Impact of Clustering on a Synthetic Instance Generation in Imbalanced Data Streams Classification

Ireneusz Czarnowski[1,[0000-0003-0867-3114]] and Denis Mayr Lima Martins[2,[0000-0002-8262-2369]]

[1] Department of Information Systems, Gdynia Maritime University
Morska 83, 81-225 Gdynia, Poland
i.czarnowski@umg.edu.pl

[2] Department of Information Systems, University of Münster, ERCIS
Leonardo-Campus 3, 48149 Münster, Germany
denis.martins@wwu.de

Abstract. The goal of the paper is to propose a new version of the Weighted Ensemble with one-class Classification and Over-sampling and Instance selection (WECOI) algorithm. This paper describes WECOI and presents the alternative approach for over-sampling, which is based on a selection of reference instances from produced clusters. This approach is flexible on applied clustering methods; however, the similarity-based clustering algorithm has been proposed as a core. For clustering, different methods may also be applied. The proposed approach has been validated experimentally using different clustering methods and shows how the clustering technique may influence synthetic instance generation and the performance of WECOI. The WECOI approach has also been compared with other algorithms dedicated to learning from imbalanced data streams. The computational experiment was carried out using several selected benchmark datasets. The computational experiment results are presented and discussed.

Keywords: classification, learning from data streams, imbalanced data, over-sampling, clustering.

1 Introduction

Data analysis is a key component of the effective decision-making processes as well as modern, innovative, and autonomous decision-making systems. Prediction is one of the main tasks of data analysis and is very often solved using machine learning tools. Prediction can also be considered through the prism of classification and regression. Considering in this paper a classification task, supervised machine learning tools train models using labelled data instances, finally obtaining a classifier that is able to work with unlabelled data and unseen data. After that, the task of the trained classifier is to assign appropriate decision classes to new instances flowing into the system [4].

A current trend in machine learning research (as well as in the field of data-driven decision-making) focuses on learning models from data streams. However, decision-

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08754-7_63

making on streaming data is not an easy task, since high data volumes are continuously flowing into the system [3], especially in cases where the machine learning model must deal with data that are imbalanced in nature.

The problem of learning from imbalanced data streams is seen to be more complicated and complex than learning from static data where all the decision classes are known and balanced in the training dataset [1]. This problem arises in many real scenarios such as social media analytics, text classification tasks, fraud detection, technical and manufacturing systems etc. However, despite the increasing number of studies addressing imbalanced data using different methods, more attention needs to be given to the problem of dealing with streaming data when the data exhibit a changing and imbalanced class ratio [1], [2].

When the imbalanced data problem is considered, the popular approach proposed to eliminate an unfavorable imbalance between instances from different decision classes is to apply the SMOTE (Synthetic Minority Over-sampling Technique) algorithm [5]. The SMOTE algorithm produces new, synthetic instances located between random instances of the minority class. Alternatively, for generating synthetic instances that are closer to the model's decision boundary between the minority instances and other classes (i.e., the majority class), the borderline-SMOTE algorithm has been proposed [12]. These two algorithms are well-known examples among a myriad of the existing algorithms addressing imbalanced data and belong to the family of over-sampling algorithms. Opposite to the SMOTE class of algorithms, which create instances from the least frequent class only, under-sampling techniques for removing instances from the most frequent class have also been proposed. In both cases, the aim is to reach a good balance between instances belonging to all considered classes.

In general, the approaches for solving (i.e., eliminating) the class imbalance problem are divided into the following categories: data-level approaches, algorithm-level approaches, and hybrid approaches. A comprehensive discussion on this matter is included in [6].

The problem of imbalanced data needs more attention when the data streams are considered. A comparative study of selected algorithms dedicated to solving a class imbalance problem has been presented in [7]. More updated reviews of different techniques for imbalanced data streams are included in [8] and [9]. In [8] the authors underlined that the approaches that are able to learn from imbalanced data streams can be divided into two groups: passive and active approaches. This taxonomy is based on the possibility of the algorithms to detect drift. Among the discussed algorithms are: RLSACP, ONN, ESOS-ELM, an ensemble of neural networks, OnlineUnderOverBagging, OnlineSMOTE-Bagging, OnlineAdaC2, OnlineCSB2, OnlineRUSBoost and OnlineSMOTEBoost, ARFRE, RebalanceStream, OOB, UOB, WEOB1 and WEOB2. As an alternative, the meta-strategy Continuous-SMOTE (C-SMOTE) is proposed and compared with four other strategies which are able to deal with class imbalance in data streams, including ARFRE, RebalanceStream, OOB and UOB (see [8]).

In addition to an overview of different techniques dealing with imbalanced data streams, in [9] the Online-MC-Queue (OMCQ) algorithm is proposed. This algorithm

DOI: 10.1007/978-3-031-08754-7_63

is based on a framework for online learning in multi-class imbalanced data streams. The algorithm utilises a queue as a resampling method. The queue is dynamically created for instances belonging to each considered class. They are updated independently for each considered class, continuously assuring a balance between the instances belonging to the considered classes.

The core of this paper is also based on the framework for online learning presented previously by the author in [10]. The framework has been designed to work with imbalanced data streams. The WECOI (Weighted Ensemble with one-class Classification and Over-sampling and Instance selection) algorithm has been based on the implementation of over- and under-sampling techniques to eliminate imbalance between the instances belonging to minority and majority classes.

This paper is an extension of the work presented in [10] and considers the problem of clustering and its impact on the quality of the over-sampling process within WECOI. In this paper WECOI has been extended in a new approach to synthetic instances generation, where they are generated with respect to the reference instances representing clusters produced on the instances belonging to the majority class. In this algorithm, the way the clustering is done may influence the synthetic instance generation. Thus, the main aim of the paper is to formulate the answers to the following questions: (1) whether the clustering technique can have an impact on the quality of the over-sampling implemented within WECOI, and (2) whether the number of clusters influences the process of synthetic instances generation.

This paper is organised as follows. In the next section the WECOI-based framework is presented. Section 3 presents in detail the process of synthetic instances generation and presents the new approach. The computational experiment and results are included in Section 4. The last section points out to a few conclusions and directions for future research.

2 A Framework for Learning from Imbalanced Data Streams

The approach discussed in this paper is dedicated to learning from data streams. This approach, originally called Weighted Ensemble with one-class Classification and Over-sampling and Instance selection (WECOI), is also based on the decomposition of a multi-class classification problem into a set of sub-problems involving one-class classification. WECOI also uses mechanisms to eliminate the negative effect of the imbalance between minority and majority instances in data streams. The framework is also based on a drift detection concept within data streams and the final decision output is produced using a weighted ensemble classification model. The basic components of the framework are shown in Fig. 1.

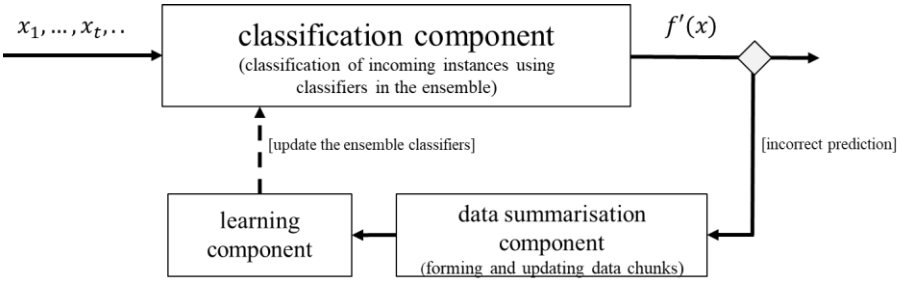
$x_1, \dots, x_t, ..$
classification component
(classification of incoming instances using classifiers in the ensemble)
$f'(x)$
[update the ensemble classifiers]
[incorrect prediction]
learning component
data summarisation component
(forming and updating data chunks)

Fig. 1. A framework for learning from imbalanced data streams [10].

The main task of the classification component is to classify the incoming instances using ensemble classifiers. The approach discussed is also based on an assumption that the correctness of the output class will be known in the future. Instances for which the decision classes have not been established in the correct way are redirected to the data summarisation component.

The data summarisation component forms data chunks from instances incoming from the classification component. The component constantly updates the data chunks, so the data chunks consist of possible current instances. Based on the data chunks, base classifiers of the ensemble are formed. Each modification (updating) of the data chunk entails an update of the base classifier, i.e. also the ensemble.

A basic assumption is that the data are considered with respect to each class independently. This means that the framework is based on decomposition of the multi-class classification problem into a set of one-class classification problems. What it also means is that each detected decision class is considered independently. From the algorithmic point of view, this means that data chunks are formed independently for each decision class. Thus, a given data chunk consists of positive instances for the considered decision class, while other data chunks are in opposition consisting of negative instances.

All the current data chunks formed by the data summarisation component are used to induce base classifiers, a process that is carried out within the learning component. The learning process is done independently for each considered decision class using the current positive and negative instances.

Each new induced base classifier replaces an older base classifier in the ensemble. This also means that the framework is based on remembering earlier-induced classifiers and the number of remembered base classifiers is a parameter of the algorithm. Finally, new incoming instances are classified using ensemble classifiers and the prediction result is determined through, for example, the weighted majority vote.

In the proposed approach, much attention is paid to updating and forming data chunks. When a new instance arrives, the current data chunks are updated in the following way:

- When the size of the data chunk is smaller than the defined threshold, then a new instance is added to the data chunk;

DOI: 10.1007/978-3-031-08754-7_63

- When the data chunk is completed, i.e. it consists of a number of instances equal to the defined threshold, the data chunk is updated.

In the second of the above cases, to decide whether instances should be added to the current set different techniques can be used. WECOI, adopting the one-class classification problem, uses two different nearest neighbour-based methods, i.e. CNN-d and ENN-d. Both of these algorithms are dedicated to instance selections between instances belonging to the same decision class and independently from other considered decision classes. The pseudocode for CNN-d and ENN-d is given in [11]. In other words, the process of updating the data chunks is carried out using under-sampling techniques.

WECOI is not free from additional challenges. When the learning component starts the induction of the classifier, the existing problem of imbalanced data in the data chunks must be eliminated. This is especially the case when, while the system is working, more of the incoming instances belong to one decision class rather than the others. This means that the number of instances included in the available data chunks is not equal. To sum up, in the considered system the problem of imbalanced data must be eliminated and for this an over-sampling procedure is applied.

In the next section of the paper the procedure for over-sampling is discussed in more detail, and a new approach for reducing the negative effect of the imbalance between minority and majority instances in data streams is also proposed. The further discussion is limited to the case of a binary classification, i.e., where only two decision classes in streams are possible.

3 Over-sampling Procedure for Imbalanced Data Streams

3.1 Basic Over-sampling Procedure

The over-sampling procedure described in [10] and implemented in the WECOI starts by clustering of the instances belonging to the data chunk consisting of the majority class. The centres of the produced clusters are next used as reference instances and, for each of them, nearest neighbours belonging to the minority are selected. The number of the neighbours in this case is a parameter of the algorithm. Next, a synthetic instance is generated between all the selected nearest neighbours. This procedure is repeated until a balance between the minority and majority classes has been reached.

Originally, the algorithm was implemented using k-means, and for a small number of clusters only. Based on the experiment results (see [11]), we observe that such an approach can be promising and ensure acceptable results for the over-sampling, but not for all possible cases. An example of such a case is when only one of the produced clusters exists close to the decision borderline and the others are distant from this decision boundary or even behind the first cluster. In this case, the reference instances belonging to the remote clusters have no real impact on the process of synthetic instance generation, which means that they have no significant impact on the target process. The basic procedure can have also an negative impact on the computational cost, when the clustering is repeated without influence on the quality of the synthetic instance generation.

The above observations have been included in the modified version described in the next subsection.

3.2 A Modified Procedure for Over-sampling

In this paper we propose an alternative approach to the generation of synthetic instances within a minority set of instances and to eliminate imbalanced data within the data chunks used by WECOI. In comparison to the basic procedure described in 3.1, this procedure assures the generation of synthetic instances with respect to each reference instance from the majority set of instances independently. The extended modified procedure gives more competence to reference instances located closer to the decision borderline than to reference instances located farther away. In other words, the reference instances lying closer to the decision borderline have more real impact on the process of the elimination of the imbalance between instances from the minority and majority class.

The pseudocode of the discussed over-sampling procedure is given in Algorithm 1.

Algorithm 1 Over-sampling procedure for the minority set of instances

```
Input: $T$, $Y$ – data chunks; $k_1, k_2$ – parameters of the procedure;
Begin
  Let $T$ be a set of minority instances;
  Let $Y$ be a set of majority instances;
  Set $q = false$;
  Map instances from $Y$ into clusters using a selected clustering algorithm;
  Let $Y_i$ $(i = 1, \dots, n)$ denote the obtained clusters and $n$ denote the number of clusters;
  For each $Y_{i:i=1,..,n}$ find its centres, called reference instances;
  Let $y_{i:i=1,..,n}$ denote the reference instances for $Y_{i:i=1,..,n}$;
  Let $Y^*$ denote a set of the reference instances;
  Let $t$ denote the reference instance for $T$;
  Repeat
  For $i$:=1 to $n$ do
    Select from $(Y^* \cup \{t\}) \backslash \{y_i\}$ $k_1$-nearest reference instances of $y_i$;
    Let $Y_i^*$ denote a set of the nearest reference instances calculated for $y_i$;
    Let $P_i$ be a subset of $Y_i^*$ such that $P_i \subset T$;
    If $|P_i| \neq \emptyset$ then
      Select $k_2$-nearest neighbours of $y_i$ from $T$;
      Let $P_i^*$ denote a set of the nearest neighbours of $y_i$ from $T$;
      Generate randomly a synthetic instance $x_a$ located between instances from $P_i^*$;
      $T = T \cup \{x_a\}$;
```

```
            $q = true$;
          End if
        End for
        If $q = false$ then $k_1$++;
        Until $|T| \cong |Y|$;
      End
```

The procedure of synthetic instance generation, shown as Algorithm 1, has implemented an adaptation mechanism, which extends the possibilities of the procedure in the event of an unfavourable data distribution and excludes the complete lack of the possibility of selecting so-called nearest reference instances from a class of minority instances. Nevertheless, the number of nearest reference vectors k_1 and the number of nearest neighbours k_2 remain the input parameters of the procedure.

The modified procedure of over-sampling is independent of the clustering algorithm and gives more freedom to decide which clustering algorithm to use. However, it is best to use a clustering algorithm that is free from any parameterization and that can produce clusters in an automatic way.

Representative here may be the similarity-based clustering algorithm (SCA) (see for example [13]). The SCA starts the clustering from the calculation of a value of the similarity coefficient independently for each instance from the data set. The number of clusters is determined by the value of the similarity coefficient. It is done based on the assumption that the similarity coefficient of all instances within a cluster is equal. In such an approach, the clusters and their number are set automatically.

Algorithm 2 SCA clustering procedure

```
Input: $X$ – data set;

Begin
  Calculate for each instance from $X$ the similarity coefficient $s_{i:i=1,\dots,N}$, where $N$ is a number of instances;
  Map instances from $X$ into $k$ clusters denoted as $X_i, i = 1, \dots, k$ in such way that each cluster contains instances with an identical value of the similarity coefficient $s_i$, where $k$ is the number of different values of $s_i$;

End
```

4 Computational Experiment

The aim of the computational experiment was to evaluate the new approach for the balancing of instances between minority and majority classes in a data stream.

The research question was: whether the choice of the clustering algorithm and selection of other clustering parameters influence the quality of the over-sampling and performance of WECOI, and whether WECOI can be considered as a competitive approach compared to other tools dedicated to learning from imbalanced data streams.

WECOI based on the new over-sampling procedure has been evaluated using: SCA (as a basic clustering approach), k-means, XMeans and Kernel-based fuzzy C-means (KFCM). SMOTE and borderline-SMOTE algorithms have also been applied for balancing instances between minority and majority data chunks within the WECOI framework. WECOI based on the new over-sampling procedure has been denoted as WECOI', WECOI based on SCA as WECOI'$_{SCA}$, WECOI'$_{k\text{-means}}$ – for k-means clustering etc. When WECOI uses a SMOTE version for generating synthetic instances, it is denoted as WECOI'$_{SMOTE}$, when the borderline-SMOTE algorithm it is noted as WECOI'$_{BR\text{-}SMOTE}$. Results provided based on using of the primary over-sampling procedure are denoted as WECOI.

For all versions of WECOI to generate the base classifiers a POSC4.5 algorithm has been applied. In the computational experiment discussed, as a base approach to under-sampling within WECOI, the ENN method has been used. All WECOI versions have been run with the number of nearest reference vectors k_1 set as two.

The results obtained have also been compared with other algorithms proposed for data streams: Oversampling and Undersampling Online Bagging (OUOB) [19], the ensemble learning algorithm based on Online Bagging (OB) [20], and the Learn++.NIE algorithm [21]. The algorithms were implemented as extensions of the Massive Online Analysis (MOA) package [18] within the WEKA environment [22]. For OUOB, OB and Learn++.NIE, the Hoeffding Tree [23] has been selected for the base classifiers and the base classifier pool equals 10. The results have been also compared with results obtained by the Accuracy Weighted Ensemble (AWE) [24], using only the Hoeffding Option Tree (HOT), iOVFDT (Incrementally Optimised Very Fast Decision Tree) [18], [23].

The performance was measured based on the classification accuracy, defined as the percentage of all instances correctly identified and using the test-then-train paradigm. All experiments were repeated 30 times on each data benchmark and the results are shown as the mean from these repetitions.

The computational experiment was performed using synthetic and real data sets. Among the data benchmarks of real data streams were: electricity, airlines, and an ozone and gas sensor array. A summary of the characteristics of the real data sets is presented in Table 1. The synthetic data streams were generated using the MOA framework [18] using: SEA [16], HYPERPLANE [17] and AGRAWAL [16]. In general, the SEA generator was used based on 10% of noise, a sudden concept drift and without class balancing. The HYPERPLANE (Hyp) generator was used with standard MOA parameters; however, a rotation of the decision boundary was set for each concept, assuring also incremental concept drift (i.e. -*t* 0.01). The AGRAWAL (Agr) generator was set to obtain a sudden concept drift and without class balancing. The number of instances was set to 10,000,000 with two decision classes and the number of attributes was set to ten with all attributes with drift. Table 1 also shows the size of the threshold which defined the size of the data chunk for the following datasets. In the case of the synthetic data, the size equals 1000 instances.

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08754-7_63

The values of the classification accuracy obtained by WECOI (and all its considered versions) as well as the other algorithms compared are shown in Table 2[1]. For WECOI the results are presented as a best obtained by the algorithm independently on parameter settings.

Table 1. Real data streams and their characteristics

Dataset	Source	#instances	#attributes	#classes	Threshold (as # of instances)
Electricity	[14]	45312	8	2	1000
Airlines	[14]	539383	7	2	1000
Ozone	[15]	2534	72	2	250

Table 2. Average accuracy for all data set (in %)

Algorithm	Electricity	Airlines	Ozone	Gas	Hyp	Agr	SEA
WECOI	76,5	64,82	81,56	84,62	85,78	91,68	88,45
WECOI’$_{\text{SCA}}$	76,7	63,48	82,21	84,7	**85,91**	92,71	88,79
WECOI’$_{\text{k-means}}$	76,31	64,15	81,04	83,45	84,17	92,24	86,5
WECOI’$_{\text{Xmeans}}$	73,51	62,48	78,21	82,01	79,48	87,45	82,14
WECOI’$_{\text{KFCM}}$	77,2	64,86	81,74	83,42	84,62	92,37	87,68
WECOI’$_{\text{SMOTE}}$	73,45	62,48	79,41	80,87	79,3	83,12	81,64
WECOI’$_{\text{BR-SMOTE}}$	75,8	63,57	80,92	**84,94**	80,62	87,84	86,4
OUOB	75,42	65,42	**82,31**	83,42	72,27	92,48	86,69
OB	**77,66**	63,21	76,56	82	84,7	81,67	83,74
Learn++.NIE	70,7	**66,8**	82,1	83,5	84,51	92,42	**89,3**
AWE	71,06	63,4	66,54	82,42	70,15	80,2	77,81
HOT	74,02	62,34	81,03	82,6	80,07	**94,99**	88,03
iOVFDT	72,52	63,52	81,25	83,54	81,56	93,67	83,02

The results in Table 2 demonstrate that none of the evaluated approaches is best for all the datasets. However, true is also conclusion that WECOI’ is a competitive algorithm to other compared, including also algorithms proposed for learning from data streams. When the WECOI’ is compared with WECOI, i.e. with algorithm,

[1] The best solution obtained by the compared algorithms is indicated in bold. The underline indicates the best solution obtained by the WECOI’ and its considered versions.

where the primary over-sampling procedure has been used, it can be notice, that more advanced approach for over-sampling assure better results. It is also important to notice that WECOI' based on SCA (WECOI'$_{SCA}$) is more accurate than others its versions based on another clustering algorithms. WECOI'$_{SCA}$ assured the best results in four cases. Alternative clustering approach is KFCM, assuring also competitive classification results. Both algorithms, that is SCA and KFCM, outperform - in terms of the resulting accuracy, two remaining clustering algorithms (k-means and XMeans).

For proper working of proposed algorithm two parameters also play an important role. The average number of nearest reference vectors k_1 during the experiments equalled 4. On the start of the algorithm the parameter was set on 2. It means that the proposed algorithm increased it to find adequate number of nearest reference instances and analysed the structure of the nearest neighbour clusters in the data space to select the best. The second parameter, i.e. the number of nearest neighbours k_2 equalled from 6 to 8. This parameter decided on number of nearest neighbours where between them a synthetic instance was generated.

The clustering approach for over-sampling is also more promising than SMOTE approach. In general, the SMOTE approach guaranteed results on the level k-means.

In one case using the borderline-SMOTE was most attractive. Although in all other cases the results produced by WECOI'BR-SMOTE have been comparable to the WECOI' based on clustering, thus it can be concluded that it is also a competitive approach for the reduction of imbalanced classes in the data stream.

One additional conclusion may be also formulated, that the proposed WECOI is flexible on the implementation of different algorithms for over-sampling.

5 Conclusions

Learning from imbalanced data streams is one of the key challenges in supervised learning, especially when data streams flow into the system and when their distribution changes over time. To address the imbalanced data in such data streams, this paper proposes a novel cluster-based approach called WECOI as an alternative approach for over-sampling. The paper discusses the proposed approach and compares it with others.

The computational experiments showed that the novel approach is competitive to other existing methods dedicated to address imbalanced data elimination in streams. The effectiveness of oversampling based on clustering needs an approach that automatically sets the appropriate number of clusters. Such properties were demonstrated by the SCA and KFCM algorithms. The paper shows also that WECOI belongs to the category of open algorithms and can be implemented using different methods for over-sampling and under-sampling.

Future research includes a more comprehensive statistical analysis of the results, particularly targeting additional datasets and benchmarks. Additionally, further research focusing on studying the influence of the size of the data chunks is required. In on-going research, the problem of diversification of selection of nearest neighbours from minority data is solved. Based on the computational experiments, it has been observed that this can have an impact on the WECOI performance. A deeper investigation of such an impact is outlined for future work.

References

1. Charte, F., Rivera, A.J., del Jesus, M.J., Herrera, F.: MLSMOTE: Approaching imbalanced multilabel learning through synthetic instance generation. Knowledge-Based Systems 89, 385–397 (2015)
2. Zyblewski, P., Sabourin, R., Woźniak, M.: Preprocessed dynamic classifier ensemble selection for highly imbalanced drifted data streams. Information Fusion 66, 138–154 (2021)
3. Sahel, Z., Bouchachia, A., Gabrys, B., Rogers, P.: Adaptive mechanisms for classification problems with drifting data. In: Apolloni, B. et al. (eds.) KES 2007, LNAI 4693, pp. 419–426. Springer-Verlag Berlin Heidelberg (2007)
4. Duda, R., Hart, P., Stork, D.: Pattern Classification. 2nd edn. John Wiley (2000)
5. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321–357 (2002)
6. Sharma, S., Gosain, A., Jain, S.: A Review of the Oversampling Techniques in Class Imbalance Problem. In: Khanna, A., Gupta, D., Bhattacharyya, S., Hassanien, A.E., Anand S., Jaiswal, A. (eds.) International Conference on Innovative Computing and Communications. Advances in Intelligent Systems and Computing, vol. 1387. Springer, Singapore (2022). doi:10.1007/978-981-16-2594-7_38
7. Nguyen, H.M., Cooper, E.W., Kamei, K.: A comparative study on sampling techniques for handling class imbalance in streaming data. In: Proceedings of the 6th International Conference on Soft Computing and Intelligent Systems, and the 13th International Symposium on Advanced Intelligence Systems, pp. 1762–1767 (2012). doi:10.1109/SCIS-ISIS.2012.6505291.
8. Bernardo, A., Gomes, H.M., Montiel, J., Pfahringer, B., Bifet, A., Valle, E.D.: C-SMOTE: Continuous Synthetic Minority Oversampling for Evolving Data Streams, In: Proceedings of 2020 IEEE International Conference on Big Data (Big Data), pp. 483–492 (2020) doi: 10.1109/BigData50022.2020.9377768.
9. Sadeghi, F., Viktor, H.L.: Online-MC-Queue: Learning from Imbalanced Multi-Class Streams. In: Proceedings of the Third International Workshop on Learning with Imbalanced Domains: Theory and Applications, PMLR 154:21–34, ECML-PKDD, Bilbao, Spain (2021)
10. Czarnowski, I.: Learning from imbalanced data streams based on over-sampling and instance selection. In: Lecture Notes in Computer Science, pp. 378–391 (2021). doi:10.1007/978-3-030-77967-2_32
11. Czarnowski, I., Jędrzejowicz, P.: Ensemble online classifier based on the one-class base classifiers for mining data streams. Cybernetics and Systems 46(1–2), 51–68 (2015). doi:10.1080/01969722.2015.1007736
12. Han, H., Wang, W.-Y., Mao, B.-H.: Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning. In: Logical Foundations of Computer Science, pp. 878–887 (2005) doi:10.1007/11538059_91
13. Czarnowski, I., Jędrzejowicz, J., Jędrzejowicz, P.: Designing RBFNs Structure Using Similarity-Based and Kernel-Based Fuzzy C-Means Clustering Algorithms. IEEE Access, 9, 4411–4422 (2021). doi: 10.1109/ACCESS.2020.3048104.

14. Harries, M.: Splice-2 comparative evaluation: Electricity pricing. Technical Report 1, University of New South Wales, Sydney, Australia (1999)
15. Asuncion, A., Newman, D.J.: UCI Machine Learning Repository. Irvine, CA: University of California, School of Information and Computer Science (2007). http://www.ics.uci.edu/~mlearn/MLRepository.html
16. Agrawal, R., Imilielinski, T., Swani, A.: Database mining: A performance perspective. IEEE Transactions on Knowledge and Data Engineering 5(6), 914–925 (1993)
17. Hulten, G., Spencer, L., Domingos, P.: Mining time-changing data streams. In: Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 97–106. ACM (2001).
18. Bifet, A., Holmes, G., Kirkby, R., Pfahringer, B.: MOA: Massive Online Analysis. Journal of Machine Learning Research 11, 1601–1604 (2010)
19. Wang, S., Minku, L.L., Yao, X.: Dealing with Multiple Classes in Online Class Imbalance Learning. In: Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI'16), July 2016
20. Oza, N.C.: Online bagging and boosting. In: Proceedings of the 2005 IEEE International Conference on Systems, Man and Cybernetics, Waikoloa, HI, USA, 10–12 October 2005; vol. 2343, pp. 2340–2345
21. Ditzler, G., Polikar, R.: Incremental Learning of Concept Drift from Streaming Imbalanced Data. IEEE Transactions on Knowledge and Data Engineering 25(10), 2283–2301 (2013). doi: 10.1109/TKDE.2012.136.
22. Frank, E., Hall, M.A., Witten, I.H.: The WEKA Workbench. Online Appendix for “Data Mining: Practical Machine Learning Tools and Techniques”. 4th edn. Morgan Kaufmann, (2016)
23. Bifet, A.: Adaptive learning and mining for data streams and frequent patterns. PhD thesis, Universitat Politecnica de Catalunya (2009)
24. Wang, H., Fan, W., Yu, P.S., Han, J.: Mining concept-drifting data streams using ensemble classifiers. In: Proceedings of 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 226-235 (2003). https://doi.org/10.1145/956750.956778