Application of Preprocessing Methods to Imbalanced Clinical Data: An Experimental Study

Szymon Wilk[1], Jerzy Stefanowski[1], Szymon Wojciechowski[1], Ken J. Farion[2], and Wojtek Michalowski[3]

[1] Poznan University of Technology, Poznan, Poland,
[2] Children's Hospital of Eastern Ontario, Ottawa, Canada,
[3] University of Ottawa, Ottawa, Canada,
szymon.wilk@cs.put.poznan.pl

Abstract. In this paper we describe an experimental study where we analyzed data difficulty factors encountered in imbalanced clinical data sets and examined how selected data preprocessing methods were able to address these factors. We considered five data sets describing various pediatric acute conditions. In all these data sets the minority class was sparse and overlapped with the majority classes, thus difficult to learn. We studied five different preprocessing methods: random under- and oversampling, SMOTE, neighborhood cleaning rule and SPIDER2 that were combined with the following classifiers: k-nearest neighbors, decision trees and rules, naive Bayes, neural networks and support vector machines. Application of preprocessing always improved classification performance, and the largest improvement was observed for random undersampling. Moreover, naive Bayes was the best performing classifier regardless of a used preprocessing method.

Keywords: clinical data, class imbalance, data difficulty factors, preprocessing methods, classification performance

1 Introduction

Clinical data pose a significant challenge for data mining due to such factors as missing or imprecise values, noisy or inconsistent observations and uneven distribution of patients from different decision classes [4, 11]. Usually, the number of patients from a critical class, who require special management, is much smaller than the number of patients from remaining classes (later in the text we refer to this critical class as to *minority* class, and to the remaining classes as to *majority* classes) – such situation is known as *class imbalance*.

Class imbalance limits the ability of constructed classifiers to make accurate predictions regardless of the applied learning method [6], and many methods for dealing with this problem have been already proposed (see [8] for their review). They can be divided into two general categories: *data-level* and *algorithm-level*.

The former methods preprocess data, e.g., by resampling, to change the distribution of classes, while the latter adjust the learning process, so they can be applied directly to imbalanced data. Although implementing modifications on the algorithmic level can potentially lead to more extensive improvement in classification performance, preprocessing methods are still a dominant approach.

However, class imbalance itself is not the only one or main problem. There are other factors related to data distribution that combined with class imbalance can seriously deteriorate classification accuracy, especially for the minority class [17]. These factors, often referred to as *data difficulty factors*, include rare subconcepts [10], overlapping regions of the minority and majority classes [20], or multiple minority class examples located inside the majority classes [17].

Napierala and Stefanowski proposed in [16] to approximate the above data difficulty factors using the local characteristic of the minority class examples. They distinguish between *safe*, *borderline*, *rare* and *outlier* examples. Safe examples lie inside the minority class and are surrounded mostly by neighbors from this class; borderline examples are located close to decision boundaries and thus their neighborhood is a mixture of the majority and minority class examples; rare examples form small islands (2-3 examples) inside the majority classes; finally outliers are isolated examples "thrown" into the majority classes. Borderline, rare and outlier examples are considered as *unsafe*, as they are more difficult to learn. The paper [16] also describes an automatic approach to label the minority class examples with their types that relies on examining the distribution of classes in their local neighborhoods.

In this paper we describe an experimental study aimed at identifying data difficulty factors in imbalanced clinical data sets, and at evaluating the impact of preprocessing methods on the performance of classifiers learned from these data sets. Specifically, this study answers the following research questions:

- What are the data difficulty factors (in terms of the types of minority class examples) encountered in the analyzed clinical data sets?
- How do the preprocessing methods improve the performance of classifiers, especially with respect to the minority class?
- What are the best (in terms of the improved performance) combinations of preprocesssing methods and classifiers?

The data sets used in this study describe pediatric patients managed for acute conditions in the emergency department (ED) of the Children's Hospital of Eastern Ontario (CHEO). Some of these conditions are highly prevalent (e.g., asthma exacerbation) and require significant resources for proper management. Others are less frequent (e.g., scrotal pain), however, their misdiagnosis may have serious health-related and legal consequences. For all conditions, quick and accurate recognition of patients from the critical class is necessary to provide appropriate and timely management. At the same time, these conditions are associated with serious class imbalance – addressing this problem together with associated data difficulty factors should help construct more accurate decision models and tools for regular use in the emergency setting.

2 Related Work

The most popular resampling methods are random *oversampling*, which replicates examples from the minority class, and random *undersampling* which randomly eliminates examples from the majority classes until a required degree of balance between classes has been reached. However, undersampling may potentially remove some important examples and oversampling may lead to overfitting [13]. Thus, recent research focuses on particular examples, taking into account information about their distribution in the attribute space [8].

Kubat and Matwin claim in [13] that characteristics of mutual positions of examples is a source of difficulty when learning from imbalanced data. They introduced a method called *one-sided sampling* (OSS) which filters the majority classes in a focused way [13]. It is based on distinguishing different types of learning examples: *safe*, *borderline* and *noisy*. They proposed to use Tomek links (two nearest examples having different class labels) to identify and delete the borderline and noisy examples from the majority classes. The critical analysis of OSS inspired the development of other informed preprocessing methods.

Neighborhood cleaning rule (NCR) represents another approach to the focused removal of examples from the majority classes [14]. It deals with the local data characteristics by applying the *edited nearest neighbor rule* (ENNR) to the majority classes [26]. ENNR first looks for a specific number of k-nearest neighbors ($k = 3$ is recommended in [14]) of a *seed* example, and uses their labels to predict the class label of this seed. In case of a wrong prediction, the neighbors from the majority classes are removed from the learning set.

The best known informed sampling method is *Synthetic Minority Oversampling TEchnique* (SMOTE) [3]. It is also based on the k-nearest neighborhood, however, this neighborhood is exploited to selectively oversample the minority class by creating new *synthetic* examples [3]. SMOTE treats each minority class example as seed and finds its k-nearest neighbors from the minority class. Then, according to the user-defined oversampling ratio o_r, SMOTE randomly selects o_r of these k neighbours and introduces new examples along the lines connecting the seed example with the selected neighbors. Although SMOTE was successfully applied to many problems, including medical ones (see experiments in [2, 17]), some of its new generalizations are better suited to deal with the data difficulty factors being considered [19].

SPIDER2 is a hybrid method that selectively filters out harmful examples from the majority class and amplifies difficult minority class examples [22]. In the first stage it applies ENNR to distinguish between safe and unsafe examples (depending how k neighbors reclassify the given seed example from the minority class). Outliers or neighbors from the majority classes that misclassify the seed example are either removed or relabeled. The remaining unsafe minority class examples are additionally replicated depending on the number of neighbors from the majority classes.

In all of these methods the k-nearest neighbourhood is calculated with the *Heterogeneous Value Difference Metric* (HVDM) [25]. HVDM aggregates nor-

malized distances for both numeric and nominal attributes, and for the latter it employs the *Value Difference Metric* (VDM) by Stanfill and Waltz [25].

Results of experiments presented in [17, 15] showed that when using decision trees and rule-based classifiers SPIDER2 and SMOTE were more accurate than random oversampling for data sets containing many unsafe types of the minority class examples. The most recent experimental study [16] demonstrated that undersampling like NCR was particularly useful when borderline examples dominated in the minority class.

3 Methods

3.1 Data Sets

The data sets considered in this study and their brief characteristics are given in Table 1. They describe pediatric patients with various acute conditions managed in the ED at CHEO. The AP and AE2 data sets were collected prospectively using either a mobile system or paper forms, and the other data sets were transcribed retrospective from charts. Moreover, the AE1 and AE2 data sets describe the same problem, however, they could not be combined together due to the differences in definitions of some of the attributes (e.g., their values were collected at different time points in the patients' care).

All data set are imbalanced – the class imbalance ratio, defined as the ratio of examples from the minority class to all examples, ranges from 0.09 to 0.16. The minority class is also the critical class, indicating patients who required special attention (e.g., specialist consultation, advanced laboratory investigations or intense and prolonged treatment). Initially, the data sets included three classes, and for the sake of analysis, the two classes corresponding to less urgent and intense management were combined into a single majority class.

Some of the data sets were initially fairly incomplete – especially SP and AE1 (both collected retrospectively). Therefore, before the analysis we removed attributes with more than 50% of missing values (Table 1 gives the numbers of attributes after the removal). There were 15 such attributes in SP and 10 in AE1. On the other hand the prospective data sets were complete – no attribute was removed from AE2 and a single attribute was removed from AP, confirming an advantage of prospective data collection.

In all of data sets, most attributes are nominal and correspond to symptoms and signs checked by physicians. Numeric attributes represent age and vital signs, such as temperature, heart rate or oxygen saturation.

3.2 Experimental Design

The experimental design covered two phases, corresponding to the main research questions formulated for our study. In the first phase we employed the labeling technique from [16] to establish the distribution of minority class example types in the considered data sets, and thus to identify the major data difficulty factors.

Table 1. Characteristics of the considered data sets

Data set	Clinical problem	# examples (minority)	Imbalance ratio	# attributes (numeric)
AP	abdominal pain	457 (48)	0.11	13 (3)
HP	hip pain	412 (46)	0.11	20 (4)
SP	scrotal pain	409 (56)	0.14	14 (3)
AE1	asthma exacerbations (2004)	362 (59)	0.16	32 (11)
AE2	asthma exacerbations (2007)	240 (21)	0.09	42 (9)

In the second phase we evaluated the impact of selected preprocessing methods on the performance of selected classifiers, with special focus on the minority class.

We selected the following preprocessing methods for our experiment (in brackets we give their symbols used further in the text): no preprocessing (none) as the baseline, random undersampling (RU), random oversampling (RO), SMOTE (SM), NCR, and SPIDER2 (SP2). Following our experience and suggestions from [24], the RU, RO and SM methods were set to produce a balanced distribution of classes in resulting data sets. Moreover, SM and SP2 were used with $k = 5$ nearest neighbors (this value is suggested for SM by its authors; its suitability was also confirmed by our earlier experiments with SP2), and the latter was parametrized for extended amplification of the minority class examples and for relabeling of the majority class examples instead of removing them (these options worked best in our previous studies – see [17] for details).

These methods were combined with the following classifiers frequently considered in clinical problems [1] (all implemented in WEKA[4]): k-NN with $k = 1$ and 3 neighbors (1NN and 3NN, respectively, see [16] for justification), PART decision rules (PART), C4.5 decision trees (C45), naive Bayes (NB), neural networks with radial basis functions (RBF), and support vector machines with radial basis function kernel (SVM). We selected such a kernel in SVM for consistency with our earlier experiments [16]. For PART and C45 we employed pruning – although unpruned classifiers are generally suggested for imbalanced data, our preliminary calculations revealed that unpruned decision rules and trees performed comparable or worse than pruned ones (this is consistent with [9] where the authors claim that pruning has the same effect as preprocessing). Moreover, parameters for RBF and SVM were optimized for each data set by using a simple grid search (systematic exploration of possible combinations of parameter values [21]) over original (i.e., not preprocessed) data sets. For the remaining classifiers we used default values of their parameters.

The classification performance was evaluated using the following measures, appropriate for imbalanced data: sensitivity and specificity for the minority class and their geometric mean (G-mean) that assesses their balance. We did not use the AUC (area under the Receiver Operating Characteristic curve) measure, as

[4] http://www.cs.waikato.ac.nz/ml/weka/

most of the classifiers selected in our study gave deterministic predictions. Here we should also note that by the minority classes we always considered the least prevalent class in the original data set (that was also critical from the clinical perspective) and we did not change it after preprocessing, even though several methods heavily modified the distribution of classes.

The above measures were estimated using a stratified 10-fold cross validation repeated 10 times to reduce the variance of results. In each pass of the validation loop, preprocessing methods were applied only to the learning set – the testing set was not modified. Finally, to gain better insight into how individual preprocessing methods affected the performance of considered classifiers, we applied non-parametric Friedman test (with $\alpha = 0.05$) that globally compared the performance of multiple combinations of preprocessing methods and classifiers over multiple data sets (see [16] for more details).

4 Results

Table 2 shows the distributions of the minority class example types in the original data sets. In all these data sets there was a large portion of borderline (24–63%) and outlier examples (25–52%), which indicates the minority class examples were "mixed" with the majority class ones. This, combined with small and very small portions of safe examples (14 and 29% for SP and AP, respectively; 2–7% for HP, SP and AE1 – much lower than the portion of rare examples), implies the considered data sets were difficult for learning (see the discussion in [16]).

Table 2. Distribution of example types [%]

Data set	Safe	Borderline	Rare	Outlier
AP	29	38	8	25
HP	7	28	15	50
SP	4	53	11	32
AE1	2	63	10	25
AE2	14	24	10	52

Due to space limitations, we focus on sensitivity when presenting detailed results from the second phase of the experiment; later we comment on the other measures (specificity and G-mean). Table 3 presents sensitivity for combinations of considered preprocessing methods and classifiers. In each row of this table the best value is marked with bold font, and the second best – with italics. Below we summarize our most important observations:

– Performance on the AP data set, which was predicted to be the easiest to learn due to the largest proportion of safe examples (see Table 2), was indeed better than the performance on the other data sets. This is especially visible for the baseline. Preprocessing diminished this difference, however, it still

could be observed. This confirms the usefulness of the labeling technique from [16] to distinguish between less and more difficult data sets.
- Application of preprocessing methods always improved performance classification performance in comparison to the baseline. The extent of improvement was dependent on the classifier – large improvement was demonstrated for the 1NN, 3NN, C45 and PART classifiers, moderate improvement for RBF, and only small improvement for NB. Poor improvement for NB was related to its stellar baseline performance that narrowed the margin for improvement in comparison with other classifiers.
- As already stated, NB was the best performing classifier. The only one serious competitor was RBF. On the one hand, its good performance could be attributed to optimization of parameters. On the other hand, parameters were optimized on the original data sets, and RBF also performed well on preprocessed data, which confirms the robustness of such optimization.
- RU was clearly the best preprocessing method. Only in a few cases was it superseded by SP2. Interestingly, these were cases where SP2 was combined with NB, which may indicate synergy between these two methods. We should also note that in cases where RU led to the best classification performance, the difference between RU and the second best preprocessing method was much larger than the difference between the best method and RU, where RU was second or third.

Table 4 reports the results of the Friedman test based on sensitivity. For each classifier, the *null* hypothesis (i.e., that all preprocessing methods performed equally) was rejected. According to the critical difference between ranks from the Nemeneyi test (3.4), the differences between best performing methods were not significant; however, the best methods were significantly better than the worst ones. Table 4 also confirms our earlier observations. RU was the best method for all classifiers, except NB. Other promising methods were SP2 (which performed best with NB) and RO – both were awarded with high average ranks for all classifiers. Here we should also note surprisingly poor performance of SM in comparison to results reported in [16]. This could be explained by the prevalence of symbolic attributes in the considered data sets – while SM is able to handle symbolic data, it does it in a very simple way (i.e., by using most frequent local values in newly created examples), thus limiting the chances for improvement.

For completeness, we also applied the Friedman test to rank classifiers applied to original data sets (no preprocessing) and to classifiers combined with RU (the best preprocessing method). The results are given in Table 5 and again they confirm our initial observation – the top ranks were assigned to NB and RBF respectively. Moreover, the ordering of classifiers according to ranks is quite stable – the only difference could be observed for 1NN an 3NN that swapped their positions.

Observations for G-mean are similar to those for sensitivity. Obtained results emphasized benefits of preprocessing (as for sensitivity, it always led to improvement in comparison to the baseline) and identified RU and NB as the

Table 3. Sensitivity for combinations of preprocessing methods and classifiers

Data set	Classifier	None	RU	RO	SM	NCR	SP2
AP	1NN	0.4300	**0.7500**	0.4300	0.5220	*0.5635*	0.5005
	3NN	0.4385	**0.7390**	*0.6495*	0.5365	0.5330	0.6230
	C45	0.3680	**0.7610**	0.5140	0.5005	0.5455	*0.5710*
	PART	0.4375	**0.7595**	0.5170	0.5255	*0.5340*	0.5325
	NB	0.7160	*0.7990*	0.7875	0.6770	0.7490	**0.8135**
	RBF	0.5130	**0.7860**	0.7645	0.6535	0.6685	*0.7405*
	SVM	0.5020	**0.7935**	0.7880	0.6150	0.5770	*0.7640*
HP	1NN	0.2035	**0.6035**	0.2035	*0.3315*	0.3040	0.2035
	3NN	0.1205	**0.6025**	*0.4300*	0.3630	0.2095	0.4280
	C45	0.2690	**0.7170**	*0.4965*	0.3865	0.3365	0.4780
	PART	0.2875	**0.6955**	*0.5115*	0.3585	0.3370	0.4840
	NB	0.7535	0.8480	*0.8510*	0.5645	0.7660	**0.8615**
	RBF	0.5475	**0.7920**	0.7145	0.4245	0.5865	0.6840
	SVM	0.5100	**0.7210**	0.4985	0.4445	0.5340	0.4970
SP	1NN	0.2743	**0.6307**	0.2743	0.3950	*0.4743*	0.2793
	3NN	0.2440	**0.6590**	*0.5553*	0.5240	0.4617	0.5513
	C45	0.3990	**0.6203**	0.5523	0.3950	0.4550	*0.5883*
	PART	0.3893	**0.6637**	0.5487	0.3597	0.4683	*0.5760*
	NB	0.4343	**0.7797**	0.7203	0.4077	0.5187	*0.7220*
	RBF	0.3913	**0.6977**	0.4920	0.4070	0.4743	*0.5220*
	SVM	0.3293	**0.6597**	0.3813	0.3350	*0.4163*	0.3947
AE1	1NN	0.2743	**0.5903**	0.2743	*0.4570*	0.3957	0.2760
	3NN	0.1623	**0.6327**	0.5097	*0.5277*	0.3163	0.4860
	C45	0.1847	**0.6080**	*0.3910*	0.2913	0.3097	0.3617
	PART	0.2553	**0.6330**	0.3723	0.2823	0.3497	*0.3953*
	NB	0.4897	*0.7143*	0.6833	0.4680	0.5803	**0.7167**
	RBF	0.4343	*0.6940*	0.6683	0.4763	0.5203	**0.7080**
	SVM	0.3217	**0.6170**	0.3147	0.3583	*0.4080*	0.3720
AE2	1NN	0.1000	**0.5867**	0.1000	*0.3217*	0.1317	0.1000
	3NN	0.0900	**0.7133**	0.4200	*0.4417*	0.1500	0.3750
	C45	0.1733	**0.6733**	*0.3933*	0.1500	0.2683	0.3300
	PART	0.2617	**0.6767**	*0.3767*	0.2817	0.3483	0.3400
	NB	0.7117	**0.7967**	0.7267	0.2400	0.7467	*0.7533*
	RBF	0.5317	**0.7917**	0.7367	0.2500	0.6800	*0.7533*
	SVM	0.4117	**0.5950**	0.3200	0.3433	*0.3533*	0.2900

most promising preprocessing method and classifier, respectively. However, unlike for sensitivity, the synergy between SP2 and NB was less visible, and NCR became a viable alternative for SP2 when combined with NB. Observations for specificity are complementary to those for sensitivity and G-mean. In most cases, application of preprocessing methods deteriorated the performance; only in a few cases was it preserved, in comparison to the baseline. As expected, the largest deterioration was observed for those methods that led to the best improvement of sensitivity, especially for RU (however, observed improvement of G-mean for

Table 4. Ranking of preprocessing methods for specific classifiers (based on sensitivity)

(a) 1NN

Method	Avg. rank
RU	6.0
SM	4.6
NCR	4.4
SP2	2.0
RO	2.0
None	2.0

(b) 3NN

Method	Avg. rank
RU	6.0
RO	4.6
SM	3.8
SP2	3.6
NCR	2.0
None	1.0

(c) C45

Method	Avg. rank
RU	6.0
SP2	4.4
RO	4.4
NCR	3.0
SM	1.8
None	1.4

(d) PART

Method	Avg. rank
RU	6.0
SP2	4.2
RO	4.0
NCR	3.4
SM	2.2
None	1.0

(e) NB

Method	Avg. rank
SP2	5.6
RU	5.2
RO	4.0
NCR	3.2
None	2.0
SM	1.0

(f) RBF

Method	Avg. rank
RU	6.0
SP2	4.8
RO	4.4
NCR	3.0
SM	1.6
None	1.4

(g) SVM

Method	Avg. rank
RU	6.0
NCR	4.2
SP2	3.0
RO	2.8
None	2.6
SM	2.4

these methods implied the loss on specificity was compensated by the gain on sensitivity). From this perspective, NCR and SP2 performed better than RU, as their negative impact on specificity was smaller (this more "balanced" behavior was consistent with our results reported in [17]).

Table 5. Ranking of classifiers methods for selected preprocessing methods (based on sensitivity)

(a) None

Classifier	Avg. rank
NB	7.0
RBF	5.8
SVM	4.6
PART	3.6
C45	3.0
1NN	2.4
3NN	1.6

(b) RU

Classifier	Avg. rank
NB	7.0
RBF	5.8
SVM	4.0
PART	4.0
C45	2.8
3NN	2.8
1NN	1.6

Finally, we checked how preprocessing methods changed the class imbalance ratio in the considered data sets. Specifically, we focused on NCR and SP2, as the other methods were set to produce a balanced class distribution. The results are given in Table 6. SP2 introduced more extensive changes than NCR – specifically it "strengthened" the minority class by introducing copies of existing examples and by relabeling examples from the majority class, which led to either almost

balanced (for AP and AE2) or balanced (for HP, SP and AE1) distribution of classes. This behavior of SP2 corresponds to suggestions in [24] and confirms the validity of our parametrization of RU, RO and SMOTE that enforced balanced distributions in preprocessed sets.

Table 6. Impact of preprocessing methods on the size and class imbalance of the considered data sets

Data set	NCR		SP2	
	# examples (minority)	Imbalance ratio	# examples (minority)	Imbalance ratio
AP	397 (48)	0.12	601 (201)	0.33
HP	367 (46)	0.13	670 (305)	0.46
SP	344 (56)	0.16	665 (316)	0.48
AE1	294 (59)	0.20	620 (323)	0.52
AE2	207 (21)	0.10	344 (126)	0.37

5 Conclusions

In the study described in this paper we analyzed data difficulty factors encountered in five imbalanced clinical data sets (all describing pediatric acute conditions managed in the ED) and examined how they could be addressed with common data preprocessing methods. Specifically, we evaluated how application of these methods affected the performance of classifiers often recommended for clinical problems [1].

All the considered data sets were characterized by large portions of borderline and outlier examples in the minority class (which was clinically also the critical class). This implies the minority class was dispersed and mixed with the majority class, thus difficult to learn, what was confirmed by the poor performance of classifiers on original (not processed) data sets.

Experimental evaluation demonstrated that preprocessing always improved classification performance (in terms of sensitivity and G-mean); the largest gain was observed for classifiers combined with random undersampling. Moreover, the best performance on both original and preprocessed data sets was observed for the naive Bayes classifier. These findings are consistent with our experience with using naive Bayes alone [7], combining random undersampling with naive Bayes [11, 12], as well as with other experimental and theoretical studies on random undersampling [6, 5, 23] and naive Bayes [18].

Here we would like to briefly comment on the viability of the naive Bayes classifier in clinical applications. On the one hand, it represents captured knowledge in the form of a priori probabilities that are definitely less comprehensible than decision rules or trees. On the other hand, Bayesian reasoning forms the foundations of clinical decision making (see, for example, [18]), and as such should be familiar to clinicians.

There were several limitations associated with our study. First, the small number of data sets affected the results of Friedman test and weakened its conclusions. Moreover, the data sets described problems from a single (and specific) clinical domain, and results may not be directly transferable to other clinical domains (e.g., chronic diseases in elderly). However, given the consistently good performance of random undersampling and naive Bayes, we think this combination offers a reasonable starting point to analyze other imbalanced clinical data sets.

Acknowledgement. The first three authors would like to acknowledge support by the Polish National Science Center under Grant No. DEC-2013/11/B/ST6/00963

References

1. Bellazzi, R., Zupan, B.: Predictive data mining in clinical medicine: current issues and guidelines. Int. J. Med. Inf. **77**(2) (2008) 81–97
2. Chawla, N.: Data mining for imbalanced datasets: An overview. In: Maimon, O., Rokach, L., eds.: The Data Mining and Knowledge Discovery Handbook. Springer (2005) 853–867
3. Chawla, N., Bowyer, K., Hall, L., Kegelmeyer, W.: SMOTE: synthetic minority over-sampling technique. J. Artif. Intell. Res. **16** (2002) 341–378
4. Cios, K., Moore, G.: Uniqueness of medical data mining. Artif. Intell. Med. **26** (2002) 1–24
5. Drummond, C., Holte, R.: C4.5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling. In: Proc. of the Workshop on Learning from Imbalanced Data Sets, ICML 2003. (2003) 1–8
6. Drummond, C., Holte, R.: Severe class imbalance: Why better algorithms aren't the answer. In: Proc. of the 16th European Conference ECML 2005, Springer (2005) 539–546
7. Farion, K., Wilk, S., Michalowski, W., O'Sullivan, D., Sayyad-Shirabad, J.: Comparing predictions made by a prediction model, clinical score, and physicians: pediatric asthma exacerbations in the emergency department. Appl. Clinic. Inform. **4**(3) (2013) 376–391
8. He, H., Ma, Y.: Imbalanced Learning: Foundations, Algorithms and Applications. Wiley (2013)
9. Hoens, T., Chawla, N.: Imbalanced datasets: from sampling to classifiers. In He, H., Ma, Y., eds.: Imbalanced Learning: Foundations, Algorithms and Applications. Wiley (2013) 43–59
10. Japkowicz, N.: Class imbalance: Are we focusing on the right issue. In: Proc. of the 2nd Workshop on Learning from Imbalanced Data Sets, ICML 2003. (2003) 17–23
11. Klement, W., Wilk, S., Michalowski, M., Farion, K., Osmond, M., Verter, V.: Predicting the need for CT imaging in children with minor head injury using an ensemble of naive bayes classifiers. Artif. Intell. Med. **54**(3) (2012) 163–170
12. Klement, W., Wilk, S., Michalowski, W., Matwin, S.: Classifying severely imbalanced data. In: Proc. of the 24th Canadian Conference on Artificial Intelligence, Canadian AI 2011, Springer (2011) 258–264

13. Kubat, M., Matwin, S.: Addressing the curse of imbalanced training sets: one-sided selection. In: Proc. of the 14th International Conference ICML 1997. (1997) 179–186
14. Laurikkala, J.: Improving identification of difficult small classes by balancing class distribution. In: Proc. of the 8th Conference AIME 2001. Volume 2101 of LNCS., Springer (2001) 63–66
15. Napierala, K., Stefanowski, J.: The influence of minority class distribution on learning from imbalance data. In: Proc. of the 7th Conference HAIS 2012. Volume 7209 of LNAI., Springer (2012) 139–150
16. Napierala, K., Stefanowski, J.: Types of minority class examples and their influence on learning classifiers from imbalanced data. J. Intell. Inform. Syst. (2016, to appear)
17. Napierala, K., Stefanowski, J., Wilk, S.: Learning from imbalanced data in presence of noisy and borderline examples. In: Proceedings of the 7th International Conference RSCTC 2010. Volume 6086 of LNAI., Springer (2010) 158–167
18. Sajda, P.: Machine learning for detection and diagnosis of disease. Annu. Rev. of Biomed. Eng. (8) (2006) 537–565
19. Saez, J., Luengo, J., Stefanowski, J., Herrera, F.: Addressing the noisy and borderline examples problem in classification with imbalanced datasets via a class noise filtering method-based re-sampling technique. Inform. Sci. **291** (2015) 184–203
20. Sanchez, V.G.J., Mollineda, R.: An empirical study of the behavior of classifiers on imbalanced and overlapped data sets. In: Proc. of the 12th Iberoamerican Conference on Progress in Pattern Recognition, Image Analysis and Applications, Springer (2007) 397–406
21. Staelin, C.: Parameter selection for support vector machines. Technical Report HPL-2002-354 (R.1). HP Laboratories, Israel (2003)
22. Stefanowski, J., Wilk, S.: Selective pre-processing of imbalanced data for improving classification performance. In: Proc. of the 10th International Conference DaWaK 2008. Volume 5182 of LNCS., Springer (2008) 283–292
23. Wallace, B., Small, K., Brodley, C., Trikalinos, T.: Class imbalance, redux. In: Proc. of the 11th IEEE International Conference on Data Mining. (2011) 754–763
24. Wei, Q., Dunbrack, R.: The role of balanced training and testing data sets for binary classifiers in bioinformatics. PLoS ONE **7**(8) (2013) e67863
25. Wilson, D., Martinez, T.: Improved heterogeneous distance functions. J. Atif. Intell. Res. **6** (1997) 1–34
26. Wilson, D., Martinez, T.: Reduction techniques for instance-based learning algorithms. Mach. Learn. J **38** (2000) 257–286