Evaluating the Interestingness of Characteristic Rules

Micheline Kamber
School of Computing Science
Simon Fraser University
Burnaby, BC V5A 1S6, Canada
kamber@cs.sfu.ca

Rajjan Shinghal
Dept. of Computer Science
Concordia University
Montreal, QC H3G 1M8, Canada
shinghal@cs.concordia.ca

Abstract

Knowledge Discovery Systems can be used to generate classification rules describing data from databases. Typically, only a small fraction of the rules generated may actually be of interest. Measures of rule *interestingness* allow us to filter out less interesting rules.

Classification rules may be discriminant ($e \to h$) or characteristic ($h \to e$), where e is evidence, and h is an hypothesis. For discriminant rules, e distinguishes h from $\neg h$. For characteristic rules, e summarizes one or more properties common to all instances of h. Both rule types can contribute insight into the data under analysis.

In this paper, we first expand on the rule interestingness measure principles proposed by Piatetsky-Shapiro (1991) and by Major and Mangano (1993) by adding a principle which, unlike the others, considers the difference between discriminant and characteristic rules.

We establish experimentally that the three popular interestingness measures for discriminant rules found in the literature do not fully serve their purpose when applied to characteristic rules. To our knowledge, no interestingness measures for characteristic rules have been published. We propose IC^{++}, an interestingness measure for characteristic rules based on necessity and sufficiency (Duda, Gaschnig, & Hart 1981). IC^{++} obeys each of the rule interestingness principles, unlike the other measures studied. If a given characteristic rule is found to be uninteresting by IC^{++}, three additional measures, which we present, can be used to derive other useful information regarding h and e.

Introduction

Knowledge Discovery Systems (KDSs) can be used to generate patterns describing the data in a given database. Typically, the number of patterns generated is very large, only a small fraction of which may actually be of interest to the KDS user or data analyst. Measures are therefore essential for the ranking of discovered patterns according to their degree of *interestingness*, thereby allowing the filtering out of less useful information.

Such interestingness measures can be objective or subjective. *Objective* measures are based on the structure of the discovered patterns and the statistics underlying them. Such factors include coverage, confidence, statistical significance, strength, simplicity, and redundancy (Agrawal, Imielinski, & Swami 1993; Piatetsky-Shapiro 1991; Major & Mangano 1993; Smyth & Goodman 1992; Hong & Mao 1991; Srikant & Agrawal 1995). *Subjective* measures are based on user beliefs regarding relationships in the data, recognizing that a pattern of interest to one user may not be so to another. Subjective measures evaluate the interestingness of discovered patterns with respect to their unexpectedness or actionability (Silberschatz & Tuzhilin 1995; Piatetsky-Shapiro & Matheus 1994). Here a pattern that contradicts a user expectation, or that a user may act on to his or her advantage, is deemed interesting. Objective and subjective measures are both useful in assessing the interestingness of data patterns. Ideally, given a set of discovered patterns, objective measures can be applied first to filter out those patterns that do not meet structural and statistical requirements. Subjective measures can then be applied to identify the patterns most interesting to the given user.

Different KDSs can generate different types of patterns (e.g., classification rules, association rules, time series patterns, etc.). This paper is concerned with the interestingness of classification rules. Classification rules may be discriminant ($e \to h$) or characteristic ($h \to e$), where e is evidence (typically a conjunction of database attribute-value conditions), and h is an hypothesis. For discriminant rules, e distinguishes h from $\neg h$. For characteristic rules, e summarizes one or more properties common to all instances of h. In medicine, for example, discriminant rules may be used to summarize the symptoms sufficient to distinguish one disease from another, whereas characteristic rules may be used to summarize the symptoms necessary for a given disease. Both types of rules can be of interest, contributing insight into the data under analysis.

Principles that all rule interestingness measures should follow have been proposed (Piatetsky-Shapiro 1991; Major & Mangano 1993). In this paper, we first expand on these by proposing an additional principle which, unlike the others, considers the difference between discriminant and characteristic rules.

The measures of Rule Interest, or RI (Piatetsky-Shapiro 1991), J (Smyth & Goodman 1992), and certainty (Hong & Mao 1991) are three popular interestingness measures from the literature for classification rules. Although these measures help assess the interestingness of discriminant rules, we establish experimentally that they do not fully serve their purpose when applied to characteristic rules. To the best of our knowledge, no interestingness measures for characteristic rules have been published. We propose IC^{++}, an objective interestingness measure for characteristic rules, based on technical definitions of necessity and sufficiency (Duda, Gaschnig, & Hart 1981). IC^{++} obeys each of the rule interestingness principles. We also present three other rule interestingness measures. If a given characteristic rule, $h \rightarrow e$, is deemed uninteresting by IC^{++}, these additional measures can be used to derive other useful information regarding h and e.

Rule Interestingness Measure Principles

Given a rule $e \rightarrow h$ or $h \rightarrow e$ we say that the rule is X% *complete* if e satisfies (i.e., is true for, or covers) X% of the tuples satisfying h. A rule is Y% *discriminant* if e satisfies (100-Y)% of the $\neg h$ tuples (i.e., tuples for which h is false). Principles that all rule interestingness measures, I, should follow have been proposed by Piatetsky-Shapiro (1991) and by Major and Mangano (1993). The principles consider rule completeness and discriminability. Let $|S|$ denote the size of set S (e.g., $|h \wedge e|$ is the number of tuples satisfying h and e). The principles are:

1. $I = 0$ if h and e are statistically independent of each other (Piatetsky-Shapiro 1991).
2. I increases monotonically with $|h \wedge e|$ when $|h|$, $|\neg h|$, and $|e|$ remain the same (Piatetsky-Shapiro 1991). In other words, the more reliable a rule is, the more interesting it is, provided all other parameters remain fixed. The reliability of a discriminant rule is $|h \wedge e|/|e|$, and is also known as rule confidence or certainty factor. $|e|$ is known as the rule cover.
3. I decreases monotonically with $|h|$ (or $|e|$) when $|h \wedge e|$, $|e|$ (or $|h|$), and $|\neg h|$ remain the same (Piatetsky-Shapiro 1991). That is, the more complete a rule is, the more interesting it is, provided its discriminability remains constant. Similarly, the more discriminant a rule is, the more interesting it is, provided its completeness remains fixed.
4. I increases monotonically with $|e|$ when the rule reliability, $|h|$, and $|\neg h|$ remain the same (Major & Mangano 1993). Essentially, given two rules having the same reliability, but where one rule has a larger cover than the other then the rule with the larger cover is the more interesting of the two.

The above principles hold for both discriminant and characteristic rules. Recall that in discriminant rules, e is intended to discriminate h from $\neg h$, while in characteristic rules, e should be *necessary* for h, i.e., 100% complete for h. However, as noted by philosophers (e.g., Wittenstein), it frequently happens that not all members of a given class have some property in common. Moreover, this can be the case when dealing with real-world databases, owing to exceptions, noisy, or missing data. Therefore, when mining characteristic rules from real-world databases, the stipulation that e must cover *all* training examples of h should be relaxed (Han, Cai, & Cercone 1993). Hence, our definition of characteristic rules allows for incompleteness.

Rule	e	h	Discrim.	Completeness
A	Fever	Flu	80%	30%
B	Sneezing	Flu	30%	80%

Table 1: Discriminability vs. Completeness.

The interestingness of a rule may be defined as the product of its goodness (reflecting the goodness of fit between the rule and data), and utility (Smyth & Goodman 1992). For characteristic rules, $h \rightarrow e$, utility is $P(h)$; for discriminant rules, $e \rightarrow h$, it is $P(e)$. For characteristic rules, multiplying by utility biases the interestingness function towards the more frequent classes (or the more frequent evidence, for discriminant rules). Depending on the data analyst or application, such a bias may be undesirable. For example, a strong rule found for diagnosing a rare disease may be just as important as an equally strong rule for diagnosing a common disease. (One might argue that recognizing the rare disease is even more important!) Hence, a user may wish to evaluate the interestingness of a rule based on it goodness alone, or on its goodness and utility.

Consider Table 1 and assume for the moment that each class, or piece of evidence, is considered equally important. Suppose that rules A and B are discriminant (i.e., respectively, *Fever*$\rightarrow$*Flu*, and *Sneezing*$\rightarrow$*Flu*). The object of discriminant rules for the class Flu is to distinguish instances of Flu from ¬Flu. Rule A is 80% discriminant, while rule B is 30% discriminant. Therefore, as discriminant rules, rule A is more interesting than rule B. Suppose, instead, that rules A and B are characteristic (i.e., respectively, *Flu* $\rightarrow$ *Fever*, and *Flu* $\rightarrow$ *Sneezing*). The object of characteristic rules for Flu is to cover all, or as many as possible, of the Flu tuples. Since rule B is 80% complete and rule A is only 30% complete, as characteristic rules, rule B is the more interesting of the two. Hence, we see a difference in the assessment of interestingness for discriminant and characteristic rules. Thus, we propose a fifth principle for rule interestingness measures.

5. Given two rules where one rule is as complete as the other is discriminant, and vice versa, then for discriminant (characteristic) rules, the more discriminant (complete) rule has greater potential interestingness. If each class or evidence is considered equally important, then the rule with greater potential interestingness is deemed the more interesting. Otherwise, the potential interestingness

(Pot_I) may be multiplied by the rule utility in order to bias interestingness towards the more frequent classes (for characteristic rules) or the more frequent evidence (for discriminant rules). Stated formally: Let r_1 and r_2 be rules where r_1 is $X\%$ discriminant and $Y\%$ complete, and r_2 is $Y\%$ discriminant and X% complete, where $X \neq Y$, $Pot\_I(r_1) \neq 0$, and $Pot\_I(r_2) \neq 0$ where $Pot\_I(r)$ is a measure of the goodness of rule r. If $X > Y$ and r_1 and r_2 are discriminant, then $Pot\_I(r_1) > Pot\_I(r_2)$. If $X < Y$ and r_1 and r_2 are discriminant, then $Pot\_I(r_1) < Pot\_I(r_2)$. If $X > Y$ and r_1 and r_2 are characteristic, then $Pot\_I(r_2) > Pot\_I(r_1)$. If $X < Y$ and r_1 and r_2 are characteristic, then $Pot\_I(r_2) < Pot\_I(r_1)$. If all events are considered equally important, then the interestingness of rule r may be assessed as $I(r)=Pot\_I(r)$. Otherwise, $I(r)=Pot\_I(r) \times Utility(r)$ where $Utility(r)$ is $P(e)$ for discriminant rule, $e \rightarrow h$, or $P(h)$ for characteristic rule, $h \rightarrow e$.

This new principle, unlike the others, considers the difference between discriminant and characteristic rules.

Proposed Interestingness Measures for Characteristic Rules

Let c^{++} be the characteristic rule $h \rightarrow e$, and d^{++} be the discriminant rule $e \rightarrow h$. Our proposed interestingness measures for characteristic rules are based on the measures of necessity and sufficiency defined as $Nec(d^{++}) = P(\neg e|h)/P(\neg e|\neg h)$ and $Suf(d^{++}) = P(e|h)/P(e|\neg h)$ (Duda, Gaschnig, & Hart 1981). The probabilities can be estimated from the given data. Nec can be used to assess the goodness of c^{++}. For example, if $\text{Nec}(d^{++}) = 0$, then $\neg e$ invalidates h, meaning $\neg e \rightarrow \neg h$, $e \vee \neg h$, that is $\underline{h \rightarrow e}$. Here, e is *necessary* for h, i.e., e covers all training examples of h, and so the characteristic rule, c^{++}, is certain. If $0<\text{Nec}(d^{++})<1$, then $\neg e$ discourages h, meaning that $\neg e$ makes h less plausible. The closer $\text{Nec}(d^{++})$ is to 0, the more certain is $h \rightarrow e$. Hence we propose IC^{++} as an interestingness measure for c^{++}, where c^{++} can be any characteristic rule, $h \rightarrow e$ (which may contain negative literals):

$$IC^{++} = \begin{cases} (1 - Nec(d^{++})) \times P(h) & 0 \leq Nec(d^{++}) < 1 \\ 0 & \text{otherwise.} \end{cases}$$

Moreover, if c^{++} is found to be uninteresting by IC^{++}, one may still use Nec and Suf to draw useful conclusions regarding h and e by evaluating the interestingness of other forms of c^{++}, such as $c^{+-}=h \rightarrow \neg e$, $c^{-+}=\neg h \rightarrow e$, or $c^{--}=\neg h \rightarrow \neg e$. If $\text{Nec}(d^{++}) \rightarrow \infty$, then $\neg e$ validates h, that is $\neg e \rightarrow h$, $e \vee h$, and $\underline{\neg h \rightarrow e}$ (i.e., c^{-+}). If $1<\text{Nec}(d^{++})<\infty$, then $\neg e$ encourages h, meaning $\neg e$ makes h more plausible, and the greater $\text{Nec}(d^{++})$ is, the more certain is $\neg h \rightarrow e$. Similarly, the smaller $\text{Suf}(d^{++})$ is, the more e discourages h, and the more certain $\underline{h \rightarrow \neg e}$ (i.e., c^{+-}) is. The larger $\text{Suf}(d^{++})$ is, the more e encourages h, and the more certain $\underline{\neg h \rightarrow \neg e}$ (i.e., c^{--}) is. Thus we propose the following additional interestingness measures for characteristic rules:

$$IC^{+-} = \begin{cases} (1 - Suf(d^{++})) \times P(h) & 0 \leq Suf(d^{++}) < 1 \\ 0 & \text{otherwise.} \end{cases}$$

$$IC^{-+} = \begin{cases} (1 - 1/Nec(d^{++})) \times \text{P}(\neg h) & 1 < Nec(d^{++}) < \infty \\ 0 & \text{otherwise.} \end{cases}$$

$$IC^{--} = \begin{cases} (1 - 1/Suf(d^{++})) \times \text{P}(\neg h) & 1 < Suf(d^{++}) < \infty \\ 0 & \text{otherwise.} \end{cases}$$

For example, if $c^{++} = Flu \rightarrow Headache$ is found to be uninteresting by IC^{++}, then IC^{-+} can be used to assess the interestingness of $c^{-+} = \neg Flu \rightarrow Headache$. Since $\text{Nec}(d^{++})$ was already computed for the evaluation of IC^{++}, then the evaluation of IC^{-+} is trivial, yet may provide useful information regarding Flu and Headache. Each of the proposed measures lies in the range $[0, 1]$, with 0 and 1 representing the minimum and maximum possible interestingness, respectively. In practice, the maximum interestingness value is equal to the probability of the most frequent hypothesis or class. So as to avoid the problems that can occur when probabilities evaluate to 0 (e.g., the denominator in an equation evaluating to 0), we use Bayesian probability estimates (Shinghal 1992).

By definition, $\text{Nec}(d^{++}) = 1$ iff $\text{Suf}(d^{++}) = 1$, in which case h and e are independent. One example of such a case is a rule having 0% discriminability and 100% completeness. The Nec and Suf values cause each of the IC measures to return an interestingness value of 0, indicating that any characteristic rule form involving h and e is uninteresting. One may argue, however, that if a characteristic rule is 100% complete, then it should be interesting no matter how well it can discriminate. We, however, adopt the view of Michalski (1983) that the most interesting of characteristic rules for h are also intended to *discriminate* h from $\neg h$. We feel that a characteristic rule is uninteresting if it cannot at all discriminate. This reasoning agrees with rule interestingness principle 1 (Piatetsky-Shapiro 1991). Rather, we feel that what is interesting is the *metafact* that h and e are independent. The four new measures presented here for characteristic rules obey rule interestingness principles 1 to 4 (Piatetsky-Shapiro 1991; Major & Mangano 1993), as well as the proposed principle 5.

Preliminary Results and Analysis

Preliminary results were obtained for characteristic rules mined from a synthetic database of 4000 tuples described by 10 attributes. Rule interestingness was assessed using the proposed IC^{++} measure for characteristic rules, as well as the RI (Piatetsky-Shapiro 1991), J (Smyth & Goodman 1992), and certainty, or 'CE' (Hong & Mao 1991) measures, the last three being popular objective rule interestingness measures from the literature. The measures were also analyzed with respect to principles 1 to 5. The results are summarized below (proofs not shown owing to limited space).

#	$\lvert h \wedge \neg e \rvert$	$\lvert h \wedge e \rvert$	$\lvert \neg h \wedge e \rvert$	$\lvert \neg h \wedge \neg e \rvert$	Nec(d^{++})	Suf(d^{++})	RI(c^{++})	J(c^{++})	CE(c^{++})	IC^{++}	IC^{+-}
1	700	1300	1300	700	1.0000	1.0000	0	0	0.504	0	0
2	0	2000	200	1800	0.0006	9.9552	900.000	0.1289	0.801	0.4997	0
3	200	1800	0	2000	0.1004	1801.0000	900.000	0.0982	0.754	0.4498	0
4	2000	0	1800	200	9.9552	0.0006	-900.000	0.1289	0.000	0	0.4997

Table 2: Characteristic rules evaluated for interestingness.

Only IC^{++} obeys all 5 principles. CE does not obey principle 1, as shown in the scenario representing characteristic rule 1 (Table 2) where h and e are independent. RI does not obey principle 5. For example, from Table 2, rule 2 is as complete (100%) as rule 3 is discriminant, and as discriminant (90%) as rule 3 is complete. Here h and $\neg h$ are equally likely or important. According to principle 5, rule 2 is therefore the more interesting characteristic rule. The IC^{++}, J, and CE measures have ranked the rules accordingly, but RI has not.

The J measure can give misleading results. For example, the J measure finds characteristic rule 4 of Table 2 to be as interesting as rule 2, even though rule 4 fails to identify any of the h tuples, and mistakes 90% of the $\neg h$ tuples for h. Furthermore, the J measure does not obey principle 2, as illustrated by rules 3 and 4 of Table 2.

If IC^{++} finds that a given characteristic rule, c^{++}, is uninteresting, other potentially useful conclusions regarding h and e may be drawn. If h and e are independent (as indicated by a Nec and Suf value of 1), then this knowledge can be reported. Suppose that rule 4 is $Flu \rightarrow \neg Headache$. From the necessity and sufficiency values computed for rule 4 (Table 2), one cannot conclude that Flu and $\neg Headache$ are independent. A user may wish to see if another version of characteristic rule 4 is interesting. The IC^{+-} measure finds that the c^{+-} version of the given rule, i.e., $Flu \rightarrow Headache$ is indeed interesting, with an interestingness value of 0.4997 (Here, IC values close to 0.5 indicate high interestingness). If the necessity and sufficiency values are precomputed, then the application of the remaining IC measures is trivial, yet it can reveal additional useful conclusions regarding the data.

Conclusions

We have proposed IC^{++}, an interestingness measure for characteristic rules. We have also proposed a rule interestingness principle in addition to the four presented by (Piatetsky-Shapiro 1991; Major & Mangano 1993) which, unlike the others, considers the difference between discriminant and characteristic rules. In comparison with the RI, J, and CE measures, IC^{++} is the only one that obeys all five principles. If a given characteristic rule, $h \rightarrow e$, is deemed uninteresting by IC^{++}, three additional measures which we also presented here can be used to derive other useful information regarding h and e. Our future work involves using the measures to guide the discovery of characteristic rules from databases.

Acknowledgements

We thank R.F. Hadley and J. Han for very helpful feedback, as well as the Natural Sciences & Engineering Research Council of Canada for financial support.

References

Agrawal, R.; Imielinski, T.; & Swami, A. 1993. Database mining: A performance perspective. *IEEE Trans. Knowledge & Data Eng.* 5(6):914–925.

Duda, R.O.; Gaschnig, J.; & Hart, P.E. 1981. Model design in the Prospector consultant system for mineral exploration. In B.L. Webber & N.J. Nilsson, eds., *Readings in Artificial Intelligence*, 334–348. Palo Alto, CA: Tioga.

Han, J.; Cai, Y.; & Cercone, N. 1993. Data-Driven Discovery of Quantitative Rules in Relational Databases. *IEEE Trans Knowledge & Data Eng.* 5(1):29–40.

Hong, J. & Mao, C. 1991. Incremental discovery of rules and structure by hierarchical and parallel clustering. In G. Piatetsky-Shapiro & W. J. Frawley, eds., *Knowledge Discovery in Databases*, 177-194. Menlo Park, CA: AAAI/MIT Press.

Major, J.A.; & Mangano, J. 1993. Selecting among rules induced from a hurricane database. In G. Piatetsky-Shapiro, ed., *Proc. AAAI-93 Workshop on Knowledge Discovery in Databases*, 28–44.

Michalski, R.S. 1983. A theory and methodology of inductive learning. *Artificial Intelligence* 20:111–161.

Piatetsky-Shapiro, G. 1991. Discovery, analysis, and presentation of strong rules. In G. Piatetsky-Shapiro & W.J. Frawley, eds., *Knowledge Discovery in Databases*, 229-248. Menlo Park, CA: AAAI/MIT Press.

Piatetsky-Shapiro, G. & Matheus, C.J. 1994. The interestingness of deviations. In *Proc. AAAI-94 Workshop on Knowledge Discovery in Databases*, 25–36.

Shinghal, R. 1992. *Formal Concepts in Artificial Intelligence.* London: Chapman & Hall.

Silberschatz, A. & Tuzhilin, A. 1995. On subjective measures of interestingness in knowledge discovery. In *Proc. 1st International Conf. on Knowledge Discovery in Databases*, 275–281.

Smyth, P. & Goodman, R.M. 1992. An information theoretic approach to rule induction from databases. *IEEE Trans. Knowledge & Data Eng.* 4(4):301–316.

Srikant, R. & Agrawal, R. 1995. Mining generalized association rules. In U. Dayal, P.M.D. Gray, & S. Nishio, eds., *Proc. 21st VLDB Conf.*, 407–419.