Proposed Interestingness Measure for 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

Introduction

Knowledge discovery systems can be used to generate rules describing data from databases. Typically, only a small fraction of the rules generated are of interest. Measures of rule interestingness are hence essential for filtering out useless information. Such measures have been predominantly objective, based on statistics underlying the discovered rules, or patterns. Examples include the J-measure (Smyth & Goodman 1992), rule strength (Piatetsky-Shapiro 1991), and certainty (Hong & Mao 1991). Although these measures help assess the interestingness of discriminant rules, they do not fully serve their purpose when applied to characteristic rules (Kamber & Shinghal 1996). Discriminant rules describe how objects of a class differ from objects of other classes. Such rules take the form $e \rightarrow h$, where e is the evidence (typically a conjunction of attribute-value conditions) and h is the hypothesis (predicting the class of objects satisfying e). Characteristic rules are of the form $h \rightarrow e$. They describe the characteristics common to all objects in a given class, although this constraint may be relaxed slightly in order to deal with real-world noisy data. Both types of rules can be of interest. We propose an interestingness measure for characteristic rules, based on the technical definition of sufficiency (Duda, Gaschnig, & Hart 1981).

Proposed Measure

The interestingness of characteristic rule $r = h \rightarrow e$ may be defined as the product of its utility and goodness (Smyth & Goodman 1992). Let $P(h)$ represent the utility of r, i.e., the probability that r will be used. The goodness of r can be assessed as a function of $Suf(r)$, the sufficiency of r, which measures the influence of h on e. This is defined as $Suf(r) = P(h|e)/P(h|\neg e)$, where the probabilities can be estimated from the given data (Duda, Gaschnig, & Hart 1981). Suf(r) lies in the range $[0, \infty]$. If Suf(r)$\rightarrow\infty$ then h invalidates $\neg e$, meaning $h \rightarrow e$ is certain. If $1<$Suf(r)$<\infty$, then the larger Suf(r) is, the more certain $h \rightarrow e$ is. If $0\leq$Suf(r)≤ 1 then r is invalid. We therefore propose the following interestingness measure for characteristic rule, r:

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

IC(r) lies in the range [0,1] with 0 and 1 representing the minimum and maximum possible interestingness, respectively. Note that IC(r) increases monotonically with Suf(r).

Conclusions

To our knowledge, no interestingness measures exist for characteristic rules. We propose an interestingness measure for such rules based on sufficiency. The measure was applied to order a number of characteristic rules according to decreasing interestingness (not shown here due to limited space). Although further testing is required, the measure's performance was promising, placing the more complete and accurate rules towards the top of the list. Our present work focuses on using the measures to constrain the search space for characteristic rule discovery from databases.

Acknowledgments

We thank R. Hadley for very helpful feedback.

References

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, Tioga, Palo Alto, CA.

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, AAAI/MIT Press, Menlo Park, CA.

Kamber, M. & Shinghal, R. 1996. Evaluating the interestingness of characteristic rules. Forthcoming.

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, AAAI/MIT Press, Menlo Park, CA.

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