Rule evaluation measures: a unifying view

Nada Lavrač[1], Peter Flach[2], Blaz Zupan[3,1]

[1] Dept. of Intelligent Systems, J. Stefan Institute, Ljubljana, Slovenia
[2] Dept. of Computer Science, University of Bristol, United Kingdom
[3] Fac. of Computer and Information Sciences, Univ. of Ljubljana, Slovenia

Abstract

Numerous measures are used for performance evaluation in machine learning. In predictive knowledge discovery, the most frequently used measure is classification accuracy. With new tasks being addressed in knowledge discovery, new measures appear. In descriptive knowledge discovery, where induced rules are not primarily intended for classification, new measures used are novelty in clausal and subgroup discovery, and support and confidence in association rule learning. Additional measures are needed as many descriptive knowledge discovery tasks involve the induction of a large set of redundant rules and the problem is the ranking and filtering of the induced rule set. In this paper we develop a unifying view on existing measures for predictive and descriptive induction. We provide a common terminology and notation by means of contingency tables. We demonstrate how to trade off most of these measures, by using what we call *weighted relative accuracy*. The paper furthermore demonstrates that many rule evaluation measures developed for predictive knowledge discovery can be adapted to descriptive knowledge discovery tasks.

Keywords: *inductive learning, heuristics, rule evaluation*

Email address of contact author: nada.lavrac@ijs.si

Phone number of contact author: ++ 386 61 177 3272

Electronic version: gzipped PostScript emailed to icml99@ijs.si

1 Introduction

Numerous measures are used for performance evaluation in machine learning and knowledge discovery. In classification-oriented *predictive induction*, the most frequently used measure is classification accuracy. Other standard measures include precision and recall in information retrieval, and sensitivity and specificity in medical data analysis. With new tasks being addressed in knowledge discovery, new measures need to be defined, such as novelty in clausal and subgroup discovery, and support and confidence in association rule learning. These new knowledge discovery tasks belong to what is called *descriptive induction*. Descriptive induction also includes other knowledge discovery tasks, such as learning of properties, integrity constraints, and attribute dependencies.

This paper provides a systematic analysis of rule evaluation measures used in machine learning and knowledge discovery. The analysis applies to cases where single rules have to be ranked according to how well they are supported by the data. It also applies to both predictive and descriptive induction. As we argue in this paper, the right way to use standard rule evaluation measures is relative to some threshold, e.g., relative to the trivial rule 'all instances belong to this class'. We thus introduce relative versions of these standard measures, e.g., relative accuracy. We then show that relative measures provide a link with descriptive measures estimating novelty. Furthermore, by taking a weighted variant of such relative measures we show that we in fact obtain a trade-off between several of them by maximizing a single measure we call *weighted relative accuracy.*

The outline of the paper is as follows. We first introduce some well known measures used in predictive induction, applying to binary classification tasks and using a simplified terminology suggested by the confusion matrix notation. In Section 3 we introduce the terminology and notation used in this paper. In particular, we introduce the contingency table notation that will be put to use in Section 4, where we reformulate predictive and descriptive measures found in the literature in this framework. Our main results concerning unifications between different predictive measures, and between predictive and descriptive measures, are presented in Section 5. In Section 6 we support our theoretical analysis with some preliminary empirical evidence. Finally, in Section 7 we discuss the main contributions of this work.

2 Standard rule evaluation measures

Predictive induction deals with learning of rules aimed at prediction and/or classification tasks. The inputs to predictive learners are classified examples, and the outputs are prediction or classification rules. These rules can be induced by propositional or by first-order learners. In both cases they have a simple *if-then* format, with the *if*-part being a conjunction of attribute tests or first-order literals, and the *then*-part assigning a class to examples satisfying the *if*-part.

Consider a binary classification problem (given only two classes: positive and negative) and prediction rules of the form *if Conditions then TargetClass*. In this case, a *confusion matrix* is used for computing rule evaluation measures.

2.1 Confusion matrix

In the confusion matrix shown in Table 1 the following notation is used. P_a denotes the number of positive examples, N_a the number of negative examples, P_p the examples covered by the rule and thus predicted as positive, and N_p the examples not covered by the rule and therefore predicted as negative. The fields of the confusion matrix contain the numbers of examples of the following four subsets (between brackets the symbol denoting the number of examples in each subset is indicated):

True positives (TP): True positive answers of a rule denoting correct classifications of positive cases.

True negatives (TN): True negative answers denoting correct classifications of negative cases.

False positives (FP): False positive answers denoting incorrect classifications of negative cases into class positive.

False negatives (FN): False negative answers denoting incorrect classifications of positive cases into class negative.

In the fields of the confusion matrix, for the convenience of computation, the absolute numbers may be replaced by the relative frequencies, e.g., TP by $\frac{TP}{N}$, and P_a by $\frac{P_a}{N}$. This is more convenient when relative frequencies are used as probability estimates, which is the case in this paper.

	predicted positive	predicted negative	
actual positive	TP	FN	P_a
actual negative	FP	TN	N_a
	P_p	N_p	N

Table 1: A confusion matrix.

2.2 Standard measures for evaluating predictive rules

In this section we recall the standard measures that are used to evaluate rules in predictive induction. We start with *rule accuracy*, which is well-known in rule induction and inductive logic programming. It is defined as the fraction of predicted positives that are true positives:

$$Acc(R) = \frac{TP}{TP + FP} = \frac{TP}{P_p}$$

This same measure is called *precision* in information retrieval. *Accuracy error*, derived from accuracy, is defined as $Err(R) = 1 - Acc(R) = \frac{FP}{P_p}$. Rule accuracy can also be used to measure the *reliability* of the rule in the prediction of positive cases since it measures the correctness of returned results. The reliability of negative predictions, on the other hand, is

$$NegRel(R) = \frac{TN}{TN + FN} = \frac{TN}{N_p}$$

Sensitivity is a measure frequently used in medical applications. It measures the fraction of actual positives that are correctly classified. In medical terms, maximizing sensitivity means detecting as many ill patients as possible.

$$Sens(R) = \frac{TP}{TP + FN} = \frac{TP}{P_a}$$

This measure is identical to *recall* known from information retrieval (recall of positive cases).

Specificity is also a measure frequently used in medical applications. Specificity can be interpreted as recall of negative cases:

$$Spec(R) = \frac{TN}{TN + FP} = \frac{TN}{N_a}$$

Maximizing specificity is equivalent to minimizing the false alarm rate, where $FalseAlarm(R) = 1 - Spec(R) = \frac{FP}{TN+FP}$. In medicine, this measure is aimed at minimizing the fraction of healthy patients declared as ill.

3 Unifying terminology and notation

In this section we introduce a common terminology and notation used throughout the paper. The main point here is that, unlike in the previous section, we are not restricted to predictive induction, and consequently the rules we consider have a more general format. Such rules do not have a single classification literal in the conclusion part, and thus the notions of positive and negative example have to be generalised. Below we only assume that induced rules are implications with a head and a body (Section 3.1). Predicted positives/negatives are then those instances for which the body is true/false, and actual positives/negatives are instances for which the head is true/false. This gives rise to a generalisation of the confusion matrix called a contingency table, as explained in Section 3.2.

3.1 Rules

We restrict attention to learning systems that induce rules of the form

$$Head \leftarrow Body$$

In propositional *predictive rules*, *Body* is (typically) a conjunction of attribute-value pairs, and *Head* is a class assignment. In first-order learning, frequently referred to as *inductive logic programming*, predictive rules are Prolog clauses, where *Head* is a single positive literal and *Body* is a conjunction of positive and/or negative literals. The important difference with propositional predictive rules is that first-order rules contain variables that are shared between literals and between *Head* and *Body*.

Descriptive induction deals with learning of rules aimed at knowledge discovery tasks other than classification tasks. Those include learning of properties, integrity constraints, functional dependencies, as well as the discovery of interesting subgroups, association rule learning, etc. The input to descriptive learners are unclassified instances, i.e. descriptive induction is unsupervised. In comparison with propositional prediction rules, in which *Head* is a class assignment, *association rules* allow the *Head* to be a conjunction of attribute

	B	$\overline{B}$	
H	$n(HB)$	$n(H\overline{B})$	$n(H)$
$\overline{H}$	$n(\overline{H}B)$	$n(\overline{HB})$	$n(\overline{H})$
	$n(B)$	$n(\overline{B})$	N

Table 2: A contingency table.

tests. Propositional association rules have recently been upgraded to the first-order case [2]. Descriptive first-order rules also include *general clauses*, which allow for a disjunction of literals to be used in the *Head*.

In the abstract framework of this paper, rules are binary objects consisting of *Head* and *Body*. Rule evaluation measures are intended to give an indication of the strength of the (hypothetical) association between *Body* and *Head* expressed by such a rule. We assume a certain unspecified language bias that determines all possible heads and bodies of rules. We also assume a given set of instances, i.e. classified or unclassified examples, and we assume a given procedure by which we can determine, for every possible *Head* and *Body*, whether or not it is true for that instance. We say that an instance is *covered* by a rule *Head* ← *Body* if *Body* is true for the instance. In the propositional case, an instance is covered when it satisfies the conditions of a rule (all the conditions of a rule are evaluated true given the instance description). In the first-order case, the atom(s) describing the instance are matched with the rule head, thus determining a substitution θ by which the variables in the rule head are replaced by the terms (constants) in the instance description. The rule covers the instance iff $Body\theta$ is evaluated as true.

3.2 Contingency table

Given the above concepts, we can construct a *contingency table* for an arbitrary rule $H \leftarrow B$, which is a generalisation of the confusion matrix introduced in Section 2.1. In Table 2, B denotes the set of instances for which the body of the rule is true, and $\overline{B}$ denotes its complement (the set of instances for which the body is false); similarly for H and $\overline{H}$. HB then denotes $H \cap B$, $\overline{H}B$ denotes $\overline{H} \cap B$, and so on.

We use $n(X)$ to denote the cardinality of set X, e.g. $n(\overline{H}B)$ is the number of instances for which H is false and B is true (i.e., the number of instances erroneously covered by the rule). N denotes the total number of instances in the sample. The relative frequency $\frac{n(X)}{N}$ associated with X is denoted by $p(X)$.[1] All rule evaluation measures considered in this paper are defined in terms of frequencies from the contingency table only.

4 Rule evaluation measures for knowledge discovery

In this section, rule evaluation measures that can be found in the literature are formulated in the contingency table terminology, which is the first step towards the unifying view developed in Section 5. The definitions are given in terms of the relative frequencies derived from the contingency table. Furthermore, since our framework is not restricted to predictive induction, we also elaborate the novelty-based measures found in the knowledge discovery literature. The usefulness of this unifying framework is then demonstrated in Section 5, where we point out the many relations that exist between weighted and relative variants of these measurs.

Definition 1 (Rule accuracy) $Acc(H \leftarrow B) = p(H|B)$

Definition 2 (Negative reliability) $NegRel(H \leftarrow B) = p(\overline{H}|\overline{B})$

Definition 3 (Sensitivity) $Sens(H \leftarrow B) = p(B|H)$

Definition 4 (Specificity) $Spec(H \leftarrow B) = p(\overline{B}|\overline{H})$

Accuracy of rule $R = H \leftarrow B$, here defined as the conditional probability that H is true given that B is true, indeed measures the fraction of predicted positives that are true positives in the case of binary classification problems: $Acc(R) = \frac{TP}{TP+FP} = \frac{n(HB)}{n(HB)+n(\overline{H}B)} = \frac{n(HB)}{n(B)} = \frac{\frac{n(HB)}{N}}{\frac{n(B)}{N}} = \frac{p(HB)}{p(B)} = p(H|B)$. As already pointed out, rule accuracy is also called precision in information retrieval. Furthermore, accuracy error $Err(H \leftarrow B) = 1 - Acc(H \leftarrow B) = p(\overline{H}|B)$. Given our more general knowledge discovery framework, it can now also be seen that rule accuracy is in fact the same as *confidence* in association rule

[1] In this paper we are not really concerned with probability estimation, and we interpret the sample relative frequency as a probability.

learning. We make further connections between predictive and descriptive induction below.

The reader can easily verify that also Definitions 2-4 correspond to the definitions given in Section 2.2. As remarked before, sensitivity is identical to recall (of positive cases) used in information retrieval, and specificity corresponds to recall of negative cases. Sensitivity can also be interpreted as the accuracy of the rule $B \leftarrow H$, which in logic programming terms is the *completion* of the rule $H \leftarrow B$. Also, notice that negative reliability measures the correctness of negative predictions.

After having re-defined standard rule evaluation measures in our more general framework, we now introduce other measures that are used to develop our unifying view in the next section.

Definition 5 (Coverage) $Cov(H \leftarrow B) = p(B)$

Definition 6 (Support) $Sup(H \leftarrow B) = p(HB)$

Coverage measures the fraction of instances covered by the body of a rule. As such it is a measure of *generality* of a rule. *Support* of a rule is a related measure known from association rule learning, also called *frequency*. Notice that, unlike the previous measures, support is symmetric in H and B.

The next measure aims at assessing the novelty, interestingness or unusualness of a rule. Novelty measures are used, e.g., in the MIDOS system for subgroup discovery [6], and in the PRIMUS family of systems for clausal discovery [3]. Here we follow the elaboration of the PRIMUS novelty measure, because it is formulated in the more general setting of clausal discovery, and because it is clearly linked with the contingency table framework.

Consider again the contingency table in Table 2. We define a rule $H \leftarrow B$ to be *novel* if $n(HB)$ cannot be inferred from the marginal frequencies $n(H)$ and $n(B)$; in other words, if H and B are not statistically *independent*. We thus compare the *observed* $n(HB)$ with the *expected* value under independence $\mu(HB) = \frac{n(H)n(B)}{N}$. The more the observed value $n(HB)$ differs from the expected value $\mu(HB)$, the more likely it is that there exists a real and unexpected association between H and B, expressed by the rule $H \leftarrow B$. Novelty is thus defined as the relative difference between $n(HB)$ and $\mu(HB)$.

Definition 7 (Novelty) $Nov(H \leftarrow B) = p(HB) - p(H)p(B)$

Notice that $p(HB)$ is what is called *support* in association rule learning. The definition of novelty states that we are only interested in high support if that couldn't be expected from the marginal probabilities, i.e. when $p(H)$ and/or $p(B)$ are relatively low. It can be demonstrated that $-0.25 \leq Nov(R) \leq 0.25$: a strongly positive value indicates a strong association between H and B, while a strongly negative value indicates a strong association between $\overline{H}$ and B.[2]

In the MIDOS subgroup discovery system this measure is used to detect unusal subgroups. For selected head H, indicating a property we are interested in, body B defines an unusal subgroup of the instances satisfying H if the distribution of H-instances among B-instances is sufficiently different from the distribution of H-instances in the sample. In situations like this, where H is selected, this definition of novelty is sufficient. However, notice that $Nov(H \leftarrow B)$ is symmetric in H and B, which means that $H \leftarrow B$ and $B \leftarrow H$ will always carry the same novelty, even though one of them may have many more counter-instances (satisfying the body but falsifying the head) than the other.

To distinguish between such cases, PRIMUS additionally employs the measure of satisfaction, which is the relative decrease in accuracy error between the rule $H \leftarrow true$ and the rule $H \leftarrow B$. It is a variant of rule accuracy which takes the whole of the contingency table into account — it is thus more suited towards knowledge discovery, being able trading off rules with different heads as well as bodies. We omit the details for lack of space.

5 A unifying view

In the previous section we formulated selected rule evaluation measures in our more general knowledge discovery framework. In this section we show the usefulness of this framework by establishing a synthesis between these measures. The main inspiration for this synthesis comes from the novelty measure, which is *relative* in the sense that it compares the support of the rule with the expected support under the assumption of statistical independence (Definition 7).

[2] Since negative novelty can be transformed into positive novelty associated with the rule $\overline{H} \leftarrow B$, systems like MIDOS and PRIMUS set $Nov(H \leftarrow B) = 0$ if $p(HB) < p(H)p(B)$. The more general expression of Definition 7 is kept because it allows a more straightforward statement of our main results.

Definition 8 (Relative accuracy) $RAcc(H \leftarrow B) = p(H|B) - p(H)$

Relative accuracy of a rule $R = H \leftarrow B$ is the accuracy gain relative to the fixed rule $H \leftarrow true$. The latter rule predicts all instances to satisfy H; a rule is only interesting if it improves upon this 'default' accuracy. Another way of viewing relative accuracy is that it measures the utility of connecting body B with a given head H.

Similarly, we define relative versions of other rule evaluation measures.

Definition 9 (Relative negative reliability)

$$RNegRel(H \leftarrow B) = p(\overline{H}|\overline{B}) - p(\overline{H})$$

Definition 10 (Relative sensitivity) $RSens(H \leftarrow B) = p(B|H) - p(B)$

Definition 11 (Relative specificity) $RSpec(H \leftarrow B) = p(\overline{B}|\overline{H}) - p(\overline{B})$

Like relative accuracy, relative negative reliability measures the utility of connecting body B wih a given head H. The latter two measures can be interpreted as sensitivity/specificity gain relative to the rule $true \leftarrow B$, i.e. the utility of connecting a given body B with head H. Notice that this view is taken in rule construction by the CN2 algorithm [1], which first builds a rule body and subsequently assigns an appropriate rule head.

To repeat, the point about relative measures is that they give more information about the utility of a rule than absolute measures. For instance, if in a prediction task the accuracy of a rule is lower than the relative frequency of the class it predicts, then the rule actually performs badly, regardless of its absolute accuracy.

There is however a problem with relative accuracy as such: it is easy to obtain high relative accuracy with highly specific rules, i.e. rules with low generality $p(B)$. To this end, a weighted variant is introduced, which is the key notion in this paper.

Definition 12 (Weighted relative accuracy)

$$WRAcc(H \leftarrow B) = p(B)(p(H|B) - p(H))$$

Weighted relative accuracy trades off generality and relative accuracy. It is known in the literature as a *gain* measure, used to evaluate the utility of a *literal* L considered for extending the body B of a rule: $\frac{p(BL)}{p(B)}(p(H|BL) - p(H|B))$. Its introduction as a *rule* evaluation measure is new.

We now come to a result, which — although technically trivial — provides a significant contribution to our understanding of rule evaluation measures.

Theorem 1 $WRAcc(R) = Nov(R)$

Proof. $WRAcc(H \leftarrow B) = p(B)(p(H|B) - p(H)) = p(B)p(H|B) - p(H)p(B) = p(HB) - p(H)p(B) = Nov(H \leftarrow B)$. ∎

Theorem 1 has the following implications.

1. Rules with high weighted relative accuracy also have high novelty, and *vice versa.*

2. High novelty is achieved by trading off generality and rule accuracy gained in comparison with a trivial rule $H \leftarrow true$. This also means that having high relative accuracy is not enough for considering a rule to be interesting, since the rule needs to be general enough as well.

This link between predictive and descriptive rule evaluation measures has — to the best of our knowledge — not been published before.

We proceed to show that weighted relative accuracy is one of the most fundamental rule evaluation measures, by showing that it also provides a trade-off between accuracy and other predictive measures such as sensitivity. To do so, we first define weighted versions of the other relative measures defined above.

Definition 13 (Weighted relative negative reliability)

$$WRNegRel(H \leftarrow B) = p(\overline{B})(p(\overline{H}|\overline{B}) - p(\overline{H}))$$

The weight $p(\overline{B})$ is motivated by the fact that overly general rules trivially have a high negative reliability.

Definition 14 (Weighted relative sensitivity)

$$WRSens(H \leftarrow B) = p(H)(p(B|H) - p(B))$$

Definition 15 (Weighted relative specificity)

$$WRSpec(H \leftarrow B) = p(\overline{H})(p(\overline{B}|\overline{H}) - p(\overline{B}))$$

Again, the weights guard against trivial solutions.

This leads us to establising a trade-off between the four standard predictive rule evaluation measures, by relating them through their weighted relative variants.

Theorem 2 *$WRAcc(R) = WRSens(R) = WRSpec(R) = WRNegRel(R)$.*

Proof. $WRAcc(H \leftarrow B) = p(B)(p(H|B) - p(H)) = p(HB) - p(H)p(B) = p(H)(p(B|H) - p(B)) = WRSens(H \leftarrow B)$.
$WRAcc(H \leftarrow B) = p(B)(p(H|B) - p(H)) = p(HB) - p(H)p(B) = (1 - p(\overline{H}B) - p(H\overline{B}) - p(\overline{HB})) - (1 - p(\overline{B}))(1 - p(\overline{H})) = (1 - p(\overline{H}) - p(\overline{B}) + p(\overline{HB})) - (1 - p(\overline{H}) - p(\overline{B}) + p(\overline{H})p(\overline{B})) = p(\overline{HB}) - p(\overline{H})p(\overline{B}) = p(\overline{H})(p(\overline{B}|\overline{H}) - p(\overline{B})) = WRSpec(H \leftarrow B)$.
$WRSpec(H \leftarrow B) = p(\overline{H})(p(\overline{B}|\overline{H}) - p(\overline{B})) = p(\overline{HB}) - p(\overline{H})p(\overline{B}) = p(\overline{B})(p(\overline{H}|\overline{B}) - p(\overline{H})) = WRNegRel(H \leftarrow B)$. ▮

We have thus established a complete synthesis between different predictive rule evaluation measures, and between these measures and the descriptive notion of novelty, by demonstrating that there is a single way in which all these measures can be combined and thus traded off in a principled way.

6 Rule evaluation measures in practice

In the previous section we have shown that a single measure, weighted relative accuracy, can be used to trade off different evaluation measures such as accuracy, sensitivity, and novelty. In this section we further support this claim with some preliminary empirical evidence. First, we describe an experiment in which weighted relative accuracy correlates better with an expert's intuitive understanding of "reliability" and "interestingness" than standard rule evaluation measures. Secondly, we show the utility of weighted relative accuracy as a filtering measure in database dependency discovery.

6.1 An experiment

The purpose of this experiment was to find out whether rule evaluation measures as discussed in this paper really measure what they are supposed to measure. To this end we compared an expert's ranking of a number of rules on two dimensions with the rankings given by four selected measures. We have used a CAR data set (see UCI Machine Learning Repository [5]), which includes 1728 instances that are described with six attributes and a corresponding four-valued class. The attributes are multi-valued and include buying price, price of maintenance, number of doors, capacity in terms of persons to carry, and estimated safety of the car.

An ML* Machine Learning environment was used to generate association rules from the CAR dataset. The designer of the experiment has semi-randomly chosen ten rules that he though may be of different quality in respect to the measures introduced in this text. Note that none of the rules, however, was explicitly measured at this stage.

The rules where then shown to the domain expert, who was asked to rank them according to their "reliability" and "interestingness". We chose these non-technical terms to avoid possible interference with any technical interpretation; neither term was in any way explained to the expert.[3] The domain expert first assigned qualitative grades to each rule (-,$\circ$,$\oplus$,+), and then chose a final rank from these grades. The results of the ranking are shown in Table 3. Note that some rules are ranked equally (e.g. the first two rules for reliability), and in such cases a rank is represented as an interval. The correlation between the expert's rankings and ranks obtained from the rule evaluation measures are given in Table 4.

Although the correlations in Table 4 are quite low, the tentative conclusion is that *WRAcc* correlates best with *both* intuitive notions of reliability and interestingness. This provides some preliminary empirical support for the idea that *WRAcc* provides the right trade-off between predictive and descriptive rule evaluation measures.

[3] During the experiment, the expert expressed some of his intuitions regarding these terms: "reliability measures how reliable the rule is when applied for a classification";
"an interesting rule is the one that I never thought of when building a classification model, e.g., those without the class (car) in the head";
"an interesting rule has to tell me something new, but needs to be reliable as well (it would help me if I would somehow know the reliability first before ranking on interestingness)";
"a highly reliable rule which is at the same time unusual is interesting";
"a rule is interesting if it tells me something new, but it's not an outlier".

Rule	Expert				Rule evaluation measures			
	Rel	#	Int	#	*Acc*	*Sens*	*Spec*	*WRAcc*
buying=med car=good → maint=low	-	7-10	○	6	1.000	0.053	1.000	0.010
buying=low car=v-good → lugboot=big	-	7-10	-	7-10	0.615	0.042	0.987	0.006
safety=low → car=unacc	+	1	-	7-10	1.000	0.476	1.000	0.100
persons=2 car=unacc → lugboot=big	-	7-10	-	7-10	0.333	0.333	0.667	0.000
lugboot=big car=good → safety=med	○	5-6	○	5	1.000	0.042	1.000	0.009
car=v-good → lugboot=big	⊕	3	+	2	0.615	0.069	0.978	0.011
car=unacc → buying=v-high	⊕	4	+	3	0.298	0.833	0.344	0.033
car=v-good → safety=high	+	2	+	1	1.000	0.113	1.000	0.025
persons=4 → lugboot=big car=unacc	-	7-10	-	7-10	0.153	0.239	0.641	-0.020
persons=4 safety=high → car=acc	○	5-6	○	4	0.563	0.281	0.938	0.038

Table 3: Ten rules ranked by a domain expert on reliability (Rel) and interestingness (Int), and corresponding rule evaluation measures.

	Acc	*Sens*	*Spec*	*WRAcc*
expert's Rel	0.150	0.152	0.116	0.323
expert's Int	0.067	-0.006	0.029	0.177

Table 4: Rank correlations between two measures elicited from the expert and four rule evaluation measures.

6.2 Rule filtering

The measures discussed in this paper are primarily intended for ranking and filtering rules output by an induction algorithm. This is particularly important in descriptive induction tasks such as association rule learning and database dependency discovery, since descriptive induction algorithms typically output several thousands of rules. We briefly describe some preliminary experience with rule filtering using the functional dependency discovery tool `fdep` [4].

We ran `fdep` on some of the UCI datasets [5], and then used $WRAcc$ to rank the induced functional dependencies. Below we give some of the highest ranked rules in several domains. They have the form $A_1, \ldots, A_n \rightarrow A$, meaning "given the values of attributes $A_1, \ldots, A_n$, the value of attribute A is fixed"; see [4] for details of the transformation into $H \leftarrow B$ form.

Lymphography:

```
[block_lymph_c,regeneration,lym_nodes_enlar,no_nodes]->[block_lymph_s]
[lymphatics,by_pass,regeneration,lym_nodes_enlar]->[lym_nodes_dimin]
```

Primary tumor:

```
[class,histologic_type,degree_of_diffe,brain,skin,neck]->[axillar]
[class,histologic_type,degree_of_diffe,bone_marrow,skin,neck]->[axillar]
[class,histologic_type,degree_of_diffe,bone,bone_marrow,skin]->[axillar]
```

Hepatitis:

```
[liver_firm,spleen_palpable,spiders,ascites,bilirubin]->[class]
[liver_big,liver_firm,spiders,ascites,varices,bilirubin]->[class]
[anorexia,liver_firm,spiders,ascites,varices,bilirubin]->[class]
```

Wisconsin breast cancer:

```
[uni_cell_size,se_cell_size,bare_nuclei,normal_nucleoli,mitoses]->[class]
[uni_cell_shape,marginal_adhesion,bare_nuclei,normal_nucleoli]->[class]
[uni_cell_size,marginal_adhesion,se_cell_size,bare_nuclei,normal_nucleoli]->[class]
```

Our experience with rule filtering in these domains suggested that $WRAcc(R)$ would drop quite sharply after the first few rules. Notice that in the last two domains the induced functional dependencies determine the class attribute.

7 Summary and discussion

In this paper we have provided a systematic analysis of rule evaluation measures used in machine learning and knowledge discovery. We have argued that, generally speaking, these measures should be used relative to some threshold, e.g., relative to the situation where this particular rule head is **not** connected to this particular rule body. Furthermore, we have proposed a single measure that can be interpreted in at least 5 different ways: as weighted relative accuracy, as weighted relative sensitivity, as weighted relative precision, as weighted relative negative reliability, and as novelty. We believe this to be a significant contribution to the understanding of rule evaluation measures, which could be obtained because of our unifying contingency table framework.

Further work includes the generalization to rule *set* evaluation measures. These differ from rule evaluation measures in that they treat positive and negative examples symmetrically, e.g. rule set accuracy would be defined as $RuleSetAcc(H \leftarrow B) = p(HB) + p(\overline{HB})$. Another extension of this work would be to investigate how some of these measures can be used as *search heuristics* rather than filtering measures. Finally, we would like to continue empirical evaluation of $WRAcc(R)$ as a filtering measure in various domains such as association rule learning and first-order knowledge discovery.

Acknowledgements

We are grateful to Dr. Marko Bohanec for his help in the expert evaluation of rules in the experimental rule evaluation, and to Dr. Iztok Savnik for his assistance in functional dependency filtering. This work was partially supported by the Esprit Long Term Research Project 20237 (*Inductive Logic Programming 2*), a Joint Project with Central/Eastern Europe funded by the Royal Society, and the Slovenian Ministry of Science and Technology.

References

[1] P. Clark and T. Niblett (1989). The CN2 induction algorithm. *Machine Learning* 3, pp. 261–284.

[2] L. Dehaspe and L. De Raedt (1997). Mining association rules with multiple relations. In N. Lavrac and S. Džeroski (Eds.), *Proc. 7th Int. Workshop on Inductive Logic Programming (ILP'97)*, pp. 125–132, LNAI 1297, Springer-Verlag.

[3] P.A. Flach and N. Lachiche (1997). *Cooking up integrity constraints with PRIMUS (preliminary report).* Technical Report CSTR-97-009, Department of Computer Science, University of Bristol, December 1997.

[4] P.A. Flach and I. Savnik (1999). Database dependency discovery: a machine learning approach. *AI Communications*, to appear.

[5] P.M. Murphy and D.W. Aha (1994). UCI Repository of machine learning databases [http://www.ics.uci.edu/~mlearn/mlrepository.html]. Irvine, CA: University of California, Department of Information and Computer Science.

[6] S. Wrobel (1997). An algorithm for multi-relational discovery of subgroups. In J. Komorowski and J. Zytkow (Eds.), *Proc. First European Symposium on Principles of Data Mining and Knowledge Discovery PKDD'97*, Springer-Verlag.