SONAR: System for Optimized Numeric Association Rules

Takeshi Fukuda
IBM Tokyo Research Laboratory
fukudat@trl.ibm.co.jp

Yasuhiko Morimoto
IBM Tokyo Research Laboratory
morimoto@trl.ibm.co.jp

Shinichi Morishita
IBM Tokyo Research Laboratory
morisita@trl.ibm.co.jp

Takeshi Tokuyama
IBM Tokyo Research Laboratory
ttoku@trl.ibm.co.jp

Recent progress in technologies for data input have made it easier for finance and retail organizations to collect massive amounts of data and to store them on disk at a low cost. Such organizations are interested in extracting from these huge databases previously unnoticed information that inspires new marketing strategies. In this demonstration, we introduce *SONAR*, a system for mining optimized association rules from databases with numeric data as well as Boolean data.

An example of an association rule is

$$(Balance \in I) \Rightarrow (ServiceX = yes),$$

which implies that bank customers whose balances fall in a range I are likely to use Service X with a certain probability. The above rule is interesting only if the number of customers whose balances are contained in I (called the *support* of I) is sufficient, and also if the probability (called the *confidence ratio*) is much higher than the average probability of the condition being met. SONAR focuses on computing an *optimized support (confidence) range* that maximizes the support (confidence) on the condition that the confidence (support) ratio is at or above a given threshold.

SONAR also generates a two-dimensional version of association rules such as

$$((Age, Balance) \in P) \Rightarrow (ServiceX = yes),$$

which implies that a bank customer whose age and balance fall in a planar region P tends to use Service X with a certain probability. We consider two classes of regions, rectangles and *admissible* (i.e. connected and x-monotone [1]) regions. For each class of regions, an *optimized support region* and an *optimized confidence region* can be defined.

SIGMOD '96 6/96 Montreal, Canada
© 1996 ACM 0-89791-794-4/96/0006...$3.50

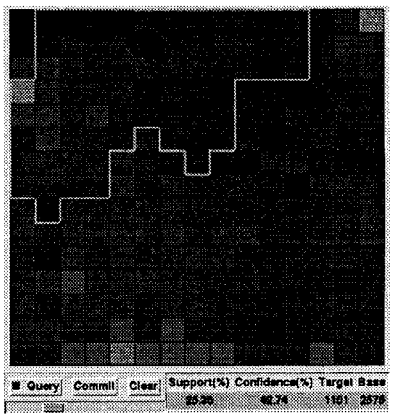

Figure 1: Optimized Region

In companion papers [1, 2], we present linear-time algorithms for computing optimized ranges [1] and efficient algorithms for optimized regions [2]. Tests show that our implementation is fast both in theory and in practice.

One defect of our approach is that the optimized region generated is sometimes hard to describe. However, we can describe the region by visualizing it. For instance, in Figure 1, the region enclosed in thick lines presents the optimized confidence region P in the rule $((Age, Balance) \in P) \Rightarrow (ServiceX = yes)$ for 25% of minimum support threshold. The shape of the region (the roughly triangular region) tells us that customers who use Service X are relatively young people whose balance is high. By moving the slider at the bottom line, we can control the trade-off between the support and the confidence: if the user moves the slider to the left, the support increases, while the confidence decreases.

References

[1] T. Fukuda, Y. Morimoto, S. Morishita, and T. Tokuyama. Mining optimized association rules for numeric attributes. In *Proc. of ACM PODS*, June 1996.

[2] T. Fukuda, Y. Morimoto, S. Morishita, and T. Tokuyama. Data mining using two-dimensional optimized association rules: Scheme, algorithms, and visualization. In *Proc. of ACM SIGMOD*, June 1996.