Two Ways of Extending BRACID Rule-based Classifiers for Multi-class Imbalanced Data

Maria Naklicka MCNAKLICKA@GMAIL.COM
F33, Walnut Creek, California, United States of America

Jerzy Stefanowski JSTEFANOWSKI@CS.PUT.POZNAN.PL
Institute of Computing Science, Poznan University of Technology, 60-965 Poznań, Poland

Editors: Nuno Moniz, Paula Branco, Luís Torgo, Nathalie Japkowicz, Michał Wozniak and Shuo Wang.

Abstract

The number of rule-based classifiers specialized for imbalanced data is quite small so far. In particular, there is no such classifier dedicated for multi-class imbalance data. Thus, in this work we considered two ways of extending BRACID, which is the effective algorithm for binary data. In the first approach, BRACID was used in the OVO ensemble along with modifications of the prediction aggregation strategy. The second approach modifies an induction of rules for multiple classes simultaneously, additionally combined with their post-pruning. Experiments showed that both approaches outperformed the baselines. Moreover, the second approach turned out to be better than OVO with respect to predictive results and producing a smaller number of rules.

Keywords: Multiple imbalanced data, rule classifier, binary decomposition ensembles

1. Introduction

Despite the development of many methods for improving classification of imbalanced data, still little interest has been paid to research on induction of rules. Recall that learning classification rules is one of classical and well - studied tasks in machine learning and data mining (Fürnkranz et al., 2012). As the rules directly provide a symbolic representation of knowledge discovered from data, they are considered to be more comprehensible and human readable than other learnt models. In particular single rules could be individually analyzed as the local patterns corresponding to specific learning examples, which is exploited in the descriptive perspective in data mining. Furthermore, rules are particularly useful for explanation of classifiers' predictions, which is important in explainable AI.

The standard rule classifiers are quite sensitive to imbalanced data, see e.g. experiments in (Grzymala-Busse et al., 2005). Their poor performance could be attributed both to the algorithmic limitations and data difficulty factors. A fairly small number of rule classifiers specialized for imbalanced data has been introduced so far, for their review see e.g. (Napierala and Stefanowski, 2012). However, the most of them address only single algorithmic or data factors. In (Napierala and Stefanowski, 2012) authors introduced a new algorithm, called BRACID (the acronym of Bottom-up induction of Rules And Cases for Imbalanced Data), which attempts to handle more of these factors and deals with complexity of imbalanced data in a more comprehensive way. According to its experimental evaluation BRACID outperformed several rule classifiers specialized for imbalanced tasks and standard ones integrated with resampling techniques (Napierała and Stefanowski, 2016). BRACID was further extended by rule post-pruning to select smaller sets of rules having both good predictive and descriptive characteristics (Napierala et al., 2016).

Nevertheless BRACID is designed for working with binary imbalanced classes (single minority vs. majority class) only. Note that recently, more and more researchers began to take interest in studying imbalanced data with multiple classes (Fernández et al., 2018; Krawczyk et al., 2015). These problems are generally assessed as more difficult than the binary ones (Krawczyk, 2016). As simple adaptations of methods designed for binary imbalanced problems do not lead to satisfactory improving the recognition of multiple classes simultaneously, new more specialized solutions are needed.

Therefore, in this paper we will consider a still open question how to generalize BRACID to deal with *multi-class imbalanced data*. We will study two ways: (1) Exploitation of binary decomposition ensemble frameworks; (2) Generalization of BRACID with a new scheme for inducing a combined set of rules from multiple classes.

The first direction follows inspirations from current literature on the decomposition of multi-class tasks into a set of binary ones and solving them in an ensemble framework of binary classifiers, such as *one against one class* (Galar et al., 2011). Here BRACID could be used in an original form, however techniques for the final aggregation of votes should be studied. The second direction will follow the sequential strategy of inducing rules from all, multiple classes and requires new solutions taking into account the types of classes.

The main aims of our paper are: to present the formulations of both ways to generalize BRACID for dealing with multi-class imbalanced data; to carry out their experimental evaluation and comparison over 20 benchmark datasets. In particular we are interested in checking which of these ways is better. Finally, their best performing variants will be compared against baseline multi-class rule based classifiers.

2. Rule Induction from binary imbalanced data with BRACID

2.1. BRACID algorithm

Following (Napierala and Stefanowski, 2012) BRACID exploits a **hybrid representation of rules and single instances**, where more general rules cover larger, homogeneous data regions (with more examples) and instances should handle non-linear class borders and rare cases or outliers. Rules are represented as: IF $(conditionpart)$ $THEN$ $(targetclass)$, where a condition part is a conjunction of attribute based conditions. For nominal attributes the conditions are formed as $(x_i = v_{ij})$, where x_i represents i-th attribute and v_{ij} is a single value from its domain. Conditions for numeric attributes are represented as closed intervals $(v_{i,lower} \leq x_i \leq v_{i,upper})$, where $v_{i,lower} \leq v_{i,upper}$ are values belonging to the domain of the attribute. In BRACID, instances are treated as maximally specific rules.

The rules are induced in a special **bottom-up rule induction** strategy. The algorithm starts from the set of the most specific rule (single, *seed learning example*) and in the next iteration it tries to generalize its conditions 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. The distance between rules and examples is calculated with the HVDM

(Heterogeneous Value Distance Metric) distance metric (Wilson and Martinez, 1997)[1]. The procedure is repeated until no rule can be further generalized. The candidate rules are evaluated with the appropriate measures for imbalanced data, either F-score or G-mean. They are estimated by a special effective variant of the internal leaving one out procedure (Napierala and Stefanowski, 2012).

Furthermore, the **type of difficulty of learning examples** is evaluated with the method (Napierała and Stefanowski, 2016)[2], which leads to labeling them as safe or unsafe examples. The generalization of rules depends on the *type of the seed example*. The minority examples tagged as *unsafe* ones (i.e., belonging to the class overlapping region or being rare cases inside the majority class) are allowed to be generalized into more than one rule, which lessen the dominance of the majority class in their neighbourhood. On the other hand, single majority examples present inside the minority class regions are treated as noise which may hinder the induction of general minority rules. BRACID has an embedded mechanism for detecting and removing them from the learning data set.

Finally, BRACID employs a **less biased classification strategy** on new instances based on the *nearest rules* to diminish the domination of strong majority rules during solving conflict situations while a new instance matches condition parts of several rules. The distance between rules and the classified instance is calculated with the same HVDM metric as in the learning phase. If more than one rule is equally distant from the instance, the sum of *rule supports*[3] is calculated and the instance is assigned to the class with the largest total sum of supports. Due to the increased number of rules for the minority class, coupled with their higher supports, minority rules are more likely to win with a more general majority rules in difficult regions of the minority class in the data.

2.2. Weighted rule coverage post-pruning

Although BRACID is an accurate classifier, it can produce quite a large number of rules. Therefore, the authors of BRACID added in (Napierała and Stefanowski, 2016) a special post-pruning strategy which aims at selecting sufficiently strong and diversified rules, i.e. covering diverse subsets of examples corresponding to sub-regions of each class.

In order to identify the rules to be selected, a product aggregation of two following rule evaluation measures was considered: *rule support* and *Bayesian confirmation measure* N (Napierala et al., 2016). The confirmation measures evaluate differences between the conditional probability of rule conclusion C_k given its satisfied condition part E and the prior probabilityof the rule conclusion class. It should be higher than zero. N measure is defined as $P(C_k|E) - P(C_k|\neg E)$. Following (Napierala et al., 2016) this measure is useful for imbalanced data.

[1] It aggregates both numeric and nominal attributes using different distance functions. The normalized Euclidean distance is applied to numeric attributes, while the distance between values of nominal attributes is defined with the differences of their class conditional probabilities.

[2] The method is based on analysing labels of k nearest neighbours of the example x. The fraction of n_{Cx}, i.e. number of neighbours with the same class label as x, to all k neighbours is the *safe level* of x. If its value is higher than 0.67 the example is treated as safe, otherwise it is considered an unsafe one. For more details of studying it see (Napierała and Stefanowski, 2016).

[3] The rule support is a number of learning examples covered by the rule.

Algorithm 1: The weighted coverage algorithm. Input: S - the rule set for the given class C_k and NR the required number of rules. Output: FR pruned set of rules from class C_k.

```
1. Delete rules with too low confirmation measure from $S$ and set $FR = \emptyset$
2. For each example $x \in C_k$ do $c(x)$= 1
3. Repeat
    (a) For each rule $R \in S$ evaluate($R$)
    (b) Select $R_b$ with the best value of evaluate($R$)
    (c) For each example $e$ covered by $R_b$ do $c(x) = c(x)$ +1
    (d) Remove $R_b$ from $S$ and $FR = FR \cup S$
4. until size of rule set $FR = NR$
```

The rules are selected by a *weighted coverage algorithm* in an iterative way focusing on examples still not covered by previous rules. Its pseudocode is presented in Algorithm 1. Note that it is run for each class separately. Firstly, it initializes each example weight $c(x)$ to 1 and calculates both rule evaluation measures with the inverse weights (e.g., the rule support is a sum of $1/c(x)$ for examples covered by this rule). As $c(x)$ of all examples covered by this rule are incremented by an unit, they contribute less to rule weights and evaluation measures in the subsequent iterations. It promotes rules covering examples from other subregions (still not covered) and follows some inspirations from (Gamberger and Lavrac, 2002). The algorithm is run until covering a sufficient percentage of learning examples or producing a required number of rules NR.

The experiments presented in (Napierała and Stefanowski, 2016; Napierala et al., 2016) showed that this algorithm could strongly reduce the number of rules in binary imbalanced data (leaving between $10\% - 50\%$ of the original BRACID rule sets) and improve at the same time the average values of interestingness measures for selected rules, without considerable decreases of predictive abilities (with respect to the G-mean, F1-score or sensitivity).

3. Binary decomposition techniques for multi-class data

The number of methods for multiple imbalanced classes is much smaller than designed for the binary problem. For their comprehensive review, see e.g., (Fernández et al., 2018). The most popular approaches are decomposition ones transforming the multi-class problem into a series of binary problems and solving them with the ensemble framework. Such methods have initially been proposed for standard classification tasks to enable binary classifiers to handle multi-class problems (Kuncheva, 2014). The most related frameworks are:

One-versus-all ensemble (OVA) constructs binary classifiers to recognize a particular class against the remaining ones aggregated into one class (Galar et al., 2011). During the 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, also called pairwise coupling) constructs binary classifiers for all pairs of classes (for C classes, the number of base classifiers is $C(C\text{-}1)/2$). The training set for a particular base classifier contains learning examples from the selected pair of classes only. Note that binary problems usually have smaller imbalance ratios than the original problem and could be easier for learning. The final decision for the new instance can be taken by majority voting of base classifiers' predictions, weighted voting with confidence scores, or more complex aggregation functions(Kuncheva, 2014).

Since these approaches train classifiers on binary problems, they can be easily used in combination with the standard techniques for binary imbalanced data. OVO and OVA are often used with oversampling or undersampling approaches applied on each binary training set (Fernández et al., 2013). Experiments of (Galar et al., 2011) with several multi-class data demonstrated that OVO outperformed OVA ensemble, which guided our further approaches.

4. Extending BRACID for multi-class imbalanced data

We will present two ways of extending BRACID for multi-class imbalanced data. The first one follows one versus one (OVO) framework while keeping binary BRACID. The other, called multi-class BRACID requires more modifications within the rule generation schema. In both ways, we assume that it is indicated which classes are minority and which are majority ones.

4.1. Exploiting OVO framework

As BRACID in its original formulation generates rules from two classes only, we have decided to adapt OVO framework following inspirations from related works and consider all pairs of classes in the given dataset.

4.1.1. Aggregation of base classifier predictions

Firstly, we consider BRACID in its original form to generate binary classifiers in OVO and to aggreaget their predictions with the standard majority voting, i.e. the class most often indicated by the base classifiers was assigned to the classified instance - this variant is abbreviated as **MV** in further experiments.Then, we extend the prediction aggregation phase by two different solutions:

1. Prediction Weighted Voting: Weights are assigned to each induced BRACID classifier depending on the evaluation of its predictive accuracy. They are real numbers in the range [0;1] and are used as votes in the final voting for the classified instances. Their values are estimated when using the given classifier to re-classify examples from the corresponding pair of classes in the dataset. Due to class imbalance, the predictive ability of the binary classifiers in OVO is evaluated by the standard measures for classification of imbalanced data, i.e. either by F1-score or by G-mean (of sensitivity and specificity). These variants of weighted voting will be further denoted as **WVF** or **WVG**, respectively.

2. Competent Classifier Selection: The new instance will be classified by only these base classifiers in OVO which were trained on examples from classes related to this instance and skipping other classifiers in the voting. This dynamic selection of competent classifiers is based on the analysis of the neighbourhood of the classified instance in the original dataset. If neither of classes for which a classifier was trained belongs to this example's neighbourhood, then such a classifier is not taken into account in the aggregation. Here, we follow proposals of (Galar et al., 2013), where the authors recommended to use the neighbourhood size equal to $3C$ examples, where C denotes the number of classes in the dataset. If all neighbours belong to the same class, the neighbourhood of double size ($6C$) is additionally analyzed. The approach will be further denoted as **CCS**.

4.1.2. Modifying rule generalization in BRACID

Recall that the original BRACID treats minority and majority classes differently during the generalization from the seed example to rules. For *safe* majority examples, only one generalization to a single nearest example is considered. For *unsafe* majority examples and *safe* minority examples, k nearest neighbours are analysed and only one best generated rule is chosen. For *unsafe* minority examples the generalization to k nearest neighbours is carried and all rules are added, if they improve the evaluation measure for the candidate rule set (Napierala and Stefanowski, 2012). As a result it may lead to increasing the number of minority rules and improving their support.

This solution worked effectively for binary imbalanced data, where one class was minority and the other majority. However, other combinations of types of classes (e.g. majority vs. majority) occur in OVO and it should affect the induction of rules without excessively increasing their number for the majority classes. This is why we modify this rule generalization with respect to types of classes considered for a particular binary classifier. More precisely,

- If **two majority classes** are considered, then (1) the generalization is done to a single nearest example for the *safe* seed examples and (2) to one, best of rules induced by generalization to k nearest examples for an *unsafe* example.
- For **two minority classes** following generalizations are applied: to one, best of rules induced by generalization to k nearest examples – for the *safe* seed examples, and up to k best rules examples induced by generalization to k nearest examples – for the *unsafe* examples.

4.2. Extending BRACID to sequential inducing rules from many classes

This extension of BRACID constructs a single classifier based on the combined sets of rules from all classes. It partly follows a sequential class covering schema for an iterative induction of rules from successive classes (Fürnkranz et al., 2012). The number of iterations is equal to the number of classes. In each iteration, the temporary training dataset is constructed. It contains positive examples from the class considered in this iteration and the negative examples from all other classes. BRACID algorithm is run on these examples and only rules describing the considered positive class are added to the final set of rules, while the other class rules are discarded. This is repeated for each class in the input data and the final set contains rules from all classes.

The next important modification of the rule generalization takes into account the type of the class considered in the given iteration. If the class is treated as a majority one, the rule generalization from seeds majority examples is modified as it was described in the previous subsection 4.1.2. The rule generalization from the minority class is done as in the original BRACID. This modification limits the number of produced rules from majority classes. If the number of rules is still too high, then the weighted coverage algorithm could be used to post-prune them - see Section 2.2.

The classification strategy is based on matching the instance description to the nearest rules as it is done in the original BRACID algorithm (Napierala and Stefanowski, 2012).

5. Experiments

5.1. Aims and experimental setup

The experiments are planned to check the following points:

- Which way of generalizing BRACID for multi-class imbalanced data is more effective?
- The usefulness of different extensions of OVO frameworks with BRACID.
- The impact of rule post-pruning in multi-class BRACID.
- The comparison of best variants of both ways of the generalization of BRACID against standard rule classifiers for multi-class data.

All experiments are carried out on 20 popular multi-class imbalanced datasets, which were previously studied in (Janicka et al., 2019). Nearly all of them are benchmark data from UCI ML Repository, which were also considered in many earlier experiments concerning methods for multi-class imbalanced data, see e.g. (Fernández et al., 2013). They differ in terms of the overall number of classes, the proportion of minority vs majority classes, the number of attributes, the data imbalance ratios between classes and data difficulty (measured by the average safe level). For more details on the data characteristics and information which classes are identified as minority ones see the online appendix [4]. Following (Janicka et al., 2019) the minority classes are those that do not contain more examples than 25% of the size of the entire dataset.

BRACID algorithm is always run with the following parameters: k=5 neigjbourhood for estimating safe or unsafe types of examples; F1-score is used as the internal measure to evaluate rules generalizations.

Predictive performance of all classifiers is evaluated by the following measures specialized for multiple classes: *Class sensitivities* (TPR or Recalls) and their aggregation with *G-mean*; *Average minority* – i.e. an arithmetic mean of sensitivities of the minority classes; *Mean Minority F-measure* – i.e. an arithmetic means of F1 scores for the minority classes. The values of all measures were estimated by 5 runs of 10 fold stratified cross validation.

5.2. Studying extensions of OVO based framework

Firstly, we compared the standard OVO frameworks (with the basic BRACID and the majority voting) and considered extensions of the aggregation of classifier predictions (weighted WVF and WVG + competent classifiers selection CCS). We also studied the usefulness of

[4] Due to page limits we present information on datasets and detailed results of experiments in an online appendix http://www.cs.put.poznan.pl/jstefanowski/pub/append-bracid.pdf

Table 1: G-mean of rule based classifiers obtained by all OVO extensions.

dataset	**MV**	**WVF**	**WVG**	**MB**	**MB**	**MV**	**CCS**
				MV	**WVF**	**WVG**	
art1	0.9675	0.9675	0.96747	0.9675	0.9675	0.9675	0.9675
art2	0.7590	**0.7596**	**0.7596**	0.7575	0.7591	0.7582	**0.7596**
art3	0.5180	0.5186	0.5175	0.5205	**0.5220**	**0.5220**	0.5175
art4	0.8317	0.8317	0.8317	0.8317	0.8317	0.8317	0.8317
autos	0.6851	0.6892	0.6881	0.6779	0.6757	0.6762	**0.6921**
balance-scale	0.4076	0.4086	0.4086	0.4099	**0.4114**	**0.4114**	0.3208
car	**0.9102**	0.9022	0.9090	0.8707	0.9012	0.9067	0.9048
cleveland	0.1716	0.1750	0.1759	0.1187	0.1342	0.1342	**0.1865**
cmc	0.5009	0.4987	0.4994	0.4969	0.4959	0.4946	**0.5022**
dermatology	0.9471	0.9204	0.9204	0.9572	0.9578	0.9578	**0.9580**
ecoli	0.7641	0.7404	0.7403	0.7405	0.7430	0.7593	**0.7725**
flare	0.4484	0.3904	0.1374	0.4958	0.5003	**0.5017**	0.1500
glass	0.0470	0.0883	0.0873	0.0728	**0.1114**	0.0979	0.0963
hayes-roth	0.7837	**0.7849**	**0.7849**	0.7784	0.7785	0.7785	0.7760
led7digit	0.7690	**0.7736**	0.7592	0.7664	0.7701	0.7557	0.7217
nursery	**0.9797**	0.9796	0.9796	0.9795	0.9794	0.9794	0.9796
new-thyroid	0.9684	0.9684	0.9684	0.9684	0.9684	0.9684	0.9684
vehicle	0.9218	0.9222	0.9222	0.9222	0.9222	0.9222	**0.9241**
winequal-red	0.4472	0.4530	0.4729	0.4653	0.4933	**0.5120**	0.4434
yeast	0.0531	**0.0808**	0.0581	0.0556	0.0565	0.0560	0.0673

modifying rule generalization in BRACID (abbreviated as MB). The results of G-mean for all variants of OVO classifiers are presented in Table 1, where bold fonts indicate the best result for a given datasets[5].

Note that there is no clear winning variant, depending on the particular datasets some extensions lead to slightly higher values. Moreover, for some datasets the differences in the results are almost imperceptible. This is supported by performing the Friedman's ranked tests. For G-mean the average ranks of OVO classifiers (the lower, the better) are following: MV 4.24; WVF 3.77; WVG 4.09; MB MV 4.88; MB WVF 3.71; MB WVG 3.82 and CCS 3.53. However, due to high p-values $= 0.602$ the null hypothesis on similar performance of multiple classifiers cannot be rejected. Average ranks calculated with Average minority are the following: MV 4.73; WVF 3.05; WVG 3.5; MB MV 5.4; MB WVF 3.3; MB WVG 3.53 and CCS 4.43 with p-values $= 0.021$. Therefore the null hypothesis is rejected, but the critical difference $CD = 2.12$ show that the differences between the best classifier variants are insignificant. Quite similar rankings of OVO variants are obtained for F1-score.

We interpret these results that considered extensions of OVO aggregation or modifications of BRACID rules generalization are slightly better than standard OVO MV version. However, these extensions perform quite similarly with some difference with respect to

[5] Results of other measures are provided in the online appendix.

Table 2: G-mean results obtained by modified BRACID with complete rule set without pruning (CRS) and its extensions with various pruning - the percentage of discarded rules is presented as $p\%$.

dataset	**CRS**	**PR CONF**	**PR 30%**	**PR 50%**	**PR 70%**	**PR 90%**
art1	**0.9688**	0.9678	0.9677	0.9640	0.9563	0.9363
art2	0.7555	0.7500	0.7521	0.7591	0.7618	**0.8012**
art3	0.5298	0.5191	0.5137	0.5184	0.5253	**0.5752**
art4	0.8353	0.8258	0.8333	0.8285	0.8420	**0.8818**
autos	**0.6791**	0.6457	0.6460	0.5932	0.5801	0.4369
balance-scale	**0.5111**	0.3457	0.2002	0.1078	0.1692	0.3212
car	0.9000	0.9008	0.8977	0.8842	**0.9108**	0.8212
cleveland	0.1367	0.1857	0.1452	0.1683	0.1861	**0.2322**
cmc	0.4681	0.4676	0.4699	0.4699	0.4723	**0.4830**
dermatology	0.9082	**0.9120**	0.9082	0.9071	0.9040	0.8803
ecoli	0.7876	0.7908	0.7940	**0.7951**	0.7868	0.7814
flare	0.4631	**0.4661**	0.4639	0.4340	0.4132	0.3269
glass	0.4053	0.3725	0.4156	**0.4517**	0.3720	0.2473
hayes-roth	**0.7401**	0.7371	0.7257	0.6944	0.6680	0.4605
led7digit	0.7209	0.7669	0.7771	0.7839	**0.7861**	0.7716
nursery	0.9754	**0.9755**	0.9710	0.9566	0.9149	0.8410
new-thyroid	**0.9758**	0.9559	0.9706	0.9681	0.9358	0.9551
vehicle	0.9232	**0.9260**	0.9240	0.9235	0.9212	0.9084
winequality-red	**0.5361**	0.5027	0.5338	0.5156	0.5132	0.4901
yeast	**0.0758**	0.0544	0.0758	0.0544	0.0541	0.0672

evaluation measures. If one focus on improving the recognition of the minority classes (F-measure and Average minority sensitivities) Weighted Voting with F-measure is the first in average ranking (there is no big difference if BRACID is used with modified generalization or not). On the other hand, if one also analyse improvements for the majority class (higher G-mean) competent classifier selection and modified BRACID are slightly better.

Nevertheless, for the final comparison we choose OVO with weighted voting by F1-score as it was the best for the minority classes and still highly placed in the G-mean ranking.

5.3. Multi–class BRACID with post-pruning

We evaluated the multi-class generalization of BRACID, which generates the complete set of rules from all classes (denoted as CRS) and its usage with different post-pruning. In order to limit the size of the complete rule set (see results for CRS in Table 5) we applied a weighted coverage algorithm discussed in subsection 2.2. Following (Napierala et al., 2016) we used the percentage of rules in each class as the stopping conditions and tested several values (e.g. CRS PR 70% denotes that 70% of rules in each class was discarded). Moreover, we checked a simpler pruning with confirmation measures only (PRCONF) - see the first line of Algorithm 1. The obtained results for G-mean are presented in Table 2.

We carried out the Friedman's ranked test to compare all pruning versions of BRACID. As it gives p-value $= 0.091$, we cannot reject the null hypothesis for α equal to 0.05. However, looking at results in Table 2 and other measures presented in the WWW appendix one can notice that performance of extensions depend on particular datasets. To better interpret it, we calculated *safe levels of examples* in all classes according to the method of (Napierala and Stefanowski, 2012) and then averaged them in the minority classes and majority ones, separately (see their values in the online appendix). The lower values of safe levels indicate the higher difficulty of the dataset. It seems that more difficult datasets (e.g. balance scale, hayes roth or yeast) are associated with no pruning, while for easier datasets the amount of pruning may be much higher.

Table 3: G-mean obtained by J48, PART, OVO WVF and CRS ADPR algorithms.

dataset	**J48**	**PART**	**OVO WVF**	**CRS ADPR**
art1	0.9408	0.9366	0.9675	**0.9688**
art2	0.6933	0.6717	0.7596	**0.8012**
art3	0.3416	0.3995	0.5186	**0.5752**
art4	0.8040	0.7535	0.8317	**0.8818**
autos	**0.8044**	0.7062	0.6892	0.6791
balance-scale	0.0099	0.3136	0.4086	**0.6790**
car	0.7925	0.8950	**0.9022**	0.9000
cleveland	0.0848	0.0597	0.1750	**0.2322**
cmc	0.4728	0.4431	**0.4987**	0.4681
dermatology	0.8860	0.8944	**0.9204**	0.9082
ecoli	0.6413	0.6374	0.7404	**0.7876**
flare	0.0442	0.1717	0.3904	**0.4639**
glass	0.3174	0.3155	0.0883	**0.4156**
hayes-roth	**0.8318**	0.8080	0.7849	0.7401
led7digit	**0.7919**	0.7112	0.7736	0.7771
nursery	0.9084	**0.9922**	0.9796	0.9754
new-thyroid	0.8778	0.8948	0.9684	**0.9758**
vehicle	0.9160	0.9147	0.9222	**0.9232**
winequality-red	0.2837	0.2917	0.4530	**0.5338**
yeast	0.0112	0.0312	**0.0808**	0.0758

Assuming that user may be interested in the smaller rule sets, we propose to consider an adaptive selection of the rule pruning percentage. It is done heuristically – select the percentage of pruning according to the difference between the average safe level in the majority and minority classes. If this difference is smaller than 0.16, then BRACID rules are not pruned. The resulting sets of rules will be further denoted as CRS ADPR; see e.g. the obtained numbers of rules in Table 5.

Then, we compared this variant CRS ADPR of BRACID and the best OVO ensemble with the weighted voting (OVO WVF) against two popular rule and tree classifiers PART and J.48 (implementation classes available in Java for WEKA). Although these algorithms were not designed for imbalanced data, they are good representatives of standard and accurate multi-class approaches offering also the rule representation. Moreover, they demonstrated good competitive performance for binary imbalanced problems in earlier BRACID papers, see experiments of (Napierala and Stefanowski, 2012). Both algorithms were run with standard parameters, except using the unpruned mode as it works better for imbalanced data. The results of G-mean and Average Minority are presented in Tables 3 and 4 respectively.

Table 4: Average Minority obtained by J48, PART, OVO WVF and CRS ADPR algorithms.

dataset	**J48**	**PART**	**OVO WVF**	**CRS ADPR**
art1	0.9308	0.9383	0.9700	**0.9733**
art2	0.6283	0.5900	0.7113	**0.7742**
art3	0.3346	0.3633	0.4267	**0.5092**
art4	0.7642	0.6871	0.8213	**0.9046**
autos	**0.8472**	0.7544	0.7372	0.7617
balance-scale	0.0040	0.1480	0.2230	**0.7340**
car	0.7426	0.8588	**0.8852**	0.8762
cleveland	0.2160	0.1969	0.2976	**0.3493**
cmc	0.4263	0.3889	**0.5218**	0.5136
dermatology	0.8800	0.9000	0.9700	**1.0**
ecoli	0.6889	0.7034	0.7809	**0.8196**
flare	0.1293	0.1708	**0.5055**	0.3852
glass	0.6333	0.6089	0.4344	**0.6500**
hayes-roth	**1.0**	0.9617	0.9550	0.9933
led7digit	**0.7848**	0.6912	0.7718	0.7832
nursery	0.7315	**0.9867**	0.9564	0.9727
new-thyroid	0.8642	0.8817	0.9800	**0.9900**
vehicle	0.9168	0.9103	0.9445	**0.9539**
winequality-red	0.3183	0.3141	0.4592	**0.5218**
yeast	0.3933	0.3678	0.4362	**0.4475**

According to the Friedman's test, the differences between these classifiers are significant: p-value = 0.0002 for G-mean and p-value = 0.000026 for Average Minority. The average ranks for G-mean are as follows: J.48 3.05, PART 3.2, PVO WVF 2.05 and CRS ADPR 1.7; and for Average Minority: J.48 3.1, PART 3.3, PVO WVF 2.25 and CRS ADPR 1.35. The critical difference with post-hoc Nemenyi CD is 0.93. So both ways of extending BRACID are significantly better than baseline classifiers. Although this post-hoc analysis does not show the significant difference between OVO WVF and CRS ADPR, we decided to analyse them more precisely with the signed one-tailed Wilcoxon test. For $\alpha = 0.05$, CRS ADPR algorithm is better than OVO WVF (The critical value = 60, while the value of T statistics are 53 for G-mean and 23 for Average Minority).

Finally, we calculated the number of rules produced by BRACID in different extensions, see Table 5. As one could expect, the way of changing rule induction in single BRACID classifier always produces less rules than OVO, which needs to build several binary classifiers. Then, the adaptive pruning ADPR strongly reduces the number of rules generated in the multiple class BRACID for many datasets, see in particular art4, balance-scale, cleveland, or led7digits. For some datasets the number of rules may be acceptable for the potential human inspection.

Table 5: Numbers of rules induced by OVO WVF, CRS BASIC and CRS ADPR approaches.

dataset	**OVO WVF**	**CRS BASIC**	**CRS ADPR**
art1	231	162	162
art2	523	428	43
art3	745	574	58
art4	463	407	41
autos	292	110	110
balance-scale	757	470	123
car	754	333	333
cleveland	611	315	32
cmc	2004	1277	1277
dermatology	468	163	163
ecoli	343	210	210
flare	1558	576	404
glass	259	97	68
hayes-roth	127	108	108
led7digit	682	280	195
new-thyroid	108	42	42
nursery	4096	3584	3584
vehicle	467	359	359
winequality-red	3626	1314	920
yeast	6515	1143	1143

6. Final remarks

This paper was intended to increase interests in the problem of inducing rules from imbalanced data. Two ways of extending binary BRACID rule induction algorithm to deal with multi-class imbalanced data were presented. The experiments demonstrated that the new multi-class 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.

We also observed that the considered extensions of techniques for aggregating classifiers predictions in OVO performed quite similarly. Nevertheless, the weighted voting seems to be slightly better than the competent classifier selection, in particular with respect of measures evaluating the recognition of only minority classes. The deeper analysis of the classifiers selection showed that in some datasets all classes were quite often present in the analysed neighborhood or in some cases the class of the testing example was absent in its neighborhood. Therefore, we hypothesize that other, more specialized, dynamic classifier selection techniques should be studied. Furthermore, we observed that modifying the rule generalization in BRACID was not so profitable, in particular for data with the single majority class and many minority ones (see e.g. cleveland data). Perhaps due to characteristics of the particular pairs of minority classes it was better to additionally generalize more rules from one smaller class only.

The multi-class single BRACID produces a much smaller number of rules than the OVO ensemble. However, due to using two fold representation of rules and single instances, it can still generate relative large sets of rules. In particular, it occurred for more unsafe, difficult datasets (e.g. yeast or winequality). We showed that for many datasets their number may be reduced with the presented weighted covering algorithm without loosing predictive abilities. The more general heuristics for tuning the amount of rule pruning should be still more deeply developed. Furthermore, for the descriptive perspective or for supporting explanations of predictions the future research on new rule selection approaches is worth to be undertaken.

Acknowledgments

We are grateful to our colleague Dr Krystyna Napierala for sharing the code of the binary version of the BRACID algorithm and for discussing with us the problems of rule induction. The research of Jerzy Stefanowski was partially supported by the Polish Ministry of Education and Science, grant no. 0311/SBAD/0709.

References

Alberto Fernández, Victoria López, Mikel Galar, María José Del Jesus, and Francisco Herrera. Analysing the Classification of Imbalanced Data-sets with Multiple Classes: Binarization Techniques and Ad-hoc Approaches. *Knowledge-Based Systems*, 42:97–110, 04 2013.

Alberto Fernández, Salvador García, Mikel Galar, Ronaldo C. Prati, Bartosz Krawczyk, and Francisco Herrera. *Learning from Imbalanced Data Sets*. Springer International Publishing, 2018.

Johannes Fürnkranz, Dragan Gamberger, and Nada Lavrac. *Foundations of Rule Learning*. Cognitive Technologies. Springer, 2012.

Mikel Galar, Alberto Fernández, Edurne Barrenechea, Humberto Sola, and Francisco Herrera. An Overview of Ensemble Methods for Binary Classifiers in Multi-class Problems: Experimental Study on One-vs-one and One-vs-all Schemes. *Pattern Recognition*, 44: 1761–1776, 2011.

Mikel Galar, Alberto Fernández, Edurne Barrenechea, Humberto Sola, and Francisco Herrera. Dynamic Classifier Selection for One-vs-One strategy: Avoiding Non-competent Classifiers. *Pattern Recognition*, 46:3412–3424, 05 2013.

Dragan Gamberger and Nada Lavrac. Expert-guided subgroup discovery: Methodology and application. *Journal of Artificial Intelligence Research*, 17:501–527, Dec 2002.

Jerzy W. Grzymala-Busse, Jerzy Stefanowski, and Szymon Wilk. A comparison of two approaches to data mining from imbalanced data. *J. Intell. Manuf.*, 16(6):565–573, 2005.

Małgorzata Janicka, Mateusz Lango, and Jerzy Stefanowski. Using Information on Class Interrelations to Improve Classification of Multiclass Imbalanced Data: A New Resampling Algorithm. *Int. Journal of Applied Mathematics and Computer Science*, 29:769–781, 2019.

Bartosz Krawczyk. Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, 5(4):221–232, 2016.

Bartosz Krawczyk, Michal Wozniak, and Francisco Herrera. On the usefulness of one-class classifier ensembles for decomposition of multi-class problems. *Pattern Recognition*, 48 (12):3969–3982, 2015.

Ludmila Kuncheva. *Combining Pattern Classifiers. Methods and Algorithms.* Wiley, 2nd edition, 2014.

Krystyna Napierala and Jerzy Stefanowski. BRACID: A Comprehensive Approach to Learning Rules from Imbalanced Data. *Journal of Intelligent Information Systems*, 39(2): 335–373, 2012.

Krystyna Napierała and Jerzy Stefanowski. Types of Minority Class Examples and Their Influence on Learning Classifiers from Imbalanced Data. *Journal of Intelligent Information Systems*, 46:563–597, 2016.

Krystyna Napierala, Jerzy Stefanowski, and Izabela Szczech. Increasing the interpretability of rules induced from imbalanced data by using bayesian confirmation measures. In *5th Int. Workshop,NFMCP 2016, Held in Conjunction with ECML-PKDD 2016, Revised Selected Papers*, volume 10312 of *LNCS*, pages 84–98. Springer, 2016.

Krystyna Napierała and Jerzy Stefanowski. Post-processing of BRACID Rules Induced from Imbalanced Data. *Fundamenta Informaticae*, 148:51–64, 12 2016.

D.R. Wilson and T.R. Martinez. Improved Heterogeneous Distance Functions. *Journal of Artificial Intelligence Research*, 6:1–34, 1997.