Classification of Multi-class Imbalanced Data: Data Difficulty Factors and Selected Methods for Improving Classifiers

Jerzy Stefanowski(✉)

Institute of Computing Science, Poznan University of Technology, Poznan, Poland
jerzy.stefanowski@cs.put.poznan.pl

Abstract. The multiple class imbalanced problem is still less investigated than its binary counterpart. In particular, the sources of its difficulties have not been sufficiently studied so far. Therefore, in this paper we summarize the few literature works on the difficulty factors and present our own latest research results. The binary method for an identification of the types of minority examples is generalized for multiple imbalance classes. The second part of this paper presents three our recent methods for learning classifies from multi-class imbalanced data which exploit information on the aforementioned difficulty factors.

Keywords: Multi-class imbalanced data · Data difficulty factors · Re-sampling methods · Rule classifiers

1 Introduction

In imbalanced data at least one class, further called the minority class, contains a much smaller number of examples than other majority classes. Imbalanced classes pose serious difficulties for learning classifiers as the algorithms are biased towards the majority class examples and fail to recognize the instances from the minority class as accurate as possible [5,10].

Most of current research have been placed on constantly proposing new algorithms and less on studying why this class imbalanced problem is so difficult. However, some researchers have attempted to better understand the nature of the imbalance data and key properties of its underlying distribution. They noticed that the *class imbalance ratio* is not necessarily the only, or main, problem causing this performance decrease. Imbalanced data are often affected by other *difficulty factors*, which in turn cause the degradation of classification performance, sometimes even stronger than the global imbalance ratio [8,13,29,31]. The data difficulty factors are related to characteristics of class distribution, such as decomposition of the class into rare sub-concepts, overlapping between classes or presence of rare minority examples inside the majority class regions. With respect to data distribution characteristics Napierala et al. proposed in

S. Ramanna et al. (Eds.): IJCRS 2021, LNAI 12872, pp. 57–72, 2021.
https://doi.org/10.1007/978-3-030-87334-9_5

[27] to distinguish different *types of examples* – safe or unsafe to be learnt (e.g. borderline, rare or outliers) and present the methods for their identification.

Nevertheless this analysis and most of the methods concern binary imbalanced problems only. Despite this, in some applications it is necessary to deal with *multiple imbalanced classes* and to improve the recognition of more than one of the minority classes. Such multi-class imbalanced data occur, e.g., in medical diagnosis (where few important and rare diseases may occur), technical diagnostics with several degrees of the device failures, text categorization, etc.

The multi-class imbalanced problems are so far less investigated than their binary counterpart. The number of specialized approaches is definitely much smaller. In general, the multi-class learning problems are recognized as harder than two class ones, however the sources of these difficulties have not been sufficiently studied so far. The essential questions to be examined are as follows:

- Should the previously identified binary imbalanced data factors be adapted to multiple classes?
- Does the nature of the multi-class problems lead itself to rather new and different factors that cause deterioration of classifier's predictions?

So far, only a few hypotheses on such issues can be found in the literature. Therefore the first part of this paper is devoted to discussing the already identified data difficulty factors and presenting our own latest research results. Then, we discuss how the earlier binary method for an identification of the types of examples and their level of difficulty [27] can be generalized for multiple classes [22]. We describe the usage of the specialized grid clustering [21] to discover sub-concepts within minority classes and to find rare examples or outliers.

The other contribution includes a brief presentation of three recent methods, introduced by the author and his coauthors, for multi-class imbalanced data which exploit information on the aforementioned difficulty factors. These are SOUP resampling method [11], the rule induction multi-class BRACID algorithm [26] and a multi-class extension of Roughly Balanced Bagging ensemble [23]. The paper ends with a discussion of open problems and further perspectives for research on multi-class imbalanced problems.

2 Related Works on Classification of Multi-class Imbalanced Data

The current approaches to multi-class imbalances are usually divided into the following categories [5]:

- binary decomposition approaches that transform the multi-class problem into the set of binary ones and apply existing methods for improving binary problems,
- specialized approaches, which could be further split intro multi-class preprocessing, variants of cost-sensitive learning, algorithm modifications – including dedicated ensembles.

Following the authors of [5], the most popular are decomposition approaches which are based on the ensembles previously proposed to solve complex multi-class tasks[19]. The most often used frameworks are:

One-versus-all ensemble (OVA), which constructs binary classifiers to recognize a particular class against the remaining ones aggregated into one class [7]. During prediction, the test instance is classified by all base classifiers and is assigned to the class of the most confident base classifier.

One-versus-one ensemble (OVO), which constructs binary classifiers for all pairs of classes. The training set for a particular base classifier contains learning examples from the selected pair of classes only. The prediction for the new instance can be taken by majority voting of base classifiers' predictions or by weighted voting with confidence scores, or more complex aggregations [6,19].

These frameworks can be easily used in combination with techniques for binary imbalanced data. Moreover, they are often used with oversampling or undersampling approaches [6].

The specialized *multi-class imbalanced re-sampling* methods, e.g., oversampling Static-SMOTE, Global-CS or MDO [1,35], attempt to increase the cardinalities of minority class towards the size of the biggest class.

The selective hybrid re-sampling is done in SPIDER3 [33], where relations between classes are captured by pre-defined misclassification costs. *SMOTE and Clustered Undersampling Technique* (SCUT) [2] applies EM clustering for each majority class, and some examples are randomly removed from these clusters. The minority classes are oversampled with the standard SMOTE.

The final group of specialized methods aims at modifying neural networks or ensembles. The authors either try to integrate over-sampling in the network or propose different loss functions that direct the training the networks towards better recognition of minority classes. Boosting algorithms are also combined with specialized re-sampling, see e.g. [32].

3 Difficulty Factors in Imbalanced Data

3.1 Earlier Studies on Binary Imbalanced Classes

Imbalanced data are characterized with a *global imbalance ratio*. For binary classes it defined as a ratio of the majority class cardinality and minority one or a percentage of the minority class in all examples in the dataset. Besides this ratio researchers such as [8,13,31] noticed that other characteristics of examples distributions in the attribute space, called *data difficulty factors* also deteriorate classifier predictions. They mainly include:

- the fragmentation of the minority class into smaller, rare sub-concepts [15],
- the impact of too strong overlapping between classes,
- the presence of small, isolated groups of minority examples located deeply inside the majority class region.

The first factor comes from experimental observations that the minority class usually does not form a homogeneous region (single concept) in the attribute space, but is scattered into smaller sub-concepts spread over the space, often surrounded by examples from the majority class. Experimental studies, e.g. [15, 29] demonstrated its important impact.

The second factor corresponds to high *overlapping* between regions of minority and majority class examples in the attribute space. In particular it may occur in the complex boundary regions of both classes which are not clearly separated and contain mixed instances from minority and majority classes. Sanchez et al. also demonstrated that the local imbalanced ratio in the overlapping region is more influential than the global imbalanced ratio [8].

The third factor corresponds to *rare cases*, which are defined as isolated, very small, groups of minority examples (e.g., containing 1–3 examples) located more deeply inside regions of the other class [27]. They could be even single examples lying either inside this class or in empty regions of the attribute space. This is different from the first factor, which refers to the decomposition of the minority class into larger sub-clusters containing more examples than rare cases.

A related view on data difficulty factors leads to distinguishing different *types of minority examples*, usually called as *safe* or *unsafe*, based on the number of minority and majority class examples near them [18].

The special method for an identification of more detailed four types minority examples was proposed by Napierala and Stefanowski in [27]. It is based on analyzing class labels of examples in their local neighborhood defined either by k-nearest neighbours or by kernels. For instance, if $k = 5$ neighbourhood is considered, then the example is labelled as a safe example if all five or four its neighbors belong to its class. If three or two neighbors belong to the same class as the considered example, then it is labelled as borderline. If there are not the same class examples in k neighbourhood it is an outlier and a rare one for the remaining proportion.

Besides using labels of example types which depend on such proportions, the authors of [27] defined a coefficient expressing a *safe level* of the given example x – being a local estimator of conditional probability of its assignment to the target class as

$$p(C|x) = \frac{k_C}{k}, \tag{1}$$

where C is the class of example x, k is the number of neighbours and k_C is the number of neighbours which belongs to class C.

Experimental studies on the role of the aforementioned factors have shown that data complexities occur in imbalanced datasets, may play a key role in explaining the difference between the performance of various classifiers [27] and proposing new algorithms for improving classifiers.

3.2 Multi-class Difficulties

Researchers working on multi-class imbalanced data often argue that these data are more difficult than binary ones. However, this and some other hypothe-

ses were already considered for standard, balanced, machine learning tasks. For instance, the claim that decision boundaries between multiple classes are more complex and non-linear than simpler boundaries for binary classes follows the older works, in particular in the context of specialized ensembles such as pairwise coupling [14] (which inspires binary decomposition OVO), see e.g. a chapter in [19]. However, there are other newer hypotheses or observations resulting from experiments with multi-class imbalanced datasets. We summarize them below:

- Wang and Yao [32] analysing their experimental results made an observation that different predictive accuracy may be related to considering various configuration of types (sizes) of classes. They distinguish two configuration *multi-majority* and *multi-minority* referring to the datasets with only one majority or only one minority class and all the other classes being of the same type. Following experiments they concluded that that multi-majority class configurations were more difficult than multi-minority ones.
- Buda et al. [4] paid attention to yet another configuration of multiple classes – *gradual imbalance*, which contains classes of linearly growing sizes.
- Krawczyk claimed that a given class can be a minority class with respect to some classes, and at the same time the majority one to another subset of classes [17]. It makes re-sampling approaches difficult[1].

The needs for transferring the idea of types of minority examples [27] into multi-class data was discussed in [17,22]. Independently, the authors of [30] also adopted the binary example types to the multi-class setting, however in the simplest one-vs-all manner. They studied the performance of classifiers trained on a dataset with oversampled minority examples of one type (using the brute force strategy for testing many variants of random oversampling examples of the particular type). Their results showed improvements of classifiers for almost all datasets, although the authors did not present any methods for tuning the degrees of oversampling nor selecting the variant of example type selection.

Lango has recently carried out a comprehensive experimental study with specially generated synthetic datasets [24]. His main conclusions are as follows:

- The class overlapping was the very influential factor when combined with the higher imbalances. Changing the imbalance ratio presented a limited impact on the recognition of datasets without class overlapping or its slight amount.
- The types of the class size configurations with multiple majority classes were more difficult than multi minority ones. In the second configuration, recognition of the smallest classes were worse than in the former one. The gradual class size configuration with the intermediate classes played a special role between them depending whether the these classes are closer to minority or majority classes.

[1] In our opinion this hypothesis may be particularly interesting for the gradual imbalance configuration, where some classes may be *intermediate* ones with respect to their sizes. Furthermore we share Krawczyk's view that it may lead to ambiguities in the decision on the degree of modifications of the examples in oversampling or undersampling. It will be even more difficult when such classes overlap, which the author did not take into account.

– The analysis of interrelations between different types of classes showed that the increase of overlapping between the minority and majority classes led to the stronger deterioration of classifier performance than between minority ones. The impact of the intermediate classes depends on the direction of overlapping with other classes. Its overlapping with the minority class caused faster deterioration of the recognition of minority class than itself, so it played a similar role to majority classes.
– An increasing the number of classes was the most influential for a smaller number of classes.

4 Identifying Types of Examples in Multi-class Imbalanced Data

The generalization of types of examples for multiple classes should take into account at least some of the difficulty factors.

Napierala et al. noticed in [22] that analyzing mutual relations between classes shows that some minority classes can be treated as more closely related to each other than to the majority class. As discussed in the previous section the degree of overlapping between various classes may be different. Thus the new multi-class type of examples may also strongly depend on their relations to other classes. For instance, a given example may be of a borderline type for certain classes and at the same time a safe example for the remaining classes. However using existing binary decomposition approaches to estimate data difficulty or the similar adaptation from [30] cannot properly handle these situations.

These motivations have led Napierala et al. to model relations between multiple imbalanced classes by means of additional information about *similarity between pairs of classes*. This information could be either acquired from users - experts or more automatically estimated from class distributions in the attribute space [22]. It means that one needs information which classes can be treated as more similar to each other than to the rest of the classes. Furthermore, this class similarity may correspond to the expert's interpretation of a mutual position of examples in the neighborhood of the example from a given class. An intuition behind this neighborhood is the following: if example x from a given class has some neighbors from other classes, then neighbors from the class with higher similarity are more preferred.

Let us come back to a medical diagnosis case considered in [22]. Two classes corresponding to similar types of the same asthma should be considered as closer to each other than similar to other types of non-asthma as they need completely different therapies.

Defining it more precisely, it is assumed that for each pair of classes C_i, C_j the degree of their similarity is defined as a real valued number $\mu_{ij} \in [0; 1]$. Similarity of a class to itself is defined as $\mu_{ii} = 1$. The degree of similarity does not have to be symmetric, i.e. for some classes C_i, C_j it may happen that $\mu_{ij} \neq \mu_{ji}$.

Although the values of μ_{ij} are defined individually for each dataset, the general recommendation of [22] is to have higher similarities ($\mu_{ig} \rightarrow 1$) for other

minority classes C_g, while similarities to majority classes C_h should be rather low ($\mu_{ih} \to 0$). This recommendation follows the earlier discussed data difficulty factors, in particular on higher difficulty of the multi-majority case.

In the case of missing expert's preferences to defining these class similarities for the given dataset, the authors of [11] proposed to use the heuristics which follows class sizes as the basic symptom of class interrelations. This is defined as:

$$\mu_{ij} = \frac{\min(|C_i|, |C_j|)}{\max(|C_i|, |C_j|)} \tag{2}$$

where $|C_i|$ is the number of examples of C_i class.

These degrees of similarities are used to generalize the idea of an identification of types of examples. If one considers the k nearest neighborhood, then determining the number of examples from the majority class in the neighborhood of the example allows to assess how safe the example is, and then to establish its type. Let us start from defining the safe level for the multiple classes.

Considering a given example x belonging to the minority class C_i its safe level is defined with respect to l classes of examples in its neighborhood as:

$$safe(x_{C_i}) = \frac{\sum_{j=1}^{l} n_{C_j} \mu_{ij}}{k} \tag{3}$$

where μ_{ij} is a degree of similarity, n_{C_j} is a number of examples from class C_j inside the considered neighborhood of x and k is a total number of neighbors. The general interpretation of the safe level of the example is as follows: the lower the value, the more unsafe (difficult) is the example.

The safe levels could be exploited in two ways: either as the direct value, or by transforming the continuous levels into discrete intervals corresponding to types of example (as done in [27]). In Sect. 6 we will show how to use safe levels in SOUP preprocessing methods and how types of the examples are used inside BRACID rule induction algorithm.

In Table 1 we present some of experimental results from [22], which show that the recognition of minority classes is related to their average safe levels.

Table 1. Sensitivity of minority classes for three classifiers and average safe levels in these classes for new-thyroid, ecoli and cleveland datasets

	CART			NBayes			3NN			Average safe level		
	Min1	Min2	Min3	Min1	Min2	Min3	Min1	Min2	Min3	Min1	Min2	Min3
NT	0.94	0.83		0.94	0.86		0.71	0.89		0.77	0.78	
EC	0.60	0.85	0.78	0.68	0.30	0.90	0.48	0.75	0.84	0.57	0.91	0.82
CL	0.28	0.11	0.07	0.14	0.25	0.15	0.08	0.00	0.00	0.29	0.32	0.34

5 Discovering Split of Classes into Sub-concepts and Rare Examples

An identification of sub-concepts in the minority class is typically done by using clustering algorithms. Nearly all approaches exploit k-means algorithms which are run on examples of a single class, without analyzing their relation to remaining classes. Japkowicz et al. showed how the discovered clusters in both minority and majority classes could be used for random oversampling them [15]. The survey [31] covers other clustering approaches and presents their applications. The use of density algorithms such as DBCAN was considered much less frequently. However, the use of clustering algorithms for real-world datasets is still a non-trivial task, in particular tuning their parameters.

In [21], authors introduced a completely different *grid-based algorithm*, called *ImGrid*. The algorithm works in the following steps: 1) dividing the attribute space into grid cells, 2) joining similar adjacent cells taking into account their minority class distributions, 3) labeling examples according to difficulty factors, 4) forming minority sub-clusters.

The number of cells and the division of the attribute range into a number of intervals are estimated with a special heuristics [21]. The cells of the grid are joined based on example distributions, where each cell should contain enough examples, and by means of the statistical tests for the comparison of two discrete distributions. For binary classes it is done with *Barnard's test*. Clusters are formed after joining several cells. Each cluster is assigned one of four difficulty labels: safe, borderline, rare, or outlier, following rules developed by Napierala in [27]. To sum up, unlike other clustering ImGrid simultaneously does two things: detects clusters and categorizes them. More precisely it detects minority sub-clusters, outliers, rare cases, and class overlapping in binary imbalanced data. Furthermore, due to its small dependency on parameter tuning, ImGrid could be used to analyze real world datasets easier than previous algorithms.

Recently, it was generalized for multiple classes [16]. The main changes are the following. A special variant of Pearson χ^2 test (inspired by the solution from ChiMerge discretization) is used to evaluate similarities of class distributions in adjacent cells. Moreover, new heuristics for ordering cells are introduced in order to get larger clusters. Then, the new rules for labelling clustered cells were introduced as a multiple class generalization of intervals over the safe level, which were earlier considered in [27]. They are better suited for handling overlapping between several classes and identifying rare cases and outliers.

The multiple class ImGrid was validated on 12 synthetic datasets showing its ability to re-discover a structure of three classes.

6 Multi-class Hybrid Resampling Algorithm SOUP

Following a critical discussion of earlier resampling multi-class techniques, such as Global-CS or Static-SMOTE, the authors of [11] introduced a new method called called *Similarity Oversampling and Undersampling Preprocessing* (abbrev.

SOUP), which combines undersampling with oversampling and exploits the information about the difficulty of examples according to their safe levels.

The authors of SOUP decided that, all majority classes are undersampled and all minority classes are oversampled to the cardinality being the median of the sizes of the biggest minority and the smallest majority class. The resulting resampled dataset has a balanced class distribution, but also with a reasonable size, which is not present in other multi-class resampling methods.

The resampling is done following information on the safe levels of the examples presented in Sect. 4. The undersampling of the majority classes is performed by removing the most unsafe examples. It means that it removes the examples located near the boundaries with minority classes or inside their regions. On the other hand, the oversampling of minority classes is performed in the opposite direction, i.e. the safest examples are duplicated as firsts, enhancing the representation of clear minority concepts.

As the safe level of a particular example in the final distribution is changing while performing consecutive steps of resampling, the classes are ordered. Undersampling majority classes is done from the biggest to the smallest one while the minority classes are oversampled from the smallest to the biggest one. Moreover after each resampling step safe levels of all examples are recomputed.

The experiments [11] showed that SOUP outperformed baseline classifiers and Static-SMOTE and Global-CS – the two popular pre-processing methods for multi-class imbalances. Moreover SOUP is slightly better then OVO with re-sampling and competitive to MRBBag (discussed in Sect. 8). Selected results from [11] for using J.8 trees are presented in Table 2.

Table 2. Comparison of specialized multi-class methods vs. SOUP and multiple RBBagging – with respect to G-mean for selected real-world data sets

Dataset	Baseline tree	Global CS	Static SMOTE	OVA Oversam.	OVO Oversam.	SOUP	mRBBag
balancescale	0.0	0.340	0.080	0.302	0.526	0.614	0.683
car	0.847	0.940	0.897	0.184	0.939	0.938	0.917
cleveland	0.000	0.000	0.032	0.287	0.288	0.256	0.155
cmc	0.483	0.478	0.452	0.510	0.509	0.520	0.517
dermatology	0.945	0.952	.927	0.082	0.921	0.960	0.960
ecoli	0.728	0.719	0.738	0.000	0.805	0.0.721	0.768
flare	0.446	0.570	0.431	0.000	0.544	0.575	0.546
glass	0.625	0.715	0.699	0.000	0.698	0.667	0.405
led7digit	0.785	0.770	0.756	0.120	0.779	0.790	0.778
vehicle	0.912	0.912	0.915	0.164	0.923	0.909	0.943
winequality	0.421	0.464	0.356	0.456	0.492	0.448	0.525

7 Multi-class Variant of BRACID Algorithm

7.1 Rule Induction from Binary Imbalanced Data with BRACID

Although induction of rules from examples is one of the well studied tasks in machine learning, rule-based classifiers have not been studied in the context of imbalanced data as intensively as other algorithms. A fairly small number of rule classifiers specialized for imbalanced data has been introduced so far, for their review see e.g. [26]. BRACID (the acronym of Bottom-up induction of Rules And Cases for Imbalanced Data) is the most accurate of these algorithms,

To handle the data difficulty factors, the authors of BRACID [26] decided to use a *hybrid representation of rules and single instances*, where more general rules cover larger, homogeneous regions with more examples and instances should handle non-linear class borders and rare minority cases or outliers. The rules are induced in a special *Bottom-up rule sequential process*. It starts from the set of the most specific rule (single, seed learning examples) and in the next iteration it tries to generalize its condition in the direction of the nearest neighbour example from the same class, provided that it does not decrease the classification abilities of the whole rule set evaluated with measures specific for imbalanced data.

An exploitation of *types of difficulty* of learning examples estimated by analysis the k-nearest neighborhood of seed examples is one of the main features of BRACID. The difficult type [27] assigned to each seed example influences the rule generalization, as for the unsafe minority example it is possible to generate additional rules covering it. As a result, the number of minority class rules, as well as their support, are increased and they are more likely to win with the stronger majority rules while classifying new instances. For details see [26].

7.2 Generalizing BRACID for Multiple Imbalanced Classes

As BRACID was proposed for binary classes only. In a recent paper [25] two ways of its generalizations for multiple classes were studied: (1) exploitation of binary decomposition ensemble frameworks OVO and using the original binary BRACID within them; (2) generalization of BRACID with a new scheme for inducing a single set of rules from all multiple classes.

The second generalization partly follows a typical sequential schema for an iterative induction of rules from successive classes. In each iteration, for each class the temporary training dataset is constructed. It contains positive examples from the considered class and the negative examples from all other classes (similar to the OVA approach). BRACID algorithm is run on such data and only rules describing the considered class are added to the final set of rules, while the other class rules are discarded. At the end, the complete set contains rules from all classes. An important modification of rule generalization takes into types of classes, i.e. whether the positive class is a minority or majority one. More precisely when the majority class is considered then (1) the internal k-nearest neighbor generalization is done to a single nearest example for *safe* seed examples and (2) to one, best of rules induced by generalization to k nearest examples for

Table 3. Comparison of rule classifiers – PART, OVO Bracid and multiple BRACID – with respect to G-mean for selected real-world data sets.

Dataset	PART	OVO-B	m-BRACID
Balancescale	0.3136	0.4086	0.6789
Car	0.7925	0.9022	0.9004
Cleveland	0.0597	0.1750	0.2322
cmc	0.4431	0.4897	0.4691
Dermatology	0.8943	0.9204	0.9082
Ecoli	0.6373	0.7404	0.7976
Flare	0.1716	0.3904	0.4639
Glass	0.3174	0.1883	0.4256
Led7digit	0.7918	0.7736	0.7713
Vehicle	0.9147	0.9221	0.9323
Winequality	0.2917	0.4529	0.5338

unsafe example. For the minority class it is unchanged. This modification limits the number of produced rules for majority classes.

In [25] experiments on similar multi-class datasets as [11] were done. Their results show that this generalization of BRACID is better than the adaptive using of the binary BRACID within OVO ensemble, both with respect to higher predictive abilities and the number of rules. Some of these comparative results are shown in Table 3. In case of producing still too many rules they can be post-pruned with the special weighted coverage algorithm [28].

8 Multi-class Extension of Bagging Ensemble

Generalizations of bagging ensembles are quite effective for binary imbalanced data. Lango et al. studied in [23] *Roughly Balanced Bagging*, which is one of the most efficient under-sampling bagging for binary imbalanced classes and it often works better than generalizations of boosting. It exploits a random *under-sampling* before generating component classifiers, which reduces the presence of the majority class examples inside each bootstrap sample of the finally constructed bagging. The random number of majority examples to be sampled to the bootstrap is estimated according to the negative *binomial distribution*, while the number of sampled minority examples is equal to the size of the minority class inside the original training dataset. Finally, these numbers of examples are sampled from each class with replacement and predictions of the learned based classifiers are aggregated with the majority voting.

Lango and Stefanowski proposed in [23] its generalization to Multi-class Roughly Balanced Bagging (further abbreviated as MRBBag). The main modification concerns a construction of bootstrap samples, which is realized in the following way. The number of examples to be sampled from each class to the

bootstrap is estimated from the multinomial distribution, which is defined by the following probability mass function:

$$p(n_1, n_2, ..., n_c) = \frac{n!}{n_1!n_2! \cdots n_c!} p_1^{n_1} p_2^{n_2} \cdots p_c^{n_c}$$

where $p_1, p_2, ..., p_c$ and $n = \sum_{i=1}^{c} n_i$ are the parameters of the distribution.

The authors handle the multi-class imbalance problem by obtaining roughly balanced bootstrap samples also with respect to class probabilities, so they fix values $p_1, p_2, ..., p_c$ to the same constant value equal to $\frac{1}{c}$, such that $\sum_{i=1}^{c} p_i = 1$. This parameterizes the upper formula. After learning component, base classifier the final decision of the classifier is constructed by the majority voting. For the pseudocode of this algorithm see [23].

In [23] MRBBag algorithm, constructed with J4.8 trees, was evaluated on several artificial and UCI real-life imbalanced datasets. It outperformed other tree and general ensemble classifiers with respect to G-mean and averaged F1-score (both adapted for the multi-class evaluation). Refer also to its performance in Table 2. Moreover MRBBag was further extended to deal with feature selection for highly dimensional data, see details in [23]. This variant was successfully applied to solve the task of categorization of twitter short text messages [20].

9 Software Implementations of Specialized Algorithms for Multi-class Imbalanced Data

The methods for dealing with binary imbalanced data are already implemented in various software libraries. The representatives are: `imbalanced-learn` with scikit-learn in Python, KEEL, WEKA and its extensions in Java or several R libraries such as IRIC or ClimbR. In case of methods for multi-class imbalanced data there are nearly no public available software implementations. In the past year two open source software kits were proposed: `multi-imbalance` Python library [9], Matlab toolkit Multiple-imbalance [34].

The first library was developed by the author's co-operators and it implements state-of-the-art approaches for multi-class imbalanced problems, which are divided into three general categories: (1) binary decomposition approaches (OVO, OVA and ECOC), (2) specialized pre-processing (Global-CS, Mahalanobis Distance Oversampling (MDO), Static-SMOTE, SPIDER3 and SOUP), and (3) other ensembles (MRBagging and SOUP-bagging). So it covers methods discussed in this paper.

On the other hand, the Matlab toolkit contains 18 methods, mainly variants of Adaboost or ECOC and specialized tree classifiers.

10 Future Research Directions and Conclusions

Looking at the current literature, we could expect that many new methods will still be proposed to improve the classification of imbalanced data, including multi-class variants. It is hoped, however, that these expected proposals

will go beyond simple adaptations of known approaches or exploitation of the binary decomposition frameworks, and in particular they will take advantage of the aforementioned data difficulty factors. Below some personal opinions are expressed as to the research directions.

Cost sensitive learning for multiple classes should estimate misclassification costs for each example also with their difficulty levels. The current proposals are too much oriented to global imbalance ratios.

In terms of further research on data difficulties, it is necessary to more carefully explore the differences in the impact of overlapping between different class types and in the context of different class size configurations. In particular, this applies to a more detailed analysis of the intermediate classes in the so-called gradual configurations that appear to be more difficult than configurations with sharp changes of class size between minority and majority ones.

In particular the role of rare examples for many classes, which previously had a large impact on deteriorating the classification of imbalanced binary data, has not been sufficiently studied for multiple classes yet.

New preprocessing methods should be developed for better dealing with overlapping between various classes as they are more critical than in the binary problems. It should also be assessed to what extent changes (e.g. by resampling) in the size of overlapping classes will affect the recognition of other classes.

Other approaches for discovery sub-concepts in multiple classes could be still studied, in particular with exploiting density based clustering.

Deeper research on specialized artificial neural networks should be undertaken. The current few studies are too focused on including random re-sampling or relatively simple modifications of the optimized loss function. This is desirable given the current strong interest in image recognition or natural language processing using deep neural networks.

An open question concerns multi-class and highly dimensional datasets. Feature random sampling does not take into account internal relations between classes. Furthermore, more research is needed on the specialized construction of new features, projections of the original ones into new representation space, like in embedded representations in deep networks or similarity learning.

Nearly all current research were done on static multi-class imbalanced data. On the other hand data streams with concept drifts occur in many modern applications of Big Data [12]. They are naturally imbalanced and the global imbalance ratio may vary over time. However, the data factors such as class split into factors, overlapping or presence of rare case may also change (similarly to typical drifts) and their drifts are definitely local as it was recently shown for binary imbalanced streams [3]. Their experiments demonstrated that these drifts deteriorate predictive performance of popular stream classifiers and posed needs for the developments of new specialized online algorithms. However such studies should be done with more complex multiple classifiers. Furthermore new online clustering algorithms for detection of the class split, their changes over time and appearance of new classes in the streams are necessary.

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. Agrawal, A., Herna, L.V., Paquet, E.: SCUT: multi-class imbalanced data classification using SMOTE and cluster-based undersampling. In: International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K), vol. 01, pp. 226–234 (2015)
3. Brzezinski, D., Minku, L.L., Pewinski, T., Stefanowski, J., Szumaczuk, A.: The impact of data difficulty factors on classification of imbalanced and concept drifting data streams. Knowl. Inf. Syst. **63**(6), 1429–1469 (2021)
4. Buda, M., Maki, A., Mazurowski, M.A.: A systematic study of the class imbalance problem in convolutional neural networks. CoRR abs/1710.05381 (2017)
5. Fernández, A., García, S., Galar, M., Prati, R.C., Krawczyk, B., Herrera, F.: Learning from Imbalanced Data Sets. Springer, Heidelberg (2018)
6. Fernández, A., López, V., Galar, M., Del Jesus, 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)
7. Galar, M., Fernández, A., Barrenechea, E., Sola, H., Herrera, F.: An overview of ensemble methods for binary classifiers in multi-class problems: experimental study on one-vs-one and one-vs-all schemes. Pattern Recogn. **44**, 1761–1776 (2011)
8. Garcia, V., Sanchez, J., Mollineda, R.: An empirical study of the behaviour of classifiers on imbalanced and overlapped data sets. In: Rueda, L., Mery, D., Kittler, J. (eds) Progress in Pattern Recognition, Image Analysis and Applications. CIARP 2007. Lecture Notes in Computer Science, **4756**, 397–406. Springer, Berlin, Heidelberg (2007). https://doi.org/10.1007/978-3-540-76725-1_42
9. Grycza, J., Horna, D., Klimczak, H., Lango, M., Plucinski, K., Stefanowski, J.: multi-imbalance: open source python toolbox for multi-class imbalanced classification. In: Dong, Y., Ifrim, G., Mladenić, D., Saunders C., Van Hoecke S. (eds) Machine Learning and Knowledge Discovery in Databases. Applied Data Science and Demo Track - European Conference, ECML PKDD, Proceedings, Part V. Lecture Notes in Computer Science, **12461**, 546–549. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-67670-4_36
10. He, H., Ma, Y.: Imbalanced Learning: Foundations, Algorithms, and Applications. Wiley-IEEE Press, Hoboken (2013)
11. Janicka, M., Lango, M., Stefanowski, J.: Using information on class interrelations to improve classification of multiclass imbalanced data: a new resampling algorithm. Int. J. Appl. Math. Comput. Sci. **29**, 769–781 (2019)
12. Japkowicz, N., Stefanowski, J.: A machine learning perspective on big data analysis. In: Japkowicz, N., Stefanowski, J. (eds) Big Data Analysis: New Algorithms for a New Society. Studies in Big Data, **16**, 1–31. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-26989-4_1
13. Japkowicz, N., Stephen, S.: The class imbalance problem: a systematic study. Intell. Data Anal. **6**(5), 429–449 (2002)
14. Jelonek, J., Stefanowski, J.: Experiments on solving multiclass learning problems by n2-classifier. In: Nédellec, C., Rouveirol, C. (eds) Machine Learning: ECML-1998. ECML 1998. Lecture Notes in Computer Science (Lecture Notes in Artificial Intelligence), **1398**. LNCS(LNAI), 172–177. Springer, Berlin, Heidelberg (1998). https://doi.org/10.1007/BFb0026687

15. Jo, T., Japkowicz, N.: Class imbalances versus small disjuncts. SIGKDD Explor. **6**(1), 40–49 (2004)
16. Kocur, Z.: Clustering algorithm for multi-class imbalanced data to improve classification quality. Ph.D. thesis, Poznan University of Technology (2020)
17. Krawczyk, B.: Learning from imbalanced data: open challenges and future directions. Prog. Artif. Intell. **5**(4), 221–232 (2016)
18. Kubat, M., Matwin, S.: Addressing the curse of imbalanced training sets: one-side selection. In: Proceedings of the 14th International Conference on Machine Learning ICML-1997, pp. 179–186 (1997)
19. Kuncheva, L.: Combining Pattern Classifiers. Methods and Algorithms, 2nd edn. Wiley, Hoboken (2014)
20. Lango, M.: Tackling the problem of class imbalance in multi-class sentiment classification: an experimental study. Found. Comput. Decis. Sci. **44**, 151–178 (2019)
21. Lango, M., Brzezinski, D., Firlik, S., Stefanowski, J.: Discovering minority subclusters and local difficulty factors from imbalanced data. In: Discovery Science - 20th International Conference, DS 2017, Proceedings, pp. 324–339 (2017)
22. Lango, M., Napierała, K., Stefanowski, J.: Evaluating difficulty of multi-class imbalanced data. In: Proceedings of 23rd International Symposium on Methodologies for Intelligent Systems, pp. 312–322 (2017)
23. Lango, M., Stefanowski, J.: Multi-class and feature selection extensions of roughly balanced bagging for imbalanced data. J. Intell. Inf. Syst. **50**(1), 97–127 (2018)
24. Lango, M., Stefanowski, J.: What makes multi-class imbalanced problems difficult? (2021). (manuscript under review)
25. Naklicka, M., Stefanowski, J.: Two ways of extending Bracid rule-based classifiers for multi-class imbalanced data. In: Nuno, M., Paula, B., Luis, T., Nathalie, J., Michal, W., Shuo, W. (eds) Proceedings of the Third International Workshop on Learning with Imbalanced Domains: Theory and Applications, co-located with ECML–PKDD 2012, Proceedings of Machine Learning Research (2021)
26. Napierala, K., Stefanowski, J.: BRACID: a comprehensive approach to learning rules from imbalanced data. J. Intell. Inf. Syst. **39**(2), 335–373 (2012)
27. Napierała, K., Stefanowski, J.: Types of minority class examples and their influence on learning classifiers from imbalanced data. J. Intell. Inf. Syst. **46**, 563–597 (2016)
28. Napierala, K., Stefanowski, J., Szczech, I.: Increasing the interpretability of rules induced from imbalanced data by using Bayesian confirmation measures. In: Appice, A., Ceci, M., Loglisci, C., Masciari, E., Raś, Z. (eds) New Frontiers in Mining Complex Patterns. NFMCP 2016. Lecture Notes in Computer Science, **10312**84–98. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-61461-8_6
29. Prati, R., Batista, G., Monard, M.: Class imbalance versus class overlapping: an analysis of a learning system behavior. In: Proceedings of 3rd Mexican International Conference on Artificial Intelligence, pp. 312–321 (2004)
30. Seaz, J., Krawczyk, B., Wozniak, M.: Analyzing the oversampling of different classes and types in multi-class imbalanced data. Pattern Recogn. **57**, 164–178 (2016)
31. Stefanowski, J.: Dealing with data difficulty factors while learning from imbalanced data. In: Mielniczuk, J., Matwin, S. (eds.) Challenges in Computational Statistics and Data Mining, **605**, 333–363. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-18781-5_17
32. Wang, S., Yao, X.: Mutliclass imbalance problems: analysis and and potential solutions. IEEE Trans System Man Cybern. Part B. **42**(4), 1119–1130 (2012)

33. Wojciechowski, S., Wilk, S., Stefanowski, J.: An algorithm for selective preprocessing of multi-class imbalanced data. In: Proceedings of the 10th International Conference on Computer Recognition Systems CORES 2017, Polanica Zdroj. Advances in Intelligent Systems and Computing, vol. 578, pp. 238–247 (2017)
34. Zhang, C., Bi, J., Xu, S., Ramentol, E., Fan, G., Qiao, B., Fujita, H.: Multi-imbalance: an open-source software for multi-class imbalance learning. Knowl. Based Syst. **174**, 137–143 (2019)
35. Zhou, Z.H., Liu, X.Y.: On multi-class cost sensitive learning. Comput. Intell. **26**(3), 232–257 (2010)