Prog Artif Intell (2016) 5:221–232
DOI 10.1007/s13748-016-0094-0

REVIEW

Learning from imbalanced data: open challenges and future directions

Bartosz Krawczyk[1]

Received: 5 January 2016 / Accepted: 11 April 2016 / Published online: 22 April 2016

Abstract Despite more than two decades of continuous development learning from imbalanced data is still a focus of intense research. Starting as a problem of skewed distributions of binary tasks, this topic evolved way beyond this conception. With the expansion of machine learning and data mining, combined with the arrival of big data era, we have gained a deeper insight into the nature of imbalanced learning, while at the same time facing new emerging challenges. Data-level and algorithm-level methods are constantly being improved and hybrid approaches gain increasing popularity. Recent trends focus on analyzing not only the disproportion between classes, but also other difficulties embedded in the nature of data. New real-life problems motivate researchers to focus on computationally efficient, adaptive and real-time methods. This paper aims at discussing open issues and challenges that need to be addressed to further develop the field of imbalanced learning. Seven vital areas of research in this topic are identified, covering the full spectrum of learning from imbalanced data: classification, regression, clustering, data streams, big data analytics and applications, e.g., in social media and computer vision. This paper provides a discussion and suggestions concerning lines of future research for each of them.

Keywords Machine learning · Imbalanced data · Multi-class imbalance · Big data · Data streams · Imbalanced clustering · Imbalanced regression

✉ Bartosz Krawczyk
bartosz.krawczyk@pwr.edu.pl

[1] Department of Systems and Computer Networks, Wrocław University of Technology, Wyb. Wyspiańskiego 27, 50-370 Wrocław, Poland

1 Introduction

Canonical machine learning algorithms assume that the number of objects in considered classes is roughly similar. However, in many real-life situations the distribution of examples is skewed since representatives of some of classes appear much more frequently. This poses a difficulty for learning algorithms, as they will be biased towards the majority group. At the same time usually the minority class is the one more important from the data mining perspective, as despite its rareness it may carry important and useful knowledge. Therefore, when facing such disproportions one must design an intelligent system that is able to overcome such a bias. This domain is known as learning from imbalanced data [29]. This problem has been widely discussed in the last two decades. One of the first deeper analyses of learning from skewed data was related to convergence rates of backpropagation-trained neural networks [2]. Since then a plethora of methods were developed, usually concentrating on balancing data via preprocessing or modifying and adapting existing classifiers.

Several excellent surveys and paper collections that capture recent advances in imbalanced learning field were published during last years. He and Garcia [24] provided a systematic review of metrics and algorithm-level approaches. In the same year another survey by Sun et al. [51] was published, concentrating on the classification aspect of imbalanced learning. A more recent book in a form of paper collection was edited by He and Ma [25], covering such important issues as sampling strategies, active learning and streaming data. A book by García et al. [21] discusses the topics of data preprocessing, among which a reasonable amount of space is dedicated to preparing, sampling and cleaning imbalanced datasets. A more global review on learning from skewed data was proposed by Branco [5] and concentrates

on a more general issue of imbalanced predictive modeling. Among more specialized discussions on this topic a thorough survey on ensemble learning by Galar et al. [17], an in-depth insight into imbalanced data characteristics by López et al. [36] and discussion on new perspectives for evaluation classifiers on skewed datasets [42] deserve mentioning.

Contrary to mentioned works this paper is not an exhaustive review of existing methodologies. Instead it aims to discuss open challenges and future directions in learning from imbalanced data. It points to important issues that are yet to be addressed in order to gain a deeper understanding of this vast field. It discusses emerging topics and contemporary applications that require new methods for managing data imbalance.

The scope of this paper is not limited to classification problems that seem to capture the majority of focus in the imbalanced domain. Instead it discusses the varied forms of learning where data imbalance may be the issue. Seven vital areas are identified and open challenges in each of them are highlighted. These include classification of binary and multi-class problems, multi-label and multi-instance learning, semi-supervised and unsupervised handling of imbalanced datasets, performing regression on skewed examples, learning from imbalanced data streams in stationary and drifting environments, and finally large scale and big data cases. Along with a detailed discussion of these open topics, we present our position on promising research directions that should be explored to address these challenges.

The remaining part of this manuscript is organized as follows. Next section gives a necessary background on the imbalanced learning domain. Section 3 discusses open challenges in binary classification, while Sect. 4 expands this to multi-class problems. Going beyond these popular tasks Sect. 6 presents future directions in imbalanced regression and Sect. 7 in semi-supervised, active and unsupervised learning. Perspectives on mining imbalanced streaming and big data are given in Sects. 8 and 9. The final section concludes this paper.

2 Learning from imbalanced data: preliminaries

This section introduces basic concepts related to imbalanced learning and presents the nomenclature used during discussing open challenges and future directions in following sections. Recent and emerging applications that face the difficulty of skewed class distributions are also discussed.

2.1 Tackling imbalanced data

We may distinguish three main approaches to learning from imbalanced data:

- *Data-level methods* that modify the collection of examples to balance distributions and/or remove difficult samples.
- *Algorithm-level methods* that directly modify existing learning algorithms to alleviate the bias towards majority objects and adapt them to mining data with skewed distributions.
- *Hybrid methods* that combine the advantages of two previous groups.

Let us now present a short overview of the mentioned approaches.

Data-level methods: concentrate on modifying the training set to make it suitable for a standard learning algorithm. With respect to balancing distributions we may distinguish approaches that generate new objects for minority groups (oversampling) and that remove examples from majority groups (undersampling). Standard approaches use random approach for selection of target samples for preprocessing. However, this often leads to removal of important samples or introduction of meaningless new objects. Therefore, more advanced methods were proposed that try to maintain structures of groups and/or generate new data according to underlying distributions [9]. This family of algorithms also consists of solutions for cleaning overlapping objects and removing noisy examples that may negatively affect learners [49].

Algorithm-level methods: concentrate on modifying existing learners to alleviate their bias towards majority groups. This requires a good insight into the modified learning algorithm and a precise identification of reasons for its failure in mining skewed distributions. The most popular branch is cost-sensitive approaches [67]. Here, given learner is modified to incorporate varying penalty for each of considered groups of examples. This way by assigning a higher cost to less represented set of objects we boost its importance during the learning process (which should aim at minimizing the global cost associated with mistakes). It must be noted that for many real-life problems it is difficult to set the actual values in the cost matrix and often they are not given by expert beforehand. Another algorithm-level solution is to apply one-class learning that focuses on target group, creating a data description [28]. This way we eliminate bias towards any group, as we concentrate only on a single set of objects. One, however, needs some specialized methods to use one-class learners for more complex problems [33].

Hybrid methods: concentrate on combining previously mentioned approaches to extract their strong points and reduce their weaknesses [63]. Merging data-level solutions with classifier ensembles [64], resulting in robust and efficient learners [34] is highly popular. There are some works that propose hybridization of sampling and cost-sensitive learning [57].

Table 1 A list of selected recent real-life applications with data imbalance present

Application area	Problem description
Activity recognition [19]	Detection of rare or less-frequent activities (multi-class problem)
Behavior analysis [3]	Recognition of dangerous behavior (binary problem)
Cancer malignancy grading [30]	Analyzing the cancer severity (binary and multi-class problem)
Hyperspectral data analysis [50]	Classification of varying areas in multi-dimensional images (multi-class problem)
Industrial systems monitoring [44]	Fault detection in industrial machinery (binary problem)
Sentiment analysis [65]	Emotion and temper recognition in text (binary and multi-class problem)
Software defect prediction [48]	Recognition of errors in code blocks (binary problem)
Target detection [45]	Classification of specified targets appearing with varied frequency (multi-class problem)
Text mining [39]	Detecting relations in literature (binary problem)
Video mining [20]	Recognizing objects and actions in video sequences (binary and multi-class problem)

2.2 Real-life imbalanced problems

Developments in learning from imbalanced data have been mainly motivated by numerous real-life applications in which we face the problem of uneven data representation. In such cases the minority class is usually the more important one and hence we require methods to improve its recognition rates. This is closely related with important issues like preventing malicious attacks, detecting life-threatening diseases, managing atypical behavior in social networks or handling rare cases in monitoring systems.

A list of selected recent real-life problems that have embedded the class imbalance difficulty is presented in Table 1.

3 Binary imbalanced classification

Binary classification problems can be considered as the most developed branch of learning from imbalanced data [51]. This task originates from various real-life applications, like medicine (sick vs. healthy), computer security (valid activity vs. unauthorized or malicious one), or computer vision (target object vs. background). In such a scenario the relationship between classes is well defined: one of them is a majority, while the other is a minority group. This allows for many straightforward approaches to balance the distributions or shift classifiers towards the minority class. Despite numerous works on this topic there are still many open challenges that need to be addressed. We identify the following future directions of research.

3.1 Analyzing the structure of classes

One of the most interesting directions in binary imbalanced classification is the notion that imbalance ratio is not the sole source of learning difficulties. Even if the disproportion is high, but both classes are well represented and come from non-overlapping distributions we may obtain good classification rates using canonical classifiers. Degradation of performance may also be caused by the presence of difficult examples, especially within the minority class [35]. Recent work of Napierala and Stefanowski [40] proposed to analyze the neighborhood of each minority class example and assign it to one of four predefined groups: safe, borderline, rare and outliers. This allows to shed new light onto imbalanced datasets and analyze their difficulty from the minority class structure perspective. Such a proposal interestingly expands the understanding of difficulties embedded in the nature of data. This can be seen as a starting point for promising future research.

Following directions should be taken in order to fully understand and exploit the role of minority class structure:

- It is important to propose new classifiers that can either directly or indirectly incorporate this background knowledge about objects into their training procedure. Thus, when designing an efficient classifier one must not only alleviate the bias towards the majority class, but also pay attention to difficulties of individual minority examples. Preliminary works present in the literature show the high potential of this solution [4,32].
- The same idea can be carried over to data preprocessing approaches. Here established structure of the minority class would allow for selecting important or difficult samples to concentrate on. This way we may vary the level of oversampling according to example types (first idea can be tracked to ADASYN method that ignored safe examples [23]) or supervise the undersampling procedure in order not to discard the important minority representatives.
- Separate studies should be done on the role of noisy / outlier samples in minority class. Recent work by Sáez et al. [47] suggests to remove such examples. However, when dealing with small sample sizes further reduc-

tions may be dangerous. Additionally, how one can be sure that given object is an actual noise or outlier and not an inappropriately sampled minority representative? Thus removing such example may lead to wrong classification of potential new objects to appear in its neighborhood.
– The current labeling method for identifying types of minority objects relies on k-nearest neighbors (with $k = 5$) or kernel methods. This, however, strongly implies uniform distribution of data. It seems promising to propose adaptive methods that will adjust the size of analyzed neighborhood according to local densities or chunk sizes.

3.2 Extreme class imbalance

Another important issue is related to the disproportion between classes. Most of the contemporary works in class imbalance concentrate on imbalance ratios ranging from 1:4 up to 1:100. However, there is a lack of studies on the classification of extremely imbalanced datasets. In real-life applications such as fraud detection [61] or cheminformatics [12] we may deal with problems with imbalance ratio ranging from 1:1000 up to 1:5000. This poses new challenges to data preprocessing and classification algorithms, as they must be adjusted to such extreme scenarios.

Following directions should be taken in order to gain a better insight into classification of extremely imbalanced datasets:

– In cases with such a high imbalance the minority class is often poorly represented and lacks a clear structure. Therefore, straightforward application of preprocessing methods that rely on relations between minority objects (like SMOTE) can actually deteriorate the classification performance. Using randomized methods may also be inadvisable due to a high potential variance induced by the imbalance ratio. Methods that will be able to empower the minority class and predict or reconstruct a potential class structure seem to be a promising direction.
– Another possible research track to be envisioned relies on decomposition of the original problem into a set of subproblems, each characterized by a reduced imbalance ratio. Then canonical methods could be used. This approach, however, requires a two-way development: (i) algorithms for meaningful problem division and (ii) algorithms for reconstruction of the original extremely imbalanced task.
– Third challenge lies in efficient extraction of features for such problems. Internet transactions or protein data are characterized by potentially high-dimensional and sparse feature spaces. Furthermore one can predict emergence of such problems in social networks and computer vision. This calls for development of new approaches for representing such data that will allow at the same time for an efficient processing and boosting discrimination of the minority class.

3.3 Classifier's output adjustment

As mentioned in Sect. 2 methods for addressing class imbalance are either concentrating on modifying the learning algorithm or the training set. However, as noted by Provost [43], simply changing the data distribution without considering the imbalance effect on the classification output (and thus adjusting it properly) may be misleading. Recent studies show that weighting or putting a threshold on continuous output of a classifier (known also as support functions or class probability estimates) can often lead to better results than data resampling and may be applied to any conventional classifier [31,66].

Following directions should be taken in order to further develop classifier's output compensation for imbalanced data:

– Currently the output is being adjusted for each class separately, using the same value of compensation parameter for each classified object. However, from our previous discussion we may see that the minority class usually is not uniform and difficulty level may vary among objects within. Therefore, it is interesting to develop new methods that will be able to take into consideration the characteristics of classified example and adjust classifier's output individually for each new object.
– A drawback of methods based on output adjustment lies in possibility of overdriving the classifier towards the minority class, thus increasing the error on the majority one. As we may expect that the disproportion between classes will hold also for new objects to be classified, we may assume that output compensation will not always be required (as objects will predominantly originate from majority class distribution). Techniques that will select uncertain samples and adjust outputs only for such objects are of interest to this field. Additionally, dynamic classifier selection between canonical and adjusted classifiers seems as a potential useful framework.
– Output adjustment is considered as an independent approach. Yet from a general point of view it may be fruitful to modify outputs even when data-level or algorithm-level solutions have been previously utilized. This way we may achieve class balancing on different levels, creating more refined classifiers. Analyzing the output compensation may also bring new insight into supervising undersampling or oversampling to find balanced performance on both classes.

3.4 Ensemble learning

Ensemble learning is one of the most popular approaches for handling class imbalance [4,17,34]. Hybridization of Bagging, Boosting and Random Forests with sampling or cost-sensitive methods prove to be highly competitive and robust to difficult data. However, most of these approaches used are heuristic based and still there is a lack of proper insight into the performance of classifier committees with skewed classes.

Following directions should be taken in order to expand the branch of imbalanced ensemble learning:

- There is a lack of good understanding of diversity in imbalanced learning. What actually contributes to this notion? Is diversity on majority class as important as on minority class? Undersampling-based ensembles usually maintain the minority class intact or introduce some small variations to it. Therefore, their diversity should not be very high which intuitively is a significant drawback. How then we can explain their excellent classification performance? A detailed study of this is required to better understand this phenomenon.
- There are no clear indicators on how large ensembles should be constructed. Usually their size is selected arbitrarily, which may result in some similar classifiers being stored in the pool. It would be beneficial to analyze relations betweens characteristics of imbalanced dataset and the number of classifiers required to efficiently handle it, while maintaining their individual quality and mutual complementarity. Additionally, ensemble pruning techniques dedicated specifically to imbalanced problems should be developed [18].
- Finally, most of imbalanced ensemble techniques use majority voting combination method. In standard scenarios this is a simple and effective solution. However, we may ask ourselves if this is really suitable for imbalanced learning, especially in case of randomized methods. It seems reasonable to assume that base classifiers trained using sampling methods will differ in their individual qualities due to being based on examples with varying difficulties. Therefore, this will propagate into their individual competencies that should be exploited during the combination phase.

4 Multi-class imbalanced classification

Multi-class imbalanced classification is not as well-developed as its binary counterpart. Here we deal with a more complicated situation, as the relations among the classes are no longer obvious. A class may be a majority one when it is compared to some other classes, but a minority or well-balanced for the rest of them [57]. When dealing with multi-class imbalanced data we may easily lose performance on one class while trying to gain it on another [14]. Considering this problem there are many issues that must be addressed by novel proposals. A deeper insight into the nature of the class imbalance problem is needed, as one should know in what domains does class imbalance most hinder the performance of standard multi-class classifiers when designing a method tailored for this problem. While most of the challenges discussed in Sect. 3 can also be transferred to multi-class problems, there is a number of topics specific just to them. We identify the following vital future directions of research.

4.1 Data preprocessing

The role of data preprocessing may be of even higher importance here than in case of binary problems. One may easily identify possible difficulties: class overlapping may appear with more than two groups, class label noise may affect the problem and borders between classes may be far from being clearly defined. Therefore, proper data cleaning and sampling procedures that take into account the varying characteristics of classes and balanced performance on all of them must be proposed.

Following directions should be taken for introducing new methods dedicated to multi-class imbalanced preprocessing:

- It is interesting to analyze the type of examples present in each class and their relations to other classes. Here it is not straightforward to measure the difficulty of each sample, as it may change with respect to different classes. For example, a given object may be of borderline type for some groups and at the same time a safe example when considering remaining classes. Therefore, a new and more flexible taxonomy should be presented. Our initial works on this topic indicate that analysis of example difficulty may significantly boost the multi-class performance [46].
- New data cleaning methods must be developed to handle presence of overlapping and noisy samples that may additionally contribute to deteriorating classifier's performance. One may think of projections to new spaces in which overlapping will be alleviated or simple removal of examples. However, measures to evaluate if a given overlapping example can be discarded without harming one of classes are needed. In case of label noise it is very interesting to analyze its influence on actual imbalance between classes. Wrongly labeled samples may increase the imbalance (when actually it's ratio is lower) or mask actual disproportions (wrongly re-balancing the classes). Such scenarios require dedicated methods for detecting

and filtering noise, as well as strategies for handling and relabeling such examples.
– New sampling strategies are required for multi-class problems. Simple re-balancing towards the biggest or smallest class is not a proper approach [1,15]. We need to develop dedicated methods that will adjust the sampling procedures to both individual properties of classes and to their mutual relations. Hybrid approaches, utilizing more than one method seem as an attractive solution.

4.2 Multi-class decomposition

An intuitive approach for handling multi-class imbalanced datasets is to apply a decomposition strategy and reduce it to a set of binary problems that can be solved by one of existing techniques [14]. Advantages of such an approach include simplified subproblems and alleviation of some data-level difficulties (like overlapping and class noise). However, one must be aware of possible drawbacks such as loss of balanced performance on all of classes or rejecting the global outlook on the multi-class problem. Nevertheless, this direction can be considered as highly promising one, yet still requiring a significant development.

Following directions should be taken when designing decomposition strategies for multi-class imbalanced problems:

- Decomposition methods used so far [14] apply identical approach for each pair of classes. This seems as a slightly naive solution as the pairwise relationships may highly vary. Adjusting used solution to individual pairwise problems seems a highly more flexible solution. Calculating the cost penalties or oversampling ratios individually for each pairs is a good starting direction, but the true challenge lies in proposing a framework that will be able to select specific data or algorithm-level solution on the basis of subproblem characteristics.
- So far only binary decomposition in form of one-vs-one and one-vs-all was considered. However, there is a number of different techniques that may achieve the same goal, while alleviating the drawbacks of binary solutions (like high number of base classifiers or introducing additional artificial imbalance). Hierarchical methods seem as a promising direction. Here we need solutions to aggregate classes according to their similarities or dissimilarities, preprocess them at each level and then use a sequential, step-wise approach to determine the final class. Alternatively, decomposition with one-class classifiers can be considered, as they are robust to class imbalance and can serve as an efficient tool for dealing with difficult multi-class datasets [33].
- Finally, using decomposition methods require dedicated combination strategies that are able to reconstruct the original multi-class problem. As ones canonically used in this field were designed for roughly balanced scenarios, it seems worthwhile to design new fusion approaches suitable for cases with skewed distributions. This way it may be possible to compensate for the imbalance both on decomposed class level and on final output combination level.

4.3 Multi-class classifiers

High potential lies in the design of multi-class classifiers that are skew-insensitive. They will allow to handle multi-class problems without referring to decomposition or resampling strategies, while using algorithm-level solutions to counter the class imbalance. Recently a Hellinger-distance modification of decision trees [10] and neural networks [66] were proposed and proved to work highly efficiently. Therefore, one may wonder if it is possible to adapt other popular classifiers to this scenario.

Following issues should be considered when designing multi-class classifiers in presence of class imbalance:

– A deeper insight is required into how multiple skewed distributions affect the forming of decision boundaries in classifiers. As Hellinger distance was proven to be useful in class imbalance cases, it should be incorporated to other distance-based classifiers. Other solutions with potential robustness to imbalance, like density-based methods, must be explored. When combined with approaches for reducing class overlapping and label noise (two phenomenons present in multi-class imbalanced data that may severely degrade the density estimation process) they should provide a powerful tool for classification.
– Ensemble solutions should be investigated. They may offer ways to rebalance the distributions in varied ways. Exploring local competencies of classifiers and creating sectional decision areas may also alleviate significantly the difficulty of the problem. Here once again rises the issue of maintaining the diversity in ensemble systems and proper selection of useful base learners.

5 Multi-label and multi-instance imbalanced classification

Multi-label and multi-instance learning are specific branches of pattern classification problems, dealing with structured outputs and inputs. In the first case a single example can be characterized by more than one class label. In second case we work with bags of objects and labels are provided only for bags, not for objects within. One must note that having assigned a class label to a bag does not imply that

this bag consists only of objects from a given class. Both of these problems became very popular in machine learning communities in recent years. However, little attention was paid to imbalanced learning in their context, despite the fact that these areas suffer from it. In multi-label learning so far measures of imbalance and SMOTE-based oversampling approach have been proposed [8]. In multi-instance learning cost-sensitive and standard resampling have been used to counter skewed distributions with regard to number of bags and number of instances within bags [38,59]. Therefore, imbalanced learning from multi-label and multi-instance data still requires a significant development and presents many open challenges lying before future researchers.

Following challenges are to be faced in the field of learning from imbalanced multi-label and multi-instance data:

- In multi-label learning there is a need for skew-insensitive classifiers that do not require resampling strategies. It seems promising to use existing multi-label methods (such as hierarchical multi-label classification tree or classifier chains) and combine them with skew-insensitive solutions available in multi-class classification domain. An ideal goal would be development of such multi-label classifiers that display similar performance to canonical methods on balanced multi-label problems, while being at the same time robust to presence of imbalance.
- Another interesting direction is to investigate the possibilities of using decomposition-based solutions. Binary relevance is a popular method, transforming a multi-label problem into a set of two-label subproblems. Hence, balancing label distributions seems more straightforward when applied to each subproblem individually. This problem could also be approached as an aggregation, to create balanced super-classes and then solve the problem in an iterative divide-and-conquer manner.
- In multi-instance learning when resampling bags we must take into consideration that they may be characterized by a different level of uncertainty. Bags that have a higher probability of consisting purely of objects from the same class should be prioritized, as they carry a better representation of the target concept. This requires a development of new measures for assessing the quality of training bags and selecting the most useful ones.
- Current sampling approaches for multi-instance learning work on the level of either bags or objects within bags. However, difficulties may arise from both of these situations occurring at the same time. Firstly, it is important to establish how each of those types of sample imbalance affects a classifier's performance and is any of them more harmful to it. Secondly, global schemes for tackling these two types of imbalance at once must be proposed. They should identify, which type of imbalance is predominant and adapt their solution based on the characteristics of analyzed problem.

6 Regression in imbalanced scenarios

Another branch of learning algorithms that is yet to be explored from the imbalanced perspective is regression. Many important real-life applications like economy, crisis management, fault diagnosis, or meteorology require predicting rare and extreme values of continuous target variable. These values are often accompanied by an abundance of standard or common values that model normal behavior of the analyzed system. This leads to creation of an imbalanced learning scenario. Despite frequent presence of this phenomenon in various problems so far little attention was paid to it. Main works done on this topic include proposal of evaluation metrics that take into account varying importance of observations [54] and adaptation of undersampling and SMOTE to continuous output prediction problems [53]. This shows that the research community took only its first step into the problem of imbalanced regression and further works on this topic are of vital importance.

Following open issues are crucial when developing novel methods for imbalanced regression:

- Development of cost-sensitive regression solutions that are able to adapt the penalty to the degree of importance assigned to rare observations. Not only existing methods should be modified to include object-related cost information, but also methodologies for assigning such costs must be developed. It seems interesting to investigate the possibility of adapting the cost not only to the minority group, but to each individual observation. This would allow for a more flexible prediction of rare events of differing importance.
- Methods that will allow to distinguish between minority and noisy samples must be proposed. In case of small number of minority observations the potential presence of noisy ones may significantly affect both preprocessing and regression itself. Therefore, a deeper insight into what makes a certain observation a noisy outlier and what a valuable minority example is required.
- As in classification problems, ensemble learning may offer significant improvement in both robustness to skewed distributions and in predictive power. This direction is especially interesting, as the notion of ensemble diversity is better explored from theoretical point of view in regression than in classification [6]. This should allow to developing efficient ensemble regression systems that will have directly controlled diversity level on minority observations.

7 Semi-supervised and unsupervised learning from imbalanced data

Previous four sections concentrated on imbalanced scenarios in supervised tasks. While this is the most popular area, it is at the same time not the only one where skewed distributions may affect the learning process. This phenomenon appears often in semi-supervised [56], active [68] and unsupervised learning [41], especially in clustering. Despite numerous solutions dedicated to this problem, most of them display reduced effectiveness when true underlying groups of data have highly varying sizes. This is due to the so-called uniformity effect that causes these algorithms to generate clusters of similar sizes. This is especially vivid in case of centroid-based approaches [60], while density-based ones seem to display some robustness to it [52].

Clustering imbalanced data can be seen from various perspectives: as a process of group discovery on its own, as a method for reducing the complexity of given problem, or as a solution to analysis of the minority class structure. Such knowledge discovery approach is important in many aspects of learning from imbalanced data.

Following open issues are to be faced when developing novel methods for clustering imbalanced datasets:

- There exist plethora of cluster validity indexes designed for evaluating and selecting proper clustering models. However, none of them takes into account the fact that actual clusters may be of highly varying sizes. Therefore, existing methods should be modified and new indexes should be proposed that measure how well discovered groups reflect the actual skewed distributions in analyzed dataset.
- Popular centroid-based clustering approaches should be adjusted to imbalanced scenarios, allowing more flexible cluster adjustment. A high potential lies in hybridization with density-based methods for analyzing the local neighborhood of centroid, or in local differing of similarity measures that are used to assign given object to certain cluster.
- It is very interesting to apply clustering only on the minority class. This way we would be able to discover substructures within that can be utilized in further learning steps. Extracting additional knowledge about the examples within the minority class would also allow for a more detailed insight into difficulties embedded in the nature of data. However, for this we need refined clustering solutions that would be able to analyze the structure of this class, while paying attention to different possible types of examples that may appear. Here an algorithm cannot be to generative (not to lose small groups of objects) or to prone to overfitting (which can be an issue in case of small-sample sized minority classes).
- Clustering can also be used as a method for fitting more specialized classifiers. Hence, we would be interested in methods that are able to divide the original problem into a set of atomic subproblems and identify difficult areas in the decision space. Such clusters would allow to decompose the classification problem and analyze it independently for each part of the decision space, thus being able to locally adapt the solution to the level and type of difficulty present.
- Important issue is how to detect the underlying class imbalance in semi-supervised and active learning. Here the initial batch of examples may be fairly balanced, but original distributions can be significantly skewed. On the other hand, the starting set of objects may display imbalance that do not reflect the original one (majority class may be in fact the minority one or there may be in fact no imbalance at all). To answer such questions novel unsupervised methods for assessing the distributions and potential difficulty of unlabeled objects must be introduced. Additionally, novel active learning strategies that can point out to the most difficult objects that will have highest effect on learned decision boundaries are of interest to machine learning community.

8 Learning from imbalanced data streams

Mining data streams is one of the important contemporary topics in machine learning [16,62]. Dynamical nature of data that arrive either in batches or online poses new challenges when imbalanced distributions are to be expected [26]. Whether dealing with stationary or evolving streams we require adaptive methods that are able to deal with skewed objects coming in real time. Additionally, in case of changing streams the relationships between classes are no longer permanent. We may face a situation in which the imbalance ratio, as well as which class is the minority one, changes with stream progress. Despite intense works on this subject in recent years with single model approaches [22] and ensemble learning [58] there are still numerous open challenges to be addressed.

Following open issues are of high importance when designing new algorithms for learning from imbalanced data streams:

- Most of existing works assume that we deal with a binary data stream for which the relationships between classes may change over time: one of the classes may obtain increased number of samples in given time window, other class may have reduced number of new instances, or classes may reach equilibrium. However, the imbalanced problem may occur in data streams from various reasons. A very important aspect that requires a proper attention

is connected with the phenomenon of new class emergence and/or disappearance of the old ones. When a new class appears it will be naturally underrepresented with respect to the ones present so far. Even with increased number of examples from it the imbalance still will be present, as classifiers have processed a significant number of objects before this new class emerged. It is interesting how to handle this bias originating from classifier's history. Another issue is how to handle fading classes. When objects from given distribution will become less and less frequent should we increased their importance along with increasing imbalance ratio? Or on the contrary, we should reduce it as this class becomes less representative to the current state of the stream?

– Another vital challenge is connected with the availability of class labels. Many contemporary works assume that class label is available immediately after the new sample is being classified. This is far from real-life scenarios and would impose tremendous labeling costs on the system. Therefore, method for streaming data that reduce the cost of supervision (e.g., by active learning) is currently of crucial importance in this field [69]. This raises an open issue on how to sample imbalanced data streams? Can the active learning approach be beneficial to reducing the bias towards the majority class by intelligent selection of samples? On the other hand, one must develop labeling strategies that will not overlook minority representatives and adjust their labeling ratio to how well the current state of minority class is captured.
– In many real-life streaming applications (like computer vision or social networks) the imbalance may be caused by reappearing source. This leads to recurring drift that affects proportions between class distributions. Thus, it seems worthwhile to develop methods dedicated to storing general solutions to such scenarios instead of reacting to each reappearance of class imbalance anew. This requires development of algorithms to extract drift's templates and use them to train specific classifiers that will be stored in repository. When a similar change in class distributions will appear we may use available classifier instead of training a new one from the scratch. To obtain best flexibility and account for the fact that even in recurring drifts the properties of individual imbalanced cases may differ, one needs fast adaptation methods that will use the stored classifiers as a starting point.
– Using characteristics and structure of minority class is a promising direction for static imbalanced learning. But is it possible to adapt this approach to streaming data? The analysis of minority example types poses a high computational cost, but one may use hardware speed-up to alleviate this problem. Additionally, even if samples arrive online their influence on minority class is local which may be an useful hint for designing such systems. Algorithms dedicated to tracking and analyzing the history of changes in minority class structure may also provide us with valuable insight into the evolution of imbalance problem in given stream over time.

9 Imbalanced big data

The final open challenge discussed in this paper is connected with increasing complexity of data. Modern systems generate massive amounts of information that forces us to develop computationally effective solutions for processing them. Big data can also be affected by class imbalance, posing increased challenge to learning systems [55]. Not only the increasing data volume can become prohibitive for existing methods, but also the nature of problem can cause additional difficulties. Big imbalanced data can originate from various specific areas like social networks [27] or computer vision [11] that forces us to work with specific types of data like graphs, tensors or video sequences. This requires not only scalable and efficient algorithms, but also methods for handling heterogeneous and atypical data. Additional challenges are posed by computing environments like Spark or Hadoop that were initially not developed for handling skewed datasets.

Following open issues must be analyzed in order to gain better understanding of how to tackle imbalanced big data:

– SMOTE-based oversampling methods applied in distributed environments such as MapReduce tend to fail [13]. This can be caused by a random partitioning of data for each mapper and thus introducing artificial samples on the basis of real objects that have no spatial relationships. Therefore, to apply SMOTE-based techniques for massive datasets one either require new global-scale and efficient implementations, data partitioning methods that preserve relations between examples, or some global arbitration unit that will supervise the oversampling process.
– When mining big data we are interested in value we can extract from this process. A promising direction lies in developing methods that by analyzing the nature of imbalance will allow us to gain a deeper understanding of given problem. What is the source of imbalance, which types of objects are the most difficult ones, where overlapping and noise occurs and how does this translate to its business potential? Additionally, interpretable classifiers that can handle massive and skewed data are of interest.
– We must develop methods for processing and classifying big data in form of graphs, xml structures, video sequences, hyperspectral images, associations, tensors etc [7,37]. Such data types are becoming more and more frequent in both imbalanced and big data analytics and impose certain restrictions on machine learning systems.

Instead of trying to convert them to numerical values it seems valuable to design both preprocessing and learning algorithms that will allow a direct handling of massive and skewed data represented as such complex structures.
– When dealing with imbalanced big data we face one of two possible scenarios: when majority class is massive and minority class is of a small sample size and when imbalance is present but representatives from both classes are abundant. First issue is related directly to the problem of extreme imbalance discussed in Sect. 3.2 and solutions proposed there should be adjusted to large-scale analytics. Second issue is related to the observations that imbalance ratio may not be the main source of learning difficulties. This requires an in-depth analysis of the structure of minority class and examples present there. But it also raises a question: is the taxonomy discussed in Sect. 3.1 still valid when facing massive datasets? Big data imbalance may cause the appearance of new types of examples or changes in properties of already described types. Additionally, we deal with a much more complex scenarios that would require local analysis of each difficult region and fitting solutions individually to each of them.

10 Conclusions

In this paper, we have discussed current research challenges standing before learning from imbalanced data that have roots in contemporary real-world applications. We analyzed different aspects of imbalanced learning such as classification, clustering, regression, mining data streams and big data analytics, providing a thorough guide to emerging issues in these domains. Despite intense works on imbalanced learning over the last two decades there are still many shortcomings in existing methods and problems yet to be properly addressed.

In summary the research community should consider the following directions when further developing solutions to imbalanced learning problems:

– Focus on the structure and nature of examples in minority classes in order to gain a better insight into the source of learning difficulties.
– Develop methods for multi-class imbalanced learning that will take into account varying relationships between classes.
– Propose new solutions for multi-instance and multi-label learning that are based on specific structured nature of these problems.
– Introduce efficient clustering methods for unevenly distributed object groups and measures to properly evaluate and select partitioning models in such scenarios.
– Consider imbalanced regression problems and develop methods for deeper analysis of individual properties of rare examples.
– Analyze the nature of class imbalance in data streams beyond simple notion of two classes with shifting distributions.
– Gain a deeper insight into the potential value that can be extracted from interpretable analysis of imbalanced big data.

This paper showed that there are many challenges in the vast field of imbalanced learning that require attention from the research community and intensive development. There are still many unaudited directions to be taken in this branch of machine learning. And this is what makes it still fresh and exciting. Let us hope that all of issues and challenges highlighted in this paper will be addressed in future and that this will lead to advancing our understanding of the imbalance phenomenon in learning systems.

Acknowledgments This work is supported by the Polish National Science Center under the Grant no. DEC-2013/09/B/ST6/02264.

References

1. Abdi, L., Hashemi, S.: To combat multi-class imbalanced problems by means of over-sampling techniques. IEEE Trans. Knowl. Data Eng. **28**(1), 238–251 (2016)
2. Anand, R., Mehrotra, K.G., Mohan, C.K., Ranka, S.: An improved algorithm for neural network classification of imbalanced training sets. IEEE Trans. Neural Netw. **4**(6), 962–969 (1993)
3. Azaria, A., Richardson, A., Kraus, S., Subrahmanian, V.S.: Behavioral analysis of insider threat: a survey and bootstrapped prediction in imbalanced data. IEEE Trans. Comput. Soc. Syst. **1**(2), 135–155 (2014)
4. Blaszczynski, J., Stefanowski, J.: Neighbourhood sampling in bagging for imbalanced data. Neurocomputing **150**, 529–542 (2015)
5. Branco, P., Torgo, L., Ribeiro, R.P.: A survey of predictive modelling under imbalanced distributions. CoRR. arXiv:1505.01658 (2015)
6. Brown, G., Wyatt, J.L., Tiňo, P.: Managing diversity in regression ensembles. J. Mach. Learn. Res. **6**, 1621–1650 (2005)
7. Brzezinski, D., Piernik, M.: Structural XML classification in concept drifting data streams. New Generat. Comput. **33**(4), 345–366 (2015)
8. Charte, F., Rivera, A.J., del Jesús, M.J., Herrera, F.: MLSMOTE: approaching imbalanced multilabel learning through synthetic instance generation. Knowl. Based Syst. **89**, 385–397 (2015)
9. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: synthetic minority over-sampling technique. J. Artif. Intell. Res. **16**, 321–357 (2002)

10. Cieslak, D.A., Hoens, T.R., Chawla, N.V., Kegelmeyer, W.P.: Hellinger distance decision trees are robust and skew-insensitive. Data Min. Knowl. Discov. **24**(1), 136–158 (2012)
11. Cyganek, B.: Object Detection and Recognition in Digital Images:Theory and Practice. Wiley, New York (2013)
12. Czarnecki, W.M., Rataj, K.: Compounds activity prediction in large imbalanced datasets with substructural relations fingerprint and EEM. In: 2015 IEEE TrustCom/BigDataSE/ISPA, Helsinki, Finland, August 20–22, 2015, vol. 2, p. 192 (2015)
13. del Río, S., López, V., Benítez, J.M., Herrera, F.: On the use of mapreduce for imbalanced big data using random forest. Inform. Sci. **285**, 112–137 (2014)
14. Fernández, A., López, V., Galar, M., del Jesús, M.J., Herrera, F.: Analysing the classification of imbalanced data-sets with multiple classes: binarization techniques and ad-hoc approaches. Knowl. Based Syst. **42**, 97–110 (2013)
15. Fernández-Navarro, F., Hervás-Martínez, C., Gutiérrez, P.A.: A dynamic over-sampling procedure based on sensitivity for multi-class problems. Pattern Recognit. **44**(8), 1821–1833 (2011)
16. Gaber, M.M., Gama, J., Krishnaswamy, S., Gomes, J.B., Stahl, F.T.: Data stream mining in ubiquitous environments: state-of-the-art and current directions. Wiley Interdiscip. Rev. Data Min. Knowl. Discov. **4**(2), 116–138 (2014)
17. Galar, M., Fernández, A., Barrenechea, E., Bustince, H., Herrera, F.: A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. IEEE Trans. Syst. Man Cybern. Part C **42**(4), 463–484 (2012)
18. Galar, M., Fernández, A., Barrenechea, E., Bustince, H., Herrera, F.: Ordering-based pruning for improving the performance of ensembles of classifiers in the framework of imbalanced datasets. Inform. Sci. **354**, 178–196 (2016)
19. Gao, X., Chen, Z., Tang, S., Zhang, Y., Li, J.: Adaptive weighted imbalance learning with application to abnormal activity recognition. Neurocomputing **173**, 1927–1935 (2016)
20. Gao, Z., Zhang, L., Chen, M.-Y., Hauptmann, A.G., Zhang, H., Cai, A.-N.: Enhanced and hierarchical structure algorithm for data imbalance problem in semantic extraction under massive video dataset. Multimed. Tools Appl. **68**(3), 641–657 (2014)
21. García, S., Luengo, J., Herrera, F.: Data Preprocessing in Data Mining, In: Intelligent Systems Reference Library, vol. 72. Springer, Berlin (2015)
22. Ghazikhani, A., Monsefi, R., Yazdi, H.S.: Online cost-sensitive neural network classifiers for non-stationary and imbalanced data streams. Neural Comput. Appl. **23**(5), 1283–1295 (2013)
23. He, H., Bai, Y., Garcia, E.A., Li, S.: Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In: Proceedings of the International Joint Conference on Neural Networks, 2008, part of the IEEE World Congress on Computational Intelligence, 2008, Hong Kong, China, June 1–6, 2008, pp. 1322–1328 (2008)
24. He, H., Garcia, E.A.: Learning from imbalanced data. IEEE Trans. Knowl. Data Eng. **21**(9), 1263–1284 (2009)
25. He, H., Ma, Y.: Imbalanced Learning: Foundations, Algorithms, and Applications, 1st edn. Wiley-IEEE Press, New York (2013)
26. Hoens, T.R., Polikar, R., Chawla, N.V.: Learning from streaming data with concept drift and imbalance: an overview. Progress AI **1**(1), 89–101 (2012)
27. Hurtado, J., Taweewitchakreeya, N., Kong, X., Zhu, X.: A classifier ensembling approach for imbalanced social link prediction. In: 12th International Conference on Machine Learning and Applications, ICMLA 2013, Miami, FL, USA, December 4–7, 2013, vol. 1, pp. 436–439 (2013)
28. Japkowicz, N., Myers, C., Gluck, M.: A novelty detection approach to classification. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence, vol. 1, pp. 518–523, Morgan Kaufmann Publishers Inc, San Francisco (1995)
29. Japkowicz, N., Stephen, S.: The class imbalance problem: a systematic study. Intell. Data Anal. **6**(5), 429–449 (2002)
30. Krawczyk, B., Galar, M., Jelen, Ł., Herrera, F.: Evolutionary undersampling boosting for imbalanced classification of breast cancer malignancy. Appl. Soft Comput. **38**, 714–726 (2016)
31. Krawczyk, B., Woźniak, M.: Cost-sensitive neural network with roc-based moving threshold for imbalanced classification. In: Intelligent Data Engineering and Automated Learning—IDEAL 2015—16th International Conference Wroclaw, Poland, October 14–16, 2015, Proceedings, pp. 45–52 (2015)
32. Krawczyk, B., Woźniak, M., Herrera, F.: Weighted one-class classification for different types of minority class examples in imbalanced data. In: 2014 IEEE Symposium on Computational Intelligence and Data Mining, CIDM 2014, Orlando, FL, USA, December 9–12, 2014, pp 337–344 (2014)
33. Krawczyk, B., Woźniak, M., Herrera, F.: On the usefulness of one-class classifier ensembles for decomposition of multi-class problems. Pattern Recognit. **48**(12), 3969–3982 (2015)
34. Krawczyk, B., Woźniak, M., Schaefer, G.: Cost-sensitive decision tree ensembles for effective imbalanced classification. Appl. Soft Comput. **14**, 554–562 (2014)
35. Kubat, M., Matwin, S.: Addressing the curse of imbalanced training sets: One-sided selection. In: In: Proceedings of the 14th International Conference on Machine Learning, pp. 179–186. Morgan Kaufmann (1997)
36. López, V., Fernández, A., García, S., Palade, V., Herrera, F.: An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. Inform. Sci. **250**, 113–141 (2013)
37. Mardani, M., Mateos, G., Giannakis, G.B.: Subspace learning and imputation for streaming big data matrices and tensors. IEEE Trans. Signal Process. **63**(10), 2663–2677 (2015)
38. Mera, C., Arrieta, J., Orozco-Alzate, M., Branch, J.: A bag oversampling approach for class imbalance in multiple instance learning. In: Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications—20th Iberoamerican Congress, CIARP 2015, Montevideo, Uruguay, November 9–12, 2015, Proceedings, pp. 724–731 (2015)
39. Munkhdalai, T., Namsrai, O.-E., Ryu, K.H.: Self-training in significance space of support vectors for imbalanced biomedical event data. BMC Bioinform. **16**(S–7), S6 (2015)
40. Napierala, K., Stefanowski, J.: Types of minority class examples and their influence on learning classifiers from imbalanced data. J. Intell. Inform. Syst. (2015). doi:10.1007/s10844-015-0368-1
41. Nguwi, Y.-Y., Cho, S.-Y.: An unsupervised self-organizing learning with support vector ranking for imbalanced datasets. Expert Syst. Appl. **37**(12), 8303–8312 (2010)
42. Prati, R.C., Batista, G.E.A.P.A., Silva, D.F.: Class imbalance revisited: a new experimental setup to assess the performance of treatment methods. Knowl. Inform. Syst. **45**(1), 247–270 (2015)
43. Provost, F.: Machine learning from imbalanced data sets 101. In: Proceedings of the AAAI 2000 workshop on imbalanced data sets, pp. 1–3 (2000)
44. Ramentol, E., Gondres, I., Lajes, S., Bello, R., Caballero, Y., Cornelis, C., Herrera, F.: Fuzzy-rough imbalanced learning for the diagnosis of high voltage circuit breaker maintenance: the SMOTE-FRST-2T algorithm. Eng. Appl. AI **48**, 134–139 (2016)
45. Razakarivony, S., Jurie, F.: Vehicle detection in aerial imagery: a small target detection benchmark. J. Vis. Commun. Image Represent. **34**, 187–203 (2016)
46. Sáez, J.A., Krawczyk, B., Woźniak, M.: Analyzing the oversampling of different classes and types of examples in multi-class imbalanced datasets. Pattern Recognit. doi:10.1016/j.patcog.2016.03.012 (2016)
47. Sáez, J.A., Luengo, J., Stefanowski, J., Herrera, F.: SMOTE-IPF: addressing the noisy and borderline examples problem in imbal-

anced classification by a re-sampling method with filtering. Inform. Sci. **291**, 184–203 (2015)
48. Siers, M.J., Islam, M.Z.: Software defect prediction using a cost sensitive decision forest and voting, and a potential solution to the class imbalance problem. Inform. Syst. **51**, 62–71 (2015)
49. Stefanowski, J.: Dealing with data difficulty factors while learning from imbalanced data. In: Challenges in Computational Statistics and Data Mining, pp. 333–363 (2016)
50. Sun, T., Jiao, L., Feng, J., Liu, F., Zhang, X.: Imbalanced hyperspectral image classification based on maximum margin. IEEE Geosci. Remote Sens. Lett. **12**(3), 522–526 (2015)
51. Sun, Y., Wong, A.K.C., Kamel, M.S.: Classification of imbalanced data: a review. Int. J. Pattern Recognit. Artif. Intell. **23**(4), 687–719 (2009)
52. Tabor, J., Spurek, P.: Cross-entropy clustering. Pattern Recognit. **47**(9), 3046–3059 (2014)
53. Torgo, L., Branco, P., Ribeiro, R.P., Pfahringer, B.: Resampling strategies for regression. Expert Syst. **32**(3), 465–476 (2015)
54. Torgo, L., Ribeiro, R.P.: Precision and recall for regression. In: Discovery Science, 12th International Conference, DS 2009, Porto, Portugal, October 3–5, 2009, pp. 332–346 (2009)
55. Triguero, I., del Río, S., López, V., Bacardit, J., Benítez, J.M., Herrera, F.: ROSEFW-RF: the winner algorithm for the ecbdl'14 big data competition: an extremely imbalanced big data bioinformatics problem. Knowl. Based Syst. **87**, 69–79 (2015)
56. Triguero, I., García, S., Herrera, F.: SEG-SSC: a framework based on synthetic examples generation for self-labeled semi-supervised classification. IEEE Trans. Cybern. **45**(4), 622–634 (2015)
57. Wang, S., Li, Z., Chao, W.-H., Cao, Q.: Applying adaptive oversampling 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)
58. 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)
59. Wang, X., Liu, X., Japkowicz, N., Matwin, S.: Resampling and cost-sensitive methods for imbalanced multi-instance learning. In: 13th IEEE International Conference on Data Mining Workshops, ICDM Workshops, TX, USA, December 7–10, 2013, pp. 808–816 (2013)
60. Wang, Y., Chen, L.: Multi-exemplar based clustering for imbalanced data. In: 13th International Conference on Control Automation Robotics & Vision, ICARCV 2014, Singapore, December 10–12, 2014, pp. 1068–1073 (2014)
61. Wei, W., Li, J., Cao, L., Ou, Y., Chen, J.: Effective detection of sophisticated online banking fraud on extremely imbalanced data. World Wide Web **16**(4), 449–475 (2013)
62. Woźniak, M.: A hybrid decision tree training method using data streams. Knowl. Inform. Syst. **29**(2), 335–347 (2011)
63. Woźniak, M.: Hybrid Classifiers—Methods of Data, Knowledge, and Classifier Combination. In: Studies in Computational Intelligence, vol. 519. Springer, Berlin (2014)
64. Woźniak, M., Grana, M., Corchado, E.: A survey of multiple classifier systems as hybrid systems. Inform. Fusion **16**(1), 3–17 (2014)
65. Xu, R., Chen, T., Xia, Y., Lu, Q., Liu, B., Wang, X.: Word embedding composition for data imbalances in sentiment and emotion classification. Cogn. Comput. **7**(2), 226–240 (2015)
66. Yu, Hu, Sun, C., Yang, X., Yang, W., Shen, J., Qi, Y.: Odoc-elm: optimal decision outputs compensation-based extreme learning machine for classifying imbalanced data. Knowl. Based Syst. **92**, 55–70 (2016)
67. Zhou, Z.-H., Liu, X.-Y.: On multi-class cost-sensitive learning. Comput. Intell. **26**(3), 232–257 (2010)
68. Zieba, M., Tomczak, J.M.: Boosted SVM with active learning strategy for imbalanced data. Soft Comput. **19**(12), 3357–3368 (2015)
69. Zliobaite, I., Bifet, A., Pfahringer, B., Holmes, G.: Active learning with drifting streaming data. IEEE Trans. Neural Netw. Learn. Syst. **25**(1), 27–39 (2014)