Computational Intelligence, Volume 17, Number 3, 2001

RULE QUALITY MEASURES FOR RULE INDUCTION SYSTEMS: DESCRIPTION AND EVALUATION

AIJUN AN AND NICK CERCONE
Department of Computer Science, University of Waterloo
Waterloo, Ontario N2L 3 G1, Canada

A rule quality measure is important to a rule induction system for determining when to stop generalization or specialization. Such measures are also important to a rule-based classification procedure for resolving conflicts among rules. We describe a number of statistical and empirical rule quality formulas and present an experimental comparison of these formulas on a number of standard machine learning datasets. We also present a meta-learning method for generating a set of formula-behavior rules from the experimental results which show the relationships between a formula's performance and the characteristics of a dataset. These formula-behavior rules are combined into formula-selection rules that can be used in a rule induction system to select a rule quality formula before rule induction. We will report the experimental results showing the effects of formula-selection on the predictive performance of a rule induction system.

Key words: rule quality measures, rule induction, meta-learning.

1. INTRODUCTION

A rule induction system generates decision rules from a set of training data. The set of decision rules determines the performance of a classifier that exploits the rules to classify unseen objects. It is therefore important for a rule induction system to generate decision rules that have high predictability or reliability. These properties are commonly measured by a function called rule quality.

A rule quality measure is needed in both the rule induction and classification processes. A rule induction process is usually considered as a search over a hypothesis space of possible rules for a decision rule that satisfies some criterion. The possible rules, in this case, are those rules that are defined by a concept description language, such as propositional rules. In the rule induction process that is based on general-to-specific search (such as CN2 (Clark and Boswell 1991), HYDRA (Ali and Pazzani 1993)), a rule quality measure can be used as an evaluation heuristic to select attribute-value pairs in the rule specialization process; and/or it can be employed as a significance measure to stop further specialization. The main reason to focus special attention on the stopping criterion can be found in the studies on *small disjunct problems* (Holte, Acker, and Porter 1989; Ting 1994). The studies indicated that small disjuncts, which cover a small number of training examples, are much more error prone than large disjuncts that cover a large amount of training examples. To prevent small disjuncts, a stopping criterion based on rule consistency (i.e., the rule is consistent with the training examples) is not suggested for use in rule induction. Other criteria, such as the G2 likelihood ratio statistic as used in CN2 (Clark and Niblett 1989) and the degree of logical sufficiency as used in HYDRA (Ali and Pazzani 1993), have been proposed to "pre-prune" a rule to avoid overspecialization of the rule. Some rule induction systems, such as C4.5 (Quinlan 1993) and ELEM2 (An and Cercone 1998), use an alternative strategy to prevent the small disjunct problem. In these systems, the rule specialization process is allowed to run to completion (i.e., it forms a rule that is consistent with the training data or as nearly consistent as possible) and "post-prunes" overfitted rules by

Address correspondence to the authors at University of Waterloo, Ontario N2L 3G1 Canada; e-mail: {aan, ncercone}@uwaterloo.ca.

removing components that are deemed unreliable. Similar to pre-pruning, a criterion is needed in post-pruning to determine when to stop this generalization process.

A rule quality measure is also needed in the classification process. It is possible that an unseen example satisfies multiple decision rules that are assigned to different classes. In this situation, some conflict resolution scheme must be applied to assign the unseen object to the most appropriate class. It is therefore useful for each rule to be associated with a numerical factor which can represent its classification power, its reliability, etc.

This paper consists of three parts. First, we briefly survey a number of rule quality measures that have appeared in the literature. The measures encompass statistical or empirical formulas, most of which have been discussed by Bruha (1993, 1996) and An and Cercone (1999). Second, we describe our experiments that evaluate these rule quality formulas and report the evaluation results in terms of predictive accuracy of a learning system that uses these formulas on a number of standard data sets. We compare each pair of the formulas by indicating the significance level of the difference between the two formulas. The learning system we used in our evaluations is ELEM2 (An and Cercone 1998) which induces decision rules from a set of training data and uses the induced rules to classify (new) examples. In our earlier work (An and Cercone 1999), we evaluated some of these formulas on a smaller collection of data sets. One contribution of this paper is to include more formulas in our experiments and the tests also go beyond our earlier tests by including more data sets in the experiments. Finally, we present a meta-learning method that discovers relationships between the characteristics of a data set and the performance of a formula on the data set. For each formula, the learned relationships are represented by a set of formula-behavior rules. The formula-behavior rules for all the tested formulas are further screened and combined into formula-selection rules. These formula-selection rules are a set of meta-rules that can be employed by ELEM2 to select a rule quality formula before inducing rules from a dataset. We report the experimental results showing the effects of formula-selection on ELEM2's predictive performance.

2. RULE QUALITY MEASURES

Many rule quality measures are derived by analyzing the relationship between a decision rule R and a class C. The relationship can be depicted by a 2×2 *contingency table* (Arkin and Colton 1990; Bruha and Kockova 1993) which consists of a cross-tabulation of categories of observations with the frequency for each cross-classification as shown in Table 1 where n_{rc} is the number of training examples covered by rule R and belonging to class C; $n_{r\bar{c}}$ is the number of training examples covered by R but not belonging to C, etc.; N is the total number of training examples; n_r, $n_{\bar{r}}$, n_c and $n_{\bar{c}}$ are marginal totals, e.g., $n_r = n_{rc} + n_{r\bar{c}}$, which is the number of examples covered by R. The

TABLE 1. Contingency Table with Absolute Frequencies

	Class C	Not class C	
Covered by rule R	n_{rc}	$n_{r\bar{c}}$	n_r
Not covered by R	$n_{\bar{r}c}$	$n_{\bar{r}\,\bar{c}}$	$n_{\bar{r}}$
	n_c	$n_{\bar{c}}$	N

TABLE 2. Contingency Table with Relative Frequencies

	Class C	Not class C	
Covered by rule R	f_{rc}	$f_{r\bar{c}}$	f_r
Not covered by R	$f_{\bar{r}c}$	$f_{\bar{r}\,\bar{c}}$	$f_{\bar{r}}$
	f_c	$f_{\bar{c}}$	1

contingency table can also be presented using relative rather than absolute frequencies as shown in Table 2 where $f_{rc} = \frac{n_{rc}}{N}$, $f_{r\bar{c}} = \frac{n_{r\bar{c}}}{N}$, and so on.

2.1. Empirical Formulas

Some rule quality formulas represent an ad hoc approach to the definition of rule quality. Bruha (1993) refers to these formulas as *empirical* formulas because they are not necessarily backed by statistical or information theories, but rather by intuition. We describe two empirical formulas that combine two basic characteristics of a rule: consistency and coverage. Using the elements of the contingency table, the consistency (also called apparent accuracy, i.e., accuracy over the training examples) of a rule R can be defined as $cons(R) = n_{rc}/n_r$ and its coverage as $cover(R) = n_{rc}/n_c$. Both consistency and coverage are important indicators of a rule's reliability. Rules that cover a large number of positive examples of a class may also cover negative examples well. On the other hand, only considering consistency may lead to generation of rules covering few examples which in turn results in poor predictive performance since the rule may overfit the data. Two formulas that are based on consistency and coverage are described below.

Weighted Sum of Consistency and Coverage. Michalski (1990) proposes to use the weighted sum of the consistency and coverage as a measure of rule quality as follows:

$$Q_{WS} = w_1 \times cons(R) + w_2 \times cover(R),$$

where w_1 and w_2 are user-defined weights with their values belonging to (0, l) and summed to 1. This formula is applied in an incremental learning system YAILS (1993). The weights in YAILS are specified automatically as:

$$w_1 = 0.5 + \frac{1}{4}cons(R) \text{ and } w_2 = 0.5 - \frac{1}{4}cons(R).$$

The weights here are dependent on consistency. The larger the consistency, the more influence consistency has on rule quality.

Product of Consistency and Coverage. Brazdil and Torgo (1990) propose to use a product of consistency and coverage as rule quality:

$$Q_{Prod} = cons(R) \times f(cover(R)),$$

where f is an increasing function. The authors conducted a large number of experiments and chose to use the following form of f: $f(x) = e^{x-1}$. This setting of f makes the difference in coverage have smaller influence on rule quality, which results in the rule quality formula to prefer consistency.

2.2. Measures of Association

A measure of association indicates a relationship between the classification for the columns and the classification for the rows in the 2×2 contingency table. Two statistics can be used to measure the association.

Pearson χ^2 *Statistic*. The χ^2 statistic is based on the assumption: if the classification for the columns is independent of that for the rows, the frequencies in the cells of the contingency table should be proportional to the marginal totals. The χ^2 value is given by

$$\chi^2 = \sum \frac{(n_o - n_e)^2}{n_e},$$

where n_o is the observed absolute frequency of examples in a cell, and n_e is the expected absolute frequency of examples for the cell. For example, for the upper-left cell, $n_o = n_{rc}$ and $n_e = n_r n_c / N$. The value $(n_o - n_e)^2 / n_e$ is computed for each cell of the table individually and the values for all cells are added to yield the value of χ^2. A computational formula for χ^2 can be obtained using only the values in the contingency table with absolute frequencies (Bruning and Kintz 1997):

$$\chi^2 = \frac{N(n_{rc} n_{\bar{r}\bar{c}} - n_{r\bar{c}} n_{\bar{r}c})^2}{n_c n_{\bar{c}} n_r n_{\bar{r}}}.$$

This value measures whether the classification of examples by the rule R and one by the class C are related, i.e., whether the rule R does affect the class C. The lower the χ^2 value, the more likely it is that the correlation between R and C is due to chance.

G^2 *Likelihood Ratio Statistic*. The G^2 likelihood ratio measures the distance between two distributions: the observed frequency distribution of examples among classes satisfying the rule R and the expected frequency distribution of the same number of examples under the assumption that the rule R selects examples randomly. The value of this statistic can be obtained using the absolute frequencies in the contingency table as follows:

$$G^2 = 2\left(\frac{n_{rc}}{n_r} \ln \frac{n_{rc} N}{n_r n_c} + \frac{n_{r\bar{c}}}{n_r} \ln \frac{n_{r\bar{c}} N}{n_r n_{\bar{c}}} \right).$$

The lower the G^2 value, the more likely it is that the apparent association between the two distributions is due to chance. Both the χ^2 and the likelihood ratio statistics are distributed asymptotically as χ^2 with one degree of freedom.

2.3. Measures of Agreement

A measure of agreement concerns the association of the elements of a contingency table on its main diagonal only (Bruha and Kockova 1993). The following two measures of agreement are used in our experiments.

Cohen's Formula. We can measure the actual agreement by simply summing up the main diagonal using the relative frequencies: $f_{rc} + f_{\bar{r}\bar{c}}$. A *chance* agreement occurs if the row variable is independent of the column variable, which can be measured by $f_r f_c + f_{\bar{r}} f_{\bar{c}}$. Cohen (1960) suggests to compare the actual agreement with the chance

agreement by using the normalized difference of the two which we can use as a rule quality measure:

$$Q_{Cohen} = \frac{f_{rc} + f_{\bar{r}\bar{c}} - (f_r f_c + f_{\bar{r}} f_{\bar{c}})}{1 - (f_r f_c + f_{\bar{r}} f_{\bar{c}})}.$$

When both elements f_{rc} and $f_{\bar{r}\bar{c}}$ are reasonably large, Cohen's statistic gives a higher value which indicates the agreement on the main diagonal.

Coleman's Formula. Coleman (Bishop, Fienbehg, and Holand 1991; Bruha and Kockova 1993) defines a measure of agreement that indicates an association between the first column and any particular row in the contingency table. Bruha (1993) suggests using a modified version of Coleman's measure for the purpose of rule quality definition, which actually responds to the agreement on the upper-left element of the contingency table. The formula is also derived by normalizing the difference between the actual and chance agreement as follows:

$$Q_{Coleman} = \frac{f_{rc} - f_r f_c}{f_r - f_r f_c}.$$

C_1 and C_2 Formulas. Both Coleman's formula and Cohen's formula can be represented using consistency and coverage of a rule (R) as follows:

$$Q_{Coleman} = \frac{cons(R) - f_c}{1 - f_c}, \quad Q_{Cohen} = \frac{cons(R) - f_c}{\frac{1}{2}\left(1 + \frac{cons(R)}{cover(R)}\right) - f_c}.$$

one can find that Coleman's formula does not properly comprise the coverage (i.e, completeness) of rule. On the other hand, Cohen's statistic is more completeness-based. Therefore, Bruha (1996) modified Coleman's formula in two ways, which yields formulas C_1 and C_2:

$$Q_{C1} = Q_{Coleman} \times \frac{2 + Q_{Cohen}}{3}, \quad Q_{C2} = Q_{Coleman} \times \frac{1 + cover(R)}{2},$$

where the coefficients 2, 3 and 1, 2 are used for the normalization purpose.

2.4. Measure of Information

The measure of information is another statistical measure that can be used to define rule quality. Given a class C, the amount of information necessary to correctly classify an instance into class C whose prior probability is $P(C)$ is defined as $-\log_2 P(C)$ (Kononenko and Bratko 1991). Now given a rule R, the amount of information we need to correctly classify an instance into class C is $-\log P(C|R)$, where $P(C|R)$ is the posterior probability of C given R. Therefore, the amount of information obtained by the rule R is

$$-\log P(C) + \log P(C|R).$$

Kononenko and Bratko (1991) call the value of this formula the *information score*, which measures the amount of information the rule R contributes. Using frequencies to estimate the probabilities, the formula can be written as

$$Q_{IS} = -\log \frac{n_c}{N} + \log \frac{n_{rc}}{n_r}.$$

2.5. Measure of Logical Sufficiency

The logical sufficiency measure is a standard likelihood ratio statistic, which have been applied to measure rule quality (Duda, Gaschnig, and Hart 1979; Ali and Pazzani 1993). Given a rule R and a class C, the degree of logical sufficiency of R with respect to C is defined by

$$Q_{LS} = \frac{P(R|C)}{R(R|\overline{C})},$$

where P denotes probability. A rule for which Q_{LS} is large means that the observation of R is encouraging for the class C—in the extreme case of Q_{LS} approaching infinity, R is sufficient to establish C in a strict logical sense. On the other hand, if Q_{LS} is much less than unity, then the observation of R is discouraging for C. Using frequencies to estimate the probabilities, the formula can be expressed as

$$Q_{LS} = \frac{n_{rc}/n_c}{n_{r\bar{c}}/n_{\bar{c}}}.$$

2.6. Measure of Discrimination

Another statistical rule quality formula is the measure of discrimination, which is applied in the ELEM2 rule induction system (An and Cercone 1998). The formula was inspired by a query term weighting formula used in the probability-based information retrieval. The formula measures the extent to which a query term can discriminate between relevant and non-relevant documents (Robertson and Sparck Jones 1976). If we consider a rule R as a query term in an information retrieval setting, positive examples of a class C as relevant documents, and negative examples as non-relevant documents, then the following formula can be used to measure the extent to which the rule R can discriminate between the positive and negative examples of the class C:

$$Q_{MD} = \log \frac{P(R|C)(1 - P(R|\overline{C}))}{P(R|\overline{C})(1 - P(R|C))},$$

where P denotes probability. The formula represents the ratio between the rule's positive and negative odds and can be estimated using the frequencies as

$$Q_{MD} = \log \frac{n_{rc}/n_{\bar{r}c}}{n_{r\bar{c}}/n_{\bar{r}\bar{c}}}.$$

3. EXPERIMENTS WITH RULE QUALITY MEASURES

3.1. The Learning System

To evaluate rule quality measures, ELEM2 (An and Cercone 1998) is used as the rule induction system in our experiments. Given a set of training data, ELEM2 sequentially learns a set of rules for each of classes in the data set. To induce rules for a class C, ELEM2 conducts general-to-specific heuristic search over a hypothesis space to generate a disjunctive set of propositional rules. ELEM2 uses a *sequential covering* learning strategy; it reduces the problem of learning a disjunctive set of rules to a sequence of simpler problems, each requiring that a single conjunctive rule be learned that covers a

subset of positive examples. The learning of a single conjunctive rule begins by considering the most general rule precondition, i.e., the empty test that matches every training example, then greedily searching for an attribute-value pair that is most relevant to the class C according to the following attribute-value pair evaluation function:

$$SIG_C(av) = P(av)(P(C|av) - P(C)),$$

where av is an attribute-value pair and P denotes probability.[1] The selected attribute-value pair is then added to the rule precondition as a conjunct. The process is repeated by greedily adding a second attribute-value pair, and so on, until the hypothesis reaches an acceptable level of performance. In ELEM2, the acceptable level is based on the consistency of the rule: it forms a rule that is as consistent with the training data as possible. Since this "consistent" rule may be a small disjunct that overfits the training data, ELEM2 "post-prunes" the rule after the initial search for this rule is complete.

To post-prune a rule, ELEM2 first computes a rule quality value according to the formula of measure of discrimination Q_{MD} (section 2.6). It then checks each attribute-value pair in the rule in the reverse order in which they were selected to see if removal of the attribute-value pair will decrease the rule quality value. If not, the attribute-value pair is removed and the procedure checks all the other pairs in the same order again using the new rule quality value resulting from the removal of that attribute-value pair to see whether another attribute-value pair can be removed. This procedure continues until no pair can be removed.

After rules are induced for all the classes, the rules can be used to classify new examples. The classification procedure in ELEM2 considers three possible cases when a new example matches a set of rules.

1. *Single match*. The new example satisfies one or more rules of the same class. In this case, the example is classified to the class indicated by the rule(s).
2. *Multiple match*. The new example satisfies several rules that indicate at least two different classes. In this case, ELEM2 activates a conflict resolution scheme for the best decision. The conflict resolution scheme computes a decision score for each of the matched classes as follows:

$$DS(C) = \sum_{i=1}^{k} Q_{MD}(r_i),$$

where r_i is a matched rule that indicates C, k is the number of this kind of rules, and $Q_{MD}(r_i)$ is the rule quality of r_i. The new example is then classified into the class with the highest decision score.
3. *No match*. The new example is not covered by any rule. Partial matching is considered where some attribute-value pairs of a rule match the values of corresponding attributes in the new example. If the partially-matched rules do not agree on the classes, a partial matching score between an example e and a partially-matched rule r_i with n attribute-value pairs, m of which match the corresponding attributes of e, is computed as $PMS(r_i) = \frac{m}{n} \times Q_{MD}(r_i)$. A decision score for a class C is computed as

$$DS(C) = \sum_{i=1}^{k} PMS(r_i),$$

[1]See An and Cercone (1998) for discussion of this formula.

where k is the number of partially-matched rules indicating class C. In decision making, the new example is classified into the class with the highest decision score.

We can see that the rule quality measure Q_{MD} is used in both the post-pruning and classification processes of ELEM2.

3.2. Experimental Design

We evaluate the rule quality formulas described in Section 2 by determining how different rule quality formulas affect the predictive performance of ELEM2. In our experiments, we run versions of ELEM2, each of which uses a different rule quality formula. The formulas: Q_{MD}, Q_{Cohen}, $Q_{Coleman}$, Q_{C1}, Q_{C2}, Q_{IS}, Q_{LS}, Q_{WS}, and Q_{Prod} are used exactly as described in section 2, while the χ^2 and G^2 likelihood statistics are applied as follows. The χ^2 statistic is used in two ways, in both of which the χ^2 formula is used as the ELEM2 rule quality measure. They differ in the method to post-prune a generated rule.

1. $Q_{\chi^2_{.05}}$ In post-pruning, the removal of an attribute-value pair depends on whether the rule quality value after removing an attribute-value pair is greater than $\chi^2_{.05}$ i.e., the tabular χ_2 value for the significance level of 0.05 with one degree of freedom. If the calculated value is greater than tabular $\chi^2_{.05}$, then remove the attribute-value pair; otherwise check other pairs or stop post-pruning if all pairs have been checked.
2. $Q_{\chi^2_{.05+}}$ In post-pruning, an attribute-value pair is removed if and only if the rule quality value Q_{after} after removing an attribute-value pair is greater than $\chi^2_{.05}$ and Q_{after} is no less than the rule quality value before removing the attribute-value pair.

The G^2 statistic, denoted $Q_{G2_{.05+}}$, is used in the same way as $Q_{\chi^2_{.05+}}$, i.e., a pair is removed in post-pruning if and only if the value of $Q_{G2_{.05+}}$ is greater than $\chi^2_{.05}$ and the removal does not cause the rule quality value to decrease.

Our experiments are conducted using 27 benchmark datasets obtained from the UCI Repository of Machine Learning database. The datasets represent a mixture of characteristics described in Table 3. The current version of ELEM2 removes all the examples that contain missing values. For the datasets that contain missing values (such as "crx" and "post-operative"), the number of examples shown in Table 3 is the number of examples after the removal.

3.3. Results

On each dataset, we conduct the ten-fold evaluation of a rule quality measure using ELEM2. The results in terms of predictive accuracy mean on each dataset for each formula are shown in Figure 1. The average of the accuracy means for each formula over the 27 datasets is shown in Table 4, where the rule quality formulas are listed in decreasing order of average accuracy means. Whether a formula with a higher average is significantly better than a formula with a lower average is determined by paired t-tests using the S-Plus statistics software. The t-test results in terms of p-values are reported in Table 5. A small p-value indicates that the null hypothesis (the difference between the two formulas is due to chance) should be rejected in favor of the alternative at any significance level above the calculated value. For example, the p-value from comparing

TABLE 3. Description of Datasets

	Datasets	Number of classes	Number of attributes	Number of examples	Class Distribution	Domain
1	abalone	3	8	4177	Even	Prediction of the age of abalone from physical measurements
2	australia	2	14	690	Even	Credit card application approval
3	balance-scale	3	4	625	Uneven	Balance scale classification
4	breast-cancer	2	9	683	Uneven	Medical diagnosis
5	bupa	2	6	345	Uneven	Liver disorder database
6	car	4	6	1728	Uneven	Car evaluation
7	crx	2	15	653	Uneven	Credit card applications
8	diabetes	2	8	768	Uneven	Medical diagnosis
9	ecoli	8	7	336	Uneven	Prediction of protein localization sites
10	flag	8	28	194	Uneven	National flags classification
11	german	2	20	1000	Uneven	Credit database to classify people as good or bad credit risks
12	glass	6	9	214	Uneven	Glass identification for criminological investigation
13	heart	2	13	270	Uneven	Heart disease diagnosis
14	ionosphere	2	33	351	Uneven	Classification of radar returns
15	iris	3	4	150	Even	Iris plant classification
16	lenses	3	4	24	Uneven	Database for fitting contact lenses
17	optdigits	10	64	3823	Even	Optical recognition of handwritten digits
18	page-blocks	5	10	5473	Uneven	page-blocks classification
19	pendigits	10	16	7494	Even	Pen-based recognition of hand-written digits
20	post-operative	3	8	87	Uneven	Postoperative Patient Data
21	segment	7	18	2310	Even	image segmentation
22	spambase	2	57	4601	Uneven	Email classification: spam or non-spam
23	tae	3	5	151	Even	Teaching performance evaluation
24	tic-tac-toe	2	9	958	Uneven	Tic-Tac-Toe Endgame database
25	wine	3	13	178	Uneven	Wine recognition data
26	yeast	10	8	1484	Uneven	Prediction of protein localization sites
27	zoo	7	16	101	Uneven	Animal classification

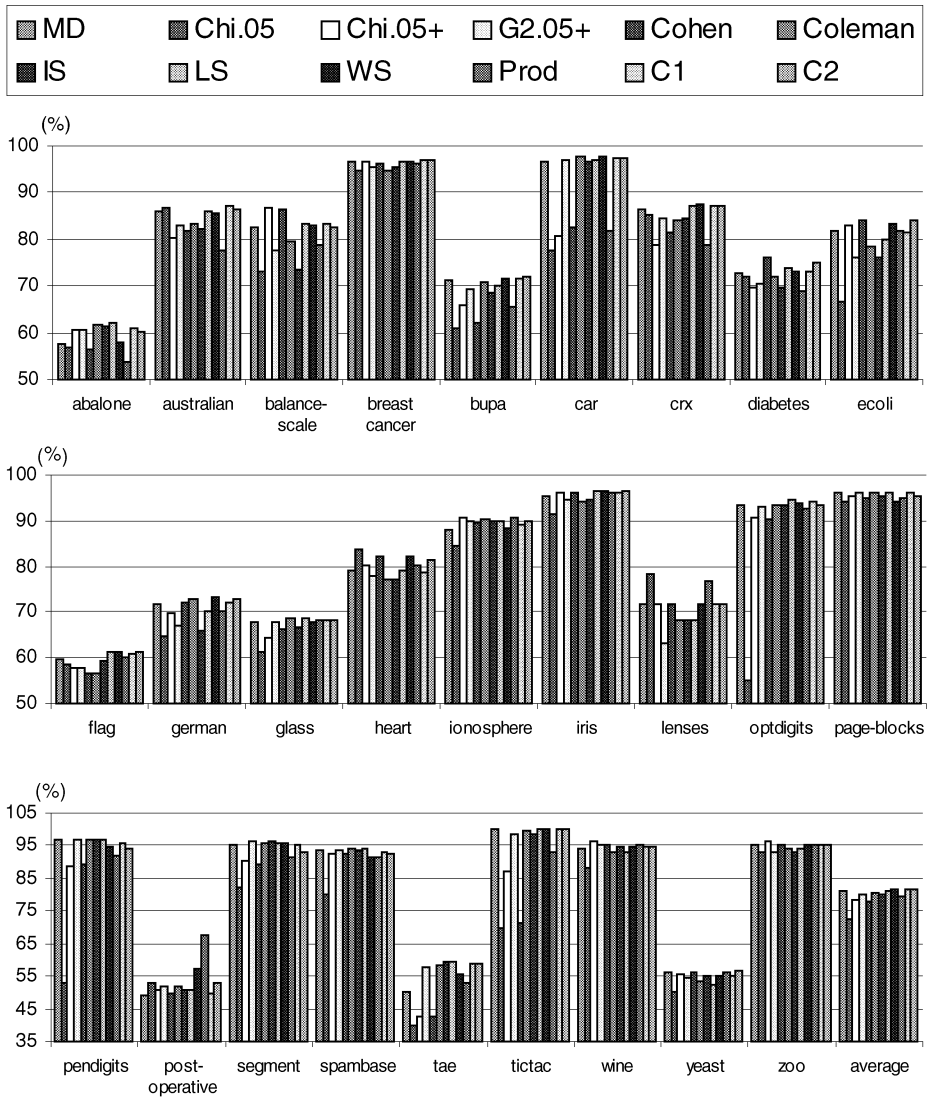

FIGURE 1. Accuracy means of different rule quality formulas on the 27 datasets.

Q_{WS} with $Q_{Coleman}$ is 0.047, which means that we would reject the null hypothesis at the 5% significance level, but would not reject it at the 1% significant level. In Table 5, the p-values that are smaller than 0.05 are underlined to indicate that the formula with higher average is significantly better than the formula with the lower average at the 5% significance level.

Generally speaking, we can say that, in terms of predictive performance, Q_{C2}, Q_{WS}, Q_{C1}, Q_{LS} and Q_{MD} are comparable even if their performance may not agree on a particular dataset. The same for $Q_{Coleman}$, $Q_{G2_{\cdot 05+}}$, Q_{IS} and Q_{Prod}, and $Q_{\chi^2_{\cdot 05+}}$ and Q_{Cohen}. The performance of $Q_{G2\cdot 05+}$ and Q_{IS} are not only comparable, but also similar on each particular dataset (seen from Figure l), which indicates that the two formulas have similar trends with regard to n_{rc}, n_r, n_c, and N in the contingency table.

TABLE 4. Average of Accuracy Means for Each Formula over the Datasets

	*C*2	*WS*	*C*1	*LS*	*MD*	*Coleman*	$G2_{.05+}$	*IS*	*Prod*	$\chi^2_{.05+}$	*Cohen*	$\chi^2_{.05}$
Average	81.89	81.71	81.61	81.38	80.95	80.65	79.94	79.87	79.59	78.44	78.08	72.42

TABLE 5. Significance Levels (p-values from Paired t-test) of Improvement

	*C*2	*WS*	*C*1	*LS*	*MD*	*Coleman*	$G2_{.05+}$	*IS*	*Prod*	$\chi^2_{.05+}$	*Cohen*	$\chi^2_{.05}$
*C*2	NA	0.548	0.278	0.165	0.026	0.008	0.001	0.002	0.033	0.002	0.007	0.000
WS	—	NA	0.807	0.487	0.072	0.047	0.005	0.009	0.029	0.003	0.011	0.000
*C*1	—	—	NA	0.333	0.072	0.009	0.001	0.001	0.087	0.005	0.016	0.001
LS	—	—	—	NA	0.367	0.028	0.001	0.002	0.134	0.012	0.028	0.001
MD	—	—	—	—	NA	0.544	0.091	0.104	0.223	0.012	0.034	0.001
Coleman	—	—	—	—	—	NA	0.038	0.063	0.357	0.061	0.091	0.002
$G2_{.05+}$	—	—	—	—	—	—	NA	0.823	0.765	0.210	0.230	0.006
IS	—	—	—	—	—	—	—	NA	0.806	0.253	0.263	0.006
Prod	—	—	—	—	—	—	—	—	NA	0.221	0.243	0.003
$\chi^2_{.05+}$	—	—	—	—	—	—	—	—	—	NA	0.625	0.007
Cohen	—	—	—	—	—	—	—	—	—	—	NA	0.008
$\chi^2_{.05}$	—	—	—	—	—	—	—	—	—	—	—	NA

4. META-LEARNING FROM THE EXPERIMENTAL RESULTS

From the experimental results, we posit that, even if on some datasets (such as the *breast cancer* dataset) the performance of the learning system is not very sensitive to the rule quality formula used, the performance greatly depends on the formula on most of the other datasets. It would be desirable that we can apply a "right" formula that gives the best performance among other formulas on a particular dataset. For example, even though the formula $Q_{\chi^2_{.05}}$ is not a good formula in general, it performs better than other formulas on some datasets such as *heart* and *lenses*. If we can find the conditions under which each formula leads to a good performance of the learning system, we can select "right formulas" for different datasets and can improve the predictive performance of the learning system further.

To find out this regularity, we use ELEM2 to meta-learn "formula selection rules" from the experimental results shown in the last section. The learning problem is divided into (1) learning the rules for each rule quality formula that describe the conditions under which the formula produces "very good," "good," "medium," or "bad" results, and (2) combining the rules for all the formulas that describe the conditions under which the formulas produce the "very good" results. The resulting set of rules is the formula-selection rules that can be used by the ELEM2 classification procedure to perform formula selection. This meta-learning process is described in Figure 2.

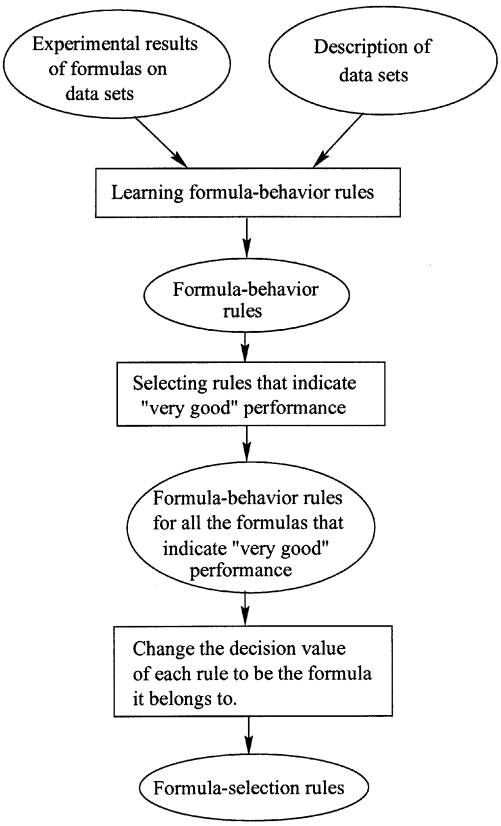

FIGURE 2. The meta-learning process for generating formula-selection rules.

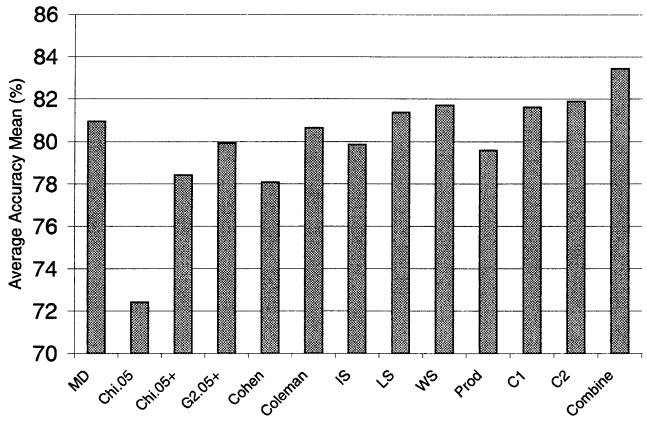

FIGURE 3. Average of accuracy means of each formula on the 27 datasets.

4.1. Data Representation

For the purpose of learning formula-behavior rules, i.e., the rules that describe the conditions under which a formula leads to "very good," "good," "medium," or "bad" performance, we construct training examples from Figure 1 and Table 3. First, on each dataset, we decide the relative performance of each formula as "very good," "good," "medium," or "bad." For example, on the *balance-scale* dataset, we say that the formulas whose accuracy mean is above 85% produce "very good" results; the formulas whose accuracy mean is between 80% and 85% produce "good" results; the ones with the mean between 75% and 80% are "medium" and other formulas give "bad" results. Then, for each formula, we construct a training data set in which a training example describes the characteristics of a dataset and also a description in term of whether the formula produces "very good," "good," "medium," or "bad" result on this dataset. Thus, to learn the rules for each formula, we have 27 training examples. The characteristics of a dataset is described in terms of number of examples, number of attributes, number of classes and the class distribution. A sample of training examples for learning the behavior rules of the formula Q_{IS} is shown in Table 6.

4.2. The Meta-learning Results

ELEM2 with its default rule quality formula (Q_{MD}) is used to learn the "behavior" rules from the training dataset constructed for each formula. Table 7 lists some of these

TABLE 6. Sample of Training Examples for Learning the Behavior of a Formula

Examples	Number of Attributes	Classes	Class Distribution	Performance
4177	8	3	Even	Very Good
690	14	2	Even	Medium
625	4	3	Uneven	Bad
683	9	2	Uneven	Medium
1728	6	4	Uneven	Good

TABLE 7. Some Formula Behavior Rules

Formula	Condition	Decision	Rule Quality	No. of Support Datasets
Q_{C2}	$(768 < N \leq 1728)$	Very Good	1.30	4
	$(N \leq 653)$and$(NofA > 10)$and$(NofC \leq 7)$	Good	1.36	5
Q_{WS}	$(625 < N \leq 1728)$and$(NofA > 8)$and$(ClassDistr! = Even)$	Very Good	1.48	4
	$(N > 336)$and$(NofC > 5)$	Good	1.38	4
Q_{C1}	$(N > 270)$and$(8 < NofA \leq 15)$	Very Good	1.66	5
	$(15 < NofA \leq 57$	Good	1.43	7
Q_{LS}	$(N > 2310)$	Very Good	1.45	5
	$(N \leq 87)$	Bad	2.41	2
Q_{MD}	$(N > 768)$and$(8 < NofA \leq 16)$	Very Good	2.04	3
	$(351 < N \leq 4601)$and$(NofA > 13)$	Good	1.23	6
$Q_{Coleman}$	$(N > 958)$and$(NofC \leq 5)$	Very Good	1.79	5
	$(N \leq 87)$	Bad	1.51	2
$Q_{G2_{.05+}}$	$(N > 101)$and$(10 < NofA \leq 18)$and$(NofC > 2)$	Very Good	2.04	3
	$(270 < N \leq 690)$and$(NofA \leq 15)$	Medium	2.25	6
Q_{IS}	$(N > 150)$and$(NofC > 2)$and$(ClassDistr = Even)$	Very Good	2.13	4
	$(N \leq 101)$	Bad	1.50	3
Q_{Prod}	$(N \leq 214)$and$(NofA > 7)$and$(NofC \leq 6)$	Very Good	1.80	3
	$(N > 768)$and$(8 < NofA \leq 57)$	Medium	1.70	6
$Q_{\chi^2_{.05+}}$	$(N \leq 178)$and$(NofA > 9)$	Very Good	1.67	2
	$(N \leq 214)$and$(4 < NofA \leq 9)$	Bad	1.39	3
Q_{Cohen}	$(345 < N \leq 1484)$and$(NofA \leq 8)$	Very Good	1.80	3
	$(4 < NofA \leq 6)$	Bad	1.63	3
$Q_{\chi^2_{.05}}$	$(9 < NofA \leq 14)$and$(NofC \leq 2))$	Very Good	1.91	2
	$(N > 24)$	Bad	0.98	20

TABLE 8. Significance Levels of the Improvement of "Combine" over Individual Formulas

	C2	*WS*	*C1*	*LS*	*MD*	*Coleman*	$G2_{.05+}$	*IS*	*Prod*	$\chi^2_{.05+}$	*Cohen*	$\chi^2_{.05}$
p-value	0.014	0.001	0.015	0.008	0.003	0.001	0.000	0.000	0.000	0.000	0.001	0.000

behavior rules for each formula, where N stands for the number of examples, NofA is the number of attributes, NofC is the number of classes, and "No. of Support Datasets" means the number of the datasets that support the corresponding rule. These rules serve two purposes. First, we can summarize the predictive performance of each formula in terms of characteristics of datasets. Second, we can create a set of formula-selection rules by combining all the "very good" rules, i.e., the rules that predicts "very good" performance for each formula, and use them to select a "right" rule quality formula for a (new) dataset. For formula selection, we can use the ELEM2 classification procedure that takes formula-selection rules to classify a data set into a class of using a particular formula.

5. CONCLUSIONS

We have described and experimented with various statistical and empirical formulas for defining rule quality measures. All formulas are applicable to a rule induction system for the purpose post-pruning and classification, but their performance varies among the datasets. The empirical formulas, especially Q_{WS}, work very well even if they are not backed by statistical theories. Among statistical formulas, Q_{C2}, Q_{C1}, Q_{LS}, and Q_{MD} work the best on the tested dataset and are comparable with Q_{WS}.

To determine the regularity of the rule quality formula's performance in terms of dataset characteristics, we used our learning system to induce formula-behavior rules from a dataset constructed from the experimental results for different formulas. These rules provided ideas about the situations in which a formula leads to very good, good, medium or bad performance. These rules can also be combined and used to automatically select a rule quality formula before rule induction begins. Our experiment showed that this selection of rule quality formula can lead to significant improvement over the rule induction system using a single rule quality formula. Future work includes testing our conclusions on more datasets to obtain more reliable formula-behavior rules. With more datasets available, we will test the formula-selection rules on the datasets that are different from the datasets used for generating the rules.

ACKNOWLEDGMENT

The authors are members of the Institute for Robotics and Intelligent Systems (IRIS) and wish to acknowledge the support of the Networks of Centres of Excellence of the Government of Canada, the Natural Sciences and Engineering Research Council, and the participation of PRECARN Associates, Inc.

REFERENCES

ALI, K., and M. PAZZANI. 1993. HYDRA: A noise-tolerant relational concept learning algorithm. Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence (IJCAI'93), Chambery, France. Morgan Kaufmann.

AN, A., and N. CERCONE. l998. ELEM2: A Learning System for More Accurate Classifications. Lecture Notes in Artificial Intelligence **1418**.

AN, A., and N. CERCONE. 1999. An Empirical Study on Rule Quality Measures. Proceedings of the Seventh International Workshop on Rough Sets, Fuzzy Sets, Data Mining, and Granular-Soft Computing, Yamaguchi, Japan.

ARKIN, H., and R. R. COLTON. 1970. Statistical Methods. Barnes & Noble Inc., New York.

BISHOP, Y. M. G, S. E. FIENBERG, and P. W. HOLAND, 1991. Discrete Multivariate Analysis: Theory and Practice. The MIT Press.

BRAZDIL, P., and L. TORGO. 1990. Knowledge acquisition via knowledge integration. *In* Current Trends in Knowledge Acquisition. IOS Press.

BRUHA, I., and S. KOCKOVA. 1993. Quality of decision rules: Empirical and statistical approaches. Informatica, **17**:233–243.

BRUHA, I. 1996. Quality of decision rules: Definitions and classification schemes for multiple rules. *In* Machine Learning and Statistics. *Edited by* G. Nakhaeizadeh, and C. C. Taylor. The Interface. John Wiley & Sons Inc.

BRUNING, J. L., and B. L. KINTZ. 1997. Computational Handbook of Statistics. Addison-Wesley Educational Publishers Inc.

CLARK, P., and T. NIBLETT. 1989. The CN2 induction algorithm. Machine Learning, **3**:261–283.

CLARK, P., and R. BOSWELL. 1991. Rule induction with CN2: Some recent improvements. Proceedings of European Working Session on Learning, Porto, Portugal, pp. 151–163.

COHEN, J. 1960. A coefficient of agreement for nominal scales. Educational and Psych Meas. **22**:37–46.

DUDA, R., J. GASCHNIG, and P. HART. 1979. Model design in the prospector consultant system for mineral exploration. *In* Expert Systems in the Micro-electronic Age, Edited by D. Michie. Edinburgh University Press, Edinburgh, UK.

HOLTE, R., L. ACKER, and B. PORTER. 1989. Concept learning and the problem of small disjuncts. Proceedings of the Eleventh International Joint Conference on Artificial Intelligence, Detroit, Michigan.

KONONENKO, I., and I. BRATKO. 1991. Information-based evaluation criterion for classifier's performance. Machine Learning, 6: 67–80.

MICHALSKI, R. S. 1990. Pattern recognition as rule-guided inductive inference. IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-2, 4.

QUINLAN, J. R. 1993. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers. San Mateo, CA.

ROBERTSON, S. E., and K. SPARCK JONES. 1976. Relevance weighting of search terms. Journal of the American Society for Information Science. **27**:129–146.

TING, K. M. 1994. The problem of small disjuncts: Its remedy in decision trees. Proceedings of the Tenth Canadian Conference on Artificial Intelligence.

TORGO, L. 1993. Controlled redundancy in incremental rule learning. ECML-93, pp. 185–195.