

On the role of cost-sensitive learning in imbalanced data oversampling

Bartosz Krawczyk¹[0000-0002-9774-0106] and Michal Wozniak²[0000-0003-0146-4205]

¹ Department of Computer Science, Virginia Commonwealth University, Richmond, VA, USA

bkrawczyk@vcu.edu

² Department of Systems and Computer Networks, Wrocław University of Science and Technology, Wrocław, Poland

michal.wozniak@pwr.edu.pl

Abstract. Learning from imbalanced data is still considered as one of the most challenging areas of machine learning. Among plethora of methods dedicated to alleviating the challenge of skewed distributions, two most distinct ones are data-level sampling and cost-sensitive learning. The former modifies the training set by either removing majority instances or generating additional minority ones. The latter associates a penalty cost with the minority class, in order to mitigate the classifiers' bias towards the better represented class. While these two approaches have been extensively studied on their own, no works so far have tried to combine their properties. Such a direction seems as highly promising, as in many real-life imbalanced problems we may obtain the actual misclassification cost and thus it should be embedded in the classification framework, regardless of the selected algorithm. This work aims to open a new direction for learning from imbalanced data, by investigating an interplay between the oversampling and cost-sensitive approaches. We show that there is a direct relationship between the misclassification cost imposed on the minority class and the oversampling ratios that aim to balance both classes. This becomes vivid when popular skew-insensitive metrics are modified to incorporate the cost-sensitive element. Our experimental study clearly shows a strong relationship between sampling and cost, indicating that this new direction should be pursued in the future in order to develop new and effective algorithms for imbalanced data.

Keywords: machine learning · imbalanced data · cost-sensitive learning · data preprocessing · oversampling · SMOTE

1 Introduction

Class imbalance occurs when the distribution of instances among classes in the training set is skewed [2]. As the training procedure of most classifiers is based on the predictive accuracy (or 0-1 loss function), an equal importance of all

training instances is inherently assumed. Therefore, learning algorithms tend to get biased towards the majority class, as it leads to overall smaller error than when trying to properly model infrequent and difficult minority class. Despite more than two decades of constant progress, learning from imbalanced data still poses a challenge for machine learning community [12]. It can be contributed to the constant emergence of new real-life problems, in which instances coming from one of the classes are much less frequent than from the others. Traditional examples of such cases include medicine, where we deal with diagnosis of a rare disease, or fraud detection systems, where we have a plethora of correct transactions versus a handful of fraudulent ones. Recent advances in machine learning and data mining brought the challenge of tackling class imbalance into new fields, such as big data [15], data stream mining [21], or structured outputs [6], among others. This creates new challenges force researchers to come up with new algorithms that are able to scale-up to ever-increasing volume and velocity of data, as well as adapt to emerging difficulties embedded in the nature of analyzed datasets.

To address the problem of imbalanced data two main approaches are used: data-level [7] and algorithm-level solutions [13]. The former ones concentrate on modifying the training set by removing or generating instances, in order to achieve rebalanced distributions. The latter ones aim at gaining an insight into what causes a given classifier to fail and modify its underlying mechanisms. Data-level solutions can be seen as more general ones, as they usually do not involve a specific classifier while performing sampling. Therefore, the processed dataset can be used by any conventional machine learning technique. The algorithm-level solutions are more specialized, usually designed for a given specific type of classifier and cannot be that easily transferred to another family of learners. At the same time, they may offer a more precise solution for tackling class imbalance.

Cost-sensitive learning is arguably the most wide-spread algorithm-level solution [8]. It assumes the modification of the standard 0-1 loss function and adding a learning penalty for misclassification of the minority class [4]. This will lead to an increased importance of the minority class instances during training and alleviation of the bias towards the better-represented majority class. It can be seen either as modifying the cost matrix for a classifier [1], or as an realization of instance weighting [23]. While this approach is efficient and many existing classifiers can be easily modified to their cost-sensitive versions [5, 9], its main limitation lies in a lack of well-defined techniques for estimating the optimal misclassification cost. When improperly set, the cost parameter may significantly deteriorate the performance of a classifier, which is a main cause of many researchers preferring the data-level solutions [14].

One should notice that in many real-life imbalanced problems the parameter cost may be obtained from a domain expert [12]. In case of medical diagnosis, it will be a cost of making a wrong prediction about a patient and thus following issues with incorrect medications. In fraud detection, it will be the cost of allowing for a adversarial transaction to take place. Despite this fact, many solutions

to these problems ignore the underlying cost and focus on data-level solutions. We will discuss here that this is not a correct approach to such applications.

In this paper, we propose to investigate a relationship between data-level algorithms and cost-sensitive learning. We argue that one cannot simply apply a sampling technique without any regard for the associated costs. Additionally, existing cost-sensitive algorithms use the cost parameter during training, but never take it into account during the evaluation phase. This leads to incorrect error estimations that may be too optimistic. Through a thorough experimental study, we investigate the interplay between varying misclassification costs and oversampling ratios used by popular data-level techniques. We show that using cost-sensitive modifications of skew-insensitive performance metrics reveals a clear correlation between these two factors that cannot be neglected. This is a starting work on proposing a new paradigm for learning from imbalanced data that combined sampling and cost-sensitive algorithms.

The contributions of this work are as follow:

- A proposal of new direction in learning from imbalanced data that uses the information from cost-sensitive learning in data-level solutions.
- A new experimental setup for imbalanced cost-sensitive learning, where misclassification cost is taken into account both during training and testing.
- A thorough experimental study investigation relationships between cost-sensitive framework and oversampling performance.

The remaining of this manuscript presents an insight into the problem of imbalanced data classification, with special emphasis on cost-sensitive solutions, discusses the relationships between cost and oversampling, depicts and discusses results of the experimental study, as well as presents lines for future research in this topic.

2 Learning from imbalanced data

Imbalanced data is widely known problem in machine learning domain, where unequal distribution of possible classes occurs in datasets [2, 12]. In this paper, we will focus on the imbalanced data problem with two-class problem being taken into account in which two classes can be specified and one of them is underrepresented. Imbalanced dataset provides insufficient or inadequate representation of one class known as *minority* class, while *majority* class refers to the one that is more representative or even overrepresented.

Due to its nature, imbalanced data is mostly characterized by its *Imbalance Ratio* (IR) as well as intrinsic characteristic like disjuncts of overlapping of classes. Imbalance Ratio is defined as a ratio between number of objects that corresponds to the majority class and number of instances of the majority class. In other words, the higher the value, the more imbalanced dataset is due to the minority class being highly underrepresented.

However, the IR is not the sole source of learning difficulties. Small sample size of the minority class may inhibit any generalization capabilities of a classifier,

while local data characteristics make some instances harder to classify than the others [18]. Such cases as borderline or noisy instances pose additional challenge to a classifier and thus should be paid special attention during the learning phase.

As a solution to class imbalance, three main groups of techniques were developed. *Preprocessing methods* refers to group of algorithms that alters inner structure of dataset by either introducing new minority class samples (*oversampling*) or removing majority class samples (*undersampling*). Oversampling technique can be done simply by randomly duplicating minority class samples or by artificially introducing new instances of minority class as it is done in quite popular SMOTE algorithm [7]. Other group of methods that are used for dealing with such problem are *algorithm level methods* which refers to the modification of base classifier in order to make it more sensitive to the imbalanced datasets [3]. Finally, *ensemble methods* involve forming a pool of classifiers and may combine its learners either both preprocessing or algorithm level methods [22].

3 On the role of misclassification cost in data oversampling

In this paper we aim to investigate if there is a connection between the performance of oversampling methods and the underlying cost associated with a given problem. As we mentioned in the previous section, sampling and cost-sensitive methods have been considered as separate approaches [20]. We propose to change this way of thinking and initiate a discussion on cost-sensitive sampling for imbalanced data. This section will focus on two core challenges in this new area: (i) how to tune oversampling methods when cost is involved; (ii) how to properly evaluate classifiers when cost is involved.

3.1 Cost-sensitive oversampling

Oversampling is one of the most efficient approaches for handling skewed data distributions, as new artificial instances are being introduced into the minority class. Regardless of the fact if a simple random oversampling or guided sampling algorithms are used, the number of introduced instances remains as an ad-hoc parameter. There are no clear rules on how to select (sub)optimal oversampling ratio, despite a crucial role of this factor [17]. Oversampling should be seen as a trade-off approach. Too small number of artificial instances will fail to adjust the class distributions properly, while too high number may lead to minority class shift and negatively impact the performance on the majority class.

It seems interesting to investigate if having an access to the cost associated with misclassification of minority instances would lead to a better control over the artificial instance generation procedure. While all data-level algorithms ignore the cost, even if it is provided by a domain expert, one may see that this leads to simply discarding useful information about the problem.

Cost may be associated to a degree in which the minority class is important for the considered problem. Higher costs of misclassification should force the

classification system to concentrate more on the minority class, even if it comes at the cost of impairing performance on the majority class. On the other hand, low misclassification cost should direct the classification system towards achieving a balanced performance on both of classes.

We propose to analyze if there is a relationship between the provided misclassification cost and performance of oversampling methods, with special emphasis put on the number of generated instances. Our hypothesis is that problems characterized by a higher cost would benefit from increasing the oversampling ratio. At the same time, for problems with a low misclassification cost the role of oversampling ratio should not be that significant. If our hypothesis is verified, then it would lead to a development of new branch of hybrid algorithms for imbalanced data that are cost-sensitive, while working on data-level.

3.2 Cost-sensitive evaluation of algorithms

Another issue related with existing cost-sensitive approaches lies in their evaluation [11]. The cost parameter is usually taken into an account during classifier training phase. During the testing phase, most of works in the literature use one of many skew-insensitive metrics, such as G-mean or F-measure [19]. While this is a proper approach from the class imbalance point of view, it neglects completely the presence of the cost parameter, as all of skew-insensitive measures assume 0-1 loss function.

Such an experimental framework is therefore flawed, as misclassification cost, if known for a given problem, should be considered during all steps of learning and evaluation. Furthermore, by neglecting the role of cost, one puts cost-sensitive methods in a disadvantaged position. There were only few efforts in the literature to propose evaluation metrics tackled specifically for cost-sensitive problems [16, 10], however they do not explicitly take into an account imbalanced data distributions. Additionally, as for imbalanced data there is already a plethora of established metrics proposed [2], it would be more interesting in adapting these metrics to cost-sensitive data, rather than adding more metrics to the stack.

In this paper, we formulate a hypothesis that misclassification cost, if known, should be taken into account during evaluation for all types of algorithms. Such an analysis would allow to gain a deeper insight into the performance of popular data- and algorithm-level solutions, as well as formulate a more realistic evaluation framework.

For the mentioned investigation of relationship between the misclassification cost and oversampling ratio, we will adopt cost-sensitive modifications of existing metrics. This will allow for a fair evaluation of the role of cost-sensitive learning in imbalanced data oversampling.

4 Experimental study

This experimental study was designed in order to answer the following research questions:

- Is there any relationship between the provided misclassification cost and performance of oversampling algorithms, with a special emphasis put on the oversampling ratio that returns the best performance.
- Is it worthwhile to use cost-sensitive modifications of popular skew-insensitive evaluation metrics and does such an evaluation leads to gaining an additional insight into evaluated algorithms.

For experimental purposes, a number of diverse benchmark datasets were selected from the public KEEL Imbalanced Data repository. Datasets related to the two-class problem were already prepared for the 5-Fold Cross Validation and selected with specific Imbalance Ratio (IR) in mind as shown in Section 4.1. Algorithms used for the evaluation purpose, as well as their implementations are covered in Section 4.2, where detailed information about evaluation methodology and metrics can be seen in Section 4.3.

4.1 Datasets

Selected datasets that were used for the experiment are shown in Table 1, sorted by the value of Imbalance Ratio. Each dataset is described by the Imbalance Ratio, number of features, instances as well by the amount of majority and minority samples.

Table 1: Selected datasets for evaluation

Dataset	IR	Feat.	Inst.	Maj.	Min.
glass1	1.82	10	214	138	76
wisconsin	1.86	10	683	444	239
pima	1.87	9	768	500	268
haberman	2.78	4	306	225	81
vehicle2	2.88	19	846	628	218
vehicle1	2.90	19	846	629	217
vehicle3	2.99	19	846	634	212
glass-0-1-2-3_vs_4-5-6	3.20	10	214	163	51
vehicle0	3.25	19	846	647	199
new-thyroid1	5.14	6	215	180	35
segment0	6.02	20	2308	1979	329
glass6	6.38	10	214	185	29
vowel0	9.98	14	988	898	90
cleveland-0_vs_4	12.31	14	173	160	13
abalone9-18	16.40	9	731	689	42
glass5	22.78	10	214	205	9
lymphography-normal-fibrosis	23.67	19	148	142	6
winequality-red-4	29.17	12	1599	1546	53
winequality-white-3_vs_7	44.00	12	900	880	20
kddcup-buffer_overflow_vs_back	73.43	42	2233	2203	30

4.2 Set-up

For experimental purposes, a framework written in *R language* with parts of code related to the *k-Nearest Neighbors* search written in *C++11* was introduced. In order to fairly assess performance of proposed solution, *5-Fold Cross Validation* (5-CV) was done on selected datasets. As a base classifier, C5.0 decision tree was used from the *C50* package. Experiment depends on two implemented oversampling techniques, more precisely Random Oversampling as well as SMOTE which allows to emphasize minority class either by duplicating instances or artificially introducing new samples respectively. Implemented SMOTE technique was used with Euclidean metric with parameter $k = 5$ which corresponds to the amount of neighbors taken into account in the neighborhood of computed instance.

4.3 Cost sensitive metrics

The basic metrics for the classifier evaluation for binary imbalanced datasets are true positive (*TP*), true negative (*TN*), false positive (*FP*) and false negative (*FN*) which can be deducted from the confusion matrix built from the predictions and reference labeling of test subset. However, aggregated measures are needed in order to compare different classifiers with or without preprocessing methods applied. For our experimental study, we will use the following ones with cost sensitivity taken into account which is applied to the false negative (*FN*) as shown in Eqn. 1. Cost sensitivity depends on the cost value provided to such metric, which varies in range $cost \in \{1, 2, 8, 16, 32, 64\}$ as shown in the results of experiment done in Section 4.4.

$$FN_{cost} = FN * cost \quad (1)$$

Information about proper classification of minority class can be obtained by *Sensitivity* metric also known as *Recall* or *True Positive Rate*, shown in Eqn. 2.

$$Sensitivity_{cost} = \frac{TP}{TP + FN_{cost}} \quad (2)$$

As the above metric takes only one class into consideration, Geometric Mean shown in Eqn. 3 is used as it balances between classification accuracy over the instances from both minority and majority classes at the same time.

$$GM_{cost} = \sqrt{\frac{TP}{TP + FN_{cost}} * \frac{TN}{FP + TN}} \quad (3)$$

F-Measure shown in Eqn. 4 can be considered as a harmonic mean of both precision and sensitivity which can measure accuracy of the test.

$$FMeasure_{cost} = \frac{2 * TP}{2 * TP + FP + FN_{cost}} \quad (4)$$

Balanced Accuracy shown in Eqn. 5 is a metric that was used for performance evaluation and can be described as an average accuracy received from both minority and majority class.

$$BAccuracy_{cost} = \frac{1}{2} \left(\frac{TP}{TP + FP} + \frac{TN}{TN + FN_{cost}} \right) \quad (5)$$

4.4 Results and discussion

Results for both, Random Oversampling and SMOTE preprocessing methods are shown in Figures 1 - 4. For each metric, averaged results on all datasets from the Section 4.1 are shown with different *Cost* as well as the *Oversampling percentage* which refers to the amount of minority samples to be introduced either by simply duplicating or artificially creating new one, relative to the reference amount of minority instances.

Presented figures should be analyzed from two levels. The individual analysis should focus on the impact of varying oversampling ratios on the performance of evaluated methods under a pre-set cost. The global analysis should focus on capturing the trends in performance related to increasing cost value and how does this affect the stability of oversampling methods.

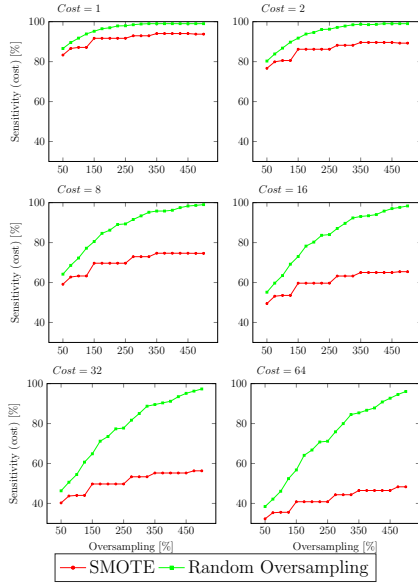


Fig. 1: Cost-sensitive Sensitivity.

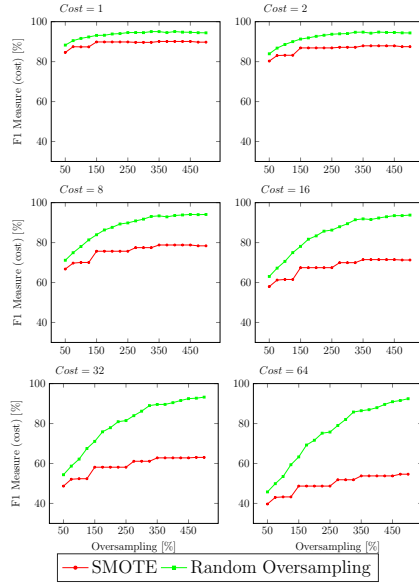


Fig. 3: Cost-sensitive F1-Measure.

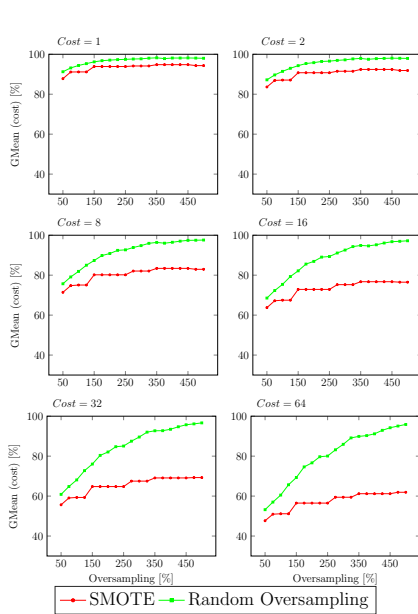


Fig. 2: Cost-sensitive G-Mean.

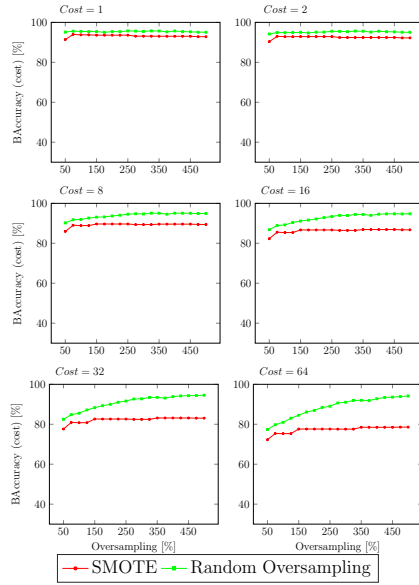


Fig. 4: Cost-sensitive Balanced Accuracy.

The obtained results allow us to draw a number of interesting conclusions. The most important one is the fact that there is a clear correlation between the cost and oversampling ratios. Regardless of the metric chosen, one can observe that for higher costs an increased oversampling ratio is preferred. When high values of cost are used (e.g., cost = 64) a high number of instances needs to be introduced in order to maximize the performance. On the other hand, for low cost values a good performance of oversampling methods is achieved even with < 100% oversampling ratio. When cost is not taken into account (i.e., cost = 1), all oversampling methods display similar performance regardless of the number of instances introduced. These observations prove our hypothesis that the underlying cost has a crucial impact on the performance of data-level solutions. It allows to better tune the balancing process and as we can see from the trends associated with the increasing cost, it is also beneficial for avoiding pitfalls related to introducing incorrect number of instances, such as data shift or increased computational complexity of the learning process. Therefore, we may conclude that cost-sensitive imbalanced data preprocessing is a direction worth pursuing.

When comparing random oversampling and SMOTE, one can see that they display different performance when combined with cost-sensitive information. SMOTE, while still strongly affected by cost values, stabilizes its performance with a lower values of oversampling ratios. This was to be expected, as SMOTE aims at introducing more meaningful instances than randomized approaches. Random oversampling is much more sensitive to cost and benefits from much higher oversampling ratios. However, especially for high cost parameter values, random oversampling easily outperforms SMOTE. This is an interesting observation, as one would expect SMOTE to be superior. It seems that by combining high misclassification costs with high oversampling ratios, random oversampling is capable for better empowering the minority class regions, thus translating to alleviated classification bias. This shows that each data-level method should be analyzed individually, in order to learn how it copes with cost-sensitive paradigm.

Finally, the results prove the usefulness of cost-sensitive metrics for gaining an insight into the nature of class imbalance learning algorithms. When no cost is taken into account (i.e., cost = 1), one cannot see significant differences between SMOTE and random oversampling. By scaling our metrics with cost value, the differences in performance between these two methods become obvious. We hope that this evaluation framework for any imbalanced algorithms will lead to better understanding which algorithms succeed and which fall under varying conditions.

5 Conclusions

In this paper, we proposed a new approach for looking at imbalanced data oversampling from a cost-sensitive perspective. We stated that when the misclassification associated with a given dataset is known, then it is beneficial to take it into an account when introducing new artificial instances to balance class dis-

tributions. Additionally, we pointed out the fact that in most works related to class imbalance the cost parameter is taken into account only during the learning phase, not during the testing phase. We argued that such an approach is incorrect, as one cannot neglect the role of associated cost when evaluating learning algorithms. Therefore, we have proposed to use cost-sensitive modifications of popular skew-insensitive metrics in scenarios where value of the cost parameter is known.

Our experimental study revealed a clear correlation between the value of cost parameter and the oversampling ratio. Higher costs, when used with cost-sensitive measures, favored higher number of artificial instances being introduced. For lower costs, the higher oversampling ratios did not contribute to the improvement of predictive power. This showed that cost-sensitive approaches may be used to tune and guide the oversampling, by allowing a more precise and automatic adaptation to a given imbalanced problem.

Obtained results encourage us to continue works in the new direction of cost-sensitive data-level solutions to class imbalance. Our next steps will be to propose an automatic way for embedding cost into oversampling methods in order to tune their parameters, and to evaluate this approach for multi-class imbalanced data scenarios.

Acknowledgement

This work was supported by the Polish National Science Centre under the grant No. 2017/27/B/ST6/01325 as well as by the statutory funds of the Department of Systems and Computer Networks, Faculty of Electronics, Wrocław University of Science and Technology.

References

1. Bernard, S., Chatelain, C., Adam, S., Sabourin, R.: The multiclass ROC front method for cost-sensitive classification. *Pattern Recognition* **52**, 46–60 (2016)
2. Branco, P., Torgo, L., Ribeiro, R.P.: A survey of predictive modeling on imbalanced domains. *ACM Comput. Surv.* **49**(2), 31:1–31:50 (2016)
3. Cano, A., Zafra, A., Ventura, S.: Weighted data gravitation classification for standard and imbalanced data. *IEEE Trans. Cybernetics* **43**(6), 1672–1687 (2013)
4. Cao, P., Zhao, D., Zaïane, O.R.: An optimized cost-sensitive SVM for imbalanced data learning. In: *Advances in Knowledge Discovery and Data Mining, 17th Pacific-Asia Conference, PAKDD 2013, Gold Coast, Australia, April 14-17, 2013, Proceedings, Part II*. pp. 280–292 (2013)
5. Castro, C.L., de Pádua Braga, A.: Novel cost-sensitive approach to improve the multilayer perceptron performance on imbalanced data. *IEEE Trans. Neural Netw. Learning Syst.* **24**(6), 888–899 (2013)
6. Charte, F., Rivera, A.J., del Jesús, M.J., Herrera, F.: Addressing imbalance in multilabel classification: Measures and random resampling algorithms. *Neurocomputing* **163**, 3–16 (2015)

7. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research* **16**(16), 321 – 357 (2002)
8. Domingos, P.M.: Metacost: A general method for making classifiers cost-sensitive. In: *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Diego, CA, USA, August 15-18, 1999. pp. 155–164 (1999)
9. Ducange, P., Lazzarini, B., Marcelloni, F.: Multi-objective genetic fuzzy classifiers for imbalanced and cost-sensitive datasets. *Soft Comput.* **14**(7), 713–728 (2010)
10. George, N.I., Lu, T., Chang, C.: Cost-sensitive performance metric for comparing multiple ordinal classifiers. *Artif. Intell. Research* **5**(1), 135–143 (2016)
11. Holte, R.C., Drummond, C.: Cost-sensitive classifier evaluation using cost curves. In: *Advances in Knowledge Discovery and Data Mining, 12th Pacific-Asia Conference, PAKDD 2008, Osaka, Japan, May 20-23, 2008 Proceedings*. pp. 26–29 (2008)
12. Krawczyk, B.: Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence* **5**(4), 221–232 (2016)
13. Ksieniewicz, P., Woźniak, M.: Dealing with the task of imbalanced, multidimensional data classification using ensembles of exposers. In: *First International Workshop on Learning with Imbalanced Domains: Theory and Applications, LIDTA@PKDD/ECML 2017, 22 September 2017, Skopje, Macedonia*. pp. 164–175 (2017)
14. López, V., Fernández, A., Moreno-Torres, J.G., Herrera, F.: Analysis of preprocessing vs. cost-sensitive learning for imbalanced classification. open problems on intrinsic data characteristics. *Expert Syst. Appl.* **39**(7), 6585–6608 (2012)
15. López, V., del Río, S., Benítez, J.M., Herrera, F.: Cost-sensitive linguistic fuzzy rule based classification systems under the mapreduce framework for imbalanced big data. *Fuzzy Sets and Systems* **258**, 5–38 (2015)
16. McDonald, R.A.: The mean subjective utility score, a novel metric for cost-sensitive classifier evaluation. *Pattern Recognition Letters* **27**(13), 1472–1477 (2006)
17. del Río, S., Benítez, J.M., Herrera, F.: Analysis of data preprocessing increasing the oversampling ratio for extremely imbalanced big data classification. In: *2015 IEEE TrustCom/BigDataSE/ISPA, Helsinki, Finland, August 20-22, 2015, Volume 2*. pp. 180–185 (2015)
18. Skryjomski, P., Krawczyk, B.: Influence of minority class instance types on SMOTE imbalanced data oversampling. In: *First International Workshop on Learning with Imbalanced Domains: Theory and Applications, LIDTA@PKDD/ECML 2017, 22 September 2017, Skopje, Macedonia*. pp. 7–21 (2017)
19. Thai-Nghe, N., Gantner, Z., Schmidt-Thieme, L.: Cost-sensitive learning methods for imbalanced data. In: *International Joint Conference on Neural Networks, IJCNN 2010, Barcelona, Spain, 18-23 July, 2010*. pp. 1–8 (2010)
20. Wang, S., Li, Z., Chao, W., Cao, Q.: Applying adaptive over-sampling technique based on data density and cost-sensitive SVM to imbalanced learning. In: *The 2012 International Joint Conference on Neural Networks (IJCNN), Brisbane, Australia, June 10-15, 2012*. pp. 1–8 (2012)
21. Wang, S., Minku, L.L., Yao, X.: Resampling-based ensemble methods for online class imbalance learning. *IEEE Trans. Knowl. Data Eng.* **27**(5), 1356–1368 (2015)
22. Woźniak, M., Graña, M., Corchado, E.: A survey of multiple classifier systems as hybrid systems. *Information Fusion* **16**, 3–17 (2014)
23. Zhao, H.: Instance weighting versus threshold adjusting for cost-sensitive classification. *Knowl. Inf. Syst.* **15**(3), 321–334 (2008)