Increasing the Interpretability of Rules Induced from Imbalanced Data by Using Bayesian Confirmation Measures

Krystyna Napierała[1,2], Jerzy Stefanowski[1], and Izabela Szczęch[1]

[1] Institute of Computing Science, Poznań University of Technology, 60-965 Poznań, Poland
[2] DATAX sp. z o.o. 53-609 Wroclaw, Poland

Abstract. Approaches to support an interpretation of rules induced from imbalanced data are discussed. In this paper, the rule learning algorithm BRACID dedicated to class imbalance is considered. As it may induce too many rules, which hinders their interpretation, their filtering is applied. We introduce three different strategies, which aim at selecting rules having good descriptive characteristics. The strategies are based on combining Bayesian confirmation measures with rule support, which have not yet been studied in the class imbalance context. Experimental results show that these strategies reduce the number of rules and improve values of rule interestingness measures at the same time, without considerable losses of prediction abilities, especially for the minority class.

Keywords: Bayesian confirmation measures, interpretability of rules, class imbalance, rule post-pruning

1 Introduction

Learning classification rules is one of mature and well studied tasks in machine learning. The popularity of rules comes from the fact that they directly provide a symbolic representation of knowledge discovered from data, which is more comprehensible and human-readable than other representations [5]. Many various algorithms for inducing rules have been already introduced (for their review see, e.g. [5]). Nevertheless, such aspects of data complexity as *class imbalance* still constitute difficulties [11]. The majority of standard rule algorithms are biased towards the majority classes and tend to neglect the minority class. Two kinds of reasons for poor performance of rule based classifiers for imbalanced data are usually pointed out – algorithmic and data level ones [11, 16].

Some extensions of rule classifiers for class imbalances have been already proposed, for their review see [16]. However, most of them address only a single or at most a few of algorithmic or data-related factors. In [16] we introduced a new rule induction algorithm, called BRACID (the acronym of Bottom-up induction of Rules And Cases for Imbalanced Data), which attempts to deal with more of the aforementioned factors. The previous comparative experiments have

clearly demonstrated that the rule classifier induced by BRACID significantly outperformed other rule classifiers generated by the best, standard rule learning algorithms as well as the rule extensions specialized to class imbalances, with respect to predictive measures [16]. On the other hand, BRACID may generate too many rules (see also experiments in Section 6). As it restricts human experts' abilities to analyze or interpret the rules, we are looking for a post-processing approach that could identify the most valuable rules. The first attempt, recently undertaken in [18], has shown that it is possible to select rules characterized by high supports and still leading to sufficient predictive performance.

Nevertheless, focusing attention on the most interesting rules should also take into account other characteristics than simply the rule support. In particular, it is important not to neglect the descriptive abilities of rules, which are often overwhelmed by the need to increase the predictive performance. Note that the predictive and descriptive aspects often stand in opposition to each other [13, 20]. However, when human experts seek for a compact knowledge representation, improving the interpretability of each single rule can even justify some loses on the predictive performance.

Establishing when rules are interesting to users touches both subjective and objective aspects [4]. In this paper we follow the latter aspect and consider *rule interestingness measures* which are often applied to filter the set of rules [7, 15]. They are calculated from learning data and aim at quantifying the relationship between a rule's premise and its conclusion. A particular group of these measures, called *Bayesian confirmation measures*, is well suited for supporting rule interpretability, as it focuses on advancing rules for which the probability of the conclusion given the premise is greater than the genuine probability of the conclusion itself [3, 10]. In other words, confirmation measures promote rules, in which the value that the premise adds to conclusion is considerably high.

Although the concept of confirmation has been firstly considered by philosophers of science in a very different context (see e.g. [2, 3, 19]), it has been adopted to rule interestingness measures, mainly for filtering association rules [8]. Nevertheless, these measures have not been considered for imbalanced data yet. Their application should turn out to be particularly useful in the context of imbalance since considering the probability of each conclusion separately, as done by confirmation measures, would be related to imbalance ratios.

For the purpose of this paper we focus on two particular confirmation measures called S [2] and N [19]. We have chosen them from a wider collection of confirmation measures discussed in the literature because of the desired properties that they possess [9, 10]. In our opinion, these measures satisfy properties that should influence the interpretability of rules [10].

The main aim of this paper is to introduce an approach that uses confirmation measures S and N to post-prune rules induced by BRACID. We focus this study on BRACID only, as experiments [16] have shown that it outperformed other best rule based classifiers over a large collection of imbalanced data. The new approach should reduce the number of its rules while improving values of rule interestingness measures at the same time, especially for the minority class.

The paper is organized as follows. Firstly, we briefly review related works in Section 2. Section 3 introduces the concept of Bayesian confirmation and defines two particularly valuable representatives of confirmation measures: measures S and N. The algorithm used for rule induction, called BRACID, is summarized in Section 4. The three new rule filtering strategies are introduced in Section 5. Their usefulness to improve BRACID rules is evaluated in several experiments, which are described in Section 6. The experimental results are discussed in Section 7. In the final section we draw lines of future research.

2 Related Works on Rule Evaluation and Filtering

Many algorithms for constructing rule based classifiers employ rule pruning. The representative approaches are Grow, IREP or RIPPER; for their review, see chapter 9 in [5]. However, these approaches follow the *classification perspective of rule induction* and pruning is oriented toward good predictive ability of the complete set of rules. Other objectives are stated in the *descriptive knowledge discovery* which aims at discovering from data information patterns and regularities (or sometimes exceptions) which are potentially *interesting* and *useful* to different kinds of users [20].

The descriptive rule discovery perspective, which is considered in this paper, requires other algorithmic strategies than in the classification perspective, e.g., classification versions of association rules, richer sets of satisfactory rules [20] or rule representations of subgroup discovery [6]. However, these algorithms often generate a too high number of rules which makes it impossible for users or domain expert to inspect them. Thus, users lose the opportunity to interpret the results, find interesting rules or to further modify them to have a more accurate classifier [12].

To help the user find relevant knowledge inside huge rule sets, the *rule interestingness measures* have been proposed (for their review see [7, 15]). They are divided into two categories: *subjective* (user-oriented) and *objective* (data-oriented) ones. The subjective measures take into account the user's goals, background knowledge or his belief on the data domain [4]. Objective measures are those that are not application- or user-specific and depend only on raw data. Many of them are defined on the basis of contingency tables summarizing the data set (see the next section). *Support* and $confidence$ are the most universal interestingness measures which are often applied in the process of rule generation (e.g., Apriori search for association rules) and sometimes in post-filtering [1]. Although they are so popular, other measures could be better suited to deal with larger sets of rules and to select the most relevant (i.e. interesting) candidates. Numerous rule interestingness measures have been proposed (lists can be found for example in [7, 12, 13, 15]) and choosing the best one for a given problem is not a trivial task.

In general, the interestingness measures are used to assess, rank (sort) and filter the rules according to various points of view [7]. For these aims, the experts either select some single measures or consider their aggregated, more complex

versions. For instance, [13] describes a case study in which several measures have been used, and the results were interpreted by an expert with a recommendation to use a weighted relative accuracy. Another, more multiple-criteria analysis has been advocated by Bayardo and Agrawal, who proposed to analyze partial ordering of the rules (instead of the typical total ordering of rules) according to different interestingness measures [1]. The authors of [12], on the other hand, discussed other related proposals and proposed a subset of measures based on specialist's preferences; see also [14]. The authors of [9] analyzed properties of the interestingness measures and showed that some measures may be preferred to others. Furthermore, other researchers looked for concise representations (e.g. closed items in associations), rule summaries, grouping of similar rules (with respect to rule condition parts or to subsets of covered examples), or developed interactive visualization tools.

Nevertheless, the choice of the interestingness measures still depends on the expert's preferences and the problem at hand. In this paper, following motivations presented in Section 1, we direct our interest to a particular class of measures based on Bayesian confirmation. Although they have been recently used to filter association rules [8], they have not been considered for classification rules in the class imbalanced tasks.

3 Bayesian Confirmation Measures

To present Bayesian confirmation measures the basic notation is introduced. Rules are consequence relations represented as *IF* (*condition part*) *THEN* (*target class*), where a condition part (premise) is a conjunction of elementary tests on values of attributes characterizing learning examples and a target class points to one of the predefined values of the decision attribute (represented in a rule conclusion). For simplicity, rules will be denoted as $E \rightarrow H$ or simpler as R.

Interestingness measures quantify the relationship between E and H, and are usually defined as functions of four non-negative values that can be gathered in a 2×2 contingency table (see Table 1). For a particular data set, a is the number of objects that satisfy both the rule's premise and its conclusion, b is the number of learning examples for which only H is satisfied, etc. For instance, the support of $E \rightarrow H$ rule is defined as $sup(H, E) = a$ and its confidence as $conf(H, E) = a/(a+c)$. Note that a, b, c and d can also be regarded as frequencies for estimating probabilities: e.g. $P(E) = (a + c)/n$ or $P(H) = (a + b)/n$.

Table 1. An exemplary contingency table of the rule's premise and conclusion

	H	$\neg H$	Σ
E	a	c	$a + c$
$\neg E$	b	d	$b + d$
Σ	$a + b$	$c + d$	n

Among many interestingness measures, we drew our attention to a particular group of *Bayesian confirmation measures* (or simply *confirmation measures*). All those measures are characterized by a feature called *property of Bayesian confirmation*, which requires that an interestingness measure $c(H, E)$ obtains: positive values when $P(H|E) > P(H)$; 0 when $P(H|E) = P(H)$; and negative values when $P(H|E) < P(H)$.

Thus, confirmation measures are designed to depict simply through their scale the confirmatory, neutral or disconfirmatory impact of the rule's premise on its conclusion. Confirmation, interpreted as an increase in the probability of the conclusion H provided by the premise E, is a desirable situation. Let us stress that basic interestingness measures such as support or confidence do not possess the property of confirmation and thus, their utility is lower for the descriptive perspective of knowledge discovery.

The difference of semantics and utility of confidence on one hand, and measure $S(H, E)$ (defined below in Equation 1) being a representative of confirmation measures on the other hand, can be shown on the following illustrative example. Consider the possible result of rolling a dice: 1, 2, 3, 4, 5, 6 points, and let the conclusion H = "*the result is divisible by* 2". Given two different potential rule premises:
E_1 = "*the result is a number from a set* {1*,* 2*,* 3}",
E_2 = "*the result is a number from a set* {2*,* 3*,* 4}"
we get, respectively: $conf(H, E_1) = 1/3$, $S(H, E_1) = -1/3$ and $conf(H, E_2) = 2/3$, $S(H, E_2) = 1/3$. This example clearly shows that the values of confirmation measures have a more useful interpretation than confidence. In particular, in the case of rule $E_1 \rightarrow H$, the premise actually disconfirms the conclusion as it reduces the probability of conclusion H from $1/2 = P(H)$ to $1/3 = P(H|E_1) = conf(H, E_1)$. This fact is expressed by a negative value of confirmation measure $S(H, E)$ (and in fact any confirmation measure), but it cannot be concluded by observing only the value of confidence.

Note that the property of confirmation leaves plenty of space for defining various, non-equivalent confirmation measures (for their review see [3, 9]). To guide the user towards the measures that reflect his expectations, researchers proposed special properties of confirmation measures. These properties express requirements for a measure behavior in certain situations. Taking into account possession of desirable properties, we focus our further interest only on two representatives of confirmation measures. The chosen measures $S(H, E)$ [2] and $N(H, E)$ [19], both ranging from -1 (showing complete disconfirmation) to $+1$ (showing complete confirmation), are defined as:

$$S(H, E) = P(H|E) - P(H|\neg E) = \frac{a}{a+c} - \frac{b}{b+d}, \tag{1}$$

$$N(H, E) = P(E|H) - P(E|\neg H) = \frac{a}{a+b} - \frac{c}{c+d}. \tag{2}$$

Among properties that valuable confirmation measures should satisfy let us mention property of monotonicity M [10] and property of *maximality/minimality*

[8]. Monotonicity M favors measures that are non-decreasing with respect to a and d, and non-increasing with respect to b and c. It is intuitively clear that we would like higher values of measures for rules that are supported by a greater number of positive examples (i.e. increase of a), and exactly the opposite when the number of counter-examples grows (i.e. increase of c). The property of *maximality/minimality* on the other hand, requires that a measure obtains its maximal value if and only if $b = c = 0$, and its minimal values if and only if $a = d = 0$. It is thus a property concentrated on the behavior of measures in the extreme cases. It was verified in [9, 10] that the measures $S(H, E)$ and $N(H, E)$ are among few confirmation measures that satisfy both monotonicity M and *maximality/minimality*.

We have focused our study on those two measures also because the interpretation of their definitions is rather straightforward (contrary to some other confirmation measures possessing M and *maximality/minimality* e.g. measure $c_3(H, E)$ [9] [3]). Measure $S(H, E)$ expresses how much more probable is H with E rather than with $\neg E$. Following some medical examples, e.g. if some symptoms occur then a certain disease is diagnosed, we could say that measure $S(H, E)$ assesses how much more probable becomes the disease when we know that the symptoms occurred (instead of knowing that the symptoms did not occur). In case of measure $N(H, E)$, we would say that is expresses how much more probable are some symptoms for a certain disease than for a case when the disease is excluded (does not occur). Measures $S(H, E)$ and $N(H, E)$ are thus somewhat complementary, as they look at rules from different perspectives: that of the rule's premise and that of the rule's conclusion.

Summing up, taking into account possession of desirable properties and interpretation of the measures' definitions, this study focuses only on application of confirmation measures $S(H, E)$ and $N(H, E)$.

4 Rule Induction with BRACID

BRACID is a specialized algorithm to learn rules from imbalanced data. For its details see [16]. Here, we summarize its main characteristics:

– **Hybrid representation of rules and instances**: BRACID tries to create a general description in regions where the examples form large disjuncts (using rules) and instances to better approximate the more difficult decision boundaries. BRACID allows some (difficult) examples to remain not generalized to rules. They can be treated as maximally specific rules.

[3] $c_3(H, E) = A(H, E)Z(H, E)$ in case of confirmation and
$c_3(H, E) = -A(H, E)Z(H, E)$ in case of disconfirmation
where
$Z(H, E) = 1 - P(\neg H|E) \div P(\neg H)$ in case of confirmation and
$Z(H, E) = P(H|E) \div P(H) - 1$ in case of disconfirmation;
$A(H, E) = [P(E|H) - P(E)] \div [1 - P(E)]$ in case of confirmation and
$A(H, E) = [P(H) - P(H|\neg E)] \div [1 - P(H)]$ in case of disconfirmation.

- **Bottom-up rule induction**: Unlike a top-down strategy typical for rule induction, BRACID follows a bottom-up (or a specific-to-general) strategy as a more appropriate for imbalanced data. It starts from the set of most specific rules each covering a single learning example – which is called a seed of the rule. Then, in every iteration each rule is generalized 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 procedure is repeated until no rule in the set can be further generalized.
- **Resignation from greedy, sequential covering technique**: As this technique, popular in typical rule learning algorithms, increases the data fragmentation and is problematic for the minority examples, BRACID takes into account all the learning examples when evaluating new rule candidate.
- **Facing borderline minority examples**: Types of learning examples are evaluated and rules are generated differently depending on the type of the seed example of a rule [17]. The minority examples belonging to the borderline region are allowed to be generalized into more than one rule, to lessen the dominance of the majority class in this region.
- **Facing noisy examples from the majority class**: Noisy majority examples, present inside the minority class regions, may hinder the induction of general minority rules. BRACID has an embedded mechanism for detecting and removing such examples from the learning data set.
- **Less biased classification strategy**: BRACID employs a classification strategy based on nearest rules to diminish the domination of strong majority rules during solving conflict situations while a new instance matches condition parts of many rules.

Note that some mechanisms employed in this algorithm lead to the increase of the number of rules (mainly a bottom-up rule induction and generation of more rules in the borderline regions). However, the increased number of rules for the minority class, coupled with an increased rule support, are beneficial for final classification. The experimental evaluation of classification performance of BRACID showed indeed that it significantly outperformed many standard rule classifiers (induced by RIPPER, PART, C4.5rules, and others) as well as other rule approaches specialized for class imbalance such as modifications of rule search and classification strategies, or the best standard algorithms (e.g., PART) combined with SMOTE methods transforming class distributions [16].

5 Selecting Rules with Respect to Confirmation

We aim to select a subset of induced rules with respect to appropriate rule evaluation measures. In [18] we have already postulated that it would be profitable to find rules which cover diverse sets of examples referring to different sub-parts of the class distribution. Focusing the expert's attention on a subset of rules having such characteristics should be particularly good for the minority class which is often decomposed into many rare sub-concepts.

Recall that several post-pruning techniques have already been proposed to order rules or to reduce their number. However, as we discussed in [18], it may not lead to diverse subsets of rules in BRACID, as e.g. high supports may characterize many rules having similar syntax and covering similar subsets of learning examples. Other post-pruning techniques considered in rule classifiers are focused on optimizing the predictive performance of the rules rather than on improving their descriptive properties [5].

Therefore, we follow a different inspiration, coming from using rules to represent patterns in *subgroup discovery*, where the task is to find subgroups of individuals that are statistically "most interesting" (e.g. covering as many examples as possible and having the most unusual statistical characteristics [5]). In our opinion these kinds of local, diverse patterns correspond to decomposition of the minority class in sub-concepts. In this paper we generalize the algorithm originally proposed in [6] to find rules describing subgroups.

Our approach to select a given number of diverse rules with respect to a given rule evaluation measure is presented in Algorithm 1. It is run for each class separately and takes as an input the set of all rules induced for this class and their required number after selection – later on we discuss how to tune it.

Algorithm 1 Rule Filtering Algorithm

Input: Set of Rules SR for class P, required NUMBER of rules; rule evaluation ev;
Output: Prunned set of rules FR

```
Delete rules with too low confirmation from SR
FR ← ∅
for every example e ∈ P do
  c(e) ← 1
end for
repeat
  for each rule R ∈ SR do
    calculate rule evaluation measure ev(R)
  end for
  Select R_max = arg max_R(ev(R))
  for each e covered by R_max do
    c(e) ← c(e) + 1
  end for
  Remove R_max from SR
  FR = FR ∪ R_max
until size of FR = NUMBER
```

Firstly, we remove all rules with the non-positive value of a selected confirmation measure (except the option where rules are evaluated with the support only). The key idea of the algorithm is to assign a weight $c(e)$ to each learning example. It is initialized with $c(e) = 1$ for all examples from the given class. When rule R is selected, then weights for examples covered by this rule are increased by adding 1. Then, while evaluating the next rule being a candidate for

selection, the example takes part in all calculation with the weight $1/c(e)$. For instance, the support of a rule is computed as a sum of $1/c(e)$ for all target class examples covered by this rule.

This weighted coverage causes that in the subsequent iterations of the algorithm, examples already covered by the selected rules contribute less to the evaluation of new rule. It promotes the rules referring to examples not yet covered and directs the search toward diverse regions of the class.

In this study we will consider three different versions of the rule evaluation $ev(R)$[4] for selecting rules:

1. a standard rule support $sup(R)$;
2. a product of support with a confirmation measure S : $sup(R) \times S(R)$;
3. a product $sup(R) \times N(R)$.

The choice of rule support $sup(R)$ results from earlier experiments in [18] and we want to consider it as a baseline. The choice of both confirmation measures S and N has been justified in Section 3. We want to aggregate them with a rule support to represent a trade off in a bi-criteria evaluation where the user is interested in sufficiently strong patterns describing the classes.

6 Experimental Evaluation

In the experiments we will verify whether the proposed post-pruning strategies select a limited number of BRACID rules having better values of interestingness measures than in case of non-pruned rules.

As the evaluation criteria we choose the average values of confirmation measures S and N, rule support and rule confidence. We consider the last two measures due to their popularity in the previous rule filtering techniques and to their easy interpretation for the users. These criteria represent descriptive properties of single rules with respect to their possible interpretability and they are treated as primary criteria in our study. As a secondary criterion, we also evaluate the predictive ability of the rule set, which will be estimated by G-mean and F-measure, both well suited for cases with imbalanced data sets. We use this criterion to control whether pruning the set of rules does not dramatically deteriorate the performance compared to all rules produced by the BRACID algorithm. The predictive measures are evaluated in a repeated stratified 10-fold cross validation procedure while rule evaluation measures are calculated for a set of rules induced from the complete data set.

We analysed previous experiments from [16] and chose 11 data sets where BRACID generated too many rules compared to other, standard rule induction algorithms. They are characterized by different imbalance ratios (from 3% to 30%), data sizes (from 155 to 1728) and types of attributes (only nominal, only numeric, or mixed). Although the imbalance ratios of some of these data sets

[4] For simplicity we will further use a notation of a rule as R instead of (H, E) in symbols of measures

Table 2. Basic characteristics of data sets

Data set	#Examples	Minority class size	Imbalance ratio [%]	#Attributes (numeric)	Minority class name
balance-scale	625	49	7.84	4 (4)	B
breast-cancer	286	85	29.72	9 (0)	rec-events
car	1728	69	3.99	6 (0)	good
cleveland	303	35	11.55	13 (6)	positive
cmc	1473	333	22.61	9 (2)	long-term
ecoli	336	35	10.42	7 (7)	imU
haberman	306	81	26.47	3 (3)	died
hepatitis	155	32	20.65	19 (6)	die
solar-flareF	1066	43	4.03	12 (0)	F
transfusion	748	178	23.80	4 (4)	yes
yeast-ME2	1484	51	3.44	8 (8)	ME2

are medium, all these data are also affected by different difficulty factors characterizing the distribution of examples from the minority class. According to experimental studies [17] these factors lead to difficulties while learning rules.

All these data sets come from the UCI repository. We analyzed them as binary problems – the minority class vs. majority one (which may aggregate others), as it is a typical view of class imbalances with focusing attention on improving recognition of the class of special importance. The basic characteristics of these data sets are presented in Table 2.

We checked that for all data sets (except cleveland and hepatitis), the BRACID rule sets contained some rules with negative values of confirmation measures. For instance, balance-scale contained 8, car 36, cmc 19, solar-flareF 18 and transfusion 14 such rules.

While using the algorithm for selecting rules we need to define a number of required rules as the stopping condition. In general, this parameter should represent the analyst's expectations and his abilities to inspect the rules. Here we recall our previous experiments [18], where we studied a wide range of values of this parameter (up to 30%). The results showed that the threshold 10% often led to rule sets having the good average rule support and comparable classification performance as the original set of BRACID rules.

Yet another option is to select all the rules which are necessary to cover all the learning examples in each class. We studied this coverage option in [18] and observed that it usually produced higher classification prediction (with respect to G-mean or sensitivity measure) than the percentage option. However, it also selected more rules than the percentage option. As in this study we aim at reducing the number of rules, we decided to consider the percentage option with the parameter tuned to 10% of the original set of rules for each class [5]

[5] More detailed experimental results, including also the coverage option are provided at the page http://www.cs.put.poznan.pl/iszczech/publications/nfmcp-2016.html.

Table 3. Characteristics of filtered rules for the minority class

Data set	Pruning	#Rules	Avg.sup	Avg.$conf$	Avg.S	Avg.N
balance-scale	none	52	2.077	0.611	0.535	0.033
	sup	5	6.000	0.266	0.192	0.056
	$sup * S$	5	2.000	0.875	0.799	0.037
	$sup * N$	5	4.600	0.317	0.243	0.065
breast-cancer	none	77	3.364	0.711	0.420	0.030
	sup	8	9.625	0.711	0.434	0.089
	$sup * S$	8	9.125	0.817	0.541	0.094
	$sup * N$	8	10.125	0.736	0.460	0.095
car	none	54	1.444	0.972	0.933	0.021
	sup	5	5.200	0.700	0.663	0.073
	$sup * S$	5	4.800	0.800	0.763	0.068
	$sup * N$	5	5.200	0.700	0.663	0.073
cleveland	none	97	5.495	0.910	0.811	0.154
	sup	10	8.300	0.864	0.773	0.232
	$sup * S$	10	7.300	0.966	0.873	0.207
	$sup * N$	10	8.600	0.864	0.774	0.240
cmc	none	354	6.588	0.723	0.500	0.016
	sup	35	14.914	0.666	0.447	0.037
	$sup * S$	35	12.686	0.782	0.562	0.033
	$sup * N$	35	18.571	0.652	0.434	0.046
ecoli	none	46	10.413	0.872	0.796	0.291
	sup	5	17.400	0.802	0.746	0.483
	$sup * S$	5	17.000	0.889	0.832	0.478
	$sup * N$	5	18.200	0.788	0.734	0.503
haberman	none	122	6.049	0.716	0.464	0.062
	sup	12	9.917	0.650	0.406	0.099
	$sup * S$	12	9.417	0.900	0.658	0.109
	$sup * N$	12	12.250	0.783	0.546	0.135
hepatitis	none	66	7.424	0.986	0.820	0.231
	sup	7	12.000	0.971	0.832	0.373
	$sup * S$	7	12.571	1.000	0.864	0.393
	$sup * N$	7	12.571	1.000	0.864	0.393
solar-flareF	none	39	3.051	0.527	0.490	0.066
	sup	4	6.750	0.362	0.327	0.142
	$sup * S$	4	4.500	0.790	0.753	0.102
	$sup * N$	4	7.750	0.382	0.348	0.164
transfusion	none	161	6.360	0.673	0.440	0.028
	sup	16	16.062	0.630	0.404	0.067
	$sup * S$	16	15.562	0.768	0.543	0.071
	$sup * N$	16	18.500	0.679	0.456	0.083
yeast-ME2	none	155	7.432	0.905	0.875	0.145
	sup	16	9.375	0.915	0.886	0.183
	$sup * S$	16	8.875	0.944	0.915	0.174
	$sup * N$	16	10.688	0.904	0.877	0.209

Table 4. Characteristics of filtered rules for the majority class

Data set	Pruning	#Rules	Avg.sup	Avg.$conf$	Avg.S	Avg.N
balance-scale	none	306	12.889	0.996	0.076	0.021
	sup	31	30.097	0.994	0.076	0.049
	$sup * S$	31	30.452	0.997	0.079	0.051
	$sup * N$	31	34.194	0.996	0.079	0.057
breast-cancer	none	75	4.973	0.959	0.261	0.022
	sup	8	11.750	0.925	0.234	0.050
	$sup * S$	8	12.500	0.994	0.304	0.061
	$sup * N$	8	13.375	0.994	0.305	0.065
car	none	69	68.478	0.924	-0.036	0.017
	sup	7	361.286	0.987	0.037	0.187
	$sup * S$	7	351.429	1.000	0.051	0.212
	$sup * N$	7	356.571	1.000	0.051	0.215
cleveland	none	94	16.426	1.000	0.123	0.061
	sup	9	53.444	1.000	0.142	0.199
	$sup * S$	9	53.444	1.000	0.142	0.199
	$sup * N$	9	54.111	1.000	0.142	0.202
cmc	none	401	7.302	0.971	0.198	0.006
	sup	40	21.725	0.975	0.204	0.017
	$sup * S$	40	21.500	0.987	0.217	0.018
	$sup * N$	40	22.975	0.986	0.216	0.019
ecoli	none	47	64.128	0.990	0.141	0.210
	sup	5	207.800	0.999	0.271	0.685
	$sup * S$	5	208.000	0.999	0.271	0.685
	$sup * N$	5	208.000	0.999	0.271	0.685
haberman	none	60	6.383	0.977	0.247	0.027
	sup	6	15.833	0.990	0.269	0.068
	$sup * S$	6	15.833	0.990	0.269	0.068
	$sup * N$	6	15.833	0.990	0.269	0.068
hepatitis	none	52	18.615	1.000	0.241	0.151
	sup	5	59.600	1.000	0.341	0.485
	$sup * S$	5	59.600	1.000	0.341	0.485
	$sup * N$	5	65.200	1.000	0.357	0.530
solar-flareF	none	64	27.781	0.957	-0.002	0.012
	sup	6	165.333	0.982	0.031	0.123
	$sup * S$	6	158.500	0.989	0.039	0.128
	$sup * N$	6	163.833	0.986	0.036	0.129
transfusion	none	118	11.720	0.965	0.206	0.016
	sup	12	51.500	0.932	0.183	0.064
	$sup * S$	12	41.750	0.959	0.209	0.060
	$sup * N$	12	51.750	0.947	0.200	0.073
yeast-ME2	none	613	204.979	1.000	0.041	0.143
	sup	61	514.000	1.000	0.055	0.358
	$sup * S$	61	566.131	1.000	0.057	0.395
	$sup * N$	61	609.197	1.000	0.059	0.425

In our study, we will examine three proposed strategies to select rules with the rule evaluation $ev(R)$ (see Section 5), defined as: (1) a standard rule support $sup(R)$; (2) a product $sup(R) \times S(R)$; and (3) a product $sup(R) \times N(R)$.

The rule characteristics with respect to considered criteria are given in Tables 3 and 4, for the minority and majority class, respectively. The column "pruning" corresponds to the selection strategy (note that results for using the standard version of BRACID without pruning is presented in the first row for each data set with an abbreviation "none").

Additionally, we constructed rule classifiers with the three filtering strategies and evaluated their classification performance. The values of G-mean and F-measure are presented in Table 5.

Table 5. G-mean and F-measure for BRACID with all rules vs. filtered rules

	G-mean				F-measure			
Data set	BRACID	sup	$sup * S$	$sup * N$	BRACID	sup	$sup * S$	$sup * N$
balance-scale	0.56	0.63	0.59	0.60	0.19	0.23	0.22	0.21
breast-cancer	0.56	0.59	0.61	0.61	0.44	0.48	0.49	0.49
car	0.88	0.60	0.61	0.64	0.73	0.41	0.42	0.44
cleveland	0.57	0.71	0.72	0.73	0.33	0.41	0.41	0.42
cmc	0.64	0.64	0.64	0.64	0.45	0.45	0.45	0.45
ecoli	0.83	0.85	0.85	0.84	0.60	0.55	0.54	0.55
haberman	0.58	0.54	0.54	0.54	0.44	0.44	0.43	0.43
hepatitis	0.75	0.76	0.75	0.74	0.60	0.59	0.57	0.54
solar-flareF	0.64	0.73	0.65	0.73	0.28	0.32	0.32	0.31
transfusion	0.64	0.63	0.63	0.65	0.47	0.47	0.46	0.48
yeast-ME2	0.71	0.72	0.73	0.77	0.42	0.40	0.40	0.38

7 Discussion of the Experiments

Each of the filtering strategies improves the interestingness measure used in the given strategy. Note that all of them improve average rule supports for both minority and majority classes. For some data sets these improvements are quite high, for instance, for cmc data the average rule supports increase from 6.59 to 18.57 examples in the minority class, and from 7.30 to 22.98 examples in the majority class. Similar high improvements also occur for car, solar flare, ecoli and transfusion data.

The third strategy (based on $sup(R)$ and $N(R)$) increases the average value of measure N for all data sets in both classes — see e.g. hepatitis data, where the improvements are from 0.23 to 0.39 for the minority class and from 0.15 to 0.49 for the majority class. Similar increases have been observed for other data. Similarly, the second strategy (based on $sup(R)$ and $S(R)$) improves the average values of the confirmation measure S – however, it is more visible for the

minority class than for the majority one, for instance changes from 0.46 to 0.65 in the minority class and from 0.25 to 0.27 in the majority one for haberman data. Note that values of the confirmation measure S are always higher than N.

It is worth observing that the proposed strategies also improve rule evaluation measures other than the ones used in each strategy. In particular, the third strategy usually provides the highest values of the average support – in the majority of data sets it is better than the first strategy that uses the support only. Although it sometimes slightly improves the confirmation measure S, it usually decreases the average confidence of rules. On the other hand, the second strategy offers the highest increases of the rule confidence. It is more visible for the minority class as the confidence of majority rules is already quite high.

What is also interesting, classification performance of such filtered rules does not decrease too much compared to the original set of rules and for few data it is even better – see results in Table 5.

The differences in results obtained by strategies using S and N measures could be explained by analyzing their formulae (see Equations 1 and 2). They exploit the contingency matrix in a different, although symmetric, way. Measure S is more focused on considering a pair of numbers (a and c) decreased by (b and d), while N aggregates a different combination. As BRACID tries to induce rules with a very high confidence (which refers to the pair a and c), it is naturally oriented on obtaining higher values of the S measure. On the other hand, as measure N exploits complementary information to the one used in BRACID rule induction process, it may better co-operate with the rule support in the pruning strategy and may lead to better descriptive rule evaluation as well as classification results.

8 Conclusions and Final Remarks

To sum up, our experiments have clearly demonstrated that all proposed filtering strategies lead to selecting a much smaller number of BRACID induced rules, which are characterized by better values of considered interestingness measures than in case of non-pruned rules.

As future research, we plan to extend the experimental evaluation with other rule classifiers specialized for class imbalances in order to show the generality of our approach. We also intend to confront our pruning strategies with a baseline approach involving a simple rule filtering. Furthermore, we plan to investigate a more local way of calculating the interestingness measures, which will be based on the analysis of neighbor examples to the given rule rather than on all data elements as it is currently done.

Ack. The research was supported by NCN grant DEC-2013/11/B/ST6/00963.

References

1. Bayardo, R., Agrawal, R.: Mining the most interesting rules, Proc. 5th ACM SIGKDD Conf. on Knowledge Discovery and Data Mining, 145–154 (1999).
2. Christensen, D.: Measuring confirmation. Journal of Philosophy, 96, 437–461 (1999).
3. Fitelson, B.: The plurality of Bayesian measures of confirmation and the problem of measure sensitivity. Philosophy of Science, 66, 362–378 (1999).
4. Freitas, A.: On rule interestingness measures. Knowledge-Based Systems, 12, 309–315 (1999).
5. Furnkranz, J., Gamberger, D., Lavrac, N.: Foundations of Rule Learning, Springer Verlag (2012).
6. Gamberger, D., Lavrac, N. : Expert-guided subgroup discovery: methodology and application. J. Artificial Int. Research, 17(1), 501–527 (2002).
7. Geng, L., Hamilton, H.: Interestingness measures for data mining: A survey. ACM Computing Surveys, 38(3) (2006)
8. Glass, D.: Confirmation Measures of Association Rule Interestingness. Knowledge-Based Systems, 44, 65–77 (2013).
9. Greco, S., Slowinski, R., Szczech, I.: Properties of rule interestingness measures and alternative approaches to normalization of measures. Information Sciences, 216, 1–16 (2012).
10. Greco, S., Slowinski, R., Szczech, I.: Measures of rule interestingness in various perspectives of confirmation. Information Sciences 346-347C, 216–235 (2016).
11. He, H., Yungian, Ma (eds): Imbalanced Learning. Foundations, Algorithms and Applications, IEEE - Wiley (2013).
12. Heravi, M., R. Zaiane, O.: A Study On Interestingness Measures for Associative Classifiers. In Proc. of ACM-SAC 2010 Conference Track on Data Mining. 1040–1047 (2010).
13. Lavrac, N., Flach, P., Zupan, B.: Rule Evaluation Measures: A Unifying View. In Proc. of ILP-99, Springer LNAI vol. 1634, 174–185, (1999).
14. Lenca, P., Vaillant, B., Meyer, P., Lallich, S.: Associations rule interestingness measures: Experimental and theoretical studies. In Guillet, F. , Hamilton, H. J. (Ed.), Quality Measures in Data Mining, Studies in Computational Intelligence, Springer, 51–76, (2007).
15. McGarry, K.: A survey of interestingness measures for knowledge discovery. The Knowledge Engineering Review, 20(1), 39–61 (2005).
16. Napierala, K., Stefanowski, J.: BRACID: a comprehensive approach to learning rules from imbalanced data, Journal of Intelligent Information Systems, 39(2), 335–373 (2012).
17. Napierala, K., Stefanowski, J.: Types of Minority Class Examples and Their Influence on Learning Classifiers from Imbalanced Data. Journal of Intelligent Information Systems, 46(3), 563–597 (2016).
18. Napierala, K., Stefanowski, J.: Post-processing of BRACID Rules Induced from Imbalanced Data. Fundamenta Informaticae, 146 - accepted for publication (2016).
19. Nozick, R.: Philosophical Explanations, Clarendon Press, Oxford, UK (1981).
20. Stefanowski J., Vanderpooten D.: Induction of decision rules in classification and discovery-oriented perspectives. Int. Journal of Intelligent Systems. 16 (1), 13–28 (2001).