Bayesian Confirmation Measures in Rule-based Classification

Dariusz Brzezinski, Zbigniew Grudziński, and Izabela Szczęch

Institute of Computing Science, Poznan University of Technology,
ul. Piotrowo 2, 60–965 Poznan, Poland
{izabela.szczech,dariusz.brzezinski}@cs.put.poznan.pl

Abstract. With the rapid growth of available data, learning models are also gaining in sizes. As a result, end-users are often faced with classification results that are hard to understand. This problem also involves rule-based classifiers, which usually concentrate on predictive accuracy and produce too many rules for a human expert to interpret. In this paper, we tackle the problem of pruning rule classifiers while retaining their descriptive properties. For this purpose, we analyze the use of confirmation measures as representatives of interestingness measures designed to select rules with desirable descriptive properties. To perform the analysis, we put forward the CM-CAR algorithm, which uses interestingness measures during rule pruning. Experiments involving 20 datasets show that out of 12 analyzed confirmation measures c_1, F, and Z are best for general-purpose rule pruning and sorting. An additional analysis comparing results on balanced/imbalanced and binary/multi-class problems highlights also N, S, and c_3 as measures for sorting rules on binary imbalanced datasets. The obtained results can be used to devise new classifiers that optimize confirmation measures during model training.

Keywords: rule classifiers, interestingness measures, Bayesian confirmation, rule pruning

1 Introduction

Recent years have seen the rise of such terms as big data and data science, which brought many machine learning and data mining methods to public attention. This growing popularity of pattern mining methods results in numerous practical applications, such as healthcare, online education, social network analysis, or smart houses [20,18]. Many of these applications involve cooperation with human experts, who often have to understand not only direct algorithm results, but also entire learning models.

Arguably the most studied data mining task is classification [18]. Various types of classifiers have been developed over the years, however rules are continuously regarded as one of the most popular approaches to practical applications involving non-data-mining experts. It is due to the symbolic form of rules, which makes them comprehensible. Thus, when both pattern usage and understanding are key goals, rules are a common form of knowledge representation.

Nevertheless, in most studies data miners tend to focus solely on the predictive performance of learning models [13,6,2]. This is also the case of rule mining. As a result, the descriptive value that rules can carry is often neglected. Unquestionably, a compilation of good predictive and descriptive abilities of a classifier is sought for in many applications. Preferably, these abilities should also be accompanied by a compact representation. In particular, for rule-based classifiers this requirement can be achieved by limiting the number of rules, since otherwise the set of rules could exceed the human-expert's understanding capabilities. For example, in medical applications, doctors are usually interested in a reduced set of rules that describes the patients well and offers good predictions [26].

The evaluation and, thus, pruning of rule sets is usually done by interestingness measures; for a survey see e.g. [14,24]. In classification, these measures are used to improve the predictive performance of learning models, often neglecting the descriptive value of each rule. Nonetheless, many interestingness measures were designed especially for evaluating the descriptive properties of rules. In particular, Bayesian confirmation measures [12] constitute a group of measures that quantify the degree with which the rule's premise supports the conclusion. Confirmation measures obtain positive values only when the premise widens our knowledge about the conclusion, thus, they allow to swiftly choose meaningful rules and filter out the misleading ones. Additionally, the usefulness of confirmation measures in the descriptive context has been depicted with many desirable properties they possess [7,12,15,16].

In this paper, we analyze the impact of using confirmation measures in rule-based classification. For this purpose, we put forward the CM-CAR algorithm, which uses confirmation measures to sort and prune a list of rules. As a result, the proposed algorithm is capable of producing a concise set of descriptive rules, while retaining high predictive performance. Summarizing, the main contributions of this paper are as follows:

- the analysis of interestingness measures with good descriptive properties in the context of predictive classification problems;
- the proposal of the CM-CAR algorithm for discovering and pruning decision rules with high confirmation;
- a comprehensive series of experiments analyzing 12 Bayesian confirmation measures for sorting and pruning rule lists.

The remainder of this paper is organized as follows. Section 2 provides basic notation, definitions, reviews Bayesian confirmation measures, and discusses related works. Section 3 presents the CM-CAR algorithm. In Section 4, we discuss experimental results, which demonstrate the properties of the analyzed measures. Finally, Section 5 concludes the paper and draws lines of future research.

2 Preliminaries and Related Works

Among various knowledge representations, patterns in the form of rules are known and appreciated for their high comprehensibility and interpretability.

Such form of knowledge representation is often found easy to understand and use by decision makers.

Rules are usually induced from a dataset being a set of objects characterized by a set of attributes. Rules are consequence relations, denoted as $E \rightarrow H$ ("*if* E *then* H"), between the condition E and conclusion H formulas built from attribute-value pairs. The condition formulas are called the premise (or evidence) and the conclusion formulas are referred to as the conclusion (or hypothesis) of the rule. If the set of attributes that can occur in the conclusion is limited to a predefined *class attribute*, then the rule is regarded as a *decision rule*.

The evaluation of the quality and utility of rules induced from data is most commonly done by means of *interestingness measures*, which quantify the relationship between E and H. In the context of a particular dataset, interestingness measures can be usually defined on the basis of four non-negative values: a, b, c and d, briefly represented in Table 1.

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

The number of objects in a dataset that satisfy both the rule's premise and conclusion is quantified by a. The number of objects for which the premise is not satisfied, but the conclusion is, will be denoted by b, etc. This notation can be effectively used for defining such interestingness measures as, for example, confidence: $conf(H,E) = a/(a+c)$ or support: $sup(H,E) = a$.

In this paper we focus on a particular group of interestingness measures that are referred to as *Bayesian confirmation measures* (or simply *confirmation measures*). Their common feature is that they obtain:

- positive values when $P(H|E) > P(H)$,
- 0 when $P(H|E) = P(H)$,
- negative values when $P(H|E) < P(H)$.

Observe that the notation based on a, b, c, and d can also be used to estimate probabilities, e.g. $P(H) = (a+b)/n$ or $P(H|E) = a/(a+c)$. Thus, the conditions that a confirmation measure, denoted as $cm(H,E)$, must satisfy can be expressed as follows:

$$cm(H,E) \begin{cases} > 0 \text{ when } \frac{a}{a+c} > \frac{a+b}{n}, \\ = 0 \text{ when } \frac{a}{a+c} = \frac{a+b}{n}, \\ < 0 \text{ when } \frac{a}{a+c} < \frac{a+b}{n}. \end{cases} \tag{1}$$

Thus, confirmation measures quantify the degree to which E provides support *for* or *against* H [12].

Due to the fact that the above conditions do not favor any single measure as the most adequate, there are many alternative, ordinally non-equivalent measures of confirmation [7,12]. Definitions of 12 popular confirmation measures are listed in Table 2.

Table 2: Popular confirmation measures

$D(H,E) = P(H\|E) - P(H) = \frac{a}{a+c} - \frac{a+b}{n} = \frac{ad-bc}{n(a+c)}$	[11]
$M(H,E) = P(E\|H) - P(E) = \frac{a}{a+b} - \frac{a+c}{n} = \frac{ad-bc}{n(a+b)}$	[25]
$S(H,E) = P(H\|E) - P(H\|\neg E) = \frac{a}{a+c} - \frac{b}{b+d} = \frac{ad-bc}{(a+c)(b+d)}$	[5]
$N(H,E) = P(E\|H) - P(E\|\neg H) = \frac{a}{a+b} - \frac{c}{c+d} = \frac{ad-bc}{(a+b)(c+d)}$	[27]
$C(H,E) = P(E \wedge H) - P(E)P(H) = \frac{a}{n} - \frac{(a+c)(a+b)}{n^2} = \frac{ad-bc}{n^2}$	[3]
$F(H,E) = \frac{P(E\|H) - P(E\|\neg H)}{P(E\|H) + P(E\|\neg H)} = \frac{\frac{a}{a+b} - \frac{c}{c+d}}{\frac{a}{a+b} + \frac{c}{c+d}} = \frac{ad-bc}{ad+bc+2ac}$	[21]
$Z(H,E) = \begin{cases} 1 - \frac{P(\neg H\|E)}{P(\neg H)} = \frac{ad-bc}{(a+c)(c+d)} & \text{in case of confirmation} \\ \frac{P(H\|E)}{P(H)} - 1 = \frac{ad-bc}{(a+c)(a+b)} & \text{in case of disconfirmation} \end{cases}$	[7]
$A(H,E) = \begin{cases} \frac{P(E\|H) - P(E)}{1 - P(E)} = \frac{ad-bc}{(a+b)(b+d)} & \text{in case of confirmation} \\ \frac{P(H) - P(H\|\neg E)}{1 - P(H)} = \frac{ad-bc}{(b+d)(c+d)} & \text{in case of disconfirmation} \end{cases}$	[16]
$c_1(H,E) = \begin{cases} \alpha + \beta A(H,E) & \text{in case of confirmation when } c = 0 \\ \alpha Z(H,E) & \text{in case of confirmation when } c > 0 \\ \alpha Z(H,E) & \text{in case of disconfirmation when } a > 0 \\ -\alpha + \beta A(H,E) & \text{in case of disconfirmation when } a = 0 \end{cases}$	[16]
$c_2(H,E) = \begin{cases} \alpha + \beta Z(H,E) & \text{in case of confirmation when } b = 0 \\ \alpha A(H,E) & \text{in case of confirmation when } b > 0 \\ \alpha A(H,E) & \text{in case of disconfirmation when } d > 0 \\ -\alpha + \beta Z(H,E) & \text{in case of disconfirmation when } d = 0 \end{cases}$	[16]
$c_3(H,E) = \begin{cases} A(H,E)Z(H,E) & \text{in case of confirmation} \\ -A(H,E)Z(H,E) & \text{in case of disconfirmation} \end{cases}$	[16]
$c_4(H,E) = \begin{cases} min(A(H,E), Z(H,E)) & \text{in case of confirmation} \\ max(A(H,E), Z(H,E)) & \text{in case of disconfirmation} \end{cases}$	[16]

The selected confirmation measures obtain values ranging from -1 to $+1$, except for measures $D(H,E)$ and $M(H,E)$, whose values approach -1 or $+1$ only for n approaching $+\infty$. Moreover, measure $C(H,E)$ originally obtains values from $-1/4$ to $+1/4$ (regardless of n), so a simple linear transformation (a multiplication by 4) has been introduced and all further results concern the transformed $C(H,E)$. For brevity and clarity of presentation, the definitions of measures $Z(H,E)$, $A(H,E)$, $c_1(H,E)$, $c_2(H,E)$, $c_3(H,E)$ and $c_4(H,E)$ in Table 2 omit the situation of neutrality, in which the measures default to 0. Moreover, measures $c_1(H,E)$ and $c_2(H,E)$ have been computed for the values of $\alpha = \beta = 1/2$.

Our interest in confirmation measures results mostly from their valuable scale semantics. Notice, how easy it is to filter out misleading rules (i.e., those for which the premise actually disconfirms the conclusion) only by observing the value of the measure. Especially when working with imbalanced data, it is important not to give credit to rules in which the probability of the conclusion given the premise is smaller than the genuine probability of the conclusion itself. Nevertheless not all popular interestingness measures depict such situations, e.g. confidence, support. That is why, we direct our interest to confirmation measures. They have been widely studied as measures in single-rule evaluation [7,12,16] for descriptive purposes, neglecting however their potential usefulness in classifiers. Thus, our experimental study intentionally focuses only on confirmation measures, which in our opinion should gain in popularity in the context of rule-based classification.

Although classical approaches to rule classification concentrate on predictive performance and rule coverage [6,9,13,28], there have already been studies on using interestingness measures in rule-based classification. The algorithm that particularly inspired our work is CBA [23]. The Classification Based on Associations (CBA) algorithm is based on applying association rule induction approaches to finding classification rules. In CBA the classifier construction process starts by generating association rules characterized by minimal support. Next, the obtained associations are transformed to classification rules by selecting only those rules where the conclusion is the class attribute. Furthermore, these classification rules are filtered and limited only to those with confidence equal or greater than a user-defined threshold. Finally, the set of rules is ordered on the basis of their confidence, support, and length.

Other attempts to use frequent patterns/association rules in classification include the CAEP classifier [10], which is based on emerging patterns. Emerging patterns are defined as patterns whose supports increase significantly from one class to another and, as the CAEP method shows, prove to work well even with high dimensional problems [10]. Among more recent proposals, Ceci and Appice [4] focus on propositional and structural approaches to spatial classification in multi-relational data mining. This work also studies an associative classification framework, one that employs spatial association rules. Nevertheless, none of the cited works investigates the use of Bayesian confirmation measures, which are the main focus of this paper.

3 The CM-CAR Algorithm

In this paper, we analyze the potential of using confirmation measures in classification. However, existing rule classifiers [13,6,28,9] try to optimize accuracy or instance coverage rather than the descriptive value of the created rules. Therefore, we put forward a new algorithm called Confirmation Measure Class Association Rules (CM-CAR), which creates a user-defined number of decision rules based on Bayesian confirmation measures. The pseudocode of CM-CAR is presented in Algorithm 1.

Algorithm 1 CM-CAR

Input: $\mathcal{D}$: data set, *minsup*: minimal support, k: number of rules, C: class attribute, Q_s: ordered set of sorting measures, Q_p: ordered set of pruning measures
Output: $\mathcal{CAR}$: decision rule list of length k

1: $\mathcal{CAR} \leftarrow \emptyset$
2: $\mathcal{L} \leftarrow$ itemsets with support $\geq$ *minsup* ▷ Find frequent associations
3: **for all** subsets l_k of itemsets $l \in \mathcal{L}$ **do** ▷ Create decision rules
4: **if** $l - l_k = \{C\}$ **then**
5: $r \leftarrow$ decision rule $l_k \rightarrow C$
6: $\mathcal{CAR} \leftarrow \mathcal{CAR} \cup r$
7: Sort $\mathcal{CAR}$ according to Q_s ▷ Create decision list
8: Leave in $\mathcal{CAR}$ k-best rules according to Q_p ▷ Prune decision list

First, the CM-CAR algorithm finds frequent itemsets. For this purpose we use the Apriori algorithm [1], however, in practice any frequent itemset mining algorithm could be used. Next, CM-CAR creates decision rules based on those frequent sets that contain the class attribute C. Finally, two sets of interestingness measures, Q_s and Q_p, are used to sort and filter the rules, respectively. As its classification model, the algorithm outputs a list of k decision rules, where k is a user-defined value.

CM-CAR can be considered a generalization of the CBA algorithm proposed by Liu et al. [23], where instead of using support and confidence, we use arbitrary interestingness measures to create a list of rules. As in the CBA algorithm, the time performance of CM-CAR depends mostly on the frequent pattern mining phase which has a complexity of $O(2^n)$, n being the dataset size.

It is worth noting that the proposed algorithm uses two sets of measures for two distinct purposes. Q_s is a set of measures that order the rules and, therefore, decide which rule is used if more than one rule covers an example. If $Q_s = \{sup, conf\}$, rules are sorted according to their support and then, in case of ties, confidence. On the other hand, Q_p prunes the sorted rules. For example, if $Q_p = \{S, N\}$ then the rule list is limited to k best rules according to measure S and, in case of ties, N.

With two separate sets of measures, CM-CAR is capable of dividing the responsibility for the predictive (Q_s) and descriptive (Q_p) properties of its classification model. In the following section, we use this feature to compare various confirmation measures in a series of experiments.

4 Experimental Study

The goal of this paper is to perform a comparison of confirmation measures. For this purpose, we use the CM-CAR algorithm with varying values of Q_s and Q_p. The use of other rule-based classifiers is out of the scope of this study.

The experiments are divided into two groups. In the first group, we are interested in assessing confirmation measures in the context of rule pruning. Therefore, we set $Q_s = \{conf, sup, length\}$ and $Q_p = \{CM\}$, where *length* signifies the number of conditional attributes in a rule and CM is one of the 12 confirmation measures from Table 2. For reference, we also analyzed the usage of *conf* as a pruning measure. It is worth noting that using *conf* for pruning makes CM-CAR work exactly like the CBA algorithm. Therefore, *conf* can be considered a baseline against which the remaining measures can be compared. By keeping Q_s fixed in this group of experiments, we ensure that differences in model performance are only due to the measure used for pruning.

In the second group of experiments, we focus on verifying the utility of confirmation measures in the context of classification. To achieve this, we set $Q_s = \{CM, sup, length\}$ and $Q_p = \{CM\}$, making one of the 12 confirmation measures (or *conf*) a key factor responsible for the predictive and descriptive performance. As in the first group of experiments, *conf* serves as a baseline approach against which other measures can be compared.

The *minsup* parameter for frequent pattern mining was set to obtain a number of rules close to 10 000 for each dataset. Such a number was selected to ensure that it is possible to perform a long series of rule prunings. The use of each confirmation measure was evaluated on a holdout test set consisting of 34% of the original dataset using [19]:

- Balanced accuracy: $\frac{1}{2}(\text{sensitivity} + \text{specificity})$,
- G-mean: $\sqrt{\text{sensitivity} \cdot \text{specificity}}$,
- F_1-score: $2 \cdot \frac{\text{sensitivity} \cdot \text{precision}}{\text{precision} + \text{sensitivity}}$,
- AUC: area under the Receiver Operator Characteristic curve [19].

For multi-class problems, performance was calculated using macro averaging, i.e., evaluation measures where computed "one-vs-all" for each class and averaged without weighting. All four measures were chosen based on their ability to assess classifiers on imbalanced data. The CM-CAR algorithm was written in Java as part of the WEKA [17] framework.[1]

4.1 Datasets

In our study, we used 20 datasets with various numbers of classes, imbalance ratios, and containing nominal as well as numeric attributes. All of the used datasets are publicly available, mostly through the UCI machine learning repository [22]. Table 3 presents the main characteristics of each dataset.

[1] Sources available at: http://www.cs.put.poznan.pl/dbrzezinski/software.php

Table 3: Dataset characteristics

Dataset	Size	Num. Attr.	Nom. Attr.	Classes	Maj. class	Mined rules	Bala-nced	Binary
adult-census	32,561	6	8	2	75.90%	4,299	×	✓
autos	205	15	10	7	32.68%	8,109	×	×
cmc	1,473	2	7	3	43.70%	10,001	✓	×
credit-g	1,000	7	13	2	70.00%	8,540	×	✓
diabetes	768	8	0	2	64.10%	10,085	✓	✓
electricity	45,312	7	1	2	57.50%	9,210	✓	✓
hepato	536	9	0	4	33.20%	5,055	✓	×
king-and-rook	28,056	0	6	18	16.20%	10,266	×	×
kr-vs-kp	3,196	0	36	2	52.20%	10,542	✓	✓
lymph	148	3	15	4	54.73%	8,934	×	×
madelon	2,600	500	0	2	50.00%	2,431	✓	✓
mushroom	8,124	0	22	2	51.80%	6,468	✓	✓
nursery	12,960	0	8	5	33.30%	9,642	×	×
poker-hand	829,201	5	5	10	50.10%	9,267	×	×
spect	267	0	22	2	79.40%	9,290	×	✓
splice	3,190	0	61	3	51.88%	8,313	×	×
tic-tac-toe	958	0	9	2	65.34%	9,134	×	✓
vowel	990	10	3	11	9.09%	6,921	✓	×
waveform	5,000	40	0	3	33.80%	10,644	✓	×
wine	153	13	0	3	39.87%	4,697	✓	×

Out of all the datasets, 10 can be considered balanced, whereas 10 suffer from class-imbalance. Similarly, 9 datasets represent binary classification problems, while 11 have more than two classes. Most datasets have from few hundred to few thousand examples, with the notable exception of poker-hand which contains 829,201 instances. It is also worth highlighting madelon as the dataset with most descriptive attributes (500) and king-and-rook as the one with most class attribute values (18).

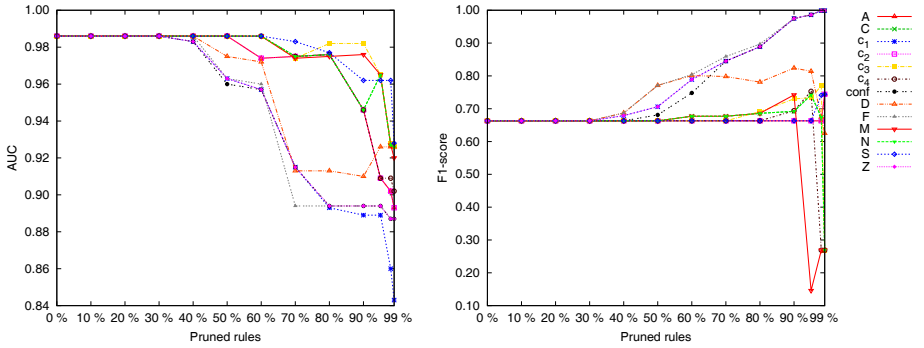
Due to the fact that CM-CAR creates rules from frequent itemsets, it requires instances described only by nominal attributes. Therefore, all numerical attributes were discretized into ten equal-frequency bins. Datasets preprocessed in this way were used in all the discussed experiments.

4.2 Rule Pruning

In this group of experiments, the generated rule set was pruned subsequently by: 0%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, 95%, 98%, 99% of the original model size. Thus, at the extremes the rule set was not pruned at all or was limited to only 1% of the initial set. Due to the large number of tested measures and datasets, we will only present the most interesting results; detailed tables and additional plots are available in the supplementary materials.[2]

[2] Supplement: http://www.cs.put.poznan.pl/dbrzezinski/software/CMCAR.html

For evaluations using the G-mean measure, it was observed that since G-mean multiplies the true positive rate of each class, in situations where the rules did not cover examples from one of the classes the reported performance was zero. This shows that for highly imbalanced data coverage should be additionally controlled. Partially due to this phenomenon, on some of the datasets (madelon, spect, tic-tac-toe, poker-hand, kr-vs-kp, king-and-rook) the differences in performance were very small and did not discriminate confirmation measures in terms of pruning capabilities. However, on the remaining data clear differences were visible, and two cohesive groups of measures were identified: 1) A and c_2; 2) F, Z, and c_1. Figure 1 presents measure performance on two datasets, which exemplify the relations between these two groups.



(a) mushroom (AUC) (b) adult-census (F_1-score)

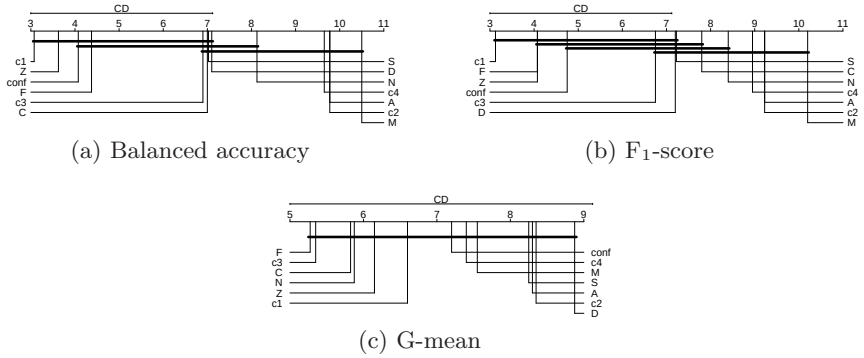
Fig. 1: CM-CAR's AUC on the mushroom dataset and F_1-score on adult-census for different pruning levels with $Q_s = \{conf, sup, length\}$ and $Q_p = \{CM\}$, where CM is one of the measures listed in the legend.

The dependency between measures A and c_2 can be explained by the fact that the value of c_2 is in some cases proportional to the value of A. Such a situation occurs in the case of confirmation and when additionally b (the number of objects not supporting the premise, but supporting the conclusion) is greater than 0. Indeed, analyzing the obtained frequent itemsets we noticed that these two requirements were met for most datasets.

The relation between measures in the second group is more difficult to explain. Under certain conditions, c_1 is proportional to Z, however the interdependence with F is not expressed in any way in the definitions of these measure. It is worth noting that all three measures were among the best performing pruning measures, when evaluated using balanced accuracy, G-mean, AUC, and F_1-score.

To verify the significance of the observed differences, we performed the non-parametric Friedman test [8]. The null-hypothesis of the Friedman test (that there is no difference between the performance of all the tested confirmation measures) can be rejected for balanced accuracy, G-mean, and the F_1-score

with $p < 0.05$, but not for AUC. To verify which confirmation measures perform better than the other, we computed the critical difference (CD) chosen by the Nemenyi post-hoc test [8] at $\alpha = 0.05$. Figure 2 depicts the results of the test for balanced accuracy, F_1-score, and G-mean by connecting the groups of measures that are not significantly different (the lower the rank the better).



(a) Balanced accuracy

(b) F_1-score

(c) G-mean

Fig. 2: Performance ranking of all measures ($Q_s = \{conf, sup, length\}$, $Q_p = \{CM\}$) averaged over all the analyzed pruning levels. Measures that are not significantly different according to the Nemenyi test (at $\alpha = 0.05$) are connected.

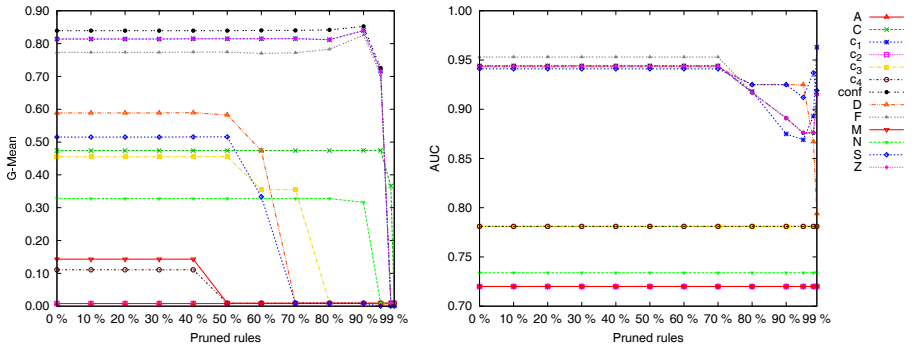
As mentioned earlier, F, Z, c_1 are among the best measures according to balanced accuracy and the F_1-score. Similar rankings were found for G-mean, however, due to the large number of compared measures the post-hoc test for these measures was unable to distinguish groups of measures performing significantly differently. For balanced accuracy and F_1-score, the test was not able to showcase a significant difference with $conf$, S, D and c_3, however, at $\alpha = 0.05$ the three highlighted measures pruned significantly better than C, N, c_4, M, c_2, and A. It is also worth noticing, that according to G-mean $conf$ performs much worse than according to balanced accuracy or F_1-score. This may suggest that $conf$ promotes focusing on overall accuracy potentially neglecting underrepresented minority classes.

4.3 Classification using Confirmation Measures

In the second group of experiments, we used confirmation measures to sort the rule list and, thus, influence the classification procedure. Tables with balanced accuracy, G-mean, AUC, and F_1-score performance for CM-CAR using each of the analyzed measures are available in the supplementary material[2], whereas below we summarize the main findings.

In terms of average predictive performance for all pruning levels, F, Z, c_1 were once again the best performing measures. It is also worth highlighting S and

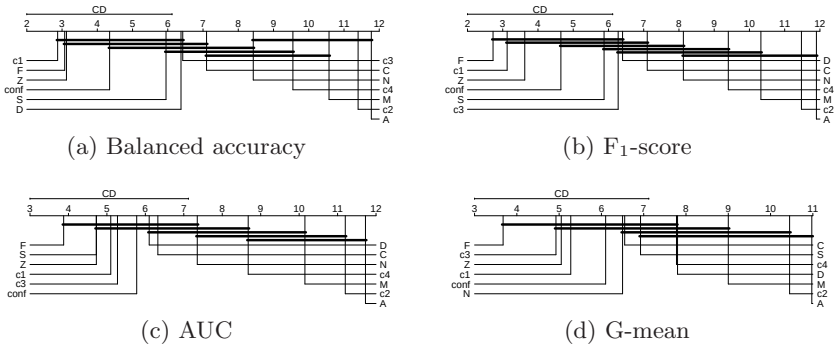
c_3, which were also among the best measures. This is particularly interesting as these measures possess desirable properties, such as minimality/maximality or evidence symmetry and evidence-hypothesis symmetry, which are not showcased by F, Z, or c_1 [16]. Another consistent observation was that of M, A, and c_2 being the worst measures for rule sorting. Two exemplary datasets where these relations can be seen are presented in Fig. 3.



(a) splice (G-mean) (b) nursery (AUC)

Fig. 3: CM-CAR's G-mean on the splice dataset and AUC on nursery for different pruning levels with $Q_s = \{CM, sup, length\}$ and $Q_p = \{CM\}$.

As in the first group of experiments, we performed the Friedman test. The null-hypothesis of the Friedman test can be rejected for all four evaluation measures (balanced accuracy, G-mean, AUC, F-score) with $p < 0.001$. Figure 4 visually presents the results of the Nemenyi test.



(a) Balanced accuracy (b) F_1-score

(c) AUC (d) G-mean

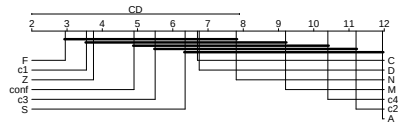
Fig. 4: Performance ranking of all measures ($Q_s = \{CM, sup, length\}$, $Q_p = \{CM\}$) averaged over all the analyzed pruning levels. Measures that are not significantly different according to the Nemenyi test (at $\alpha = 0.05$) are connected.

As the results show, F, Z, c_1 are once again the best measures, and are significantly better at rule sorting than c_4, M, c_2, and A.

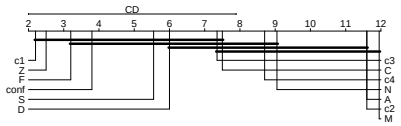
4.4 The Impact of Imbalance Data and Multiple Classes

The previous two subsections analyzed the potential of using Bayesian confirmation measures for rule list pruning and sorting. However, datasets selected for this study allow us to differentiate the performance of the measures on balanced/imbalanced and binary/multi-class problems. The last two columns of Table 3 distinguish both types of dataset categorizations.

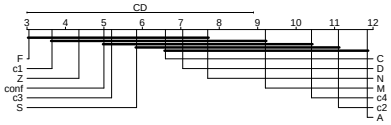
Figures 5 and 6 present the results of Nemenyi post-hoc tests at $\alpha = 0.05$, with performance on balanced/binary in the left column and imbalanced/multi-class data in the right column. Due to space limitations we only show results for strategies where the confirmation measure was used for both pruning and sorting; for additional plots please refer to the supplementary materials.[2]



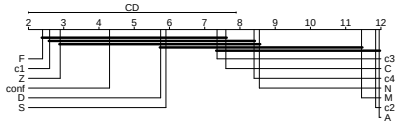
(a) Bal. accuracy (balanced datasets)



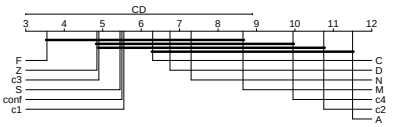
(b) Bal. accuracy (imbalanced datasets)



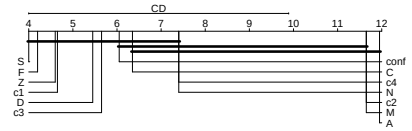
(c) F_1-score (balanced datasets)



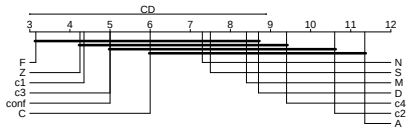
(d) F_1-score (imbalanced datasets)



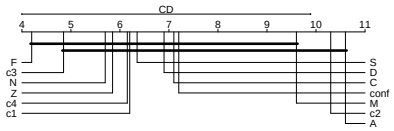
(e) AUC (balanced datasets)



(f) AUC (imbalanced datasets)

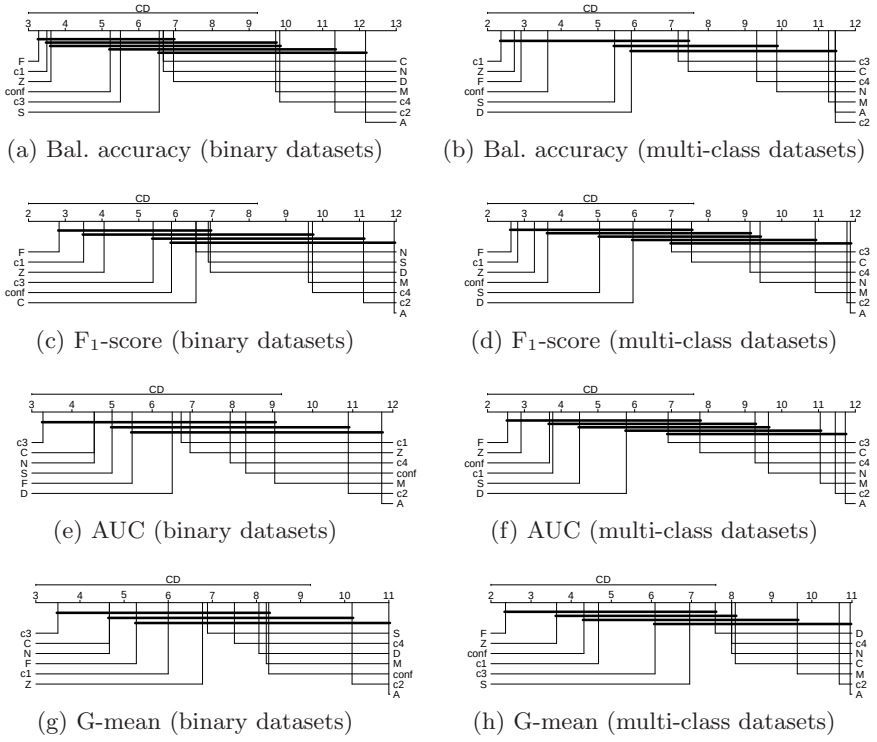


(g) G-mean (balanced datasets)



(h) G-mean (imbalanced datasets)

Fig. 5: Performance ranking of all measures ($Q_s = \{CM, sup, length\}$, $Q_p = \{CM\}$) analyzed separately for balanced and imbalanced datasets. Measures that are not significantly different according to the Nemenyi test (at $\alpha = 0.05$) are connected.



(a) Bal. accuracy (binary datasets) (b) Bal. accuracy (multi-class datasets)

(c) F_1-score (binary datasets) (d) F_1-score (multi-class datasets)

(e) AUC (binary datasets) (f) AUC (multi-class datasets)

(g) G-mean (binary datasets) (h) G-mean (multi-class datasets)

Fig. 6: Performance ranking of all measures ($Q_s = \{CM, sup, length\}$, $Q_p = \{CM\}$) analyzed separately for binary and multi-class problems. Measures that are not significantly different at $\alpha = 0.05$ are connected.

Considering balanced datasets, the results are fairly similar to those obtained when analyzing all datasets and highlight F, Z, and c_1. However, when looking at critical distance plots for AUC and G-mean it is also worth mentioning S, N, and c_3 as highly ranked measures. This is interesting as all three measures possess minimality/maximality, evidence symmetry, and evidence-hypothesis symmetry properties, mentioned previously [16].

Comparing measure rankings on binary and multi-class problems we can see that most evaluations still promote F, Z, and c_1. A slight deviation from this pattern can be seen on critical distance plots of AUC and G-mean for binary datasets, where c_3, N, and C are the three highest ranked confirmation measures.

5 Conclusions

Mining a concise set of descriptive rules that is characterized by good predictive performance is a challenging task. In this paper, to tackle this problem we proposed the CM-CAR algorithm, which uses confirmation measures to sort and

prune a list of rules. Using the proposed algorithm we reviewed the applicability of 12 confirmation measures to rule pruning and sorting.

The results of the experiments show that Bayesian confirmation measures can be successfully applied to reduce the set of rules while maintaining satisfactory predictive performance. In particular, the F, Z, c_1 measures consistently showed better performance than the popularly used *confidence* measure.

An additional analysis comparing results for balanced and imbalanced datasets highlighted N, c_3, and S as promising measures for imbalanced data. This result is particularly interesting as all three measures are well established in the field of interestingness measures and possess additional properties compared to F, Z, c_1, such as: evidence symmetry, evidence-hypothesis symmetry, or minimality/maximality. A similar analysis comparing results for binary and multiclass problems revealed that F, Z, c_1 are ranked highest on both types of problems, with the exception of AUC and G-mean results for binary datasets where c_3, N, and C were the three best confirmation measures.

The results of the research described in this paper inspire us to continue working with confirmation measures in the context of rule-based classification. In particular, we plan to analyze the impact that confirmation measures can have on the coverage of the training set of objects, as in certain applications it is advisable to propose a set of rules that covers the whole or the vast part of the training set. Moreover, based on the results of the comparison performed in this paper, we plan to use selected measures as components of more specialized rule-based classifiers. Finally, a possible extension of CM-CAR can include optimizing the set of classification association rules to those that are not contained by other discovered rules.

Acknowledgments. This work was supported by the National Science Centre grant DEC-2013/11/B/ST6/00963. D. Brzezinski acknowledges the support of an FNP START scholarship and Institute of Computing Science Statutory Fund.

References

1. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules in large databases. In: Proceedings of the 20th International Conference on Very Large Data Bases. pp. 487–499 (1994)
2. Brzezinski, D., Stefanowski, J.: Combining block-based and online methods in learning ensembles from concept drifting data streams. Inf. Sci. 265, 50–67 (2014)
3. Carnap, R.: Logical Foundations of Probability. University of Chicago Press (1962)
4. Ceci, M., Appice, A.: Spatial associative classification: propositional vs structural approach. J. Intell. Inf. Syst. 27(3), 191–213 (2006)
5. Christensen, D.: Measuring confirmation. Journal of Philosophy 96, 437–461 (1999)
6. Cohen, W.W.: Fast effective rule induction. In: Proceedings of the 12th International Conference on Machine Learning (ICML'95). pp. 115–123 (1995)
7. Crupi, V., Tentori, K., Gonzalez, M.: On bayesian measures of evidential support: Theoretical and empirical issues. Philosophy of Science 74, 229–252 (2007)

8. Demsar, J.: Statistical comparisons of classifiers over multiple data sets. J. Mach. Learn. Res. 7, 1–30 (2006)
9. Domingos, P.: The rough set based rule induction technique for classification problems. In: In Proceedings of the Sixth IEEE International Conference on Tools with Artificial Intelligence. pp. 704–707 (1994)
10. Dong, G., Zhang, X., Wong, L., Li, J.: CAEP: classification by aggregating emerging patterns. In: Proceedings of the 2nd International Conference on Discovery Science (DS'99). pp. 30–42 (1999)
11. Eells, E.: Rational Decision and Causality. Cambridge University Press (1982)
12. Fitelson, B.: The plurality of bayesian measures of confirmation and the problem of measure sensitivity. Philosophy of Science 66, 362–378 (1999)
13. Frank, E., Witten, I.H.: Generating accurate rule sets without global optimization. In: Shavlik, J. (ed.) Fifteenth International Conference on Machine Learning. pp. 144–151. Morgan Kaufmann (1998)
14. Geng, L., Hamilton, H.: Interestingness measures for data mining: A survey. ACM Computing Surveys 38(3) (2006)
15. Glass, D.H.: Confirmation measures of association rule interestingness. Knowlegde Based Systems 44, 65–77 (2013)
16. Greco, S., Słowiński, R., Szczęch, I.: Properties of rule interestingness measures and alternative approaches to normalization of measures. Information Sciences 216, 1–16 (2012)
17. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The weka data mining software: An update. SIGKDD Explor. Newsl. 11(1), 10–18 (2009)
18. Han, J., Kamber, M., Pei, J.: Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers Inc., 3rd edn. (2011)
19. Japkowicz, N.: Assessment metrics for imbalanced learning. In: He, H., Ma, Y. (eds.) Imbalanced Learning: Foundations, Algorithms, and Applications, pp. 187–206. Wiley-IEEE Press (2013)
20. Kantardzic, M.: Data Mining: Concepts, Models, Methods, and Algorithms, chap. Data-Mining Applications, pp. 496–509. John Wiley & Sons, second edn. (2011)
21. Kemeny, J., Oppenheim, P.: Degrees of factual support. Philosophy of Science 19, 307–324 (1952)
22. Lichman, M.: UCI machine learning repository (2013), http://archive.ics.uci.edu/ml
23. Liu, B., Hsu, W., Ma, Y.: Integrating classification and association rule mining. In: Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining. pp. 80–86 (1998)
24. McGarry, K.: A survey of interestingness measures for knowledge discovery. The Knowledge Engineering Review 20(1), 39–61 (2005)
25. Mortimer, H.: The Logic of Induction. Paramus, Prentice Hall (1988)
26. Napierala, K., Stefanowski, J.: Addressing imbalanced data with argument based rule learning. Expert Syst. Appl. 42(24), 9468–9481 (2015)
27. Nozick, R.: Philosophical Explanations. Clarendon Press, Oxford, UK (1981)
28. Stefanowski, J.: The rough set based rule induction technique for classification problems. In: In Proceedings of 6th European Conference on Intelligent Techniques and Soft Computing EUFIT. vol. 98 (1998)