Rule Evaluation Measures: A Unifying View

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

[1] Department of Intelligent Systems
Jožef Stefan Institute, Ljubljana, Slovenia
[2] Department of Computer Science
University of Bristol, United Kingdom
[3] Faculty of Computer and Information Sciences
University 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 some of the 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 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.

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 an analysis of selected rule evaluation measures. 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

S. Džeroski and P. Flach (Eds.): ILP-99, LNAI 1634, pp. 174–185, 1999.

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 called *weighted relative accuracy*.

The outline of the paper is as follows. In Section 2 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 3, where we formulate 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 4. In Section 5 we support our theoretical analysis with some preliminary empirical evidence. Finally, in Section 6 we discuss the main contributions of this work.

2 Terminology and Notation

In this section we introduce a terminology and notation used throughout the paper. Since we are not restricted to predictive induction, the rules we consider have a more general format than the format of prediction rules that have a single classification literal in the conclusion of a rule. Below we only assume that induced rules are implications with a head and a body (Section 2.1). Due to this general rule form, the notions of positive and negative example have to be generalized: predicted positives/negatives are those instances for which the body is true/false, and actual positives/negatives are instances for which the head is true/false. In this framework, a contingency table, as explained in Section 2.2, is used as the basis for computing rule evaluation measures.

2.1 Rules

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

$$Head \leftarrow Body$$

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 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 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* $\leftarrow$ *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*θ is evaluated as true.

2.2 Contingency Table

Given the above concepts, we can construct a *contingency table* for an arbitrary rule $H \leftarrow B$. In Table 1, 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.

Table 1. A contingency table.

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

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.

Notice that a contingency table is a generalisation of a confusion matrix, which is the standard basis for computing rule evaluation measures in binary classification problems. In the confusion matrix notation, $n(H) = P_a$ – the number of positive examples, $n(\overline{H}) = N_a$ – the number of negative examples, $n(B) = P_p$ – the number of examples covered by the rule therefore predicted as positive, $n(\overline{B}) = N_p$ – the number of the examples not covered by the rule and therefore predicted as negative, $n(HB) = TP$ – the number of true positives, $n(\overline{H}\overline{B}) = TN$ – the number of true negatives, $n(\overline{H}B) = FP$ – the number of false positives, and $n(H\overline{B}) = FN$ – the number of false negatives.

3 Selected Rule Evaluation Measures

In this section, selected rule evaluation measures are formulated in the contingency table terminology, which is the first step towards the unifying view developed in Section 4. The definitions are given in terms of relative frequencies derived from the contingency table. Since our framework is not restricted to predictive induction, we also elaborate some novelty-based measures found in the knowledge discovery literature; see [9,5] which discuss also other measures and the axioms that rule evaluation measures should satisfy. The usefulness of our unifying framework is demonstrated in Section 4, where we point out the many relations that exist between weighted and relative variants of these measures.

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, measures the fraction of predicted positives that are true positives in the case of binary classification problems:

$$Acc(R) = \frac{TP}{TP+FP} = \frac{n(HB)}{n(HB)+n(\overline{H}B)} = \frac{n(HB)}{n(B)} = \frac{\frac{n(HB)}{N}}{\frac{n(B)}{N}} = \frac{p(HB)}{p(B)} = p(H|B).$$

Rule accuracy is also called precision in information retrieval. Furthermore, accuracy error $Err(H \leftarrow B) = 1 - Acc(H \leftarrow B) = p(\overline{H}|B)$.

Our definition of rule accuracy is intended for evaluating single rules, and therefore biased towards the accuracy of positive examples. As such, it is different from what we call rule set accuracy [6], defined as $Acc = \frac{TP+TN}{N} =$

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

$p(HB) + p(\overline{HB})$, which is standardly used for evaluation of hypotheses comprised of several rules.

Given our general knowledge discovery framework, it can now also be seen that rule accuracy is in fact the same as *confidence* in association rule learning. 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 is in binary classification problems computed as follows: $NegRel(R) = \frac{TN}{TN+FN} = \frac{TN}{N_p} = \frac{n(\overline{HB})}{n(\overline{B})} = \frac{p(\overline{HB})}{p(\overline{B})} = p(\overline{H}|\overline{B})$.

Sensitivity is identical to *recall* (of positive cases) used in information retrieval. Sensitivity, here defined as the conditional probability that B is true given that H is true, measures the fraction of true positives that are correctly classified in the case of binary classification problems: $Sens(R) = \frac{TP}{TP+FN} = \frac{TP}{P_a} = \frac{n(HB)}{n(HB)+n(H\overline{B})} = \frac{n(HB)}{n(H)} = \frac{p(HB)}{p(H)} = p(B|H)$. 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 $H \leftarrow B$.

Specificity is the conditional probability that B is false given that H is false. In binary classification problems, it is equal to the recall of negative cases in information retrieval: $Spec(R) = \frac{TN}{TN+FP} = \frac{TN}{N_a} = \frac{n(\overline{HB})}{n(\overline{H})} = p(\overline{B}|\overline{H})$.

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 [8], 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 1. 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 unusual subgroups. For selected head H, indicating a property we are interested in, body B defines an unusual 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.

Definition 8 (Satisfaction). $Sat(H \leftarrow B) = \frac{p(\overline{H}) - p(\overline{H}|B)}{p(\overline{H})}$.[3]

It can be shown that $Sat(H \leftarrow B) = \frac{p(H|B) - p(H)}{1 - p(H)}$, since $p(\overline{H}) - p(\overline{H}|B) = (1 - p(H)) - (1 - p(H|B)) = p(H|B) - p(H)$. We thus see that $Sat(H \leftarrow B)$ is similar to rule accuracy $p(H|B)$, e.g., $Sat(R) = 1$ iff $Acc(R) = 1$. However, unlike rule accuracy, satisfaction takes the whole of the contingency table into account and is thus more suited towards knowledge discovery, trading off rules with different heads as well as bodies.

Finally, we mention that PRIMUS trades off novelty and satisfaction by multiplying them, resulting in a χ^2-like statistic:

$$Nov(H \leftarrow B) \times Sat(H \leftarrow B) = \frac{(Np(\overline{H}B) - \mu(\overline{H}B))^2}{\mu(\overline{H}B)}$$

This is one term in the χ^2 sum for the contingency table, corresponding to the lower left-hand cell (the counter-instances). We omit the details of the normalization.

[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.

[3] Again, in practice we put $Sat(R) = 0$ if $p(\overline{H}) > p(\overline{H}|B)$.

4 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).

Definition 9 (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 10 (Relative negative reliability).

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

Definition 11 (Relative sensitivity). $RSens(H \leftarrow B) = p(B|H) - p(B)$.

Definition 12 (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 with 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 13 (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: $\frac{p(BL)}{p(B)}(p(H|BL) - p(H|B))$.

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 14 (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 15 (Weighted relative sensitivity).

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

Definition 16 (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 establishing 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. $WRAcc(H \leftarrow B) = p(B)(p(H|B) - p(H)) = p(HB) - p(H)p(B) = p(H)(p(B|H) - p(B)) = WRSens(H \leftarrow B)$.
$WRAcc(H \leftarrow B) = p(B)(p(H|B) - p(H)) = p(HB) - p(H)p(B) = (1 - p(\overline{H}B) - p(H\overline{B}) - p(\overline{H}\overline{B})) - (1 - p(\overline{B}))(1 - p(\overline{H})) = (1 - p(\overline{H}) - p(\overline{B}) + p(\overline{H}\overline{B})) - (1 - p(\overline{H}) - p(\overline{B}) + p(\overline{H})p(\overline{B})) = p(\overline{H}\overline{B}) - p(\overline{H})p(\overline{B}) = p(\overline{H})(p(\overline{B}|\overline{H}) - p(\overline{B})) = WRSpec(H \leftarrow B)$.
$WRSpec(H \leftarrow B) = p(\overline{H})(p(\overline{B}|\overline{H}) - p(\overline{B})) = p(\overline{H}\overline{B}) - p(\overline{H})p(\overline{B}) = p(\overline{B})(p(\overline{H}|\overline{B}) - p(\overline{H})) = WRNegRel(H \leftarrow B)$. □

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.

5 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.

5.1 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 [7]), 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 were 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.[4] The domain expert first

[4] 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".

assigned qualitative grades to each rule (-,○,⊕,+), and then chose a final rank from these grades. The results of the ranking are shown in Table 2. 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 3.

Table 2. Ten rules ranked by a domain expert on reliability (Rel) and interestingness (Int), and corresponding rule evaluation measures.

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

Although the correlations in Table 3 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.

Table 3. Rank correlations between two measures elicited from the expert and four 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

5.2 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 [7], 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.

6 Summary and Discussion

In this paper we have provided an analysis of selected 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., $RuleSetAcc(H \leftarrow B) = p(HB) + p(\overline{HB})$.

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. We also acknowledge the useful comments of three anonymous reviewers. 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 (1999). A first-order approach to unsupervised learning. Submitted.
4. P.A. Flach and I. Savnik (1999). Database dependency discovery: a machine learning approach. *AI Communications*, to appear.
5. W. Klösgen (1996). Explora: A multipattern and multistrategy discovery assistant. In U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Uthurusamy (Eds.), *Advances in Knowledge Discovery and Data Mining*, 249–271, AAAI Press.
6. T. Mitchell (1997). *Machine Learning*, McGraw Hill.
7. 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.
8. 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.
9. G. Piatetsky-Shapiro (1991). Discovery, analysis and presentation of strong rules. In G. Piatetsky-Shapiro and W. Frawley (Eds.), *Knowledge Discovery in Databases*, 229–249. AAAI Press.