Quality of Association Rules by Chi-Squared Test

Wen-Chi Hou
Southern Illinois University, USA

Maryann Dorn
Southern Illinois University, USA

INTRODUCTION

Mining market basket data (Agrawal et al. 1993, Agrawal et al. 1994) has received a great deal of attention in the recent past, partly due to its utility and partly due to the research challenges it presents. Market basket data typically consists of store items purchased on a per-transaction basis, but it may also consist of items bought by a customer over a period of time. The goal is to discover buying patterns, such as two or more items that are often bought together. Such finding could aid in marketing promotions and customer relationship management. Association rules reflect a fundamental class of patterns that exist in the data. Consequently, mining association rules in market basket data has become one of the most important problems in data mining.

Agrawal et al. (Agrawal, et al. 1993, Agrawal et al. 1994) have provided the initial foundation for this research problem. Since then, there has been considerable amount of work (Bayardo et al. 1999, Bayardo et al. 1999, Brin et al. 1997, Han et al. 2000, Park et al. 1995, Srikant et al. 1995, Srikant et al. 1997, Zaki et al. 1997, etc.) in developing faster algorithms to find association rules. While these algorithms may be different in their efficiency, they all use minsup (minimum support) and minconf (minimum confidence) as the criteria to determine the validity of the rules due to their simplicity and natural appeals. Few researchers (Brin et al. 1997, Aumann et al. 1999, Elder, 1999, Tan et al. 2002) have suspected the sufficiency of these criteria. On the other hand, Chi-squared test has been used widely in statistics related fields for independence test. In this research, we shall examine the rules derived based on the support-confidence framework (Agrawal et al. 1993, Agrawal et al. 1994) statistically by conducting Chi-squared tests. Our experimental results show that a surprising 30% of the rules fulfilling the minsup and minconf criteria are indeed insignificant statistically.

BACKGROUND

The task of mining association rules is first to find all itemsets that are above a given minimum support ratio (minsup). Such itemsets are called large or frequent itemsets. Then, association rules are derived based on these frequent itemsets. For example, both {A, B, C, D} and {A, B} are frequent itemsets. The association rule, AB=>CD, is derived if at least c% of the transactions that contain AB also contain CD, where c% is a pre-specified constant called minimum confidence (minconf).

Support-Confidence Framework

We use the example in (Brin et al. 1997) to illustrate the support-confidence framework (Agrawal, et al. 1993, Agrawal et al. 1994). Suppose there are totally 100 transactions. 25 transactions buy tea and among them, 20 transactions also buy coffee. Based on the support-confidence framework, the rule 'tea=> coffee' has a support of 20% (20 / 100) and a confidence of 80% (20 / 25). Suppose minsup =5% and minconf =60%. Then, the rule is validated by the framework.

Chi-Squared Test for Independence

Chi-squared (χ^2) test is a non-parametric statistical method that can be used to test independence among attributes. Compared to the support-confidence framework, it uses more information, such as the numbers of transactions that buy tea but not coffee, buy coffee but not tea, and buy neither coffee nor tea, to determine the independence of attributes.

Table 1. 2 by 2 contingency table

	coffee	no_coffee	∑ row rorow
tea	20	5	25
no_tea	70	5	75
∑ col	90	10	100

In Table 1, we show a 2 by 2 contingency table (Glass, 1984, Gokhale, 1978), which contains more information for the independence test. The chi-square test result indicates that tea and coffee are independent variables since $\chi^2 = 3.703703$ (with degree of freedom = 1) is non-significant at 95% confidence level. In other words, tea is not a determining factor whether people will buy coffee or not, contradicting the rule derived by the support-confidence framework.

Chi-squared test is reliable under a fairly permissive set of assumptions. As a rule of thumb, Chi-squared test is recommend (Glass, 1984, Mason, et al. 1998) only if (1) all cells in the contingency table have expected value greater than 1, and (2) at least 80% of the cells in the contingency table have expected value greater than 5. For the large tables (more than four cells), the usual alternative is to combine or collapse cells (Glass, 1984, Han, et al. 2000, Mason, et al. 1998) when the cells have low values. The potential advantages of the χ^2 statistics (Brin, et al. 1997) over the commonly used support-confidence framework are:

1. The use of the Chi-squared significance test for independence is more solidly grounded in statistical theory. In particular, there is no need to choose ad-hoc values for support and confidence.
2. The Chi-squared statistic simultaneously and uniformly takes into account all possible combinations of the presence and absence of the various attributes being examined as a group.
3. Chi-squared test at a given significance level is upward closed. In other words, if an i-itemset is correlated, all its supersets are also correlated.

MAIN FOCUS

Experimental Design

Four synthetic data sets are generated using the IBM/ Quest data generator (Bayardo et al. 1999), which has been widely used for evaluating association rule mining algorithms. The data sets generated are then fed into CBA/DBII data mining system (Liu et al. 1999) to generate association rules. Finally, Chi-squares tests are conducted on the association rules generated.

The synthetic transactions are to mimic the transactions generated in the retailing environment. A realistic model will address the observation that people tend to buy a number of items together. Thus, transaction sizes are typically clustered around a mean and a few transactions have many items.

In order to model the phenomenon that frequent itemsets often have common items, some fraction of items in subsequent itemsets are chosen from the previous itemset generated. It uses an exponentially distributed random variable (Bayardo et al. 1999) with the mean equal to a given correlation level to decide this fraction for each itemset. The remaining items are picked at random. In this study, the correlation level was set to 0.25. Bayardo et al. (Bayardo et al. 1999) ran some experiments with correlation levels set to 0.5 and 0.75 but did not find much difference in the results.

Each itemset in the database has a weight associated with it, which corresponds to the probability that this itemset will be picked. This weight is picked from an exponential distribution with unit mean, and is then normalized so that the sum of the weights for all the itemsets is 1.

Four synthetic datasets were generated with slightly different parameters, such as the number of transactions, number of items, and average confidence for rules (see Table 2). The other common factors are 'average transaction length (5),' 'number of patterns (100),' 'average length of pattern (4),' and 'correlation between consecutive patterns (0.25).'

Once the data sets are generated, CBA system (Liu et al. 1999, Liu et al. 1999) is used to generate association rules that satisfy the given minsup and minconf.