A note on "Beyond Market Baskets: Generalizing Association Rules to Correlations"

Khalil M. Ahmed
CS Department
Alexandria University
Alex., Egypt 21544
drkhalil@usa.net

Nagwa M. El-Makky
CS Department
Alexandria University
Alex., Egypt 21544
nagwa@alex.eun.eg

Yousry Taha
CS Department
Alexandria University
Alex., Egypt 21544
yousry_taha@hotmail.com

ABSTRACT

In their paper [1], S. Brin, R. Matwani and C. Silverstien discussed measuring significance of (generalized) association rules via the support and the chi-squared test for correlation. They provided some illustrative examples and pointed that the chi-squared test needs to be agumented by a measure of interest that they also suggested.

This paper presents a further elaboration and extension of their discussion. As suggested by Brin et al, the chi-squared test succeeds in measuring the cell dependencies in a 2x2 contingency table. However, it can be misleading in cases of bigger contingency tables. We will give some illustrative examples based on those presented in [1]. We will also propose a more appropriate reliability measure of association rules.

1. THE SUPPORT-CONFIDENCE FRAMEWORK FOR ASSOCIATION RULES

For an association rule $t \Rightarrow c$ the support and confidence are defined as:

$$Support = P[t \wedge c]$$

$$Confidence = \frac{P[t \wedge c]}{P[t]}$$

The confidence is the conditional probability of c given t; $P[c \backslash t]$. This conditional probability is equal to the unconditional probability $P[c]$ iff t and c are independent. Accordingly the measure of confidence for association rules can be misleading since the confidence can be high simply because $P[c]$ is high while t and c are highly independent. Generalizing associations, S. Brain et al [1] suggested other measures like the chi-squared test of correlation and interest. The Chi-squared test method is quite appropriate for 2x2 contingency tables. However, we will show that it can fail in case of bigger contingency tables.

What is needed is a measure of dependency or correlation. A more appropriate measure of dependency is proposed and presented in section 4 of this note.

2. CHI-SQUARE TEST OF CORRELATION

We start by presenting the definition of the Chi-Square test for general RxC contingency tables [3; 4]. Suppose we are observing two characteristics A and B where A can only take R distinct values $a_1, a_2, \ldots, a_R$ and B can take only C distinct values $b_1, b_2, \ldots, b_c$. The variables A and B can be nominal or ordinal variables.

The variables A and B are independent if and only if $Pr[A = a_i, B = b_j] = Pr[A = a_i].Pr[B = b_j.]$

Let N_{ij} be the number of observations for the i^{th} row and j^{th} column.

$N_{i.} = \sum_j N_{ij}, N_{.j} = \sum_i N_{ij}, N = \sum_i \sum_j N_{ij}$

and let p_{ij} probability that an observation falls in category a_i and category b_j.

$p_{i.}$ probability that an observation falls in category a_i.

$p_{.j}$ probability that an observation falls in category b_j.

The estimates $\hat{p_{i.}}$ and $\hat{p_{.j}}$ of $p_{i.}$, $p_{.j}$ are given by $\hat{p_{i.}} = N_{i.}/N, \hat{p_{.j}} = N_{.j}/N$.

Now if A and B are independent, then $p_{ij} = p_{i.}.p_{.j}$ and the expected number of observations in category (a_i, b_j) is $Np_{i.}.p_{.j}$. Thus the estimate of the expected number of observations in the (i,j) cell is given by

$$E_{ij} = N_{i.}.N_{.j}/N.$$

In order to test the hypothesis of independence of A and B, we simply compare the observed numbers with the expected numbers by computing:

$$\chi^2 = \sum_i \sum_j \frac{(N_{ij} - E_{ij})^2}{E_{ij}}$$

Under the hypothesis of independence, the quantity χ^2 has a chi-square distribution with $(R-1)(C-1)$ degrees of freedom. Clearly, if χ^2 is large, it indicates that the observed and expected numbers of observations differ greatly and the hypothesis of independence should be rejected. More formally, the chi-square test calls for rejecting the hypothesis of independence at $(1-\alpha)$ level if

$$\chi^2 > \chi^2_{(R-1)(C-1)}; 1-\alpha$$

where $\chi^2_{(R-1)(C-1)}; 1-\alpha$ is the $1-\alpha$ percentile of the chi-square distribution with $(R-1)(C-1)$ degrees of freedom.

The chi-squared test of correlation applies to the entire contingency table, however, for association rules we need a single cell dependence test. For a 2x2 contingency table, these two tests are equivalent simply because if the events A and B are independent then A and $\overline{B}$, $\overline{A}$ and B, and $\overline{A}$ and

$\overline{B}$ are also independent and the chi-squared test result will support the independence hypothesis.
This is not the case in bigger RxC contingency tables with R and\or C greater than 2. The chi-squared test statistic can be significantly high and results in rejecting the independence hypothesis while some of the cells show independence. This fact is illustrated by the following two examples based on examples 1,4 presented in [1].

EXAMPLE 1. *This example is based on example 1 of [1] with the following contingency table:*

	c	$\overline{c}$	Σrow
t	*20*	*5*	*25*
$\overline{t}$	*70*	*5*	*75*
Σcol	*90*	*10*	*100*

where t and c represent tea and coffee purchases respectively.

Now let us split the set c into two mutually exclusive (and exhaustive) sets c_1 and c_2 represeting coffee purchases of brands 1 and 2 respectively.
A resulting 2x3 contingency table would be:

	c_1	c_2	$\overline{c}$	Σrow
t	5	15	5	25
$\overline{t}$	50	10	15	75
Σcol	55	25	20	100

The chi-squared values is 23.8 which is higher than the cutoff value (5.99 at 95% significance level with 2 d.f.) and hence the assumption of independence is rejected.
Now consider the events $\overline{t}, \overline{c}$;

$$P(\overline{t} \wedge \overline{c}) = 0.15$$

$$P(\overline{t}).P(\overline{c}) = (0.75).(0.2) = 0.15$$

and hence $\overline{t}\&\overline{c}$ are independent.
On the other hand the confidence of the association rule $\overline{c} \Rightarrow \overline{t}$ is given by:

$$\frac{P(\overline{t} \wedge \overline{c})}{P(\overline{c})} = \frac{15}{20} = 0.75$$

which is very high in spite of the independence of $\overline{c}, \overline{t}$ simply because $P(\overline{t})$ is high.

EXAMPLE 2. *This example is based on example 4 of [1] focused on testing the relationship between military service and age. The census data resulted in the following* 2x2 *contengincy table:*

	i_2	$\overline{i}_2$	Σrow
i_7	*17918*	*911*	*18829*
$\overline{i}_7$	*9111*	*2430*	*11541*
Σcol	*27029*	*3341*	*30370*

where the event $i_2 \equiv$ never served in the military and the event $i_7 \equiv$ age < 40 yrs.

Now let us splite i_7 into two mutually exclusive events:

$$\begin{array}{ll} i_7^{'} & age < 20 \ \ yrs. \\ i_7^{''} & 20 \leq age < 40 \end{array}$$

The resulting 3x2 contingency table can be given by:

	i_2	$\overline{i}_2$	Σrow
$i_7^{'}$	5163	304	5467
$i_7^{''}$	12755	607	13362
$\overline{i}_7$	9111	2430	11541
Σcol	27029	3341	30370

The chi-squared value is 1926 which is significant at the 95% level and hence the assumption of independence is rejected.
Now consider the events $i_7^{'}$, i_2

$$P(i_7^{'} \wedge i_2) = \frac{5163}{30370} = 17\%$$

$$and \quad P(i_7^{'}).P(i_2) = \left(\frac{5467}{30370}\right)\left(\frac{27029}{30370}\right) = 16.02\%$$

These close values indicate a high level of independence of $i_7^{'}$, i_2
Again the confidence of the association rule $i_7^{'} \Rightarrow i_2$ will be 94% which is very high simply because $P(i_2) = 89\%$ is high.
From the above two examples we conclude that both chi-squared test for contingency tables bigger than 2x2 and the confidence rule can lead to wrong conclusions concerning association rules. However, high order contingency tables can be folded down to 2x2 tables if less detailed association analysis is acceptable.

3. INTEREST AND CONVICTION MEASURES OF ASSOCIATION RULES

Brin. S, et al [1] proposed a measure of interest given by:

$$I = \frac{P(A \wedge B)}{P(A).P(B)}$$

The further this from 1 the more the dependence. This measure has the serious drawback of being *completely symmetric* that it gives the same value for the association rules $A \Rightarrow B$ and $B \Rightarrow A$.
To fill this gap, S. Brin et al [2] defined *conviction* as:

$$\frac{P(A).P(\overline{B})}{P(A \wedge \overline{B})}$$

Again this measure is *symmetric in the the complement.* It gives the same value for the association rules $A \Rightarrow B$ and $\overline{B} \Rightarrow \overline{A}$. which is a serious drawback.

4. PROPOSED MEASURE OF RELIABILITY OF ASSOCIATION RULES

We here propose a measure of the reliability of the association rule $A \Rightarrow B$ given by:
$R(A \Rightarrow B) = \mid \frac{P(A \wedge B)}{P(A)} - P(B) \mid$
which does not have the same value for $B \Rightarrow A$. The bigger R the stronger the association rule.
The values of I and R for example 2 of the preceeding section are as follows:

I	1.06
$R(i_7^{'} \Rightarrow i_2)$	0.05
$R(i_2 \Rightarrow i_7^{'})$	0.02

It is clear that $R(A \Rightarrow B)$ is the difference between the conditional probability of B given A and the unconditional probability of B. It measures the effect of the available information about A on the probability of B. The greater this (absolute) difference the stronger the association $(A \Rightarrow B)$. Positive correlation is indicated if $\frac{P(A \wedge B)}{P(A)} - P(B) \geq 0$ while negative correlation is indicated otherwise. Moreover, this proposed measure is a probability and can be subjected to classical analysis, e.g. confidence intervals and tests of hypothesis.

5. REFERENCES

[1] S. Brin, R. Motwani, and C. Silverstein. Beyond Market Baskets: Generalizing Association Rules to Correlations. In *The Proceedings of SIGMOD*, pages 265–276, AZ,USA, 1997.

[2] S. Brin, R. Motwani, J. D. Ullman, and S. Tsur. Dynamic Itemset Counting and Implication Rules for Market Basket Data. In *The Proceedings of SIGMOD*, pages 255–264, AZ,USA, 1997.

[3] W. C. Guenther. *Concepts of Statistical Inference*. Mc Graw-Hill Inc., N.Y., second edition, 1973.

[4] R. F. Woolson. *Statistical Methods for the Analysis of Biomedical Data*. John Wiley & Sons, Inc., N.Y., 1987.