Extending rule based classifiers for dealing with imbalanced data

Jerzy Stefanowski
Institute of Computing Science, Poznań University of Technology,
ul. Piotrowo 3A, 60-965 Poznań, Poland
e-mail: Jerzy.Stefanowski@cs.put.poznan.pl

Abstract

Many real world applications involve learning from imbalanced data sets, i.e. data where the minority class of primary importance is under-represented in comparison to majority classes. The high imbalance is an important obstacle for many traditional machine learning algorithms as they are biased towards majority classes. It is desired to improve prediction of interesting, minority class examples, i.e. sensitivity, even at the cost of additional errors for majority classes. In this paper, we study two different approaches to improve sensitivity of rule based classifiers. The first group of approaches is based on the modification of the structure of rule sets induced for the minority class by replacing the minimal set of rules with the subset of stronger rules. We introduce the other approach, where the rule induction is combined with the filtering phase that removes noisy and borderline majority class examples from the data. The results of experiments have confirmed the usefulness of these approaches.

Keywords: machine learning, data mining, classification, class imbalance, minority class prediction

1. Introduction

This paper concerns topics of machine learning and discovery of classification knowledge from learning data. It is usually assumed that learning data sets are balanced. However, it is not always the case in some applications where the data set might be *imbalanced*, which means that one class of examples (further called a *minority class*) includes much smaller number of examples than other classes. Such data could be often met as many processes produce certain observations with a different frequency. A good example is medicine, where databases regarding a rare, but dangerous, disease usually contain a smaller group of patients requiring a special attention while there is a much larger number of members of other classes – patients who do not require the special treatment. Similar situations occur in other domains, e.g. in technical diagnostics or continuous fault-monitoring tasks, where non-faulty examples may heavily outnumber faulty examples. Survey papers [16, 3] report other real technical or engineering problems as, e.g., detection of oil spills in satellite radar images, the detection of fraudulent telephone calls or credit card transactions, predicting telecommunication equipment failures, text categorization, information retrieval and filtering.

The high imbalance between classes is reported to be an important obstacle in inducing classifiers. Their performance is often degraded as they are biased towards recognition of majority classes examples and they usually have difficulties (or even are unable) to classify correctly new objects from the minority class. In [11] an information retrieval system is discussed, where the minority class (being of primary importance) contains only 0.2% examples. Although the classifiers achieve accuracy close to 100%, they are useless because they fail to deliver the requested documents from this class. The similar degradation of the classifier performance for the minority classes is also reported for other imbalanced data. So, the total classification accuracy (an average percentage of all testing examples correctly recognized by the classifier) is not the only and the best criterion characterizing the classifier performance for such data sets. The users may prefer *high enough recognition of the minority class* and the final decision is characterized rather by its *sensitivity* (the ratio of correctly recognized examples from the critical class) and its *specificity* (the ratio of correctly excluded examples from other classes). More attention is given to sensitivity than to specificity [6]. In general there is a kind of trade-off between these two measures and the *ROC (Receiver Operating Characteristics) curve* technique can also be used to summarize classifier performance; for more details see, e.g., [3, 16].

The small number of minority class examples is not the only source of difficulty. Several researchers discussed that the minority class may overlap heavily other, majority classes, see e.g. [9, 10]. In particular, boundaries between classes may be ambiguous. If boundaries and the inside of the minority class are affected by noisy or inconsistent examples from other classes, it may cause incorrect classification of many examples from the minority class. Moreover, it becomes even a more difficult problem when the minority class contains many small sub-clusters, which are difficult to be learned [8]. Other aspects, e.g. inappropriate evaluation measures or inductive biases of learning algorithms, are discussed in [16].

In recent years this problem has received a growing research interest in the machine learning and data mining communities and several methods have been proposed, see e.g. the review in [16]. The most common technique for dealing with imbalance data is to transform the original class distribution into a more balanced one by appropriate *sampling*. There are also focused sampling techniques - some of them are discussed in the next section. Other proposals change the induction strategies of learning algorithm to make it more specific in case of the rare classes, for a comprehensive review see [16].

The author with his co-operators have also considered this problem and introduced an approach which modifies the rule classifier structure to increase its sensitivity for recognizing difficult examples from the minority class [7, 14]. The preliminary results showed that such a rule classifier performs quite well when compared to other algorithms [7].

However, this proposal is focused mainly on transforming rule sets to handle the uneven cardinalities of decision classes.

As we discussed before it is not the only source of difficulties for learning from imbalanced data. It may be also beneficial to focus our attention on noisy majority class examples and boundary examples between classes. Therefore, in this paper we want to consider yet another approach to imbalanced data based on the rule induction algorithm MODLEM [13], which is combined with techniques for handling such examples. The other contribution of this paper is carrying out an experimental comparison of this approach against the standard rule based classifiers and other popular sampling techniques. Finally, we will summarize our experience with different rule based approaches to imbalance data and discuss new possible research directions.

2. Related works

We briefly describe only these previous methods, which are the most related to the topics of this paper, i.e. selected sampling techniques, identification of noisy or borderline examples and some rule based approaches. For more exhaustive reviews of other works, the reader can consult, e.g., [16].

First of all, we should mention that one of the most popular technique for dealing with imbalance data is to transform the original class distribution into a more balanced by *sampling* procedures. The basic approaches include either random *over-sampling* or *under-sampling*. In the former approach the minority class examples are randomly replicated until a balance with cardinalities of majority classes is obtained. Random under-sampling goes in the opposite way - the majority class examples are randomly eliminated until obtaining the same cardinality as the minority class. The experimental evaluation of these techniques with different classifiers could be find in many papers, see e.g. an interesting study on artificial domains [8]. However, drawbacks of the above simple random techniques are often reported [1, 3, 9, 16]. It is claimed that random under-sampling can discard potentially useful majority class examples that could be valuable for learning a good classifier. On the other hand, simple over-sampling introduces copies of original examples only, which may lead to overfitting a classifier. Thus, several more "focused" heuristic techniques have also been introduced.

An example of such focused undersampling is an approach called *one-side-sampling* [9], where the majority class examples are divided into the following categories: *noisy* examples located inside the minority class region, *borderline* examples (lying at or near the decision border between classes), *redundant* examples (i.e. majority class examples which are quite distant from the decision border between classes) and *safe* examples. The borderline and noisy examples from the majority class are assumed to be a main source of misclassification for minority class examples. Besides an obvious interpretation of noise, borderline examples are treated to be unsafe since a small amount of noise could make them fall on the wrong side of the decision border between classes. These examples are detected by means of, so called, Tomek links [9] and removed. Redundant majority class examples are also removed.

Another approach to removing noisy and borderline examples is *Neighborhood Cleaning Rule* introduced by Laurikkala in [10]. It is based on the Wilson's *Edited Nearest Neighbor Rule* and removes these majority class examples whose class labels differ from the class of at least two of its three nearest neighbors. Experimental studies [1, 10] showed that both above approaches provide better sensitivity and not worse total accuracy than a simple random over-sampling. According to [10] the Neighborhood cleaning rule has usually worked better than the one side sampling.

As to focused over-sampling, Chawla et al. proposed a technique, called *SMOTE*, which over-samples the minority class by creating *new synthetic* examples [3]. Its main idea is to create these new examples by interpolating several minority class examples that are close one to another. It widens decision boundaries for the minority class. Several experimental results provided in [1, 3] indicate that SMOTE is often more efficient than other sampling methods. Its mixture with elements of under-sampling may improve the ability to predict the minority class - see [1]. Furthermore, there are interesting extensions of SMOTE for multiple classifiers, see a modification of AdaBoost into the SMOTEBoost. The aspects of modifying multiple classifiers for imbalanced data are also presented in [16].

Moreover, other simpler approaches to handle borderline examples between classes were considered, see the review [16]. For instance, Kubat and Matwin proposed in SHRINK system re-labeling borderline majority class examples into the minority class for the case of detecting oil spills images.

Let us shortly refer to other approaches that try to modify the algorithms for inducing classifiers, in particular rule approaches. Here, after [16] we should remind that typical rule or tree induction algorithms exploit a *greedy search strategy* while looking for rule conjunctions or use evaluation criteria, which favor the majority class but may be ineffective in dealing with minority examples. The paper [16] discusses main proposals to avoid or reduce these limitations. The most related to our proposals seems to a Brute rule induction algorithm introduced by Segal and Etzioni, where they tried to develop *less greedy search* for rules. Weiss [16] also reviews Holte at al. modifications of the rule induction algorithm CN2 to improve its performance for small disjuncts referring to rare examples. Moreover, he describes two-phase rule induction, where one part focuses on optimizing sensitivity while the other corresponds to optimizing specificity.

Other approaches to imbalance may use knowledge about prior distribution of probabilities or transforming the task to cost sensitivity learning and to a deeper analysis of ROC convex hull, see [16] or [3].

3. Changing rule sets for the minority class

Unlike sampling approaches that modify the class distribution in the input data, in this paper we are interested in methods that change the phase of inducing the classifier. Although the main aim is to improve the minority class prediction, even at the cost of making large changes in data, we still try to maintain a comprehensive structure of the symbolic knowledge representation. This is way we work with rule induction algorithms.

Here, we should remind that the construction of rule classifiers is usually divided into two phases [13]: (1) induction of the rule sets, (2) the strategy of using them to classify new objects. In the first phase several rule induction algorithms could be applied, for reviews see e.g. [4, 13]. The typical algorithms exploit a greedy search while looking for rules, which favors generating more general rule for the majority class but may be ineffective in dealing with minority examples [7, 16]. As matching of a new object description to induced rules may be ambiguous (i.e. matching rules indicate different decisions) or the object may not be matched by any rules, different classification strategies could be used to assign the new object to one decision class. An example is the Grzymala's proposal of the *LERS* (Learning from Examples based on Rough Sets) classification strategy, where the decision to which class the objects belongs is made using two factors: *strength* and *support*. The *Strength factor* is a measure of how well the rule has performed during learning, and is defined as the total number of examples correctly classified by the rule during learning. The second factor, *support*, is related to a decision class and is defined as the sum of strengths of all matching rules from the class. The new object is finally assigned to the class receiving the largest support. If complete matching is impossible, all partially matching rules are identified and the support factors are modified by taking into account additionally the *matching ratio* for each rule - it is defined as the number o elementary conditions of a rule matching the new object to the total number of conditions in this rule. One can notice that rules induced for the majority classes are usually more general, covering learning examples while minority class rules are "weaker" considering an average strength. So, while classifying a new object rules matching it and voting for the minority class are outvoted by rules voting for the majority class and the sensitivity of the resulting classification system may be low.

Overcoming the above obstacles in creating rule classifiers for imbalanced data may touch either changing the rule induction phase or proposing a more specific classification strategy for using rules.

We are partly inspired by previous Grzymala's works [5, 6], where the classification strategy was modified. The sensitivity of LEM2 induced [4] rule classifiers was modified by *changing rule strengths* inside the classification strategy. It was done by multiplying the rule strength for all rules describing the minority class by the same real number called a *strength multiplier*. As a result the chance that a minority class is selected by a classifier is increased. The key point is tuning the optimal value of the rule strength multiplier. In [5] Grzymala proposed to stepwise check several values and choose this one which maximize a measure called *gain = sensitivity + specificity* - 1. Although the general idea is simple, experimental results presented in [5, 6, 7] have confirmed that it improves the sensitivity of the LEM2 classifiers.

In [7] we proposed another approach to improve the minority class prediction, which touches the phase of rule construction. Let us remind that the typical rule induction algorithms, like LEM2, follow a *sequential covering scheme* providing a *minimal set of rules covering learning examples* [13]. This scheme is a kind of two level *greedy heuristic* procedure and consists of creating a first rule by choosing sequentially the 'best' elementary conditions according to some heuristic criteria (which typically favor rule generality). Then, learning examples covered by this rule are removed from consideration and the search is iteratively repeated while some examples remain uncovered. The complete induction process is repeated iteratively for succeeding classes. However, as a result rules for the minority class have a lower strength than rules from other classes and have a smaller chance to predict classification for new objects.

To improve a recognition of the minority class examples, our approach is based on the idea of *replacing the rule set* for this class by another rule set (more numerous and having, on average, slightly higher strength) that improves the chance of the classification strategy to select the minority class. Although strengths of new rules may be still lower than rules from the majority class, their higher number may lead to an increased number of votes inside the classification strategy. In order to generate additional rules for this class, we apply the EXPLORE algorithm [15]. As opposed to minimal covering algorithms, EXPLORE performs less greedy search and induces *all rules* that *satisfy certain requirements*, here the strength greater than a given threshold value. The main part of the algorithm is based on the breadth-first search, where rules are generated from the shortest to the longest ones by adding the most promising conditions. Creation of a rule condition part stops as soon as a candidate rule satisfies the requirements or it is impossible to fulfill the requirements in further steps. As this algorithm is less greedy than minimal rule covering algorithms, it discovers additional "strong" rules hidden in data, which were omitted by these algorithms due to their operation of discarding learning examples covered by just generated rules.

To sum up, our approach includes two stages. In the first stage, the *minimal set of rules* covering examples from decision classes is induced by an algorithm like LEM2. In the next stage a minimal set of rules for the minority class is discarded and replaced by a new set of rules induced by the Explore algorithm with the strength greater than a certain threshold. It is necessary to tune a proper value of this threshold. The tuning procedure varies this value increasing it from the smallest value equal to the minimal strength observed for rules generated for the minority class in the first stage. Choosing the best value is done according to the same gain criterion as used in the Grzymala's increasing strength approach.

In [7] we carried out a comparative study of this approach and the Grzymala's approach against the standard LEM2 classifier on 9 imbalanced data sets, coming mainly from UCI ML Repository [2]. Shortly summarizing the results we can conclude that both approaches performed better than simple LEM2 rule classifier considering the sensitivity and gain measures without decreasing the total accuracy. Differences between both approaches depend on particular data at hand. Taking into account the global summary of all experiments, we can say that the difference in performance of both approaches is statistically insignificant. Examples of highest improvements of sensitivity for our approach are the following: breast-Wisconsin - 0.326, german credit data - 0.325, urology - 0.219 and pima - 0.177, hepatitis - 0.145, scrotal pain - 0.146.

4. Handling "uncovered" examples

Although the previous approach works good in an experimental evaluation, one should notice that it is necessary to tune a proper value of the rule strength threshold - which a time consuming procedure organized in a way avoiding overfitting of the final classifier (which can be done with extra verification sets of examples). However, one should also notice that for many values of this threshold the EXPLORE may favor inducing additional rules for regions where enough minority class examples already occur. Therefore, these examples could be described by few quite similar rules, while more difficult examples (e.g. lying in rare subregions) may remain uncovered by any rule.

Thus, it is possible to consider modifications of this approach to handle these examples. For instance, we can consider a kind of *hybrid approach*, where the first level of representation are rules and the second level is a set of examples uncovered by these rules. The rules for the minority class are again obtained by EXPLORE algorithm as it has been described in the previous section. The rules for other majority classes could also be modified to exclude the boundary or noisy examples from these classes. This could be done either by rule pruning or using EXPLORE algorithm with higher values of the strength threshold. The classification strategy for new objects is also a kind of two stage approach. The new object is first tried to be classified by rules. If there is no matching to any rule or the matching is ambiguous, the object is classified by k-nearest neighbor principle on the basis of stored "uncovered" examples.

This idea was preliminary verified on the case of analysis of business credit applications [14]. The interesting observation is that the hybrid approach led to a higher total classification accuracy 81%, while EXPLORE itself gave 76.67% and other classifiers as C4.5 or IBL, around 74%. Furthermore, it slightly increased the sensitivity for the minority class up to 0,667 - which corresponded to the most risky bank customers who caused the questionable or lost liabilities. The similar improvements were observed for a medical case study.

5. Combining rule classifiers with techniques for handling noisy and borderline examples

The above described approaches may have some limitations. One corresponds to time consuming and sophisticated approaches for tuning proper values of parameters as strength multiplier or rule strength threshold. What is even more important, in these approaches uncovered examples are equally treated regardless their real nature. Let us notice that the notion of a "difficult", uncovered example rather depends on the tuning procedure not its real character. If one determines too high minimum strength threshold, it will result in not covering larger number of examples. On the other hand, tuning it too low will result in inducing too many rules. Moreover, looking more precisely into "replacing rule" approach, we can say that for some data it may rather be oriented on strengthening some sub-regions, i.e. some of additionally generated rules correspond to similar subsets of minority class examples, i.e. although they use different conjunctions of elementary conditions, they are supported by similar examples. Additionally, tuning rule sets for all classes may cause that majority classes examples may be more frequent among uncovered examples than minority class ones.

Here, we could come back to approaches discussed in section 2, which selectively modify the learning set. Let us remind that according to some authors it is necessary to identify difficult examples, which may cause errors while classifying minority class examples. In particular, it may concern majority class examples that are either *noisy* examples or *borderline* ones. Results of experiments carried out, e.g. in [1, 9, 10], showed that it is beneficial to properly handle such examples while improving sensitivity of classifiers.

Following the above motivations we would like to verify the effect of handling such examples on the sensitivity of rule based classifiers. Thus, we attempt to include the techniques for detecting and removing these examples as an additional data filtering before inducing rules and finally constructing the classifier. We implemented a filtering phase inspired by the Laurikkala's work [10]. It cleans majority class examples on the basis of the Wilson's *Edited Nearest Neighbor Rule*, which recommends to remove these examples whose class labels differ from the class of at least two of its three nearest neighbors. This helps to identify noise examples. As it is also necessary to remove borderline examples, the filtering procedure is two stage checking of nearest neighbors, what is summarized below:

Table 1. Characteristics of evaluated data sets

Data set	Number of examples	Ratio of examples Minority	Majority
breast-cancer -Slowenia	286	29.7%	70.3%
bupa	345	42%	58%
ecoli	336	10.4%	89.6%
glass	214	7.9%	92.1%
pima	768	34.9%	65.1%
haberman	306	26.5%	73.5%
acl	140	28.6%	71.4%
breast-cancer - Winsconsin	699	34.5%	65.5%
hepatitis	147	21.1%	78.9%
breast-Poland	228	29.0%	71.0%

1. Split a learning set E into a minority class C and the rest of data R.

2. Identify noisy majority examples from R, i.e. for each example in $e_i \in R$ check: if the classification given by three nearest neighbors of e_i contradicts its original class then add it to the set A_1.

3. For each example $e_j \in C$: if its three nearest neighbors misclassify e_j, then the nearest neighbors that belong to the majority classes are added to the set A_2.

4. Remove from E these majority class examples that belong to a set $\{A_1 \cup A_2\}$.

The nearest neighbors of a given example are found as in k-NN algorithm, where k =3, using a proper distance metric. As the Euclidean distance is not the sufficient for solving real world problems with mixed data described by numeric and nominal attributes, we used *heterogeneous value difference metric* [17], which is defined as:

$$HVDM(x,y) = \sqrt{\sum_{a=1}^{m} d_a^2(x_a, y_a)}$$

where $d_a^2(x_a, y_a)$ is the distance for attribute a describing examples x, y. For numeric attributes it is defined as normalized absolute value of the distance between values of an attribute. A distance for a nominal attribute is the *value difference metric*, introduced by Stanfill and Waltz, i.e. for attribute a and its values x_a and y_a it is defined as:

$$vdm_a(x_a, y_a) = \sum_{c=1}^{K} (N_{a,x_a,c}/N_{a,x} - N_{a,y,c}/N_{a,y})^2$$

where N_{a,x_a} is the number of times attribute a had value x_a; $N_{a,x_a,c}$ is the number of times attribute a had value x_a and the output class was c. The distance metric $HVDM$ provides appropriate normalization between numeric and nominal attributes, as well as between numeric attributes of different scales. Moreover, it handles unknown attribute values by assigning them a large distance.

After removing noisy and borderline examples from the majority classes R in the above filtering procedure, the rule set is induced from remaining data. We decided to use the algorithm MODLEM [12] for inducing a minimal set of rules. It is more general approach than LEM2, as it does not require pre-discretization of numeric attributes and handles unknown attribute values. We should remark that in the current experiments we gave up from using additional techniques modifying the induced rule sets, as described in section 3, because we wanted to evaluate directly the influence of the technique for handling noisy and borderline examples.

In the experiments we compare the use of four difference classifiers:

1. the standard classifier induced by MODLEM without any additions for imbalanced data,

2. our combined approach including the proposed filtering and MODLEM,

3. the simple random under-sampling and MODLEM,

4. the simple random over-sampling and MODLEM.

The random sampling were added because of their frequent use in other studies. The results of these approaches are expressed be means of *sensitivity, specificity* and *total accuracy*. Their values are evaluated in 10-fold stratified cross validation way. All classifiers were evaluated on 10 data sets, which are either machine learning benchmarks coming from the UCI repository [2] or from the author's previous applications in medicine (some of them are described in [13]). These data were chosen to be consistent with other selective sampling studies [1, 9, 10] and on the other hand to consider different degrees of imbalance or to solve difficult classification problems as medical ones. Some of the considered data sets were originally composed of more than two decision classes, however, to simplify problems we decided to group all majority classes into one. Unlike the previous experiments [7] we analyzed the original form of data, i.e. they were neither prediscretized nor missing values were substituted. The characteristics of these data sets is listed in Table 1. Then, Table 2 presents classification results for all compared classifiers.

Table 2. Classification performance of compared classifiers: standard rule classifier, combined with under-sampling, oversampling and new approach to selective filtering

Data set	Classifier type	Minority class sensitivity	Minority class specificity	Total accuracy
breast	stand.	0.3056	0.8505	69%
cancer	under	0.5971	0.5915	59%
Slove-	over	0.4043	0.8657	73%
nia	filtering	0.6264	0.5317	56%
	stand.	0.7290	0.5450	62%
bupa	under	0.6707	0.6910	68%
	over	0.5935	0.7521	69%
	filtering	0.8767	0.3250	56%
	stand.	0.4167	0.9667	91%
ecoli	under	0.8208	0.8430	84%
	over	0.5150	0.9578	91%
	filtering	0.7750	0.9335	92%
	stand.	0.2500	0.9847	92%
glass	under	0.7800	0.6351	65%
	over	0.4050	0.9817	94%
	filtering	0.4000	0.9645	92%
	stand.	0.4962	0.8460	72%
pima	under	0.7093	0.7150	71%
	over	0.5519	0.8148	72%
	filtering	0.8098	0.6420	70%
	stand.	0.2597	0.9833	96%
haber-	under	0.5793	0.6358	62%
man	over	0.2465	0.7393	61%
	filtering	0.6639	0.5994	62%
	stand.	0.7250	0.9100	86%
acl	under	0.8485	0.8375	84%
	over	0.7840	0.8795	86%
	filtering	0.8750	0.8400	85%
breast	stand.	0.9083	0.9586	94%
cancer	under	0.9521	0.9484	95%
Winsc-	over	0.8326	0.8619	85%
onsin	filtering	0.9625	0.9652	96%
	stand.	0.4833	0.9229	83%
hepatitis	under.	0.7372	0.7126	72%
	over.	0.5447	0.8541	81%
	filtering	0.6500	0.8364	80%
breast	stand.	0.3619	0.8640	72%
cancer	under	0.6903	0.6650	67%
Pol-	over	0.4367	0.7296	64%
and	filtering	0.8095	0.5923	65%

6. Conclusion

Constructing classifiers from imbalanced data requires special extensions if it is necessary to focus attention on recognizing examples from the class of interest being a minority class in the data. In particular, it concerns rule based classifiers because of several reasons, e.g. greedy search strategies or evaluation criteria used in the rule induction. In this paper we have discussed different approaches to improve the sensitivity of rule classifiers.

The first group of approaches is based on the modification of the structure of rule sets induced for the minority class (e.g. by means of replacing the minimal set of rules with a less greedy subset of stronger rules additionally induced from data) or on increasing the strength of minority class rules inside the classification strategy.

On the other hand, the next approach takes into account the characteristics of the input data and attempts to clean it from noisy or borderline examples belonging to the majority class while keeping the contents of the minority class unchanged. This is done in a filtering phase before the rule induction. The results of experiments are discussed below.

For nearly all data sets we observed that the new filtering approach improved the *sensitivity* of rule classifiers comparing to the standard classifier. For some data sets the increases were quite high, e.g. haberman - 0.404, breast cancer Slovenia - 0.32, ecoli - 0.358, breast cancer Poland - 0.448. Considering this criterion and other approaches the new filtering is better than simple random undersampling and oversampling. Then, undersampling usually leads to higher sensitivity than oversampling.

One could also notice that the improvement of sensitivity is often associated with the decrease of the specificity, e.g. see the results of filtering approach for breast cancer Slovenia, bupa or pima. However, if we compute the gain measure this value may be still high.

The similar observation concerns decreasing the total classification accuracy while improving the sensitivity. However, for the majority of data set this decrease may be accepted. Here, we can notice the random oversampling is the most robust and maintains the accuracy.

Comparing these results with the experiments performed previously for the first group of approaches, which change rule sets [7], we should be cautious as in the previous study some data sets were modified (e.g. the use of LEM2 required pre-discretization). Generally, it seems the highest increases of sensitivity were observed for different data sets.

Therefore, we hypothesize that it could be interesting to combine the best parts of both groups of approaches, i.e. handling difficult examples with modifications of inducing rule sets - but concerning only the part of the most critical rules to focus the search around the selected examples. These and more advanced methods are the subject of ongoing research.

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] Blake C., Koegh E., Mertz C.J., Repository of Machine Learning, University of California at Irvine 1999 [URL: http://www.ics.uci.edu/ mlearn/MLRepositoru.html].

[3] Chawla N., Bowyer K., Hall L., Kegelmeyer W., SMOTE: *Synthetic Minority Over-sampling Technique*. Journal of Articifical Intelligence Research, **16**, 2002, 341–378.

[4] Grzymala-Busse J.W., LERS - *a system for learning from examples based on rough sets*. In: Slowinski R. (eds.) Intelligent decision support. Handbook of application and advances of the rough sets theory. Kluwer Academic Publishers, 1992, 3–18.

[5] Grzymala-Busse, J. W., Goodwin, L. K., and Zhang, X., *Increasing sensitivity of preterm birth by changing rule strengths*. In: Proc. of the Eigth Workshop on Intelligent Information Systems IIS'99, Ustron, Poland, June 14–18, 1999, 127–136.

[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. of the AAAI Workshop Learning from Imbalanced Data Sets at the 17th Conference on AI, AAAI-2000, Austin, 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: Proc. of the KES 2004 - 8-th Int. Conf. on Knowledge-based Intelligent Information & Engineering Systems, Springer LNCS vol. **3213**, 2004, 757-763.

[8] Japkowicz, N., *Learning from imbalanced data sets: a comparison of various strategies*. In: Proc. of Learning from Imbalanced Data Sets, AAAI Workshop at the 17th Conference on AI, AAAI-2000, Austin, July 30–31, 2000, 10–17.

[9] Kubat M., Matwin S., *Addressing the curse of imbalanced training sets: one-side selection*. In: Proc. of 14th Int. Conf. on Machine Learning, 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] Lewis D., Catlett J., *Heterogenous uncertainty sampling for supervised learning*. In: Proc. of 11th Int. Conf. on Machine Learning, 1994, 148-156.

[12] Stefanowski J., *The rough set based rule induction technique for classification problems*. In: Proc. of 6th European Conf. on Intelligent Techniques and Soft Computing EUFIT'98, Aaachen 7-10 Sept. 1998, 109-113.

[13] Stefanowski J., *Algorithims of rule induction for knowledge discovery* (In Polish). Habilitation Thesis published as Series Rozprawy no. 361, Poznan Univeristy of Technology Press, Poznan, 2001.

[14] Stefanowski J., Wilk S., *Evaluating business credit risk by means of approach integrating decision rules and case based learning*. Int. Journal of Intelligent Systems in Accounting, Finance and Management, **10**, 2001, 97–114.

[15] Stefanowski J., Vanderpooten D., *Induction of decision rules in classification and discovery-oriented perspectives*. Int. Journal of Intelligent Systems, **16**, 2001, 13–28.

[16] Weiss G., *Mining with rarity: a unifying framework*. ACM SIGKDD Explorations Newsletter, **6** (1), 2004, 7-19.

[17] Wilson D.R.,Martinez T., *Reduction techniques for instance-based learning algorithms*. Machine Learning Journal, **38**, 2000, 257-286.