Improving Rule-Based Classifiers Induced by MODLEM by Selective Pre-processing of Imbalanced Data

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,
136 Jean-Jacques Lussier Str., K1N 6N5 Ottawa, Canada
wilk@telfer.uottawa.ca

Abstract. In the paper we discuss inducing rule-based classifiers from imbalanced data, where one class (a minority class) is under-represented in comparison to the remaining classes (majority classes). To improve the ability of a classifier to recognize this class, we propose a new selective pre-processing approach that is applied to data before inducing a rule-based classifier. The approach combines selective filtering of the majority classes with focused over-sampling of the minority class. Results of a comparative experimental study show that our approach improves sensitivity for the minority class while preserving the ability of a classifier to recognize examples from the majority classes.

1 Introduction

Many real-life knowledge discovery problems involve learning from *imbalanced data*, which means that one of the classes (further called a *minority class*) includes much smaller number of examples than the others (further referred to as *majority classes*). Moreover, examples from the minority class are usually of primary interest. Such situation is typical for medical problems, where the number of patients requiring special attention (e.g., therapy or treatment) is much smaller than the number of patients who do not need it. Similar situations occur in other domains – in [4, 14] the following problems are reported: detecting fraud/intrusion, managing risk, detecting of oil spills in satellite images, predicting technical equipment failures and information filtering.

Learning methods usually do not work properly on imbalanced data as they are “somehow biased” to focus on the majority classes while “missing” examples from the minority class. As a result created classifiers are also biased toward better recognition of the majority classes and they usually have difficulties (or even are unable) to classify correctly new objects from the minority class. This problem also affects rough set rule-based classifiers as elementary sets for the minority class are “weaker” than the ones for the majority classes and consequently rules generated on their basis have a lesser chance to contribute to the final classification. Overall classification accuracy is not the only and the best criterion characterizing performance of a classifier induced from imbalanced data. Satisfactory recognition of the minority class may be often more preferred, thus, a classifier should be characterized rather by its *sensitivity* and *specificity* for the minority class (sensitivity is defined as the ratio of correctly recognized examples from the minority class and specificity is the ratio of correctly excluded examples from the majority classes).

Too small number of examples in the minority class is not the only one problem with creating classifiers from imbalanced data. Other problems are: overlapping of examples from the minority class with examples from the majority classes, noise, data fragmentation, inappropriate use of greedy search strategies or evaluation measures. A number of solutions have been proposed to solve them, for review see [5, 14]. The most common are pre-processing techniques that change the distribution of examples among classes by appropriate sampling. Other approaches modify either induction or classification strategy, assign weights to examples, and use boosting or other combined classifiers. Some researchers transform the problem of learning from imbalanced data to the problem of cost learning (although it is not the same and misclassification costs are unequal and unknown) and use techniques from the ROC curve analysis.

We also studied this problem in two different ways. In [7] we introduced an approach that modified the structure of a rule-based classifier to increase its sensitivity. Then in [13] we studied a rough set pre-processing approach, where examples from majority classes belonging to a boundary between rough approximations of the minority class were filtered. Although it improved sensitivity of rule-based classifiers, we also noticed that focusing only on inconsistent examples was not sufficient as other “difficult” examples from lower approximations may still have degraded classification performance. Therefore, now we focus our attention on recent selective methods that change the original class distribution. In particular we are interested in *Synthetic Minority Over-sampling Technique* (SMOTE) [4] and *Neighborhood Cleaning Rule* (NCR) [10]. SMOTE is based on a specialized random introducing artificial examples from the minority class in some regions of data [4]. NCR, on the other hand, removes these examples from the majority classes that are located on the border with the minority class or that may be treated as noise [10]. Although these methods perform well [2], some of their properties could be seen as shortcomings. NCR is focused mainly on improving sensitivity for the minority class what deteriorates recognition of the majority classes. In general, there is a kind of trade-off between sensitivity and specificity but too large drop of specificity may not be accepted. Random introduction of artificial examples by SMOTE may be questionable or quite difficult to interpret and to justify in some domains (e.g., in medicine).

The main goal of this paper is to introduce a new selective pre-processing approach that aims at improving sensitivity for the minority class while preserving the ability of a classifier to recognize the majority classes and keeping overall accuracy at an acceptable level. Our approach combines selective filtering of examples from the majority classes with over-sampling of the minority class, however, it should remove less examples than NCR. Moreover, unlike SMOTE, it does not introduce any artificial examples, but replicates existing examples from the minority class that are located in "difficult regions" (i.e., they are surrounded by examples from the majority classes).

The second goal is to conduct an experimental evaluation of our pre-processing approach in combination with rule-based classifiers induced by the MODLEM algorithm. We compare it to other pre-processing methods such as SMOTE and NCR also combined with MODLEM classifiers. MODLEM has been chosen for consistency with our previous research on imbalanced data [7, 13] and its usefulness in many classification problems [12].

2 Related Works

We briefly describe only these pre-processing methods, which are related to our proposal; for more extensive reviews see [5, 14]. As the uneven distribution of examples among classes makes induction of classifiers more difficult, sampling methods are used to transform it. The simplest 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. However, random under-sampling may potentially remove some important examples and simple over-sampling may also lead to overfitting. Thus, recent research on sampling suggests focusing on particular examples from the minority class or the majority classes.

In [9] Kubat and Matwin analyzed mutual positions of examples from the majority classes. They distinguish four categories of examples: *noisy* examples located inside the minority class region, *borderline* examples (i.e., these laying either on or very close to the border between the minority class and the majority classes), *redundant* examples (i.e., examples distant from the border between classes) and *safe* examples. They detect these categories by applying Hart's Condensed Nearest Neighbor rule and Tomek links (two closest examples from different classes). Following the ideas of example selection from pattern recognition they introduced *one-side-sampling* approach, where the majority classes are selectively reduced by removing noise, borderline and redundant examples while keeping the minority class unchanged.

Another approach to focused removal of noisy and borderline examples from the majority class is NCR introduced by Laurikkala in [10]. NCR uses the Wilson's Edited Nearest Neighbor rule [15] and it can be shortly summarized in the following way: for each example x , its 3 nearest neighbors are found; if x belongs to one of the majority classes and its nearest neighbors misclassify it, then x is removed; if x belongs to the minority class and its neighbors misclassify it, then the neighbors that belong to the majority classes are removed. Experimental studies [2, 9, 10] demonstrated that both above approaches provided better sensitivity than simple random over-sampling. According to [10] NCR performs better than one-side sampling and considers noisy examples more carefully.

Chawla et al. introduced SMOTE, which selectively over-samples the minority class by creating new synthetic (artificial) examples [4]. Its main idea is to consider each example from the minority class and randomly introduce new artificial examples along the lines joining it with some of its k nearest neighbors from the minority class. SMOTE can generate artificial examples with quantitative and qualitative attributes [4] and the number of nearest neighbors depends on how extensive over-sampling is required. SMOTE is claimed to reduce the danger of overfitting as it does not simple replicate quite specific border examples but increases the "density" between examples from the minority class. A combination of SMOTE with some elements of under-sampling may additionally improve the ability of induced classifiers to recognize the minority class [2, 4].

In our previous research on pre-processing [13] an approach based on rough sets was applied to imbalanced and inconsistent data. We studied two techniques to detect and process inconsistent examples from the majority classes in the boundary between the minority and majority classes. The first one removes these examples from the learning set while the other *relabels* them as the minority class. The idea of relabeling was partly inspired by other research on the Generalized Edited Nearest Neighbor algorithm by Koplowitz and Brown (its description is given in [3]). In experiments these techniques were combined with two rule induction algorithms – LEM2 and MODLEM. Both techniques cleaned the boundary region of the minority class, what allowed inducing less specific rules. Moreover, the relabeling technique by increasing the number of examples in the minority class resulted in stronger rules that in turn led to higher sensitivity [13].

3 New Approach to Selective Pre-processing

Our proposal to selective pre-processing of imbalanced data combines elements of focused removal of examples from the majority classes with over-sampling of the minority class. Although it is inspired by some ideas presented in section 2, we apply them differently. First, we think that one-side-sampling and NCR may remove too many examples from the majority classes. Such greedy “cleaning” should definitely lead to increased sensitivity for the minority class, however, too extensive changes in the majority classes may deteriorate the ability of an induced classifier to recognize examples from these classes. As stated in the introduction, we believe in many problems it is necessary not only to improve sensitivity for the minority class, but also to maintain an acceptable level of overall accuracy.

The other premise for our approach is criticism of over-sampling performed by SMOTE that comes from our experience in analyzing real-life, especially medical data. Namely, we claim that random introduction of artificial examples may be questionable in practice, e.g., artificial “non-existing” patients could be questionable for physicians. SMOTE may introduce quite a high number of such artificial examples as according to [4] it may use the majority of 5 neighbors to generate them. Moreover, the position of new examples is selected in the direction of the nearest examples from the minority class without checking their relation to the nearest examples from the majority classes. To overcome these shortcomings we check alternative over-sampling for the minority class. It identifies only those examples that are likely to be misclassified and amplifies them and does not modify these examples that are possibly correctly classified.

Our approach to selective pre-preprocessing consists of two phases. In the first phase we analyze the “internal characteristics” of examples by distinguishing between their two types – *safe* and *noisy*. *Safe* examples should be correctly classified by an induced classifier, while *noisy* are very likely to be misclassified and thus require special attention in the second phase. We discover the type of an example by applying the Nearest Neighbor rule with the heterogeneous value distance metric (HVMD) [15] that handles quantitative and qualitative attributes. An example is *safe* if it is correctly classified by its k nearest neighbors, otherwise it is *noisy*. We further divide *safe* examples into *safe-certain* and *safe-possible* depending on the characteristic of their nearest neighbors. Analogously, *noisy* examples are divided into *noisy-certain* and *noisy-possible*.

In the second phase we process examples according to their type. As we want to preserve all examples from the minority class, we assume that only examples from the majority classes may be removed or relabeled (i.e., assigned to the minority class). Unlike previous methods, we want to modify the majority classes more carefully, therefore, we preserve all *safe* examples from the majority classes (let us note that NCR removes some of them if they are too close to *noisy* examples from the minority class). We propose three different techniques of processing examples: *relabeling and amplification*, *weak amplification* and *strong amplification*. They all involve modification of the minority class, however, the degree and scope of changes varies between techniques.

The *relabeling and amplification* technique is inspired by our previous good experience from [13]. It relabels *noisy* examples from the majority classes that are located in the nearest neighborhood of *noisy* examples from the minority class. Then it amplifies those *noisy-certain* examples from the minority class that have only *safe* examples from the majority classes in their nearest neighborhood. The *weak amplification* technique amplifies all *noisy* examples from the minority class. Finally, *strong amplification* also amplifies all *noisy* examples from the majority class, however it does it more extensively. It also amplifies these *safe* examples from the minority class that have *safe* examples from the majority classes in their nearest neighborhood.

Our approach is presented below in details as pseudo-code. We use C to denote the minority class and O to denote a helper class that combines all the majority classes. We also use “flags” to indicate the types of examples , e.g., examples from C are flagged as *C-safe-certain*, *C-safe-possible*, *C-noisy-certain* and *C-noisy-possible*, similar flags are used for examples from O. Moreover, for better readability we introduce “wildcard” flags, e.g., *C-noisy-** denotes both *C-noisy-certain* and *C-noisy-possible*. Finally, we assume $classify\_knn(x, k)$ classifies x using its k nearest neighbors, $knn(x, k, f)$ finds these of k nearest neighbors of example x that are flagged as f, $count\_knn(x, k, c)$ counts how many of k nearest neighbors of x belong to class c, and $count\_knn(x, k, f)$ counts how many of k nearest neighbors of x are are flagged as f. Following [10] we set k to 3.

```
for each $x \in O$ do
  if $classify\_knn(x, 3)$ is correct then
    if $count\_knn(x, 3, O) = 3$ then
      flag $x$ as *O-safe-certain*
    else
      flag $x$ as *O-safe-possible*
  else {$classify\_knn(x, 3)$ is incorrect}
    if $count\_knn(x, 3, C) = 3$ then
      flag $x$ as *O-noisy-certain*
    else
      flag $x$ as *O-noisy-possible*
for each $x \in C$ do
  if $classify\_knn(x, 3)$ is correct then
    if $count\_knn(x, 3, C) = 3$
    or $count\_knn(x, 3, O) = count\_knn(x, 3, \textit{O-noisy-*})$ then
      flag $x$ as *C-safe-certain*
    else
      flag $x$ as *C-safe-possible*
  else {$classify\_knn(x, 3)$ is incorrect}
    if $count\_knn(x, 3, O) = count\_knn(x, 3, \textit{O-noisy-*})$ then
      flag $x$ as *C-noisy-possible*
    else
      flag $x$ as *C-noisy-certain*
D ← all $x \in O$ flagged as *O-noisy-**
if relabeling and amplification then
  for each $x$ flagged as *C-noisy-** do
    for each $y \in knn(x, 3, \textit{O-noisy-*})$ do
      relabel $y$ by changing its class from $O$ to $C$
      remove $y$ from $D$
  for each $x$ flagged as *C-noisy-certain* do
    amplify $x$ by creating its $count\_knn(x, 3, \textit{O-safe-*})$ copies
else if weak amplification then
  for each $x$ flagged as *C-noisy-** do
    amplify $x$ by creating its $count\_knn(x, 3, \textit{O-safe-*})$ copies
else {strong amplification}
  for each $x$ flagged as *C-safe-possible* do
    amplify $x$ by creating its $count\_knn(x, 3, \textit{O-safe-*})$ copies
  for each $x$ flagged as *C-noisy-** do
    if $classify\_knn(x, 5)$ is correct then
      amplify $x$ by creating its $count\_knn(x, 3, \textit{O-safe-*})$ copies
    else
      amplify $x$ by creating its $count\_knn(x, 5, \textit{O-safe-*})$ copies
remove all $x \in D$
```

The first phase of our approach (lines 1-22) starts with identifying the types of examples from the majority classes. If an example is correctly classified using its 3 nearest neighbors, then it is safe – if all its 3 nearest neighbors are also from the majority classes, then it is flagged as *O-safe-certain* (lines 3-4), otherwise it is flagged as *O-safe-possible* (line 6). If an example is misclassified (line 7) then it is noisy – if all its 3 nearest neighbors are from the minority class, then it is flagged as *O-noisy-certain* (lines 8-9), otherwise it is flagged as *O-noisy-possible* (line 11). In the similar way the types of examples from the minority class are checked (line 12). If an example is classified correctly with its 3 nearest neighbors, then it is safe – if all its 3 nearest neighbors are from the minority class or all examples from the majority classes in its 3-nearest neighborhood are noisy, then it is flagged as *C-safe-certain* (lines 14-15), otherwise it is flagged as *C-safe-possible* (line 17). If an example is misclassified (line 18), then it is noisy – if all examples from the majority classes in its 3-nearest neighborhood are noisy, then it is flagged as *C-noisy-possible* (lines 19-20), otherwise it is flagged as *C-noisy-certain* (line 22).

The second phase (lines 23-42) starts with selecting all *O-noisy-** examples into the removal set D (line 23). Further processing depends on the selected technique. If it is *relabeling and amplification* (line 24), then for each *C-noisy-** example all *O-noisy-** examples in its 3-nearest neighborhood are identified (line 26), relabeled (line 27), and removed from D (line 28). Then each *C-noisy-certain* example is amplified by creating as many of its copies as there are *O-safe-** examples its 3-nearest neighborhood (line 30). If the selected technique is *weak amplification* (line 31), then each *C-noisy-** example is amplified by creating as many of its copies as there are *O-safe-** examples in its 3-nearest neighborhood (line 33). If the selected technique is *strong amplification* (line 34), then each *C-safe-possible* example is amplified by creating as many of its copies as there are *O-safe-** examples its 3-nearest neighborhood (line 36). Then for each *C-noisy-** example we check its extended neighborhood and classify it using its 5 nearest neighbors. If an example is classified correctly, then it is amplified by creating as many of its copies as there are *O-safe-** examples in its 3-nearest neighborhood (lines 38-39). Otherwise if an example is still classified incorrectly, it is stronger amplified by creating as many of its copies as there are *O-safe-** examples in its 5-nearest neighborhood (line 41). Finally, all examples from D are removed from a data set.

The above approach could be combined with any learning algorithm. In this study we combine it with MODLEM – a rough set algorithm for inducing rule-based classifiers, which was introduced by Stefanowski [11]; see also [12] for its detailed description. Shortly speaking, MODLEM follows the idea of sequential covering of rough approximations of decision classes by a minimal set of rules. While creating elementary conditions it handles both qualitative and quantitative attributes, and selection of the best condition is controlled by a criterion based on entropy. A new example is classified by matching its description to all induced rules. As it may lead to ambiguous situations (e.g., multiple match), we employ a strategy described in [12], which uses the strength of matched rules to solve conflicts (the strength of a rule is defined as the number of learning examples that satisfy the condition and the decision part of this rule). For each class the total strength of matched rules is calculated and the example is assigned to the strongest class. If no rule matches the classified example, the nearest rules are identified using HVDM and their strengths are used to find the strongest class in the same way as for matched rules – for more details see [12].

4 Experimental Study

The aim of experiments was to evaluate classification abilities of rule-based classifiers created by combining three techniques of the the selective pre-processing (relabeling and amplification, weak amplification, strong amplification) with MODLEM. We compared them to the basic approach with classifiers induced by MODLEM directly from imbalanced data (without any pre-processing), and classifiers created by combining SMOTE and NCR with MODLEM. In order to find the best over-sampling degree for SMOTE, we tested its different values and selected the one leading to the highest sensitivity of induced classifiers. Moreover, to extend the comparison we also included MODLEM with an approach that modifies the classification strategy for a rule-based classifier induced directly from a data set (without pre-processing) [7]. This approach was originally introduced by Grzymala in [6] and it is based on the idea of multiplying the strength of all minority class rules by the same real number, called a *strength multiplier*, while not changing the strength of rules from the majority classes. As a result, during such minority class rules have a better chance to classify new objects. The value of the strength multiplier is found by maximizing the measure *gain = sensitivity + specificity*. Implementations of all methods and the MODLEM algorithm were done in Java using the WEKA environment [16].

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, $R_C = N_C/N$ – ratio of examples in the minority class)

Data set	N	N_A	C	N_C	R_C
Acl	140	6	with knee injury	40	0.29
Breast cancer	286	9	recurrence-events	85	0.30
Bupa	345	6	sick	145	0.42
Cleveland	303	13	positive	35	0.12
Ecoli	336	7	imU	35	0.10
Glass	214	9	vehicle_windows_float_processed	17	0.08
Haberman	306	3	died	81	0.26
Hepatitis	147	19	die	31	0.21
New-thyroid	260	5	hyper	35	0.13
Pima	768	8	positive	268	0.35

The experiments were carried out on 10 data sets listed in Table 1. They come either from the UCI repository [1] or from our medical partners (acl). We selected the data sets that were characterized by varying degree of imbalance (ratio of examples in the minority class) and that were used in related works [7, 10]. Several data sets originally included more than two classes, however, to simplify calculations we decided to collapse all majority classes into one.

In the experiments we evaluated sensitivity and specificity for the minority class attained by created classifiers – see Tables 2 and 3. To control the trade-off between these two measures we also calculated their geometric mean, denoted as GM – see Table 4. According to [9] this measure relates to the point on a ROC curve and besides maximizing values of both components it allows to keep them balanced. Finally, we also evaluated overall accuracy – see Table 5. All measures were estimated in the 10-fold stratified cross validation repeated 5 times.

In order to compare the performance of evaluated approaches on all data sets we used the Wilcoxon Signed Ranks Test – a nonparametric test for significant differences between paired observations (confidence $\alpha = 0.05$). Considering sensitivity, all other approaches significantly outperformed the basic approach with no pre-processing. NCR led to the highest increase of sensitivity among all evaluated approaches – the differences between NCR and all other approaches were significant. The second best were two new selective pre-processing techniques: relabeling and amplification (relabel) and strong amplification (strong) – the difference between them was not significant. The third was SMOTE and weak amplification (weak). The approach with the strength multiplier (multiplier) led to the smallest increase of sensitivity.

In case of specificity, the basic approach was significantly better than all other approaches. The differences between the remaining approaches, except NCR, were not significant. Specificity attained by NCR was the lowest. Similar observation applies to overall accuracy – the basic approach was the best, then there were three techniques of new proposed approach, SMOTE and multiplier. All of them were significantly better than overall accuracy achieved by NCR.

NCR provided good results in terms of GM for a few data sets (cleveland, ecoli, glass), where increase of sensitivity caused only slight decrease of specificity. In general, we can conclude that very high increase of sensitivity was usually connected with decrease of specificity and consequently deteriorated overall accuracy. For other data sets, the proposed selective approach often demonstrated good trade off between sensitivity and specificity - although differences were not significant, the highest GM were obtained for the relabel technique.

When comparing the new approach to SMOTE we observed that it led to higher sensitivity allowing to "maintain" similar specificity and overall accuracy. The multiplier approach was the least efficient in improving sensitivity, however, quite good in keeping specificity close to the basic approach.

Table 2. Sensitivity for evaluated approaches and data sets (basic – no pre-processing, relabel – relabeling and amplification, weak - weak amplification, strong – strong amplification, multiplier - strength multiplier)

Data set	None	SMOTE	NCR	Relabel	Weak	Strong	Multiplier
Acl	0.7350	0.7500	0.9100	0.8950	0.8900	0.8900	0.7800
Breast cancer	0.3186	0.4681	0.6381	0.5544	0.4369	0.5386	0.4508
Bupa	0.5199	0.7529	0.8734	0.8375	0.7985	0.8047	0.5973
Cleveland	0.0717	0.1967	0.2850	0.2033	0.1600	0.1883	0.0933
Ecoli	0.4400	0.6300	0.7283	0.6367	0.6233	0.6333	0.4683
Glass	0.1700	0.2800	0.3400	0.2800	0.3200	0.3100	0.1800
Haberman	0.2397	0.3139	0.6258	0.4681	0.4039	0.4828	0.4011
Hepatitis	0.3833	0.4167	0.5300	0.5250	0.4283	0.4617	0.4950
New-thyroid	0.8067	0.8950	0.8100	0.8500	0.8467	0.8883	0.8233
Pima	0.4853	0.6147	0.7933	0.7377	0.6853	0.7377	0.7050

Table 3. Specificity for evaluated approaches and data sets (basic – no pre-processing, relabel – relabeling and amplification, weak - weak amplification, strong – strong amplification, multiplier - strength multiplier)

Data set	None	SMOTE	NCR	Relabel	Weak	Strong	Multiplier
Acl	0.9220	0.9080	0.8320	0.8860	0.8860	0.8860	0.9060
Breast cancer	0.8043	0.6570	0.5227	0.6212	0.7097	0.6061	0.6866
Bupa	0.8200	0.5720	0.3080	0.3930	0.4530	0.4590	0.7690
Cleveland	0.9553	0.9017	0.9092	0.9381	0.9418	0.9412	0.9441
Ecoli	0.9714	0.9462	0.9235	0.9514	0.9641	0.9581	0.9674
Glass	0.9818	0.9788	0.9634	0.9737	0.9758	0.9778	0.9778
Haberman	0.8155	0.7720	0.6583	0.7196	0.7455	0.7127	0.7366
Hepatitis	0.9208	0.9315	0.8570	0.9147	0.9062	0.9168	0.8867
New-thyroid	0.9900	0.9844	0.9844	0.9867	0.9878	0.9856	0.9856
Pima	0.8556	0.7852	0.6580	0.7204	0.7736	0.6980	0.7092

Table 4. GM for evaluated approaches and data sets (basic – no pre-processing, relabel – relabeling and amplification, weak - weak amplification, strong – strong amplification, multiplier - strength multiplier)

Data set	None	SMOTE	NCR	Relabel	Weak	Strong	Multiplier
Acl	0.8232	0.8252	0.8701	0.8905	0.8880	0.8880	0.8406
Breast cancer	0.5062	0.5546	0.5775	0.5869	0.5568	0.5714	0.5563
Bupa	0.6529	0.6562	0.5187	0.5737	0.6014	0.6077	0.6777
Cleveland	0.2617	0.4211	0.5090	0.4367	0.3882	0.4210	0.2968
Ecoli	0.6538	0.7721	0.8201	0.7783	0.7752	0.7790	0.6731
Glass	0.4085	0.5235	0.5723	0.5221	0.5588	0.5506	0.4195
Haberman	0.4421	0.4923	0.6418	0.5804	0.5487	0.5866	0.5436
Hepatitis	0.5941	0.6230	0.6740	0.6930	0.6230	0.6506	0.6625
New-thyroid	0.8937	0.9386	0.8930	0.9158	0.9145	0.9357	0.9008
Pima	0.6444	0.6947	0.7225	0.7290	0.7281	0.7176	0.7071

Table 5. Overall accuracy for evaluated approaches and data sets (basic – no pre-processing, relabel – relabeling and amplification, weak - weak amplification, strong – strong amplification, multiplier - strength multiplier)

Data set	None	SMOTE	NCR	Relabel	Weak	Strong	Multiplier
Acl	86.86%	86.29%	85.43%	88.86%	88.71%	88.71%	87.00%
Breast cancer	65.97%	60.02%	55.62%	60.07%	62.85%	58.58%	61.62%
Bupa	69.36%	64.79%	54.54%	57.94%	59.80%	60.38%	69.65%
Cleveland	85.35%	82.05%	83.72%	85.36%	85.17%	85.43%	84.63%
Ecoli	91.62%	91.38%	90.36%	91.90%	92.87%	92.45%	91.56%
Glass	91.80%	92.43%	91.49%	91.90%	92.47%	92.55%	91.52%
Haberman	66.29%	65.04%	64.93%	65.23%	65.45%	65.12%	64.73%
Hepatitis	80.79%	82.34%	78.87%	83.19%	80.59%	82.00%	80.39%
New-thyroid	96.01%	97.03%	95.63%	96.47%	96.48%	97.03%	95.91%
Pima	72.65%	72.58%	70.55%	72.65%	74.29%	71.20%	70.77%

5 Conclusions

In the paper we introduced the new approach to selective pre-processing of imbalanced data that aims at improving sensitivity of an induced classifier, while keeping overall accuracy at an acceptable level. It combines selective filtering of the majority classes with over-sampling of the minority class. Our approach removes less examples than NCR and, unlike SMOTE, it does not introduce any artificial examples, but replicates some of existing ones. Moreover, it does not require the parameterized degree of oversampling as it identifies minority class regions difficult to classify and modify only these examples, which could be misclassified. Within the proposed approach we developed three techniques of processing these examples involving amplification of examples from the minority class and relabeling examples from the majority classes – relabeling and amplification, weak amplification and strong amplification.

Our approach was verified in the experimental study where we compared it to other pre-processing methods, the basic basic approach with no-preprocessing and the approach that changes classification strategy. All these approaches were combined with rule-based classifiers. Results of experiments supported our initial intuition for NCR as a method strongly oriented toward improvement of sensitivity by extensive “cleaning” examples from the majority classes. Such cleaning made the majority classes more difficult to classify, thus, improvement of sensitivity was at a cost of decreased accuracy for the majority classes. Our approach was a bit worse (but it was the second best among all evaluated approaches) in terms of improving sensitivity, however, it demonstrated better specificity and overall accuracy. Moreover, when comparing the three techniques within the new proposed approach we could notice that more radical techniques (relabeling and amplification, strong amplification) were more efficient than weaker changes of class distribution (weak amplification). Similar experiments were also conducted using the C4.5 algorithm and tree-based classifiers. Although relative improvements of sensitivity were smaller, general behavior of compared approaches remained unchanged. Thus, we can conclude that the new proposed selective pre-processing approach leads to improved sensitivity for the minority class while preserving overall accuracy for various types of classifiers.

References

1. Blake, C., Koegh, E., Mertz, C.J.,: Repository of Machine Learning, University of California at Irvine 1999 [URL: http://www.ics.uci.edu/ mlearn/MLRepository.html].
2. Batista, G., Prati, R., Monard, M.: A study of the behavior of several methods for balancing machine learning training data. ACM SIGKDD Explorations Newsletter, **6**(1), (2004) 20–29
3. Caballero, Y., Bello, R., Garcia, M., et al.: Using rough sets to edit training set in k-NN method. In: Proc. of 5th ISDA 2006 Conf., IEEE Press, (2006) 456-463.
4. Chawla, N., Bowyer, K., Hall, L., Kegelmeyer, W.: SMOTE: Synthetic Minority Over-sampling Technique. J. of Articifical Intelligence Research, **16** (2002) 341–378.
5. Chawla, N., Japkowicz, N., Kolcz, A.: Editorial: Special Issue on Learning from Imbalanced Data Sets. ACM SIGKDD Explorations Newsletter, **6**(1), (2004) 1–6.
6. Grzymala-Busse, J.W., Goodwin, L.K., Grzymala-Busse, W.J., Zheng X.: An approach to imbalanced data sets based on changing rule strength. In: Proc. Learning from Imbalanced Data Sets, AAAI Workshop at the 17th Conference on AI, AAAI-2000, Austin, TX, July 30–31 (2000) 69–74.
7. Grzymala-Busse, J.W., Stefanowski, J., Wilk, Sz.: A comparison of two approaches to data mining from imbalanced data. In: Proceedings of the KES 2004, 8-th International Conference on Knowledge-based Intelligent Information & Engineering Systems, Wellington, New Zealand, Springer LNCS **3213** (2004) 757–763.
8. 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, (2001) 261–265.
9. Kubat, M., Matwin, S.: Addresing the curse of imbalanced training sets: one-side selection. In: Proc. of 14th Int. Conf. on Machine Learning ICML 97, (1997) 179-186.
10. Laurikkala, J.: Improving identification of difficult small classes by balancing class distribution. Tech. Report A-2001-2, University of Tampere (2001).
11. Stefanowski, J.: The rough set based rule induction technique for classification problems. In: Proc. of 6th European Conference on Intelligent Techniques and Soft Computing EUFIT'98, Aachen 7-10 Sept. (1998) 109-113.
12. Stefanowski, J.: On combined classifiers, rule induction and rough sets. In Peters J. et al. (eds.): Transactions on Rough Sets VI, Springer LNCS **4374** (2007) 329-350.
13. Stefanowski, J., Wilk, Sz.: Rough sets for handling imbalanced data: combining filtering and rule-based classifiers. Fundamenta Informaticae Journal, **72**(1-3), (2006) 379-391.
14. Weiss, G.M.: Mining with rarity: a unifying framework. ACM SIGKDD Explorations Newsletter, **6**(1), (2004) 7–19.
15. Wilson, D.R., Martinez, T.: Reduction techniques for instance-based learning algorithms. Machine Learning Journal, **38** (2000) 257–286.
16. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann (1999).