Mining Quantitative Association Rules & From Association Mining to Correlation Analysis

22 September 2020 06:11 AM

Mining Quantitative Association Rules

"Quantitative association rules are multidimensional association rules in which the **numeric attributes are dynamically discretized** during the mining process so as to satisfy some mining criteria, such as maximizing the confidence or compactness of the rules mined"

, we focus specifically on how to mine quantitative association rules having **two quantitative attributes on the left-hand side of the rule** and **one categorical attribute on the right-hand side** of the rule

$$Aquan1 \wedge Aquan2 \Rightarrow Acat$$

here Aquan1 and Aquan2 are tests on quantitative attribute intervals (where the intervals are dynamically determined), and Acat tests a categorical attribute from the task-relevant data

Association Rule Clustering System

The following steps are involved in ARCS

Binning: Quantitative attributes can have a very wide range of values defining their domain. Just think about how big a 2-D grid would be if we plotted age and income as axes, where each possible value of age was assigned a unique position on one axis, and similarly, each possible value of income was assigned a unique position on the other axis.

To keep grids down to a manageable size, we instead partition the ranges of quantitative attributes into intervals.

The partitioning process is referred to as binning, that is, where the intervals are considered “bins.” Three common binning strategies area as follows.

Equal-width binning, where the interval size of each bin is the same
Equal-frequency binning, where each bin has approximately the same number of tuples assigned to it, *Clustering-based binning,*
where clustering is performed on the quantitative attribute to group neighboring points (judged based on various distance measures) into the same bin

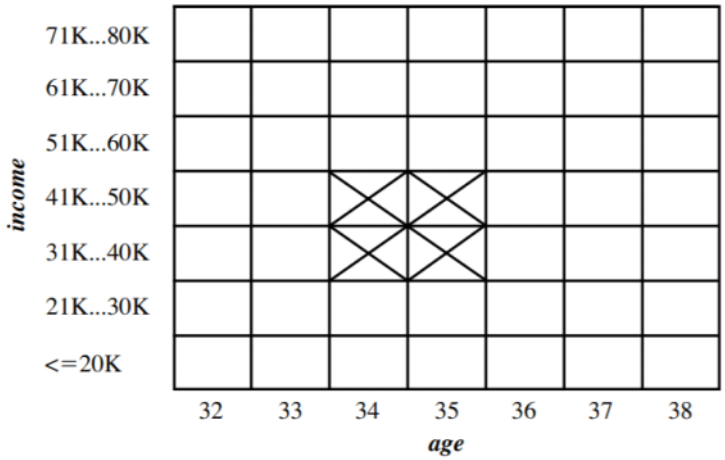
Finding frequent predicate sets: Once the 2-D array containing the count distribution for each category is set up, it can be scanned to find the frequent predicate sets (those satisfying minimum support) that also satisfy minimum confidence

Clustering the association rules: The strong association rules obtained in the previous step are then mapped to a 2-D grid. Following figure shows a 2-D grid for 2-D quantitative association rules predicting the condition buys(X, “HDTV”) on the rule right-hand side, given the quantitative attributes age and income. The four Xs correspond to the rules

age(X, 34)∧income(X, “31K...40K”)⇒buys(X, “HDTV”)
age(X, 35)∧income(X, “31K...40K”)⇒buys(X, “HDTV”)
age(X, 34)∧income(X, “41K...50K”)⇒buys(X, “HDTV”)
age(X, 35)∧income(X, “41K...50K”)⇒buys(X, “HDTV”)

Example:

age(X, “34...35”)∧income(X, “31K...50K”)⇒buys(X, “HDTV”)



A 2-D grid for tuples representing customers who purchase high-definition TVs.

$$age(X, \text{“34...35”}) \wedge income(X, \text{“31}K\text{...50}K\text{”}) \Rightarrow buys(X, \text{“}HDTV\text{”})$$

From Association Mining to Correlation Analysis

i. Strong Rules Are Not Necessarily Interesting: An Example

Ultimately, only the user can judge if a given rule is interesting, and this judgment, being **subjective**, may differ from one user to another. However, **objective interestingness measures,** based on the statistics “behind” the data.

Example:

Of the 10,000 transactions analyzed, the data show that 6,000 of the customer transactions included computer games, while 7,500 included videos, and 4,000 included both computer games and videos. cSuppose that a data mining program for discovering association rules is run on the data, using a minimum support of, say, **30% and a minimum confidence of 60%.**

buys(X, “computer games”)⇒buys(X, “videos”) [support = 40%, confidence = 66%].

It is a strong association rule and would therefore be reported, since its support value of 4,000/ 10,000 = 40% and confidence value of 4,000 / 6,000 = 66% satisfy the minimum support and minimum confidence thresholds, respectively

In fact, computer games and videos are negatively associated because the purchase of one of these items actually decreases the likelihood of purchasing the other

Ii. From Association Analysis to Correlation Analysis

A correlation measure can be used to augment the support-confidence framework for association rules.

This leads to correlation rules of the form A ⇒B [support, confidence. correlation]

Lift is a simple correlation measure that is given as follows. The occurrence of itemset A is independent of the occurrence of itemset B if P(A ∪ B) = P(A)P(B); otherwise, itemsets A and B are dependent and correlated as events.

lift(A, B) = P(AUB) / P(A)P(B) .

A 2 × 2 contingency table summarizing the transactions with respect to game and video purchases.

	game	$\overline{game}$	Σ_{row}
video	4,000	3,500	7,500
$\overline{video}$	2,000	500	2,500
Σ_{col}	6,000	4,000	10,000

The above contingency table, now shown with the expected values.

	game	$\overline{game}$	Σ_{row}
video	4,000 (4,500)	3,500 (3,000)	7,500
$\overline{video}$	2,000 (1,500)	500 (1,000)	2,500
Σ_{col}	6,000	4,000	10,000

Example Problem:

Correlation analysis using χ^2. To compute the correlation using χ^2 analysis, we need the observed value and expected value (displayed in parenthesis) for each slot of the contingency table, as shown in Table 5.8. From the table, we can compute the χ^2 value as follows:

$$\chi^2 = \Sigma\frac{(observed - expected)^2}{expected} = \frac{(4{,}000 - 4{,}500)^2}{4{,}500} + \frac{(3{,}500 - 3{,}000)^2}{3{,}000} + \frac{(2{,}000 - 1{,}500)^2}{1{,}500} + \frac{(500 - 1{,}000)^2}{1{,}000} = 555.6.$$