Original article
https://doi.org/10.21683/1729-2646-2021-21-3-54-64

Intelligent methods for improving the accuracy of prediction of rare hazardous events in railway transportation

Olga B. Pronevich[1*], Mikhail V. Zaytsev[1]
[1]*JSC NIIAS, Moscow, Russian Federation*
*****obpronevich@gmail.com*


Olga B. Pronevich


Mikhail V. Zaytsev

Abstract. *The paper* **Aims** *to examine various approaches to the ways of improving the quality of predictions and classification of unbalanced data that allow improving the accuracy of rare event classification. When predicting the onset of rare events using machine learning techniques, researchers face the problem of inconsistency between the quality of trained models and their actual ability to correctly predict the occurrence of a rare event. The paper examines model training under unbalanced initial data. The subject of research is the information on incidents and hazardous events at railway power supply facilities. The problem of unbalanced data is expressed in the noticeable imbalance between the types of observed events, i.e., the numbers of instances.* **Methods.** *While handling unbalanced data, depending on the nature of the problem at hand, the quality and size of the initial data, various Data Science-based techniques of improving the quality of classification models and prediction are used. Some of those methods are focused on attributes and parameters of classification models. Those include FAST, CFS, fuzzy classifiers, GridSearchCV, etc. Another group of methods is oriented towards generating representative subsets out of initial datasets, i.e., samples. Data sampling techniques allow examining the effect of class proportions on the quality of machine learning. In particular, in this paper, the NearMiss method is considered in detail.* **Results.** *The problem of class imbalance in respect to the analysis of the number of incidents at railway facilities has existed since 2015. Despite the decreasing share of hazardous events at railway power supply facilities in the three years since 2018, an increase in the number of such events cannot be ruled out. Monthly statistics of hazardous event distribution exhibit no trend for declines and peaks. In this context, the optimal period of observation of the number of incidents and hazardous events is a month. A visualization of the class ratio has shown the absence of a clear boundary between the members of the majority class (incidents) and those of the minority class (hazardous events). The class ratio was studied in two and three dimensions, in actual values and using the method of main components. Such "proximity" of classes is one of the causes of wrong predictions. In this paper, the authors analysed past research of the ways of improving the quality of machine learning based on unbalanced data. The terms that describe the degree of class imbalances have been defined and clarified. The strengths and weaknesses of 50 various methods of handling such data were studied and set forth. Out of the set of methods of handling the numbers of class members as part of the classification (prediction of the occurrence) of rare hazardous events in railway transportation, the NearMiss method was chosen. It allows experimenting with the ratios and methods of selecting class members. As the results of a series of experiments, the accuracy of rare hazardous event classification was improved from 0 to 70-90%.*

Key words: *machine learning, rare events, class imbalance, better accuracy of predictions, data sampling, class balancing.*

For citation: *Pronevich O.B., Zaytsev M.V. Intelligent methods for improving the accuracy of prediction of rare hazardous events in railway transportation. Dependability 2021;3: 54-64. https://doi.org/10.21683/1729-2646-2021-21-3-54-64*

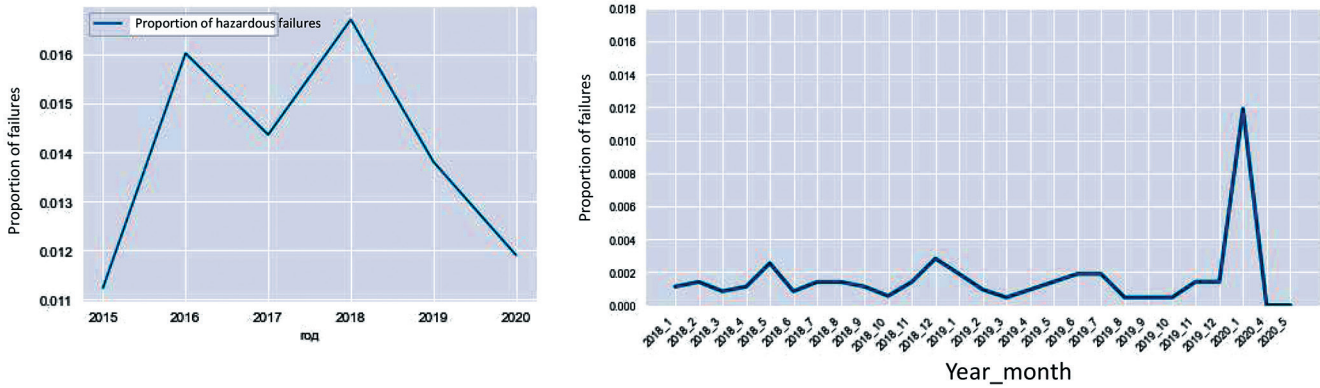
Received on: *05.07.2021* / **Revised on:** *02.08.2021* / **For printing:** *17.09.2021.*

1. The relevance of the problem of improving the quality of simulation as part of predicting rare events in railway transportation

When predicting hazardous events, researchers face a controversial problem: the fewer are the events, the better it is for the observed object, and the harder it is to train a model of the required quality. Currently, the proportion of hazardous events out of all incidents involving railway power supply facilities (RPSF) does not exceed 2% per year (Fig. 1a).

As it can be seen from the graph in Fig. 1a, since 2018, the proportion of hazardous events has been on a decline, yet the available observation period is not long enough to conclude on a steady reduction, while the data up to 2018 suggest the possibility of growing numbers of failures. Detailed monthly statistics (Fig. 1b) show the numbers of hazardous events peaking. The lack of clear seasonal patterns in the data does not allow scheduling preventive measures aimed at preventing hazardous events. A year is not an efficient observation period, as the condition of railway facilities changes greatly over a year and the planned activities may prove to be irrelevant. Therefore, of special interest are models that allow predicting hazardous events within periods of one month. At this level of observation detail, the matter of the "rarity" of the target factor becomes even more critical. On average, it accounts for less than 0.4% of cases.

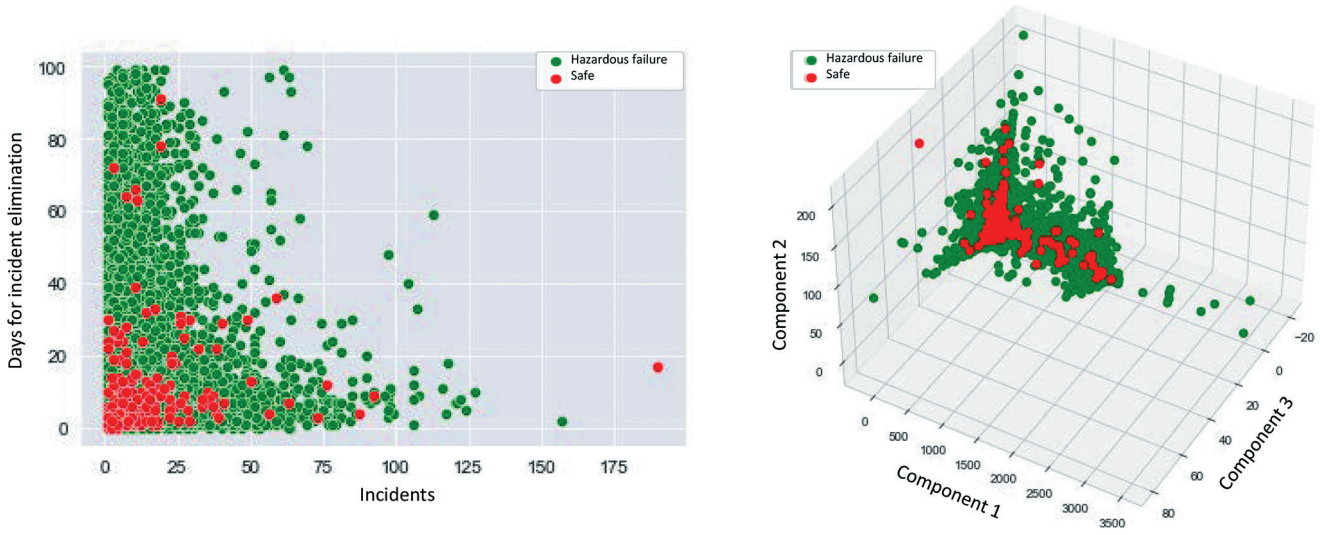
Despite the small number of hazardous events, the number of other unrelated incidents is in the thousands, which allows employing Big Data techniques. There are two main obstacles to building highly accurate models, i.e., class imbalance and data quality. In terms of data quality, the most common issues include incompleteness, duplication, inconsistent information, manual input errors. However, when working with railway facilities, the authors came to face another problem, i.e., the lack of distinct differences between the characteristics of facilities. On the monthly level, RPSF are mainly characterized by incident data. Upon data preparation, each RPSF is characterized by more than 150 features.



a) proportion of hazardous events year to year

b) proportion of hazardous events month to month

Fig. 1. The proportion of hazardous events out of incidents involving RPSF



a) Class imbalance in a plane

b) Class imbalance in 3 dimensions

Fig. 2. Class imbalance at similar facilities

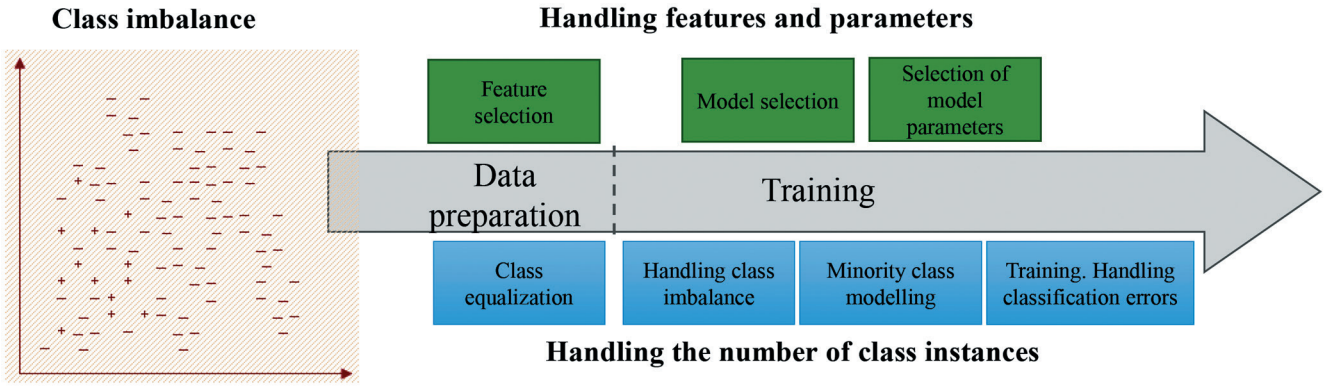


Fig. 3. Ways of improving the quality of classification

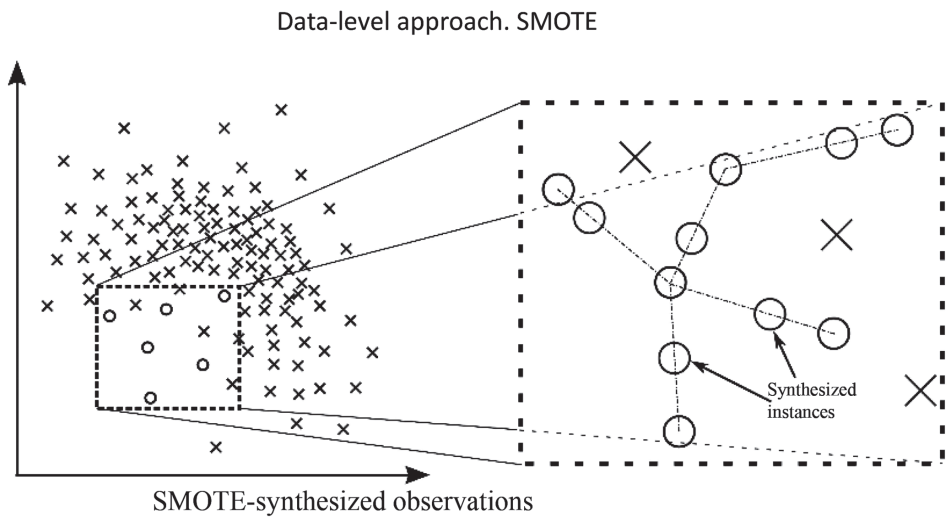


Fig. 4. Synthesized observation using *SMOTE*

Fig. 2a shows the ratio of classes on a plane with two coordinates, the number of incidents and the average number of days it takes to eliminate such incidents. It can be seen that there is no class boundary between the two features. Upon reducing the dimension using the dominant component analysis, let us move from a hundred of features to three (Fig. 2b). The graph better shows the concentration of hazardous events, yet there is still no clear boundary between classes.

Thus, improving the quality of incident classification is to involve handling class imbalance and selection of features. Otherwise, the quality of RPSF condition classification models will be unsatisfactory. (Table 1).

Table 1. Quality of classification models prior to the use of accuracy improvement techniques

Model	Accuracy of hazardous event prediction		Accuracy of safe state prediction	
	Training sample	Test sample	Training sample	Test sample
GBC	0	0.416	1	1
Logi	0	0	1	1
KNN	0	0	1	1
DTC	0	0.043	0.994	1

2. Ways of improving the accuracy of rare event prediction

The methods for improving the quality of classification can be divided into two groups: those based on features and parameters and those based on the number of class instances. Those methods can be used both at the stage of data preparation, and at the stage of training (Fig. 3).

The feature-based group includes:

1) methods of significant feature selection;
2) model selection and setting.

The methods based on the number of class instances include:

1) class equalization;
2) handling class imbalance;
3) minority class modelling;
4) selection of effective penalty function (handling classification error).

Although those methods are now widely known, there is no single algorithm of their combined application for the purpose of improved classification. The specificity and scope of data available in each particular situation force the researchers to experiment in search of an optimal combination of various methods. For instance, [1] proposed a technique that involves using a combination of classification

Table 2. Overview of the methods of handling class imbalance

Solution	Essential description of a group of methods	Advantage	Disadvantage
Data level approach			
MLSMOTE [7]	Pre-training stage: balancing by means of either undersampling, or oversampling in order to reduce the imbalance factor in the training data	Direct approach and wide application	Risk of overtraining
Diversified Sensitivity-based Undersampling (*DSUS*) [8]			
Ant Colony Optimization (*ACO)* [9]			
SMOTE [10]			
Evolutionary undersampling [11]			
Value (importance) boosting			
Cost-sensitive linguistic fuzzy rule [12]	Cost items (indicate the importance of class identification) designate the uneven importance of identification among classes. Increasing strategy may intentionally shift the training towards the classes associated with greater identification importance and ultimately improve the identification performance	A straightforward method, especially if the cost of error is known	Additional training costs due to finding an efficient cost matrix, especially when the real cost of error is unknown
Increasing cost sensitivity			
Feature selection			
Selection of minority class characteristics	Methods for selecting the features for the training	Helps solve class overlapping	Additional computational costs due to the requirement of data pre-processing
Density-based entity selection			
roc-based *FAST* generation of observations			
Correlation-based *CFS* generation of observations			
Algorithm-level approach			
Argument-based rule learning [14]	Specialized algorithms that specifically examine the distribution of class imbalance within datasets	Efficiency through modified algorithms for training solely based on the distribution of imbalance classes	It may be required to do pre-processing in order to balance-out uneven class distribution
Difference-based training [15]			
Fuzzy classifier [16]			
z-SVM [17]			
Hierarchical fuzzy rule [18]			
Distribution of conditional nearest class neighbour [19]			
k-NN sample generalization [20]			
Weighted nearest neighbour classifier [21]			
One-class training			
One-class training [22]	Classifier modelling on the minority class representation	Ease of use	Inefficient when used along with classification algorithms that are to be trained on the prevailing class
Class conditional nearest neighbour distribution (*CCNND*) [19]			
Economic training			
Bayesian SVM classifier [23]	A group of classification methods based on the cost of misclassification of both a false positive and a miss	A simple and quick processing technique	Inefficient if the actual cost is not available. Additional costs are introduced if cost investigation is required when the cost of error is unknown
Cost-sensitive *SVM* training [24]			
Cost-sensitive *NN* with *PSO* [25]			
SVM for adaptively asymmetric misclassification of cost [26]			
Ensemble-based method			
SMOTE and feature selection ensemble [27]	A group of methods based on the use of multiple classifiers that are trained directly on data and integration of the estimates for the purpose of developing a final classification solution	Multifaceted approaches	Complexity rises as the number of classifiers increases. Diversity is hard to achieve
GA ensembles [28]			
Ensembles for financial problems [29]			
Boosting in *SVM* ensemble [30]			
RUSBoost [31]			
SMOTEBoost [32]			

Solution	Essential description of a group of methods	Advantage	Disadvantage
Hybrid approach			
FTM-SVM	More than one machine learning algorithm is used in order to improve classification quality, often by combining them with other training algorithms for better results. Hybridization is used in order to simplify the sampling, selection of feature subset, optimization of the cost matrix and fine-tuning of the classical training algorithms	Symbiosis training through combination with other training algorithms is gaining popularity in class imbalance classification	A thorough project evaluation is required in order to take into account the differences between the used methods
F measure-based training [33]			
Linguistic fuzzy rule [34]			
Fuzzy classifier electronic algorithm [35]			
GA-based fuzzy rule extraction [36]			
Neuro-fuzzy [37]			
Neural network medical data [38]			
SMOTE neural networks [39]			
kNN classifier for medical data [40]			
NN trained with *BP* and *PSO* for medical data [41]			
Dependency tree kernels [42]			
Using cost sensitivity in trees [43]			
Undersampling and *GA* for *SVM* [44]			
Other methods			
ADASYN [45, 46]	Once the sample is created, it adds random small values to the points. In other words, instead of a linear correlation between the entire sample and the parent, they are a little more different	Adaptivity: the number of synthetic observations is based on the ratio of the majority to minority observations. *ADASYN* is focused on more complex data areas	The disadvantage of *ADASYN* is that it is easily affected by outliers
NearMiss [45, 47]	Random exclusion of majority class examples. When instances of two different classes are very close to each other, we remove the majority class instances in order to increase the space between the two classes. This helps the classification process	Can reduce sample overlapping between different classes	In case of undersampling, the number of insufficient samples cannot be controlled, while the mass samples that can be excluded are limited. Best used as a data cleansing method in combination with other methods
Edited nearest neighbours [45, 48]	Majority class instances are excluded if more than a half of their *K* neighbours do not belong to the majority class	Reduces misclassification of majority class instances	Cannot control the quantity of undersampling
Tomek Links [3, 45, 49]	Majority class instances are excluded that are in immediate proximity to minority class instances	All items that are immediate neighbours belong to the same class that can be better classified	
Neighbourhood cleaning rule [45]	This strategy also aims to remove those examples that negatively affect the outcome of minority class classification. For that purpose, all examples are classified according to the rule of the three nearest neighbours	Improves the accuracy of minority class instance classification	May result in a sample size that would not be sufficient for training
Condensed nearest neighbour [45, 50]	The method allows finding differences between similar examples that belong to different classes	Finding differences between similar examples that belong to different classes	

algorithms and the *RFE*, *Random Forest* and *Boruta* methods of feature selection with preliminary class balancing by means of *SMOTE* and *ADASYN* random sampling. The paper shows an increase in classification accuracy up to 98% (from 93%). However, study [2] dedicated to text classification shows that rather than trying to modify the distribution, it would be more efficient to work with decision threshold modification and the weights of errors of various kinds. Additionally, the author introduces a separation of unbalanced data into moderately unbalanced data (class ratio 7 to 1) and strongly unbalanced data (class ratio 14 to 1). Paper [2] showed that handling class ratios using strongly unbalanced data in case of some models causes improved classification accuracy. While classifying user requests, [3] identified that using class-balancing methods may not only fail to provide any results, but also cause reduced classification accuracy. Contrary to [3], [4] demonstrated the efficiency of data sampling methods.

The purpose of this paper is to practically demonstrate the effect of methods of handling class instances on the classification accuracy of hazardous events affecting RPSF.

2.1. Overview of the methods of handling class imbalance

One of the primary methods of handling class imbalance is *SMOTE*, whose algorithm was developed in early 2002 [5]. Currently, there are a number of modifications of this method. It also inspired the development of other algorithms of handling unbalanced data. *SMOTE* is at the top of the group of data sampling methods, i.e., those that involve increasing the number of minority class instances. Its basic principle is shown in Fig. 4.

Synthetic instances are generated in the "function space" rather than the "data space". Minority class samples are replenished by introducing synthetic examples along segments of the line that connect any/all of the nearest neighbours of the minority class k [5]. *SMOTE* is an oversampling method.

Another way of handling unbalanced data is to reduce the number of the majority class instances (undersampling or reindexing).

Combinations of various sampling strategies constitute hybrid methods that involve sequential application of oversampling and undersampling algorithms. A visualization of the processes and results of the above strategies is shown in Fig. 5.

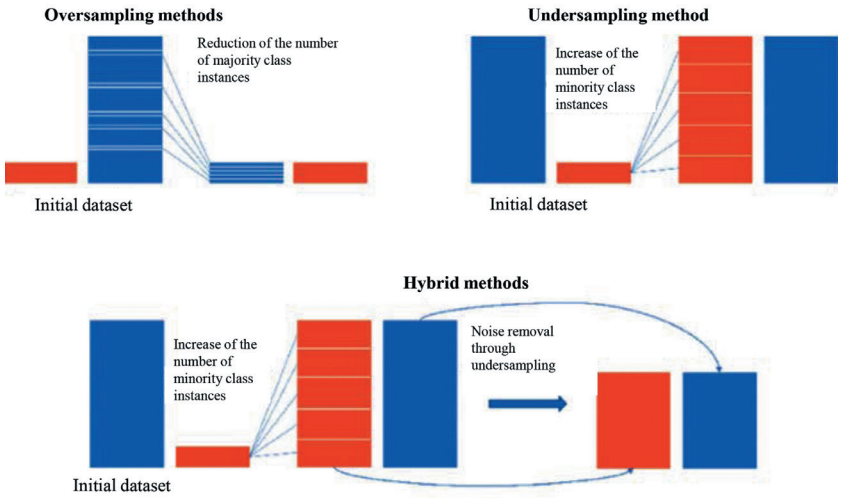
[6] made an overview of the methods of handling class imbalance. Table 2 shows information on them along with the currently-employed methods.

Based on the advantages and disadvantages of the methods examined in Table 2, as well as the experience of the researchers studying data sampling techniques [51, 52, 53, 54], *NearMiss* [47, 55] was chosen as the primary method of improving the classification of unbalanced data. The purpose of *NearMiss* is to balance the distribution of observations across minority and majority classes by estimating the distance between instances from different classes. The *NearMiss* implementation of the *imblearn Python* library (includes a set of tools for unbalanced datasets in machine learning) involves three strategies of class instance selection:

Strategy 1 (*version* = 1). Selection of observations out of the majority class, for which the average distance to k nearest observations out of the minority class is the smallest (by default $k = 3$, training variable). Only those observations without hazardous events will be retained that are the nearest to those with hazardous events;

Strategy 2 (*version* = 2). Selection of observations from the majority class, for which the average distance to k furthest observations of the minority class is the smallest (by default, $k = 3$, training variable). Only those observations without hazardous events will be retained that are at the centre of the mass of the intersection of sets of majority and minority classes;

Strategy 3 (*version* = 3). First, for each observation out of the majority class, M nearest neighbours will be retained (by default, $M = 3$, training variable). Then, observations out of the minority class are selected, for which the average



The operating principle of hybrid methods in respect to the problem of class balancing

Fig. 5. Data sampling strategies

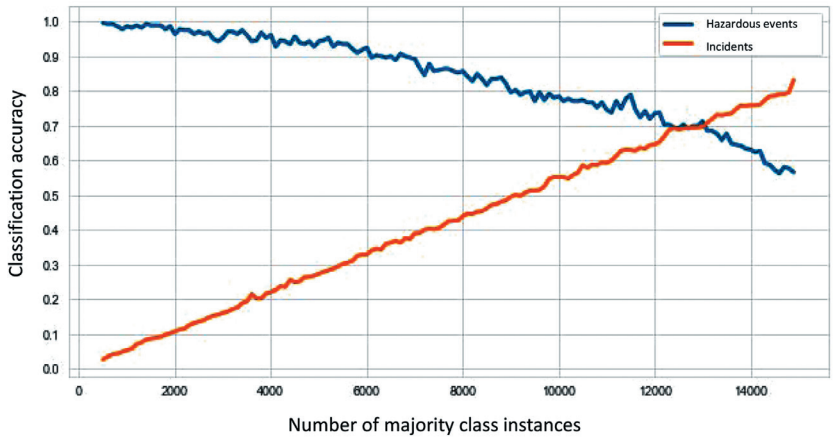


Fig. 6. Classification accuracy on main sample, *GBC*, *NearMiss* 2

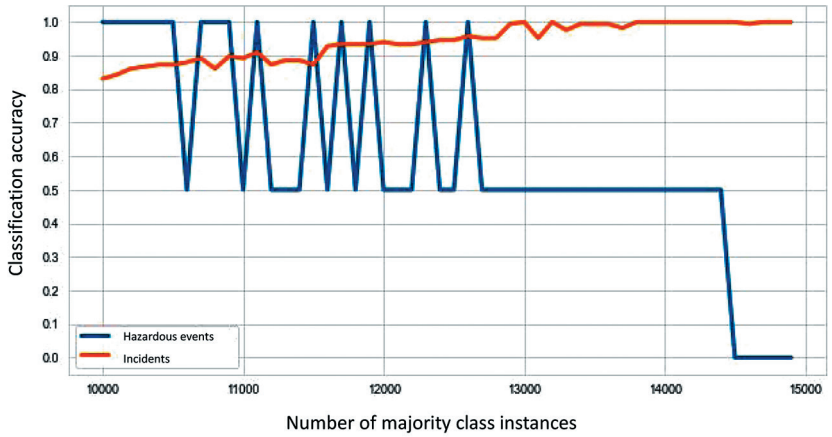


Fig. 7. Accuracy on validation sample, *GBC*, *NearMiss* 1

distance to *N* nearest neighbours is the longest (by default, *N* = 3, training variable).

Unlike in [1-4, 52, 53], the focus of our attention is not the accuracy or classification error indicator, but rather the detail of the hazardous event prediction accuracy and the incident prediction accuracy. Additionally, due to the fact that 98% of observations are in the majority class, it should be expected that the accuracy of most trained classifiers will be 98%. In this context, it is easy to conclude that class balancing is inefficient, as some researchers do.

In the course of the study, a combination of several methods of improving the quality of unbalanced data classification was used, namely *GridSearchCV* and *NearMiss*, as well as various quality functions.

The parameters that vary in the course of the experiment:

1) within the *GridSearchCV* function:

(1) GBS: random_state, tol, max_depth;

(2) Logi: tol, class_weight, max_iter, solver, random_state, C;

(3) KNN: n_neighbors, weights, metric;

(4) DTC: criterion, max_depth, min_samples_split, max_features;

2) quality functions: *max_error, balanced_accuracy,accuracy, neg_log_loss, explained_variance, neg_mean_squared_error, neg_mean_squared_log_error, neg_median_absolute_error, r2*;

3) NearMiss parameters: instance selection strategies, number of minority class instances, number of majority class instances.

Diagram of the experiment:

1) division of the sample into the master and the validation. The master sample includes observations collected before 2019, while the validation includes those collected after 2019;

2) division of the master sample into the training and test. The method is *train_test_split*, the test sample size is 20%.

The experiments have established that the best accuracy is achieved under the *neg_log_loss* quality function. Due to the extremely small number of minority class instances (less than 2%) the quality of this indicator's variation has no impact on the classification quality. In the course of most experiments the number of minority class instances was fixed and equalled the maximum possible.

Let us examine the experimental results. Fig. 6 shows the graph of accuracy of a GBC event prediction depending on the number of majority class instances on the NearMiss 2 master sample.

The initial accuracy characteristics (without NearMiss) are: accuracy on the test part of the training sample: 0.983; accuracy of hazardous event prediction: 0.416; accuracy of non-hazardous incident prediction: 1. As can be seen in Fig. 6, there is a point where the accuracy graphs of hazard-

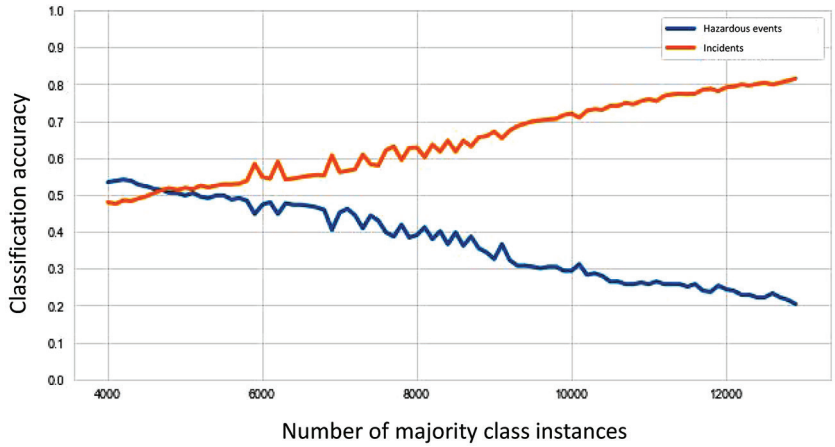


Fig. 8. Accuracy on main sample, *Logi*, *NearMiss* 1

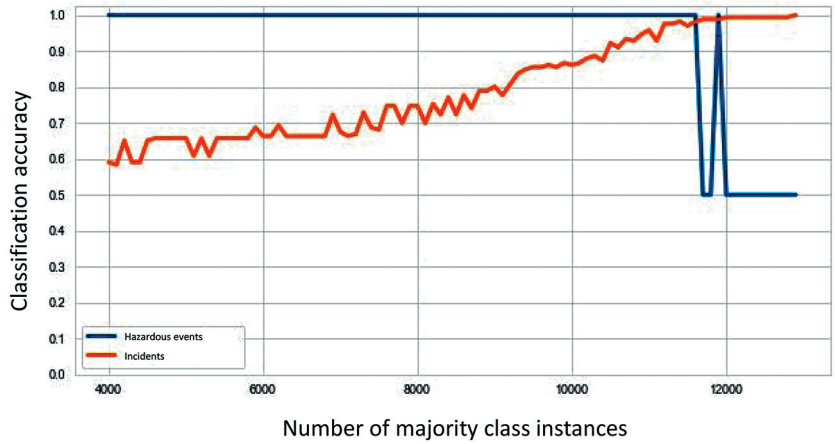


Fig. 9. Accuracy on validation sample, *Logi*, *NearMiss* 1.

ous events and incidents classification intersect. Essentially, the higher this point is, the more accurate the final model will be. Similar graphs were obtained for the *NearMiss* 1 sampling strategy. An accuracy of 70% appears to be a good result in case of the starting accuracy of hazardous event classification of 41.6%. The following experiment was conducted under selection strategy no. 2. For clarity, Fig. 7 only shows the intersection of the graphs that characterize the validation sample.

The accuracy of the classifier on the validation sample (v-sample) proved to be significantly higher than that on the master sample. According to the analysis, that was due to the size of the v-sample (the classifier "did have enough time" to make many mistakes), as well as the months covered by the prediction. The accuracy over periods similar to those of the v-sample in the master sample proved to be higher than average.

Before proceeding to the analysis of other data, let us revisit the topic of using the classifier's accuracy indicator as an efficiency characteristic. The GBC classifier accuracy on unbalanced data was 0.983. The accuracy of hazardous event prediction on the v-sample is 0. On balanced data the accuracy was 0.9811 (lower than on the original sample size), while the hazardous event prediction accuracy on the v-sample was 1.0, the accuracy of incident prediction was 0.934. Similar figures were obtained for other classifiers. Thus, when studying the classification accuracy of unbalanced data, one cannot rely on one "convoluted" quality indicator.

Fig. 8 and 9 show the accuracy graphs of the Logi model for the master sample and the NearMiss 1 v-sample.

The graphs in Fig. 8 and 9 show that the best accuracy indicators for the master and v-sample are achieved under

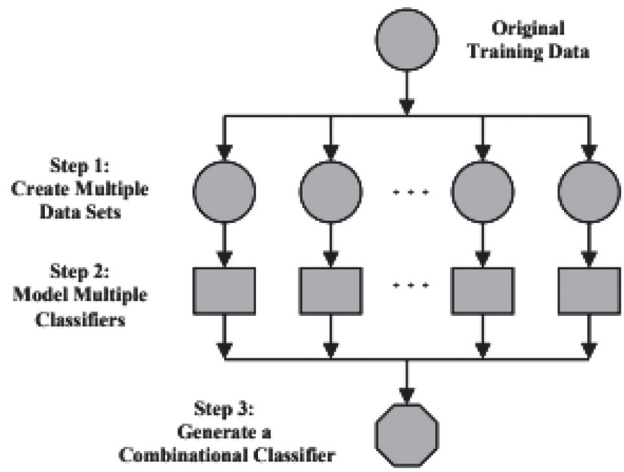


Fig. 10. General algorithm of ensemble methods

different class balances. That suggests that the models rejected on the basis of the master sample data may produce better results on new data. One of the ways of solution that problem is to use ensemble methods [56]. The key idea of studying ensembles of classifiers is to build several classifiers from the original data and then to aggregate the predictions when classifying unknown samples (Fig. 10). The application of such methods may be the subject matter of further research.

Analysis and conclusions

1. As safety systems evolve, the number of hazardous events decreases, yet the cost of potential consequences grows. Out of RPSF incidents, the proportion of hazardous events does not exceed 2% per year. For the purpose of improving the accuracy of classification and prediction of event types in case of class imbalance, sample balancing methods are to be used.

2. More than 50 methods are currently in active use that allow handling unbalanced samples. However, the publications known to the authors do not address the matter of analysing simultaneous changes in the prediction accuracy of minority and majority class instances. The paper presents graphs of classification accuracy of RPSF incidents on the training and validation sample.

3. A method of handling unbalanced data is proposed that includes a combination of several ways of improving the quality of unbalanced data classification: model parameter setting, selection of the indicator of quality and ratio of the minority to the majority class instance number using *NearMiss*.

4. The methods of dealing with class imbalance allows significantly increasing the accuracy of predicting minority class instances (hazardous events) from 0 to 70-90%. However, as the accuracy of prediction of rare events increases, the accuracy of prediction of minority class instances decreases.

5. The study may pave the way for the application of hybrid methods of classifying and predicting events, as well as for the development of a metric of training quality based on the characteristic of the intersection point of classification accuracy graphs of the instances of different classes.

References

1. Sevastianov L.A, Shetinin E.Yu. On methods for improving the accuracy of multiclass classification on imbalanced data. *Informatics and applications* 2020;14(1):63-70. (in Russ.)

2. Sadov M.A. Study of the methods of text classification for unbalanced data. *Polymathis Scientific Journal* 2016;2:28-41. (in Russ.)

3. Maslikhov S.R., Mokhov A.S., Tolcheev V.Yu. [Building balanced classes in respect to user query classification]. [Proceedings of the 5th International Science and Practice Conference Remote Education Technologies]; 2020. P. 245-248. (in Russ.)

4. Shipitsyn A.V., Zhuravleva N.V. Evaluation of online mortgage applications with machine learning algorithms. *Herald of the Belgorod University of Cooperation, Economics and Law* 2016;4(60):199-209. (in Russ.)

5. Chawla N.V., Bowyer W.B., Hall L.O. et al. Smart: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research* 2002;16:321-357.

6. Ali A., Shamsuddin S.M., Ralescu A. Classification with class impact problem: a review. *International Journal of Advances in Soft Computing* 2013;7:176-204.

7. Mladenic D., Grobelnik M. Feature selection for unbalanced class distribution and national scores. Proceedings of the Sixteenth International Conference on Machine Learning (ICML 1999). Bled (Slovenia); 1999. p. 258-267.

8. Yang T.-N., Wang S.-D. Robust algorithms for principal component analysis. *Pattern Recognition Letters* 1999;20(9):927-933.

9. Yu H., Ni J., Zhao J. ACOSampling: An ant colony optimization-based undersampling method for classifying imbalanced DNA microarray data. *Neurocomputing* 2013;101(0):309-318.

10. Chawla N.V. SMOTE: synthetic minority oversampling technique. arXiv:1106.1813; 2002.

11. García S., Herrera F. Evolutionary undersampling for classification with imbalanced datasets: Proposals and taxonomy. *Evolutionary Computation* 2009;17(3):275-306.

12. Yin L. Feature selection for high-dimensional imbalanced data. *Neurocomputing* 2013;105(0):3-11.

13. Sun Y. Cost-sensitive boosting for classification of imbalanced data. *Pattern Recognition* 2007;40(12):3358-3378.

14. Luukka P. Nonlinear fuzzy robust PCA algorithms and similarity classifier in bankruptcy analysis. *Expert Systems with Applications* 2010;37(12):8296-8302.

15. Zheng Z., Wu X., Srihari R. Feature selection for text categorization on imbalanced data. *ACM SIGKDD Explorations Newsletter* 2004;6(1):80-89.

16. Visa S., Ralescu A.L. Fuzzy Classifiers for Imbalanced Data Sets. University of Cincinnati, Computer Science Dept. Cincinnati (OH, United States); 2007.

17. Imam T., Ting K., Kamruzzaman J. z-SVM: An SVM for Improved Classification of Imbalanced Data. AI 2006: Advances in Artificial Intelligence. 19th Australian Joint Conference on Artificial Intelligence. Hobart (Australia); 2006. p. 264-273.

18. Fernández A., M.J. del Jesus, Herrera F. Hierarchical fuzzy rule based classification systems with genetic rule selection for imbalanced data-sets. *International Journal of Approximate Reasoning* 2009;50(3):561-577.

19. Kriminger E., Principe J.C., Lakshminarayan C. Nearest Neighbor Distributions for imbalanced classification. The 2012 international joint conference on neural networks (IJCNN). Brisbane (QLD, Australia): 2012. p. 1–5.

20. Li Y., Zhang X. Improving k nearest neighbor with exemplar generalization for imbalanced classification. In: Proceedings of Advances in knowledge discovery and data mining: 15th Pacific-Asia Conference, Part II. Shenzhen (China); 2011. p. 321-332.

21. Candès E.J. Robust principal component analysis. *Journal of the ACM (JACM)* 2011;58(3):11.

22. Japkowicz N., Myers C., Gluck M. A novelty detection approach to classification. *IJCAI* 1995;1:518–523.

23. Jolliffe I. Principal component analysis. Encyclopedia of Statistics in Behavioral Science. John Wiley & Sons, Ltd; 2005.

24. Cao P., Zhao D., Zaiane O. An Optimized Cost-Sensitive SVM for Imbalanced Data Learning. In: Proceedings of Advances in Knowledge Discovery and Data Mining: 17th Pacific-Asia Conference. Part II. Gold Coast (Australia); 2013. p. 280-292.

25. Cao P., Zhao D., Zaiane O. A PSO-based Cost-Sensitive Neural Network for Imbalanced Data Classification. In: Revised Selected Papers of Trends and Applications in Knowledge Discovery and Data Mining. International Workshops: DMApps, DANTH, QIMIE, BDM, CDA, CloudSD. Gold Coast (Australia); 2013. p. 452-463.

26. Wang X., Shao H, Japkowicz N, et al. Using SVM with Adaptive Asymmetric Miseducation Costs for Mine-like objects Detection. In: Proceedings of the 11th International Conference on Machine Learning and Applications. Boca Raton (Florida, USA); 2012. p. 78-82.

27. Yang P., Liu W, Zhou B.B. et al. Ensemble-based wrapper methods for feature selection and class imbalance learning. In: Proceedings of Advances in Knowledge Discovery and Data Mining: 17th Pacific-Asia Conference, Part I. Gold Coast (Australia); 2013. p. 544-555.

28. Yu E., Cho S. Ensemble based on GA wrapper feature selection. *Computers & Industrial Engineering* 2006;51(1):111-116.

29. Liao J.-J. An ensemble-based model for two-class imbalanced financial problem. *Economic Modelling* 2014;37(0):175-183.

30. Liu Y., An A., Huang X. Boosting prediction accuracy on imbalanced datasets with SVM ensembles. In: Proceedings of Advances in Knowledge Discovery and Data Mining, 10th Pacific-Asia Conference. Singapore; 2006. p. 107-118.

31. Seiffert C. RUSBoost: A hybrid approach to alleviating class imbalance. Systems, Man and Cybernetics. Part A: Systems and Humans. *IEEE Transactions* 2010;40(1):185-197.

32. Chawla N.V. SMOTEBoost: Improving prediction of the minority class in boosting. In: Proceedings of Knowledge Discovery in Databases: PKDD 2003, 7th European Conference on Principles and Practice of Knowledge Discovery in Databases. Cavtat-Dubrovnik (Croatia); 2003. p. 107-119.

33. Wasikowski M., Chen X.-W. Integrating the small sample class impact problem using feature selection. *Knowledge and Data Engineering. IEEE transactions* 2010;22(10): P.1388-1400.

34. Martino M.D. Novel Classifier Scheme for Unbalance Problems. *Pattern Recognition Letters* 2013;34(10):1146–1151.

35. Fernández A. A study of the behaviour of linguistic fuzzy rule based classification systems in the framework of imbalanced data-sets. *Fuzzy Sets and Systems* 2008;159(18):2378-2398.

36. Le X., Mo-Yuen C., Taylor L.S. Power Distribution Fault Cause Identification With Imbalanced Data Using the Data Mining-Based Fuzzy Classification E-Algorithm. *Power Systems. IEEE Transactions* 2007;22(1):164-171.

37. Soler V. Imbalanced Datasets Classification by Fuzzy Rule Extraction and Genetic Algorithms. Data Mining Workshops. ICDM Workshops. Sixth IEEE International Conference; 2006.

38. Hung C.-M., Huang Y.-M. Conflict-sensitivity contexture learning algorithm for mining interesting patterns using neuro-fuzzy network with decision rules. *Expert Systems with Applications* 2008;34(1);159-172.

39. Jeatrakul P., Wong K.W., Fung C.C. Classification of imbalanced data by combining the complementary neural network and SMOTE algorithm. Proceedings of the 17th International Conference on Neural Information Processing: Models and Applications. Part II. Sydney (Australia); Springer-Verlag. p. 152-159.

40. Malof J.M., Mazurowski M.A., Tourassi G.D. The effect of class imbalance on case selection for case-based classifiers: An empirical study in the context of medical decision support. *Neural Networks* 2012;25(0):141-145.

41. Mazurowski M.A. Training neural network classifiers for medical decision making: the effects of imbalanced datasets on classification performance. *Neural networks* 2008;21(2-3):427-436.

42. Culotta A., Sorensen J. Dependency tree kernels for relation extraction. In: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics; 2004.

43. Drummond C., Holte R.C. Exploiting the cost (in) sensitivity of decision tree splitting criteria. ICML; 2000.

44. Al-Shahib A., Breitling R., Gilbert D. Feature selection and the class imbalance problem in predicting protein function from sequence. *Applied Bioinformatics* 2005;4(3):195-203.

45. Koziarski M. Radial-Based Undersampling for imbalanced data classification. *Pattern Recognition* 2020;102.

46. He H., Bai Y., Garcia E.A. et al. ADASYN: adaptive synthetic sampling approach for imbalanced learning. In: Proceedings of IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence); 2008. p. 1322-1328.

47. Mani I., Zhang I. kNN approach to unbalanced data distributions: a case study involving information extraction. In: Proceedings of Workshop on Learning from Imbalanced Datasets. Vol. 126; 2003.

48. Wilson D.L. Asymptotic properties of nearest neighbor rules using edited data. *IEEE Trans. Syst. Man Cybern.* 1972;2(3): 408-421.

49. Tomek I. Two modifications of CNN. IEEE *Trans. Syst. Man Cybern.* 1976;6:769-772.

50. Hart P. The condensed nearest neighbor rule. *IEEE Trans. Inf. Theory* 1968;14(3):515-516.

51. Makhsotova Ts.V. [Study of classification methods in the case of class imbalance]. *Science Magazine* 2017;5(18). (accessed July 5, 2020). Available at: https://cyberleninka.ru/article/n/issledovanie-metodov-klassifikatsii-pri-nesbalansirovannosti-klassov. (in Russ.)

52. Kavrin D.A., Subbotin S.A. The methods for quantitative solving the class imbalance problem. *Radio Electronics, Computer Science, Control* 2018;1. (accessed July 6, 2020). Available at: https://cyberleninka.ru/article/n/metody-kolichestvennogo-resheniya-problemy-nesbalansirovannosti-klassov. (in Russ.)

53. Yi L., Hong G., Feldkamp L. Robust neural learning from unbalanced data samples. In: Proceedings of IEEE International Joint Conference on Neural Networks. IEEE World Congress on Computational Intelligence (Cat. No.98CH36227). Vol. 3. Anchorage (USA); 1998. p. 1816-1821.

54. Al-Stouhi S., Reddy C.K. Transfer learning for class imbalance problems with inadequate data. *Knowledge and Information Systems* 2016;48:201-228.

55. Near-Miss – version 0.9.0.dev0. API reference. (accessed July 10, 2021). Available at: https://imbalanced-learn.org/dev/references/generated/imblearn.under_sampling.NearMiss.html.

56. Sun Y. Cost-Sensitive Boosting for Classification of Imbalanced Data. *Pattern Recognition* 2007;40(12):3358-3378.

About the authors

Olga B. Pronevich, Head of Unit, JSC NIIAS, 27, bldg 1 Nizhegorodskaya St., Moscow, 109029, Russian Federation, phone: +7 (495) 786 68 57; e-mail: obpronevich@gmail.com.

Mikhail V. Zaytsev, Lead Specialist, JSC NIIAS, 27, bldg 1 Nizhegorodskaya St., Moscow, 109029, Russian Federation, phone: +7 (495) 786 68 57; e-mail: m.v.zaicev@mail.ru.

The authors' contribution

Pronevich O.B. investigated the problem of class imbalance when predicting the occurrence of hazardous events at railway power supply facilities, analysed the existing methods of unbalanced data processing, as well as conducted a series of experiments for evaluating the effect of different proportions of the instances of the minority and majority classes. The author also defined a vision for further research.

Zaytsev M.V. analysed widely used methods of event classification under class imbalance, studied the effect of various strategies of *NearMiss* selection on the quality of hazardous event classification in railway transportation.

Conflict of interests

The authors declare the absence of a conflict of interests.