Evaluating the Correlation Between Objective Rule Interestingness Measures and Real Human Interest

Deborah R Carvalho[1,3], Alex A. Freitas[2], Nelson Ebecken[3]
[1] Universidade Tuiuti do Paraná (UTP), Brazil deborah@utp.br
[2] Computing Laboratory University of Kent, CT2 7NF, UK
A.A.Freitas@kent.ac.uk
[3] COPPE/ Universidade Federal do Rio de Janeiro, Brazil
nelson@ntt.ufrj.br

Abstract. In the last few years, the data mining community has proposed a number of objective rule interestingness measures to select the most interesting rules, out of a large set of discovered rules. However, it should be recalled that objective measures are just an *estimate* of the true degree of interestingness of a rule to the user, the so-called real human interest. The latter is inherently subjective. Hence, it is not clear how effective, in practice, objective measures are. More precisely, the central question investigated in this paper is: "how effective objective rule interestingness measures are, in the sense of being a good estimate of the true, subjective degree of interestingness of a rule to the user?" This question is investigated by extensive experiments with 11 objective rule interestingness measures across eight real-world data sets.

1 Introduction

Data mining essentially consists of extracting *interesting* knowledge from real-world data sets. However, there is no consensus on how the interestingness of discovered knowledge should be measured. Indeed, most of the data mining literature still avoids this thorny problem and implicitly interprets "interesting" as meaning just "accurate" and sometimes also "comprehensible". Although accuracy and comprehensibility are certainly important, they are not enough to measure the real, *subjective* interestingness of discovered knowledge *to the user*. Consider, e.g., the classic example of the following rule: IF (patient is pregnant) THEN (patient is female). This rule is very accurate and comprehensible, but it is *not* interesting, since it represents an obvious pattern. As a real-world example, [8] reports that less than 1% of the discovered rules were found to be interesting to medical experts. It is also possible that a rule be interesting to the user even though it is not very accurate. For instance, in [9] rules with an accuracy around 40%-60% represented novel knowledge that gave new insights to medical doctors. Hence, there is a clear motivation to investigate the relationship between rule interestingness measures and the subjective interestingness of rules to the user – *an under-explored topic in the literature*.

Rule interestingness measures can be classified into two broad groups: user-driven (subjective) and data-driven (objective) measures. User-driven measures are based on comparing discovered rules with the previous knowledge or believes of the user. A rule is considered interesting, or novel, to the extent that it is different from the user's previous knowledge or believes. User-driven measures have the advantage of being a direct measure of the user's interest in a rule, but they have a twofold disadvantage. First, they require, as input, a specification of the user's believes or previous

knowledge – a very time-consuming task to the user. Second, they are strongly domain-dependent and user-dependent. To avoid these drawbacks, the literature has proposed more than 40 data-driven rule interestingness measures [5], [7], [3]. These measures estimate the degree of interestingness of a rule to the user in a user-independent, domain-independent fashion, and so are much more generic. Data-driven measures have, however, the disadvantage of being an indirect *estimate* of the true degree of interestingness of a rule to the user, which is an inherently *subjective* interestingness.

This begs a question rarely addressed in the literature: *how effective data-driven rule interestingness measures are, in the sense of being a good estimate of the true, subjective degree of interestingness of a rule to the user*? The vast majority of works on data-driven rule interestingness measures ignore this question because they do not even show the rules to the user. A notable exception is the interesting work of [5], which investigates the effectiveness of approximately 40 data-driven rule interestingness measures, by comparing their values with the subjective values of the user's interest – what they called *real human interest*. Measuring real human interest involves showing the rules to the user and ask her/him to assign a subjective interestingness score to each rule. Therefore, real human interest should not be confused with the above-mentioned user-driven rule interestingness measures.

This paper follows the same general line of research. We investigate the effectiveness of 11 data-driven rule interestingness measures, by comparing them with the user's subjective real human interest. Although we investigate a smaller number of rule interestingness measures, this paper extends the work of [5] by presenting results for eight data sets, whereas [5] did experiments with just one medical data set, a limitation from the point of view of generality of the results.

2 Objective (Data-Driven) Rule Interestingness Measures

This work involves 11 objective rule interestingness measures – all of them used to evaluate classification rules. Due to space limitations we mention here a brief definition of each of those measures – which are discussed in more detail in the literature. The measures defined by formulas (1)–(8) [5], [7] are based on the coverage and accuracy of a rule. Their formulas are expressed using a notation where A denotes the rule antecedent; C denotes the rule consequent (class); $P(A)$ denotes the probability of A – i.e., the number of examples satisfying A divided by the total number of examples; $P(C)$ denotes the probability of C; "$\neg A$" and "$\neg C$" denote the logical negation of A and C. The measures defined by formulas (9)-(11) [2] use the same notation of A and C to denote a rule's antecedent and consequent, but they also involve heuristic principles based on variables other than a rule's coverage and accuracy.

The Attribute Surprisingness measure – formula (9) – is based on the idea that the degree of surprisingness of an attribute is estimated as the inverse of its information gain. The rationale for this measure is that the occurrence of an attribute with a high information gain in a rule will not tend to be surprising to the user, since users often know the most relevant attributes for classification. However, the occurrence of an attribute with a low information gain in a rule tends to be more surprising, because this kind of attribute is usually considered little relevant for classification. In formula

(9), A_i denotes the attribute in the i-th condition of the rule antecedent A, m is the number of conditions in A, and #classes is the number of classes.

$$\Phi\text{-Coefficiente} = (P(A,C)-P(A)P(C))/\sqrt{P(A)P(C)(1-P(A))(1-P(C))} \quad (1)$$

$$\text{Odds Ratio} = P(A,C)P(\neg A,\neg C)/P(A,\neg C)P(\neg A,C) \quad (2)$$

$$\text{Kappa}=(P(A,C)+P(\neg A,\neg C)-P(A)P(C)-P(\neg A)P(\neg C)) \quad (3)$$
$$/ (1-P(A)P(C)-P(\neg A)P(\neg C))$$

$$\text{Interest} = P(A,C)/(P(A)*P(C)) \quad (4)$$

$$\text{Cosine} = P(A,C) / \sqrt{(P(A)*P(C))} \quad (5)$$

$$\text{Piatetsky-Shapiro's} = P(A,C)-P(A)P(C) \quad (6)$$

$$\text{Collective Strength} = ((P(A,C)+P(\neg A,\neg C))/(P(A)P(C)+P(\neg A)P(\neg C))) * \quad (7)$$
$$((1-P(A)P(C) - P(\neg A)P(\neg C))/(1-P(A,C)-P(\neg A,\neg C))$$

$$\text{Jaccard} = P(A,C) / (P(A)+ P(C) - P(A,C) \quad (8)$$

$$\text{Attribute Surprisingness} = 1 - ((\sum_{i=1}^{m} \text{InfoGain}(A_i) / m) / \log_2(\#\text{classes})) \quad (9)$$

$$\text{MinGen} = N / m \quad (10)$$

$$\text{InfoChange-ADT} = I^{AB1} - I^{ABo} \quad (11.1)$$

$$I^{ABo} = (- Pr(X|AB) \log_2 Pr(X|AB) + (- Pr(\neg X |AB) \log_2 Pr(\neg X |AB))) \quad (11.2)$$

$$I^{AB1} = - Pr(X|AB) [\log_2 Pr(X|A) + \log_2 Pr(X|B)] \quad (11.3)$$
$$- Pr(\neg X |AB) [\log_2 Pr(\neg X|A)+ \log_2 Pr(\neg X|B)]$$

The MinGen measure – formula 10 –considers the minimum generalizations of the current rule r and counts how many of those generalized rules predict a class different from the original rule r. Let m be the number of conditions (attribute-value pairs) in the antecedent of rule r. Then rule r has m minimum generalizations. The k-th minimum generalization of r, k=1,...m, is obtained by removing the k-th condition from r. Let C be the class predicted by the original rule r (i.e., the majority class among the examples covered by the antecedent of r) and C_k be the class predict by the k-th minimum generalization of r (i.e., the majority class of the examples covered by the antecedent of the k-th minimum generalization of r). The system compares C with each C_k, k=1,…, m, and N is defined as the number of times where C is different from C_k.

InfoChange-ADT (Adapted for Decision Trees) is a variation of the InfoChange measure proposed by [4]. Let $A \rightarrow C$ be a common sense rule and $A, B \rightarrow \neg C$ be an exception rule. The original InfoChange measure computes the interestingness of an exception rule based on the amount of change in information relative to common sense rules. In formulas (11.1), (11.2) and (11.3), I^{ABo} denotes the number of bits required to describe the specific rule $AB \rightarrow C$ in the absence of knowledge represented by the generalized rules $A \rightarrow C$ and $B \rightarrow C$, whereas I^{AB1} is the

corresponding number of bits when the relationship between C and AB is rather described by the two rules $A \rightarrow C$ and $B \rightarrow C$. One limitation of the original InfoChange measure is that it requires the existence of a pair of exception and common sense rules, which is never the case when converting a decision tree into a set of rules – since the derived rules have mutually exclusive coverage. In order to avoid this limitation and make InfoChange useful in our experiments, the new version InfoChange-ADT is introduced in this paper, as follows. A path from the root to a leaf node corresponds to an exception rule. The common sense rule for that exception rule is produced by removing the condition associated with the parent node of the leaf node. This produces a common sense rule which is "the minimum generalization" of the exception rule. Even with this modification, InfoChange-ADT still has the limitation that its value cannot always be computed, because sometimes the minimum generalization of an exception rule predicts the same class as the exception rule, violating the conditions for using this measure.

For all the 11 rule interestingness measures previously discussed, the higher the value of the measure, the more interesting the rule is estimated to be.

3 Data Sets and Experimental Methodology

In order to evaluate the correlation between objective rule interestingness measures and real, subjective human interest, we performed experiments with 8 data sets. Public domain data sets from the UCI data repository are not appropriate for our experiments, simply because we do not have access to any user who is an expert in those data sets. Hence, we had to obtain real-world data sets where an expert was available to subjectively evaluate the interestingness of the discovered rules. Due to the difficult of finding available real-world data and expert users, our current experiments involved only one user for each data set. This reduces the generality of the results in each data set, but note that the overall evaluation of each rule interestingness measure is (as discussed later) averaged over 8 data sets and over 9 rules for each data set, i.e. each of the 11 measures is evaluated over 72 rule-user pairs. The 8 data sets are summarized in Table 1. Next, we describe the five steps of our experimental methodology.

Table 1. Characteristics of data sets used in the experiments

Data Set	Nature of Data	# Examp.	# Attrib.
CNPq1	Researchers' productivity (# publications), data from the Brazilian Research Council (CNPq)	5690	23
ITU	Patients in Intensive Care Unit	7451	41
UFPR-CS	Students' performance in comp. sci. admiss. exam	1181	48
UFPR-IM	Students' performance in info. manag. admis. exam	235	48
UTP-CS	Comp. Sci. students' end of registration	693	11
Curitiba	Census data for the city of Curitiba, Brazil	843	43
Londrina	Census data for the city of Londrina, Brazil	4115	42
Rio Branco	Census data for city of Rio Branco do Ivai, Brazil	223	43

Step 1 – Discovery of classification rules using several algorithms

We applied, to each data set, 5 different classification algorithms. Three of them are decision-tree induction algorithms (variants of C4.5 [6]), and two are genetic

algorithms (GA) that discover classification rules. In the case of the decision tree algorithms, each path from the root to a leaf node was converted into an IF-THEN classification rule as usual [6]. A more detailed description of the 5 algorithms can be found in [1], where they are referred to as default C4.5, C4.5 without pruning, "double C4.5", "Small-GA", "Large-GA". The Rule Interestingness (RI) measures were applied to each of the discovered rules (after all the classification algorithms were run), regardless of which classification algorithm generated that rule.

Step 2 – Ranking all rules based on objective rule interestingness measures

For each data set, all classification rules discovered by the 5 algorithms are ranked based on the values of the 11 objective RI measures, as follows. First, for each rule, the value of each of the 11 RI measures is computed. Second, for each RI measure, all discovered rules are ranked according to the value of that measure. I.e., the rule with the best value of that RI measure is assigned the rank number 1, the second best rule assigned the rank number 2, and so. This produces 11 different rankings for the discovered rules, i.e., one ranking for each RI measure. Third, we compute an *average* ranking over the 11 rankings, by assigning to each rule a rank number which is the *average* of the 11 rank numbers originally associated with that rule. This average rank number is then used for the selection of rules in the next step.

Step 3 – Selection of the rules to be shown to the user

Table 2 shows, for each data set, the total number of rules discovered by all the 5 algorithms applied to that data set. It is infeasible to show a large number of discovered rules to the user. Hence, we asked each user to evaluate the subjective degree of interestingness of just 9 rules out of all rules discovered by all algorithms. The set of 9 rules showed to the user consisted of: (a) the three rules with the lowest rank number (i.e., rules with rank 1, 2, 3, which were the three most interesting rules according to the objective RI measures); (b) the three rules with the rank number closest to the median rank (e.g., if there are 15 rules, the three median ranks would be 7, 8, 9); and (c) the three rules with the highest rank number (least interesting rules). The selection of rules with the lowest, median and highest rank numbers creates three distinct groups of rules which ideally should have very different user-specified interestingness scores. The correlation measure calculated over such a broad range of different objective ranks is more reliable than the correlation measure that would be obtained if we selected instead 9 rules with very similar objective ranks.

Table 2. Total number of discovered rules for each data set

Data Set:	CNPq1	ITU	UFPR-CS	UFPR-IM	UTP-CS	Curitiba	Londrina	Rio Branco do Ivai
# Rules:	20,253	6,190	1,345	232	2,370	1,792	1,261	486

Step 4 – Subjective evaluation of rule interestingness by the user

For each data set, the 9 rules selected in step 3 were shown to the user, who assigned a subjective degree of interestingness to each rule. The user-specified score can take on three values, viz.: <1> – the rule is not interesting, because it represents a relationship known by the user; <2> – the rule is somewhat interesting, i.e., it contributes a little to increase the knowledge of the user; <3> – the rule is truly interesting, i.e., it represents novel knowledge, previously unknown by the user.

Step 5 – Correlation between objective and subjective rule interestingness

We measured the correlation between the rank number of the selected rules – based on the *objective* RI measures – and the *subjective* RI scores – <1>, <2>, <3> – assigned by the user to those rules. As a measure of correlation we use the Pearson coefficient of linear correlation, with a value in [–1...+1], computed using SPSS.

4 Results

Table 3 shows, for each data set, the correlation between each objective RI measure and the corresponding subjective RI score assigned by the user. These correlations are shown in columns 2 through 9 in Table 3, where each column corresponds to a data set. To interpret these correlations, recall that the lower the objective rank number the more interesting the rule is *estimated to be*, according to the objective RI measure; and the higher the user's subjective score the more interesting the rule *is to the user*. Hence, an ideal objective RI measure should behave as follows. When a rule is assigned the best possible subjective score (<3>) by the user, the RI measure should assign a low rank number to the rule. Conversely, when a rule is assigned the worst possible subjective score (<1>) by the user, the RI measure should assign a high rank number to the rule. Therefore, the closer the correlation value is to –1 the more effective the corresponding objective RI measure is in *estimating the true degree of interestingness of a rule to the user*. In general a correlation value ≤ –0.6 can be considered a strong negative correlation, which means the objective RI measure is quite effective in estimating the real human interest in a rule. Hence, in Table 3 all correlation values ≤ –0.6 are shown in bold.

In columns 2 through 9 of Table 3, the values between brackets denote the ranking of the RI measures for each data set (column). That is, for each data set, the first rank (1) is assigned to the smallest (closest to –1) value of correlation in that column, the second rank (2) is assigned to the second smallest value of correlation, etc. Finally, the last column of Table 3 contains the average rank number for each RI measure – i.e., the arithmetic average of all the rank numbers for the RI measure across all the data sets. The numbers after the symbol "±" are standard deviations.

Two cells in Table 3 contain the symbol "N/A" (not applicable), rather than a correlation value. This means that SPSS was not able to compute the correlation in question because the user's subjective RI scores were constant for the rules evaluated by the user. This occurred when only a few rules were shown to the user. In general each correlation was computed considering 9 rules selected shown to the user, as explained earlier. However, in a few cases the value of a given objective RI measure could not be computed for most selected rules, and in this case the rules without a value for an objective RI measure were not considered in the calculation of the correlation for that measure. For instance, the N/A symbol in the cell for InfoChange-ADT and data set UFPR-CS is explained by the fact that only 2 out of the 9 selected rules were assigned a value of that objective RI measure, and those two rules had the same subjective RI score assigned by the user.

As shown in Table 3, the strength of the correlation between an objective RI measure and the user's subjective RI score is quite dependent on the data set. In three data sets – namely UFPR-CS, UTP-CS and UFPR-IM – the vast majority of the

objective RI measures were quite effective, having a strong correlation (≤ –0.6, shown in bold) with the user's true degree of interestingness in the rules. On the other hand, in each of the other five data sets there was just one objective RI measure that was effective, and in most cases the effective measure (with correlation value shown in bold) was different for different data sets. Correlation values that are very strong (≤ –0.9) are rarer in Table 3, but they are found for five RI measures in the UFPR-CS data set, and for one or two RI measures in three other data sets.

Table 3. Correlations between objective rule interestingness measures and real human interest; and ranking of objective rule interestingness measures based on these correlations

Rule interestingness measure	Data Set								Avg. Rank
	ITU	UFPR-CS	UTP-CS	Curitiba	UFPR-IM	Londrina	CNPq1	Rio Bran	
Φ-Coefficient	**-0.63** (1)	**-0.91** (4)	**-0.69** (7)	-0.17 (5)	**-0.97** (2)	0.01 (4)	-0.48 (4)	0.45 (10)	4.63 ±2.8
Infochange-ADT (*)	-0.18 (10)	N/A	-0.17 (11)	**-0.70** (1)	**-1.00** (1)	-0.54 (2)	0.15 (8)	**-1.00** (1)	4.86 ±4.6
Kappa	-0.44 (6)	**-0.94** (3)	**-0.74** (5)	-0.12 (6)	**-0.87** (4)	0.12 (5)	-0.18 (7)	-0.56 (3)	4.88 ±1.5
Cosine	-0.55 (3)	**-0.79** (6)	**-0.93** (2)	-0.49 (2)	**-0.81** (7)	0.37 (8)	**-0.64** (1)	0.79 (11)	5.00 ±3.6
Piatesky Shapiro	-0.45 (5)	**-0.95** (1)	**-0.68** (8)	-0.09 (9)	**-0.87** (5)	0.19 (7)	-0.49 (3)	-0.55 (4)	5.25 ±2.7
Interest	-0.40 (8)	**-0.77** (7)	**-0.85** (3)	-0.44 (3)	**-0.87** (6)	**-0.61** (1)	0.28 (9)	-0.22 (7)	5.50 ±2.8
Collective Strength	-0.44 (7)	**-0.94** (2)	**-0.66** (9)	-0.10 (7)	**-0.88** (3)	0.19 (6)	0.35 (10)	-0.56 (2)	5.75 ±3.1
Jaccard	-0.49 (4)	**-0.69** (8)	**-0.93** (1)	-0.10 (8)	-0.30 (9)	0.41 (9)	-0.45 (5)	-0.52 (5)	6.13 ±2.9
Odds Ratio	-0.59 (2)	**-0.91** (5)	**-0.85** (4)	-0.28 (4)	N/A	0.48 (10)	0.43 (11)	0.19 (9)	6.43 ±3.5
MinGen	-0.36 (9)	**-0.60** (9)	**-0.71** (6)	0.00 (10)	0.36 (10)	-0.22 (3)	-0.53 (2)	-0.23 (6)	6.88 ±3.1
Attsurp	0.42 (11)	-0.46 (10)	-0.54 (10)	0.63 (11)	**-0.62** (8)	0.59 (11)	-0.37 (6)	-0.10 (8)	9.38 ±1.9

(*) Although InfoChange-ADT obtained the second best rank overall, it was not possible to compute the value of this measure for many discovered rules (see text).

Consider now the average rank number of each measure shown in the last column of Table 3. The RI measures are actually in increasing order of rank number, so that, overall, across the eight data sets, the most effective RI measure was the Φ-Coefficient, with an average rank of 4.63. However, taking into account the standard deviations, there is no significant difference between the average rank of Φ-Coefficient and the average rank of the majority of the measures. The only measure which performed significantly worse than Φ-Coefficient was Attribute Surprisingness, the last in the average ranking.

There is, however, an important caveat in the interpretation of the average ranking of InfoChange-ADT. As explained earlier, there are several rules where the value of

this RI measure cannot be computed. More precisely, out of the 9 rules selected to be shown to the user for each data set, the number of rules with a value for InfoChange-ADT varied from 2 to 5 across different data sets. This means that the average rank assigned to InfoChange-ADT is less reliable than the average rank assigned to other measures, because the former was calculated from a considerably smaller number of samples (rules). In particular, the correlation value of InfoChange-ADT was –1 (the best possible value) in two data sets, viz. UFPR-IM and Rio Branco, and in both data sets only 2 out of the 9 selected rules had a value for InfoChange-ADT.

5 Conclusions and Future Research

The central question investigated in this paper was: "how effective objective rule interestingness measures are, in the sense of being a good estimate of the true, subjective degree of interestingness of a rule to the user?" This question was investigated by measuring the correlation between each of 11 objective rule interestingness measures and real human interest in rules discovered from 8 different data sets. Overall, 31 out of the 88 (11 × 8) correlation values can be considered strong (correlation ≥ 60%). This indicates that objective rule interestingness measures were effective (in the sense of being good estimators of real human interest) in just 35.2% (31 / 88) of the cases. There was no clear "winner" among the objective measures – the correlation values associated with each measure varied considerably across the 8 data sets.

A research direction would be to try to predict which objective rule interestingness measure would be most correlated with real human interest for a given target data set, or to predict the real human interest in a rule using a combination of results from different objective measures. This could be done, in principle, using a meta-learning framework, mining data from previously-computed values of the correlation between objective interestingness measures and subjective human interest for a number of rules that have been previously evaluated by a given user.

References

[1] Carvalho, D.R.; Freitas, A.A. Evaluating Six Candidate Solutions for the Small-Disjunct Problem and Choosing the Best Solution via Meta Learning. *AI Review* 24(1), 61-98, 2005

[2] Carvalho, D.R.; Freitas, A.A.; Ebecken, N.F. (2003) A Critical Review of Rule Surprisingness Measures. Proc. 2003 Int. Conf. on Data Mining, 545-556. WIT Press.

[3] Hilderman, R.J.; Hamilton H.J. *Knowledge Discovery Measures of Interest*. Kluwer, 2001.

[4] Hussain, F.; Liu, H.; Lu, H. Exception Rule Mining with a Relative Interestingness Measure. PAKDD-2000, LNAI 1805, 86-96. Springer-Verlag.

[5] Ohsaki, M., Kitaguchi, S., Okamoto, K., Yokoi, H. Yamaguchi, T. Evaluation of rule interestingness measures with a clinical dataset on hepatitis. *Knowledge Discovery in Databases: PKDD 2004, LNAI 3202*, 362-373. Springer-Verlag, 2004

[6] Quinlan, J.R. *C4.5: programs for machine learning*. Morgan Kaufmann. 1993.

[7] Tan, P.N.; Kumar, V. and Srivastava, J. Selecting the right interestingness measure for association patterns. *Proc. ACM SIGKDD KDD-2002*. ACM Press, 2002

[8] Tsumoto, S. Clinical knowledge discovery in hospital information systems. *Principles of Data Mining and Knowledge Discover, PKDD-2000*, 652-656. Springer-Verlag, 2000.

[9] Wong, M.L. and Leung, K.S. *Data mining using grammar-based genetic programming and applications*. Kluwer, 2000.