variant with weak amplification is suggested. Finally, if balance between sensitivity and specificity and good overall accuracy are requested, then the variant with strong amplification is preferred. We should however note that relabeling of examples may not be accepted in some specific applications - in such cases users would have to decide between the two amplification variants.

Acknowledgment. This research supported by the grant N N519 3505 33.

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), 20–29 (2004)
2. 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, Heidelberg (2005)
3. Chawla, N., Bowyer, K., Hall, L., Kegelmeyer, W.: SMOTE: Synthetic Minority Over-sampling Technique. J. of Artifical Intelligence Research 16, 341–378 (2002)
4. Japkowicz, N., Stephen, S.: The Class Imbalance Problem: A Systematic Study. Intelligent Data Analysis 6(5), 429–450 (2002)
5. Kubat, M., Matwin, S.: Adressing the curse of imbalanced training sets: one-side selection. In: Proc. of the 14th Int. Conf. on Machine Learning, pp. 179–186 (1997)
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 1998, Aaachen, pp. 109–113 (1998)
8. Stefanowski, J., Wilk, S.: Rough sets for handling imbalanced data: combining filtering and rule-based classifiers. Fundamenta Informaticae 72, 379–391 (2006)
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, pp. 54–65 (2007)
10. Van Hulse, J., Khoshgoftarr, T., Napolitano, A.: Experimental perspectives on learning from imbalanced data. In: Proceedings of ICML 2007, pp. 935–942 (2007)
11. Weiss, G.M.: Mining with rarity: a unifying framework. ACM SIGKDD Explorations Newsletter 6(1), 7–19 (2004)
12. Wilson, D.R., Martinez, T.: Reduction techniques for instance-based learning algorithms. Machine Learning Journal 38, 257–286 (2000)