Rule Quality Measures Settings in a Sequential Covering Rule Induction Algorithm - an Empirical Approach

Marcin Michalak*, Marek Sikora*†, Łukasz Wróbel†*
*Institute of Informatics, Silesian University of Technology
ul. Akademicka 16, 44-100 Gliwice, Poland
Email: {Marcin.Michalak, Marek.Sikora}@polsl.pl
†Institute of Innovative Technologies EMAG
ul. Leopolda 31, 40-189 Katowice, Poland
Email: {Marek.Sikora, Lukasz.Wrobel}@ibemag.pl

Abstract—The paper presents the results of research related to the efficiency of the so called rule quality measures which are used to evaluate the quality of rules at each stage of the rule induction. The stages of rule growing and pruning were considered along with the issue of conflicts resolution which may occur during the classification. The work is the continuation of research on the efficiency of quality measures employed in sequential covering rule induction algorithm. In this paper we analyse only these quality measures (9 measures) which had been recognised as effective based on previously conducted research.

I. Introduction

THE SEQUENTIAL covering rule induction algorithms can be used both for classification and descriptive purposes [1, 2, 3, 4, 5, 6, 7]. The main principle of the sequential covering rule induction algorithms is unchanged, despite the development of increasingly sophisticated versions of such algorithms. It involves the induction of rules in two phases: growing phase and pruning phase. In the growing phase, the elementary conditions occurring in the rule premise are specified. In the pruning phase, some of these conditions are removed. In comparison with other induction methods, rule sets obtained by the covering algorithms are characterised by good classification as well as descriptive capabilities. Taking into consideration only the classification abilities, better results can be often obtained using other methods, for example neural-fuzzy networks [8, 9] support vector machines [10], or ensemble of classifiers [11]. Models obtained in this way, however, are characterised by much less interpretability than classification (decision) rules. In the case of rule learning for descriptive purposes, the algorithms for induction of rules satisfying certain minimum quality criteria are most commonly used [12, 13, 14]. Induction of all rules that satisfy the given quality criteria leads to the induction of a very large number of rules which then must be limited by the filtering algorithms using so-called rule interestingness measures [15, 16, 17]. If

the primary objective is to describe data, it is possible to use modified covering algorithms (so-called rule-based subgroup discovery [18, 13, 19]). These algorithms aim at the induction of statistically significant rules which cover regions in the feature space of as small union as possible. Rule sets generated in this way are characterised by worse classification abilities than in the standard covering algorithms. If the analysis objective is to create a classification system that uses an interpretable data model, application of sequential covering rule induction algorithms is the most sensible solution. The quality of the rule set obtained by the covering algorithm depends on the quality measure [20, 1, 21, 22, 23, 24, 25, 26] used in the growing and pruning phases. The used quality measure is one of the factors affecting the classification accuracy, the number of rules induced and their other characteristics (e.g. the statistical significance).

The objective of the paper is to present the results of research related to the efficiency of using the combinations of different measures at each stage of the rule induction. The sequential covering approach and the top-down method were applied as the basic induction algorithm. In this approach one can distinguish the stages of the rule growing and pruning. In these stages it is possible to use different optimisation criteria which control the rule creation process. These criteria are commonly called rule quality measures or search heuristics. Different quality measures can also be used at the stage of classification conflicts resolution - this way the mechanism of weighted voting gets modified. The paper features an analysis of the combination of 9 measures which were recognised as effective (this choice will be explained further in the paper). The analysis was conducted based on 34 data sets. In addition, the paper contains some supplementary information about the applied quality measures.

II. Related Works

Measures used for the rule evaluation are divided into two categories [16]: objective and subjective measures. Most of the objective measures are defined on the basis of the contingency

table that characterises the quality of a rule in the context of a fixed (usually training or validation) data set [27, 23]. The task of the subjective measures is to evaluate the rules according to subjective user's preferences. The subjective measures evaluate such features of rules as: unexpectedness, surprisingness, novelty, utility or actionability [16]. The method of measuring these features depends on the application domain and the purpose of analysis. Our range of interest includes objective rule quality measures, also known as evaluation or search heuristics [22, 28, 29]. At present, in the literature over 50 objective measures can be found [20, 22, 15, 23, 28, 24, 26, 29]. A lot of measures used for evaluation of decision (classification) rules have wider significance - they function also as rule interestingness (attractiveness) measures [15, 30, 16, 17, 31]. Interestingness measures are used for evaluation of already induced rules, both decision and association. In the next part of the paper we will use the common term quality measures. The quality measure applied in decision rule induction is significant for the quality of an output set of rules. This is confirmed by empirical research [32, 20, 21, 23, 25, 26].

In [22, 23] Fürnkranz and Janssen conducted an analysis of the efficiency of measures which contain one or several parameters. These authors present a methodology to adapt the values of parameters contained in quality measures to the specifics of the analysed data set. In addition, they demonstrate that, as the parameters change, certain measures begin to behave similarly to others (such an analysis was conducted for m-estimates and Klösgen measures). In our articles [26, 33], in turn, we are analysing the measures which do not contain parameters and we demonstrate their efficiency on a data set similar to the one quoted in Fürnkranz's work. However, we use an algorithm which generates the rule set instead of a rule list. This results in the necessity to use the mechanism for classification conflicts resolution, i.e. there is another place (apart from rule growing and pruning) where a rule quality measure can be applied. This way it is possible to achieve better accuracy of the classification, particularly higher specificity and sensitivity of the classifier [33].

There are also works that deal with the decision rule length. In the paper [34] the analysis of relationships between length and the coverage of the decision rule is described. Relationships are considered as the influence of the length on the rule coverage and also as the reversed: the influence of the rule coverage on its length. In the other paper [35] a Dominance Based Rough Set Approach generated rules, and their length, determines the relevance of attributes, what becomes a criterion for a feature selection. There was also an attempt to characterise rule induction as a process specified by domain ontology [36], describing all data mining algorithms.

III. RULE INDUCTION

Let us assume that a finite set Tr of examples is given. Each training example is described by a set $A \cup \{d\}$ of features, i.e. $a : Tr \rightarrow Va$ for each $a \in A \cup \{d\}$. The set Va is called the range of the attribute a. Elements of A are called conditional attributes, the variable d is called a decision

attribute, and its value is identified with the assignment of an example to a specific concept (decision class). Conditional attributes can be of symbolic (discrete-valued) or of numeric (real-valued) type. The decision attribute is of symbolic type. Each example x belonging to the Tr set can be written as a vector $x = (x_1, x_2, \ldots, x_{|A|}, y)$, where $a_i(x) = x_i$ for each $i \in \{1, 2, \ldots, |A|\}$ and $y = d(x)$. The conditional expression of the following form is called a decision rule:

$$IF \ w_1 \ AND \ w_2 \ AND \ \ldots \ AND \ w_k \ THEN \ d = v$$

An example satisfying all elementary conditions w_i is assigned to the concept indicated in the rule conclusion. Construction of elementary conditions w_i may be various and depends on a rule induction algorithm. Any elementary condition is most often the expression of the one of the two following forms: $w_i \equiv a_i \ op \ Z_i$, where a_i is the name of the conditional attribute, op is one of the relation operators $\{=, \leq, \geq, >, <\}$, Z_i is the element of the range of the attribute or just $w_i \equiv a_i \in Z_i$ where Z_i is a interval in the range Va_i: $Z_i \subset Va_i$. The positive examples are those belonging to the decision class pointed in the rule conclusion. The negative examples are the remaining ones.

Let p denote the number of positive examples covered by the rule (P stands for all positive examples in the training set), and let n denote the number of negative examples covering the rule (N stands for all negative examples). On the basis of this notation a contingency table for the rule can be built (Table I).

TABLE I
THE CONTINGENCY TABLE FOR THE RULE COVERING p POSITIVE AND n NEGATIVE EXAMPLES.

p	n	$p + n$
$P - p$	$N - n$	$P + N - p - n$
P	N	$P + N$

The aim of majority of rule induction algorithms is to find the minimal set of classification rules which cover and correctly predict decision classes of a given set of examples and which additionally are characterised by high values of p and low values of n. Finding the optimal solution for such a problem is a computationally expensive task, therefore most of the rule induction algorithms use some heuristics. One of the most common approaches, used also in our implementation, is the sequential covering (known also as separate-and-conquer) strategy [4, 6].

To put it briefly, this strategy consists in learning a rule which covers some part of training examples. Next, the examples covered by the learnt rule are removed from the training set and the rule learning process starts recursively for the remaining examples.

Our implementation of the rule induction algorithm also works in the separate and conquer fashion. The outcome of the algorithm is a set of rules describing each decision class of a training set. The process of induction of a single rule consists of two phases: growing and pruning. In the growing

```
1  RuleInduction(examples, ruleQualityMeasure)
2  the set of generated rules (initially empty)
3  ruleSet := ∅
4  decisionClass is a set of examples which have the same value of
   decision attribute
5  foreach decisionClass in examples do
6      uncoveredPositives := decisionClass
7      while uncoveredPositives ≠ ∅ do
8          rule with empty antecedent and consequent pointing current
           decision class
9          rule := ∅
10         the set of examples covered by the rule (initially empty rule
           covers all examples)
11         covered := examples
12         rule growing phase
13         do
14             conditions := PossibleElementaryConditions(covered,
               uncoveredPositives)
15             bestQuality := −∞
16             bestCondition := ∅
17             foreach c in conditions do
18                 evaluate the quality of rule with c condition added
                   to its antecedent
19                 quality := Evaluate(rule c, ruleQualityMeasure)
20                 if quality > bestQuality then
21                     bestQuality := quality
22                     bestCondition := c
23                 end
24             end
25             add selected elementary condition to the antecedent of
               the rule
26             rule := rule ∪ bestCondition
27             covered := Covered(rule, examples)
28         while until (stop criterion);
29         rule pruning phase
30         Prune(rule, ruleQualityMeasure)
31         covered := Covered(rule, examples)
32         uncoveredPositives := uncoveredPositives \ covered
33         ruleSet := ruleSet ∪ rule
34     end
35 end
36 return ruleSet
```

Algorithm 1: The algorithm of rule induction

phase, the elementary conditions are added one by one to the premise of the rule. In the case of nominal attributes, the elementary condition can take the form of $(a = v)$, and for the numerical attributes it can take one of two forms: $(a < v)$ or $(a \geq v)$. For the numerical attributes the value v is the arithmetic mean between two successive values from the range of attribute a. The set of all the possible elementary conditions which might be added to the rule is created on the basis of examples currently covered by the rule. It means that the domains of the attributes are narrowed to the values which are taken by the examples covered by the currently formed rule. Moreover, the elementary condition is tested only if its addition to the rule causes that such a refined rule covers at least one positive example not covered by rules generated so far. The refinement that has the highest value of the specified rule quality measure among all refinements possible in a single step is selected as the final one. If several refinements obtain the same value of the rule quality measure, the refinement covering more uncovered positive examples is chosen.

During the induction of rules it should be taken into account that some attribute values of individual examples are not always known. In the literature, several propositions of strategies for handling unknown values of the attributes in rule learning algorithms can be found [37]. In our implementation it is assumed that the elementary condition always returns truth value false for the unknown value of the attribute. It means that if the example has unknown values, it can be covered only by the rule which does not contain attributes with unknown values for this example.

The process of rule growing is terminated if the rule is accurate (i.e. it covers none of the negative examples) or if the addition of the next conditions to the rule no longer increases its precision (which may take place if examples with identical attributes values but different decisions exist in the training set). After the rule growing phase, the rule is pruned. The rule pruning algorithm uses a hill-climbing strategy. At each iteration, it deletes the elementary condition without which the rule has the highest improvement in the value of quality measure. After each removal, the value of rule quality measure is recalculated on the entire set of training examples and the deletion of conditions is repeated until it does not cause decrease in the current value of quality measure. The pruned rule is added to the final set of rules, and then the process of rule induction starts again for the rest of the uncovered positive examples.

The assumed heuristic of rule building therefore focuses on induction of rules characterised by high values of specified quality measure and thus constitutes a generic framework for studying behaviour of various quality measures.

IV. RULE QUALITY MEASURES

Quality measures strongly associated with decision rules evaluation are collected in surveys [27, 21, 23, 28, 29].

The papers [27, 21] focus mainly on studying measures influence' on conflicts resolution during classification. Janssen and Fürnkranz [23] present research on parametrised quality measures and the procedure of data driven selection of their parameters. Papers by Yao and Zhong [29] and Lavrac, Flach, and Zupan [28] focus on defining quality measures in the language of probability. Interestingly enough, the mentioned authors' considerations concern disjoint sets of measures.

Our choice of nine decision rule quality measures is provided in Table II. The selection is a result of a previous experiments — in already quoted works [26, 33], about 30 measures were analysed with respect to their efficiency. To check the efficiency, the measures were used in the rule induction algorithm - the same measure was used at each induction stage and during the resolution of classification conflicts. Then the classification abilities of the obtained rule classifier were checked. The classification abilities were analysed from the point of view of overall classification accuracy and balanced accuracy. In the case of two decision classes and the discrete classifier, this measure is equivalent to the AUC criterion [38]. In the case of a bigger number of classes, the balanced accuracy informs about the average accuracy of each decision class. This measure is suitable to verify the classifier which works on unbalanced data. Additionally, the number

of rules was analysed with respect to the interpretability of the obtained rule sets. Without going deeper in the measure selection methodology, which is presented for example in [33], 9 evaluation measures were selected. These measures achieved good results, at least from the point of view of one of the three above mentioned criteria - statistically they performed better than other measures in the majority of data sets.

<div align="center">

TABLE II
THE LIST OF SELECTED OBJECTIVE RULE QUALITY MEASURES.

</div>

$$\mathbf{g} = \frac{p}{p+n+2}$$

$$\mathbf{wLap} = \frac{(p+1)(P+N)}{(p+n+2)P}$$

$$\mathbf{LS} = \frac{pN}{nP}$$

$$\mathbf{Rss} = \frac{p}{P} - \frac{n}{N}$$

$$\mathbf{MS} = \frac{p}{n+P}$$

$$\mathbf{C1} = Coleman \cdot \frac{2+Cohen}{3}$$

$$\mathbf{C2} = Coleman \cdot 0.5 \left(1 + \frac{p}{P}\right)$$

$$\mathbf{corr} = \frac{pN - Pn}{\sqrt{PN(p+n)(P-p+N-n)}}$$

$$\mathbf{s} = \frac{p}{p+n} - \frac{P-p}{P-p+N-n}$$

$$Cohen = \frac{(P+N)(\frac{p}{p+n}) - P}{\frac{P+N}{2}\frac{p+n+P}{p+n} - P}$$

$$Coleman = \frac{(P+N)\frac{p}{p+n} - P}{N}$$

Now, we will focus on a brief discussion of the measures and presentation of their properties. The $g - measure$ ($g = 2$) was proposed by Fürnkranz and Flach [22] and is a simplified version of the Laplace estimate. The $g-measure$ ($g = 2$) assumes that a rule covering only one positive example is assigned true precision equal to 0.33. Many experiments (e.g. in papers [39, 23]) indicate that the precision of a rule evaluated on a training set is too optimistic. This especially concerns rules covering a small number of positive examples. If a rule covers large number of positive examples, the correction introduced by the number 2 has less and less meaning.

The $Weighted\ Laplace$ ($wLap$) measure is derived from decision tree induction algorithms and it is the Laplace estimate multiplied by $(P+N)/P$. So it takes into account the distribution of examples between the current positive class (P) and the remaining classes (N).

The $Logical\ sufficiency$ (LS) measure is applied to association rule evaluation as well as to decision rule induction [29]. It belongs to the group of measures which emphasise precision of rules during their evaluation and pay less attention to the number of examples covered by the evaluated rule.

The RSS measure is a measure equivalent, in terms of the generated rule order, to the well-known weighted relative accuracy (WRA) measure used by Lavrac, Flach, and Zupan [28], both in rule induction and subgroup discovery. However, RSS and WRA have different ranges of values. The use of RSS for classification conflicts resolving leads to better results than the use of the WRA measure.

The $MutualSupport$ (MS) measure is, in context of decision rule evaluation, presented in the paper by Yao and Zhong [29]. Let us assume that a rule of the form $\varphi \rightarrow \psi$ is given. MS measures the strength of dependencies not only between $\varphi \rightarrow \psi$, but also between $\varphi \leftarrow \psi$. It follows that we can consider it as a measure evaluating the strength of the double implication $\varphi \leftrightarrow \psi$. Therefore, the measure prefers rules characterised by high coverage.

The $Correlation$ ($Corr$) measure computes the correlation coefficient between the predicted and the target labels. It was applied to classification rule induction algorithms as well as to subgroup discovery and evaluation of association rules [15, 23, 40].

The $s - Bayesian\ confirmation$ (s) measure has been proposed by Christensen [41] and Joyce [42] as a confirmation measure. The first component of the measure evaluates the rule precision and the second is responsible for decrease in the quality of rules that cover small number of positive examples. Its application in evaluation of decision rules obtained by the rough set theory was considered in papers [43, 30].

Measures C1 and C2 have been proposed by Bruha [27, 21]. The measures are a combination of two quality measures known as the Coleman and Cohen (see Table II) measures. Bruha noticed that the Coleman measure prefers rules with high precision and small coverage, while the Cohen measure prefers rules with high coverage. Hence, C1 and C2 join the assessment made by the Coleman and Cohen measures

V. EXPERIMENTS AND RESULTS

In this paper, the performance of the rule-based classifier is evaluated by the two criteria: overall classification (Acc) accuracy and balanced accuracy ($BAcc$). The overall accuracy is the ratio of the number of correctly classified examples to the number of all examples. This is one of the most common criteria for assessing a classifier. However, in the case of unbalanced distribution of examples between decision classes, higher value of overall accuracy is often achieved at the cost of low accuracy of minority classes, therefore in such a case the balanced accuracy is more appropriate. It calculates the classification accuracy of each decision class and then takes an average over all classes. In the case of 2-class problems, balanced accuracy is equivalent to Area Under the ROC Curve (AUC) criterion [38]. One can meet with the interpretation that balanced accuracy calculated for a number of classes is a generalization of AUC for multi-class problems [44].

The used rule induction algorithm generates an unordered set of rules, therefore during the classification of examples it may happen that the test example is covered by rules pointing at different decision classes. In that case, a strategy

TABLE III
CHARACTERISTICS OF DATA SETS USED IN EMPIRICAL STUDIES.

dataset	cl.	obj.	maj. class	attributes all	nom.	dataset	cl.	obj.	maj. class	attributes all	nom.
anneal	5	898	76.17	38	32	hepatitis	2	155	79.35	19	13
audiology	24	226	25.22	69	69	horse-colic	2	368	63.04	22	15
auto-mpg	3	398	62.56	7	2	hungarian-heart-disease	2	294	63.95	13	7
autos	6	205	32.68	25	10	iris	3	150	33.33	4	0
balance-scale	3	625	33.33	4	0	mammography-masses	2	961	53.69	5	2
breast-cancer	2	286	70.28	9	9	mushroom	2	8124	51.80	22	22
breast-w	2	699	65.52	9	0	prnn-synth	2	250	50.00	2	0
car	4	1728	70.02	6	6	segment	7	2310	14.29	19	0
cleveland	5	303	54.13	13	7	sick-euthyroid	2	3772	93.88	29	22
contact-lenses	3	24	62.50	4	4	sonar	2	208	53.37	60	0
credit-g	2	1000	70.00	20	13	soybean	19	683	13.47	35	35
cylinder-bands	2	540	57.78	35	18	titanic	2	2201	67.70	3	3
diabetes	2	768	65.10	8	0	vehicle	4	846	25.77	18	0
echocardiogram	2	131	67.18	11	2	vote	2	435	61.38	16	16
ecoli	8	336	42.56	7	0	wine	3	178	39.89	13	0
flag	4	194	46.91	28	18	yeast	10	1484	31.20	8	0
heart-statlog	2	270	55.56	13	0	zoo	7	101	40.59	16	15

for resolving such conflicts has to be chosen. The most popular one is known as the voting scheme and consists in assigning a numeric value called the confidence degree to each rule. Confidence degrees of the rules covering the test example are summed up for each decision class and then the class with a maximal confidence degree is picked. In this paper the voting scheme is also used during classification. Initially, we assumed that the confidence degree of each rule is equal to its value of some quality measure used during the rule induction. If the test example is not covered by any of the rules, it is treated as wrongly classified.

All presented results stand for average values obtained based on the analysis of 34 data sets from the UCI Repository [45]. The analysis of each set was conducted with the use of the stratified 10-fold cross validation strategy. Comparisons were made on the same partitions of the data sets. The characteristics of these sets are presented in Table III. As it can be seen, the sets with different characteristics were selected for the analysis, particularly in relation to the number of decision classes, the distribution of examples between the classes and types of the attributes. The following abbreviations mean: *cl.* – number of classes, *obj.* – number of objects, *maj. class* – a majority class fraction in %, *nom.* – a number nominal attributes.

Tables IV and V contain the results of the classification of rule based classifiers which were obtained in such a way that the same quality measure was used at each stage of the rule induction and during the resolution of classification conflicts.

It can be seen that the C2 measure leads, on average, to the highest overall classification accuracy, wLap to the highest balanced accuracy, and RSS to the smallest number of rules but with significantly worse classification qualities (in the paper [33] it was proven that this worsening is statistically relevant). What is more, the Corr measure can be a reasonable compromise between the number of generated rules and their classification abilities. A more detailed analysis of the measures efficiency can be found in [26, 33]. Table V presents yet unpublished results which characterise the rule

TABLE IV
RESULTS OF THE MEASURES COMPARISON ON 34 DATA SETS.

measure	rules	$Acc[\%]$	$BAcc[\%]$
g	144	81.4	74.2
wLap	156	81.1	77.8
LS	172	80.7	75.8
RSS	34	78.2	72.3
MS	32	77.8	70.5
C1	151	82.0	73.0
C2	127	82.3	76.3
Corr	44	79.9	74.3
s	86	80.3	74.9

TABLE V
CHARACTERISTICS OF THE INDUCED RULE SETS.

Measure	Classification conflicts		Rules			
	all	wrong	number	elementary conditions	Avg Prec.	Avg Cov.
g	32	10	144	2.9	0.96	0.32
wLap	29	9	156	2.6	0.98	0.26
LS	29	8	172	2.4	0.98	0.19
RSS	58	18	34	3.6	0.77	0.74
MS	59	19	32	3.8	0.73	0.77
C1	29	8	151	2.5	0.98	0.28
C2	33	9	127	2.8	0.96	0.36
Corr	53	16	44	3.7	0.81	0.68
s	51	15	86	2.6	0.96	0.34

sets generated by means of particular measures. Here we can find the following interesting regularities:

- the highest values of Acc and $BAcc$ take (with a few exceptions) the measures leading to induction of more rules, the average precision of the rule (denoted as $AvgPrec$) obtained by such measures is greater than 0.95, the average coverage (denoted as $AvgCov$) is not greater than 0.36; if the induction of rules aims at classification, it is appropriate to use measures attaching the utmost importance to the rule precision;
- in the classifiers obtained based on measures leading

TABLE VI
AVERAGE RANKING OF DIFFERENT MEASURE APPLICATION ON THE CLASSIFICATION PHASE.

induction	criterion	C1	C2	Corr	g	LS	MS	Rss	s	wLap
g	*Acc*	5.0	4.5	5.1	5.1	4.5	4.5	4.1	5.9	6.4
	BAcc	5.0	5.1	5.0	6.3	5.2	5.5	4.9	4.2	3.8
C1	*Acc*	4.5	3.8	4.4	5.3	5.4	4.8	4.7	5.6	6.5
	BAcc	4.7	4.9	4.2	6.8	5.6	5.2	5.0	4.3	4.3
C2	*Acc*	4.1	3.9	4.6	4.9	5.9	5.9	6.0	4.7	5.1
	BAcc	4.5	4.9	4.6	6.2	5.3	6.1	6.1	4.1	3.3
Corr	*Acc*	4.3	4.1	4.7	5.0	5.0	5.6	5.7	4.4	6.1
	BAcc	4.8	4.7	4.9	6.4	3.2	6.8	5.3	5.0	4.2
wLap	*Acc*	4.5	4.3	4.1	5.3	5.0	5.0	4.8	5.8	6.3
	BAcc	4.6	5.0	4.2	6.5	5.3	5.9	5.5	4.2	3.8
LS	*Acc*	4.0	3.9	4.9	4.6	4.9	5.0	4.8	6.2	6.7
	BAcc	4.5	4.9	4.5	6.6	5.7	5.5	5.1	4.5	3.6
MS	*Acc*	4.7	4.3	4.8	4.7	5.9	4.7	5.9	4.1	5.9
	BAcc	4.6	4.7	4.9	6.1	3.5	6.5	5.3	5.4	4.1
Rss	*Acc*	4.3	4.3	4.9	5.6	3.9	5.5	5.6	4.9	6.1
	BAcc	4.9	5.1	4.9	6.6	3.2	6.2	5.0	4.9	4.3
s	*Acc*	4.3	4.4	4.8	4.6	5.4	5.4	5.8	4.5	5.9
	BAcc	4.2	4.6	5.0	6.2	4.7	6.1	5.9	4.1	4.3
avg	*Acc*	4.4	**4.2**	4.7	5.0	5.1	5.2	5.3	5.1	6.1
	BAcc	4.6	4.9	4.7	6.4	4.6	6.0	5.3	4.5	**4.0**

TABLE VII
CLASSIFICATION IMPROVEMENT COMPARISON.

Induction	Class.	Acc [%]	Wilcoxon test p-value		Induction	Class.	BAcc [%]	Wilcoxon test p-value
g	g	81.37	0.1271		g	g	74.17	0.0050
	C2	82.24				wLap	78.15	
C1	C1	81.95	0.1885		C1	C1	76.91	0.0512
	C2	82.06				wLap	77.81	
corr	corr	79.77	0.0572		C2	C2	76.22	0.0066
	C2	80.28				wLap	78.35	
wLap	wLap	81.02	0.0154		corr	corr	74.21	0.2206
	C2	82.80				wLap	76.01	
LS	LS	80.80	0.3684		LS	LS	75.67	0.0161
	C2	81.02				wLap	77.04	
MS	MS	77.90	0.5723		MS	MS	70.67	0.0090
	C2	77.14				wLap	74.57	
Rss	Rss	78.07	0.0383		Rss	Rss	72.21	0.0672
	C2	79.08				wLap	74.06	
s	s	80.39	0.9357		s	s	75.03	0.6811
	C2	79.43				wLap	74.92	

to the induction of a smaller number of more general rules (the average precision of induced rules is not higher than 0.96, the average coverage is not less than 0.34), there is a large number of classification conflicts (Classification conflicts); a large number of these conflicts is resolved incorrectly, which explains the lower quality of the classifiers;

- the larger the rule coverage, the greater the number of positive examples covered by them uniquely;
- the rule average precision and rule average coverage are correlated with the number of induced rules;
- the average number of elementary conditions (Elementary conditions) contained in the rule premises is correlated with the number of induced rules and their average coverage.

The experiments show that the rules generated by means of measures which lead to the induction of a large number of precise rules (precision > 0.95) contain fewer elementary

conditions than the rules obtained with the use of measures which promote more general rules. This means that the ranges of elementary conditions, set by means of MS and RSS measures, are wide. This, in turn, implies that each individually considered condition covers a large number of examples, including negative ones. In order to have more detailed rules with "wide" elementary conditions, it is necessary to place more conditions in the rule premise. The measures which are oriented towards the induction of precise rules (wLap, C1, C2) make conditions with narrower ranges, thus the number of conditions in the rule premise is smaller.

The first experiment to check whether it is sensible to use the combination of measures aimed at selecting a measure to resolve classification conflicts. The experiment was conducted in such a way that for the set measure (the same at the stages of growing and pruning) applied in the rule induction, the measure for resolving classification conflicts was changed. The results are presented in Table VI.

The first column of this table denotes the quality measure used in the rule induction. The second column denotes the quality criterion of a final classification phase. The following nine columns correspond to nine quality measures used during the voting in the classification. Each of these nine values is an average rank of a specific classifier (ranks for the classifiers should be read by rows) — the lower is the value of the ranking, the better are the results achieved by the measure (value equal to 1 would mean that the measure was always first in the ranking). Experiments were performed on 34 mentioned datasets in a stratified 10 fold cross validation model. A Friedman test with the significance level $\alpha = 0.05$ does not show statistical differences only for the s measure and the criterion Acc. If post–hoc analysis is performed and the critical distance is calculated it occurs that most of measures behave in a similar way: there are no significant differences between classification accuracy between different measures application. However, it is worth to notice that for the Acc criterion $C2$ measure is the right choice and for the $BAcc$ criterion the $wLap$ is the right choice. $C2$ measure (according to Acc criterion) is the best in 5 of 9 cases (measures) and is always in the best three places. $wLap$ measure (according to $BAcc$ criterion) is four time the best one and not worse than third in other cases. Additionally, an averaging row is also included in the bottom of the table. Emphasised averaged values confirm the above conclusions.

The influence of application of C2 measure during classification instead of a measure used in the rule induction is presented on the left side of the Table VII. The first column points the measure used in a rule induction, second column points the measure used during the classification and the third column means the Acc of a classification. Additional column remarks the p-value of the Wilcoxon test, which null hypothesis is that there is no statistical significant difference between the algorithm which uses the same measure during classification and during the rule induction and the algorithm which uses the C2 measure during classification. Analogical summary for the wLap measure applied in the classification with the $BAcc$ classification quality criterion is presented on the right side of the Table VII.

As it can be observed in the most of cases the final classification quality is improved as the result of application a specified rule quality measure. In the case of a C2 measure this result is rather expected as this measure is strictly dedicated for the classification conflict resolution purposes [27, 21]. wLap measure also improves classification results and the statistical significance of this improvement is very high in the most of cases. This is a new and important observation.

Having the selected rule quality measures for two criteria of classification quality (Acc and $BAcc$) we performed experiments on application of different rule quality measures during the rule growing and pruning phase. This time rankings were calculated on the basis of 81 experiments (nine measures for the growing and nine for the pruning). Results are presented on Fig. 1.

It features a matrix which reflects the ranking of the

Growing/Pruning/ Classification	Acc	Rules	Wilcoxon
1. C2-C2-C2	82.3	127	1 - 2 (+) 1 - 3 (+) 1 - 4 (+)
2. C2-Corr-C2	80.1	47	2 - 3 (-) 2 - 4 (-)
3. Corr-C2-C2	80.4	46	3 - 4 (+)
4. Corr-Corr-Corr	79.8	44	
1. wLap-wLap-C2	81.8	156	1 - 2 (+) 1 - 3 (+) 1 - 4 (+)
2. wLap-Corr-C2	79.4	47	3 - 2 (+) 2 - 4 (-)
3. Corr-wLap-C2	80.4	46	3 - 4 (+)
4. Corr-Corr-Corr	79.8	44	
1. C2-C2-C2	82.3	127	1 - 2 (+) 1 - 3 (+) 1 - 4 (+)
2. C2-RSS-C2	78.7	34	2 - 3 (-) 2 - 4 (-)
3. RSS-C2-C2	79.3	37	3 - 4 (+)
4. RSS-RSS-RSS	78.2	34	
1. wLap-wLap-C2	81.8	156	1 - 2 (+) 1 - 3 (+) 1 - 4 (+)
2. wLap-RSS-C2	78.3	30	3 - 2 (+) 2 - 4 (-)
3. RSS-wLap-C2	79.4	37	3 - 4 (+)
4. RSS-RSS-RSS	78.2	34	

given combination of measures. As it was already mentioned, 81 combinations of measures were checked. For the Acc criterion, the conflicts resolving measure was always the C2 measure, while for $BAcc$ - the wLap measure. The matrices show that it is not possible to draw straightforward conclusions from such a general comparison. It is interesting that the values presented in antidiagonal (X-X - showing the case when the same measure is used in both phases of the rule induction) are very close to the lowest values in row X and column X. A more detailed analysis was conducted for the C2 and wLap measures (as the best due to their classification abilities) as well as Corr and RSS (as the best due to the number of generated rules). In Tables VIII and IX an analysis of the combination of these measures was conducted for the Acc (Table VIII) and $BAcc$ (Table IX) criteria. The first column of both tables includes the combination of measures, while the second one - the Acc ($BAcc$) value obtained by the given combination. Finally, the third column features the achieved number of rules. The fourth column contains the results of the paired comparison of particular combinations (the Wilcoxon test was used; p-value=0.1) of four models. X-Y (+) means that the combination X is statistically better than Y. The (-) symbol means that there is no such a difference.

Let us assume that the precise measure (C2, wLap) is marked P, while the general measure (Corr, RSS) - C. Additionally, the P-C inscription means that in the phase of the rule growing the precise measure was used, while in the
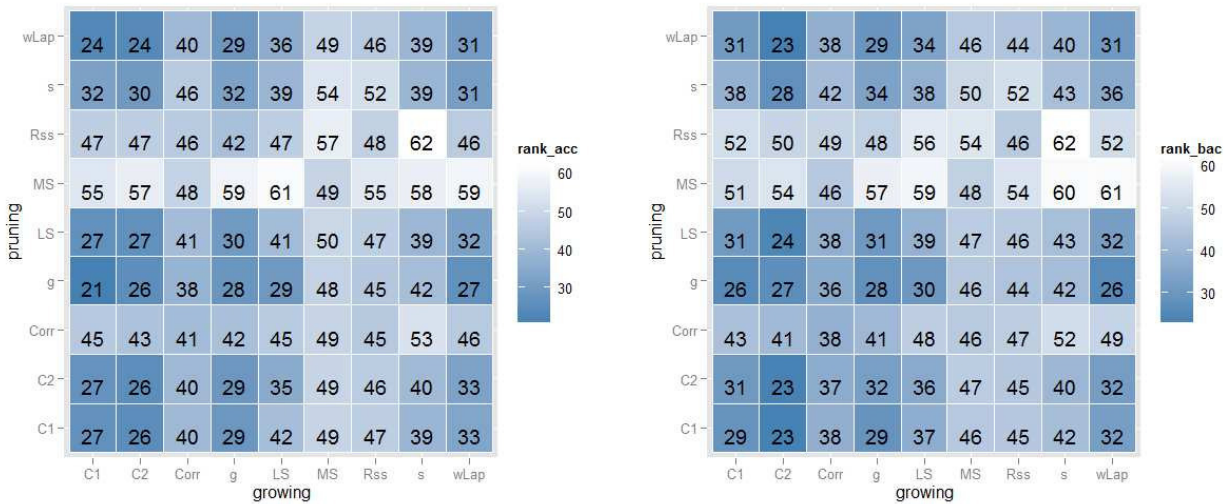
Fig. 1. Matrices of average rankings of models with different quality measures used in growing and pruning but a common classification quality: *Acc* (left) and *BAcc* (right).

TABLE IX
COMPARISON OF RESULTS OF THE COMBINATION OF QUALITY MEASURES
FOR THE *BAcc* .

Growing/Pruning/ Classification	Acc	Rules	Wilcoxon
1. wLap-wLap-wLap	77.8	156	1 - 2 (+) 1 - 3 (+) 1 - 4 (+)
2. wLap-Corr-wLap	74.7	47	3 - 2 (+) 2 - 4 (-)
3. Corr-wLap-wLap	76.1	46	3 - 4 (+)
4. Corr-Corr-Corr	74.4	44	
1.C2-C2-wLap	78.4	127	1 - 2 (+) 1 - 3 (+) 1 - 4 (+)
2. C2-Corr-wLap	75.4	47	2 - 3 (-) 2 - 4 (-)
3. Corr-C2-wLap	76.1	46	3 - 4 (+)
4. Corr-Corr-Corr	74.4	44	
1. wLap-wLap-wLap	77.8	156	1 - 2 (+) 1 - 3 (+) 1 - 4 (+)
2. wLap-RSS-wLap	72.3	30	3 - 2 (+) 2 - 4 (-)
3. RSS-wLap-wLap	74.5	37	3 - 4 (+)
4. RSS-RSS-RSS	72.3	34	
1. C2-C2-wLap	78.4	127	1 - 2 (+) 1 - 3 (+) 1 - 4 (+)
2. C2-RSS-wLap	73.2	34	2 - 3 (-) 2 - 4 (-)
3. RSS-C2-wLap	74.4	37	3 - 4 (+)
4. RSS-RSS-RSS	72.3	34	

rule pruning - the general one. Both tables demonstrate the following dependencies:

- the P-C combination, when compared with P-P, leads to a significantly worse quality of the classifier and to a decreasing number of rules,
- the C-P combination, when compared with C-C, leads to

a significant increase of the classification quality with, simultaneously, a small increase in the number of rules,
- it is not significant for the number of determined rules whether we use the C-P or P-C sequence, however it is recommended to use C-P for classification abilities as it allows to significantly raise these classification abilities (with respect to C-C).

Moreover, the C2-C2-C2 sequence allowed to achieve the highest classification accuracy. No other combination allowed to statistically improve the *Acc* value. On the other hand, the C2-C2-wLap combination leads to significantly higher *BAcc* values, than the wLap-wLap-wLap combination.

VI. CONCLUSIONS

When real problems are solved (analysed), the measure is usually selected with respect to the data set specifics. The automatic method for the measures selection is based on internal cross-validation. Our earlier works allowed to limit the number of measures which are considered by the automatic method [33], which significantly speeds up the calculations. The results of this work extend the knowledge about the combinations of measures which should be tested first in the automatic method. Our results can also be useful for the user for whom the maximisation of the classification accuracy is not the most important criterion, while it is more important to have proper balancing of the model complexity (number of rules) with its classification abilities.

Currently the RapidRuleInduction plug-in is prepared for the RapidMiner environment. The plug-in will contain the possibilities to induce the rules by means of combining different quality measures - similarly to the plug-in we developed for the R environment [46].

REFERENCES

[1] J. Błaszczyński, R. Słowiński, and M. Szeląg, "Sequential covering rule induction algorithm for variable consistency rough set approaches," *Information Sciences*, vol. 181, no. 5, pp. 987–1002, 2011. doi: 10.1016/j.ins.2010.10.030. [Online]. Available: {http://dx.doi.org/10.1016/j.ins.2010.10.030}

[2] W. Cohen, "Fast effective rule induction," in *Proc. of the 12th Int. Conference ICML95*, 1995, pp. 115–123.

[3] T. Fawcett, "Prie a system for generation rulelist to maximize roc performance," *Data Mining and Knowledge Discovery*, vol. 17, pp. 207–224, 2008. doi: 10.1007/s10618-008-0089-y. [Online]. Available: {http://dx.doi.org/10.1007/s10618-008-0089-y}

[4] J. Fürnkranz, "Separate-and-conquer rule learning," *Artificial Intelligence Review*, vol. 13, no. 1, pp. 3–54, 1999. doi: 10.1023/A:1006524209794. [Online]. Available: {http://dx.doi.org/10.1023/A:1006524209794}

[5] J. Grzymała-Busse and W. Ziarko, *Data Mining Opportunities and Challenges*. Idea Group Publishing, 2003, ch. Data mining based on rough sets, pp. 142–173.

[6] K. Kaufman and R. Michalski, "Learning in inconsistent world, rule selection in star/aq18," Machine Learning and Inference Laboratory, Tech. Rep. Report P99-2 1999, 1999.

[7] J. Stefanowski, "Rough set based rule induction techniques for classification problems," in *Proc. Of the 6th European Congress of Intelligent Techniques and Soft Computing*, 1998, pp. 107–119.

[8] E. Czogała and J. Łęski, *Fuzzy and Neuro-Fuzzy Intelligent Systems*, ser. Studies in Fuzziness and Soft Computing. Physica-Verlag, 2000, vol. 47. [Online]. Available: {http://dx.doi.org/10.1007/978-3-7908-1853-6}

[9] O. Nelles, A. Fink, R. Babuska, and M. Setnes, "Comparison of two construction algorithms for takagi-sugeno fuzzy models," *International Journal of Applied Mathematics and Computer Science*, vol. 10, no. 4, pp. 835–855, 2000.

[10] B. Boser, I. Guyon, and V. Vapnik, "A training algorithm for optimal margin classifiers," in *Proc. of the 5th Annual ACM Workshop on Computational Learning Theory*, 1992. doi: 10.1145/130385.130401 pp. 144–152. [Online]. Available: {http://dx.doi.org/10.1145/130385.130401}

[11] K. Dembczynski, W. Kotłowski, and R. Słowiński, "Ender: a statistical framework for boosting decision rules," *Data Mining and Knowledge Discovery*, vol. 21, no. 1, pp. 52–90, 2010. doi: 10.1007/s10618-010-0177-7. [Online]. Available: {http://dx.doi.org/10.1007/s10618-010-0177-7}

[12] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules in large databases," in *Proc. of the 20th VLDB Conference*, Santiago, Chile, 2004, pp. 487–499.

[13] B. Kavsek and N. Lavrac, "Apriori-sd: Adapting association rule learning to subgroup discovery," *Applied Artificial Intelligence*, vol. 20, no. 7, pp. 543–583, 2006. doi: 10.1007/978-3-540-45231-7_22. [Online]. Available: {http://dx.doi.org/10.1007/978-3-540-45231-7_22}

[14] J. Stefanowski and D. Vanderpooten, "Induction of decision rules in classification and discovery-oriented perspectives," *International Journal of Intelligent Systems*, vol. 16, no. 1, pp. 13–27, 2001.

[15] L. Geng and H. Hamilton, "Interestingness measures for data mining: A survey," *ACM Computing Surveys*, vol. 38, no. 6, p. art. no. 9, 2006. doi: 10.1145/1132960.1132963. [Online]. Available: {http://dx.doi.org/10.1145/1132960.1132963}

[16] K. McGarry, "A survey of interestingness measures for knowledge discovery," *The Knowledge Engineering Review*, vol. 20, no. 1, pp. 39–61, 2005. doi: 10.1017/S0269888905000408. [Online]. Available: {http://dx.doi.org/10.1017/S0269888905000408}

[17] S. Sahar, *Data Mining and Knowledge Discovery Handbook*. Springer-Verlag, 2010, ch. Interestingness measures - On determining what is interesting, pp. 603–612. [Online]. Available: {http://dx.doi.org/10.1007/978-0-387-09823-4_30}

[18] D. Gamberger and N. Lavrac, "Expert-guided subgroup discovery: methodology and application," *Journal of Artificial Intelligence Research*, vol. 17, no. 1, pp. 501–527, 2002. doi: 10.1613/jair.1089. [Online]. Available: {http://dx.doi.org/10.1613/jair.1089}

[19] N. Lavrac, B. Kavsek, and P. Flach, "Subgroup discovery with cn2-sd," *Journal of Machine Learning Research*, vol. 5, pp. 153–188, 2004.

[20] A. An and N. Cercone, "Rule quality measures for rule induction systems: description and evaluation," *Computational Intelligence*, vol. 17, no. 3, pp. 409–424, 2001. doi: 10.1111/0824-7935.00154. [Online]. Available: http://dx.doi.org/10.1111/0824-7935.00154

[21] I. Bruha and J. Tkadlec, "Rule quality for multiple-rules classifier: Empirical expertise and theoretical methodology," *Intelligent Data Analysis*, vol. 7, no. 2, pp. 99–124, 2003.

[22] J. Fürnkranz and P. Flach, "Roc 'n' rule learning - towards a better understanding of covering algorithms," *Machine Learning*, vol. 39–77, 2005. doi: 10.1007/s10994-005-5011-x. [Online]. Available: {http://dx.doi.org/10.1007/s10994-005-5011-x}

[23] F. Janssen and J. Fürnkranz, "On the quest for optimal rule learning heuristics," *Machine Learning*, vol. 78, pp. 343–379, 2010. doi: 10.1007/s10994-009-5162-2. [Online]. Available: {http://dx.doi.org/10.1007/s10994-009-5162-2}

[24] M. Sikora, "Rule quality measures in creation and reduction of data rule models," *Lecture Notes in Artificial Intelligence*, vol. 4259, pp. 716–725, 2006. doi: 10.1007/11908029_74. [Online]. Available: {http://dx.doi.org/10.1007/11908029_74}

[25] ——, "Decision rule-based data models using trs and nettrs - methods and algorithms," *Transaction on*

Rough Sets - Lecture Notes on Computer Science, vol. 5946, pp. 130–160, 2010. doi: 10.1007/978-3-642-11479-3_8. [Online]. Available: {http://dx.doi.org/10.1007/978-3-642-11479-3_8}

[26] M. Sikora and Ł.Wróbel, "Data-driven adaptive selection of rule quality measures for improving the rule induction algorithm," *Lecture Notes in Computer Science*, vol. 6743, pp. 279–287, 2011. doi: 10.1007/978-3-642-21881-1_44. [Online]. Available: {http://dx.doi.org/10.1007/978-3-642-21881-1_44}

[27] I. Bruha, *Machine Learning and Statistics: The Interface*. John Wiley, 1997, ch. Quality of decision rules: Definitions and classification schemes for multiple rules, pp. 107–131.

[28] N. Lavrac, P. Flach, and B. Zupan, "Rule evaluation measures: A unifying view," *Lecture Notes in Artificial Intelligence*, vol. 1634, pp. 174–185, 1999. doi: 10.1007/3-540-48751-4_17. [Online]. Available: {http://dx.doi.org/10.1007/3-540-48751-4_17}

[29] Y. Yao and N. Zhong, "An analysis of quantitative measures associated with rules," *Lecture Notes in Computer Science*, vol. 1574, pp. 479–488, 1999. doi: 10.1007/3-540-48912-6_64. [Online]. Available: {http://dx.doi.org/10.1007/3-540-48912-6_64}

[30] S. Greco, Z. Pawlak, and R. Słowiński, "Can bayesian confirmation measures be useful for rough set decision rules?" *Engineering Applications of Artificial Intelligence*, vol. 17, no. 4, pp. 345–361, 2004. doi: 10.1016/j.engappai.2004.04.008. [Online]. Available: {http://dx.doi.org/10.1016/j.engappai.2004.04.008}

[31] P. Tan, V. Kumar, and J. Srivastava, "Selecting the right interestingness measure for association patterns," in *Proc. of the 8th International Conference on Knowledge Discovery and Data Mining*, 2002. doi: 10.1145/775047.775053 pp. 32–41. [Online]. Available: {http://dx.doi.org/10.1145/775047.775053}

[32] T. Agotnes, J. Komorowski, and T. Loken, "Taming large rule models in rough set approaches," *Lecture Notes in Artificial Intelligence*, vol. 1704, pp. 193–203, 1999. doi: 10.1007/978-3-540-48247-5_21. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-48247-5_21

[33] M. Sikora and Ł.Wróbel, "Data-driven adaptive selection of rule quality measures for improving rule induction and filtration algorithm," *International Journal of General Systems*, vol. 42, no. 6, pp. 594–613, 2013. doi: 10.1080/03081079.2013.798901. [Online]. Available: {http://dx.doi.org/10.1080/03081079.2013.798901}

[34] T. Amin, I. Chikalov, M. Moshkov, and B. Zielosko, "Relationships between length and coverage of decision rules," *Fundamenta Informaticae*, vol. 129, no. 1–2, pp. 1–13, 2014. doi: 10.3233/FI-2014-956. [Online]. Available: {http://dx.doi.org/10.3233/FI-2014-956}

[35] U. Stańczyk, "Decision rule length as a basis for evaluation of attribute relevance," *Journal of Intelligent and Fuzzy Systems*, vol. 24, no. 3, pp. 429–445, 2013. doi: 10.3233/IFS-2012-0564. [Online]. Available: {http://dx.doi.org/10.3233/IFS-2012-0564}

[36] A. Ławrynowicz and J. Potoniec, "Pattern based feature construction in semantic data mining," *International Journal on Semantic Web & Information Systems*, vol. 10, no. 1, pp. 27–65, 2014. doi: 10.4018/ijswis.2014010102. [Online]. Available: {http://dx.doi.org/10.4018/ijswis.2014010102}

[37] L. Wohlrab and J. Fürnkranz, "A review and comparison of strategies for handling missing values in separate-and-conquer rule learning," *Journal of Intelligent Information Systems*, vol. 36, no. 1, pp. 73–98, 2011. doi: 10.1007/s10844-010-0121-8. [Online]. Available: {http://dx.doi.org/10.1007/s10844-010-0121-8}

[38] T. Fawcett, "An introduction to roc analysis," *Pattern Recognition Letters*, vol. 27, pp. 861–874, 2006. doi: 10.1016/j.patrec.2005.10.010. [Online]. Available: {http://dx.doi.org/10.1016/j.patrec.2005.10.010}

[39] G. Webb and D. Brain, "Generality is predictive of prediction accuracy," *Lecture Notes in Computer Science*, vol. 3755, pp. 1–13, 2006. doi: 10.1007/11677437_1. [Online]. Available: {http://dx.doi.org/10.1007/11677437_1}

[40] H. Xiong, S. Shekhar, P. N. Tan, and V. Kumar, "Exploiting a support-based upper bound of pearson's correlation coefficient for efficiently identifying strongly correlated pairs," in *In Proceedings of the 10th ACM SIGKDD*, 2004. doi: 10.1145/1014052.1014090 pp. 334–343. [Online]. Available: {http://dx.doi.org/10.1145/1014052.1014090}

[41] D. Christensen, "Measuring confirmation," *Journal of Philosophy*, vol. 96, no. 9, pp. 437–461, 1999. doi: 10.2307/2564707. [Online]. Available: {http://dx.doi.org/10.2307/2564707}

[42] J. Joyce, *The Foundations of Causal Decision Theory*. Cambridge University Press, 1999.

[43] I. Brzezińska, R. Słowiński, and S. Greco, "Mining pareto-optimal rules with respect to support and confirmation or support and anti-support," *Engineering Applications of Artificial Intelligence*, vol. 20, no. 5, pp. 587–600, 2007. doi: 10.1016/j.engappai.2006.11.015. [Online]. Available: {http://dx.doi.org/10.1016/j.engappai.2006.11.015}

[44] M. Wojnarski, A. Janusz, H. S. Nguyen, J. Bazan, and C. J. Luo, "Rsctc2010 discovery challenge: Mining dna microarrays data for medical diagnosis and treatment," *Lecture Notes in Artificial Intelligence*, vol. 6086, pp. 4–19, 2010. doi: 10.1007/978-3-642-13529-3_3. [Online]. Available: {http://dx.doi.org/10.1007/978-3-642-13529-3_3}

[45] A. Frank and A. Asuncion, "Uci machine learning repository [http://archive.ics.uci.edu/ml]."

[46] W. Malara, M. Sikora, and L. Wróbel, "An r package for induction and evaluation of classification rules," *Studia Informatica*, vol. 34, no. 2B, pp. 339–352, 2013.