Generalizing Association Rules Beyond Market Baskets

Ronen Mansuri Pavel Berengoltz

10 December 2002

Lecture Overview

In this lecture we review the concept of Association Rules. First, we define Market Basket and also introduce the standard definitions about association rules and their drawbacks. Then, we introduce the notion of dependence rules that identify statistical dependence in both the presence or absence of items in itemsets. We discuss about measuring significance of dependence via the chi-squared test for independence from classical statistics. Finally, we present an efficient algorithm for discovering dependence rules.

Definition of Market Basket

Let $I = \{i_1, ..., i_k\}$ be a set of k elements, called **items**. Then, **basket data** $B = \{b_1, ..., b_n\}$ is any collection of n subsets of I, and each subset $b_i \subseteq I$ is called a **basket** of items.

These are the standard definitions about Association Rules (fig.1):
We say there is an **Association Rule** $A \Longrightarrow B$ if:

1. A and B occurs together in at least $s\%$ of the n baskets (**Support**).
2. All the baskets containing A, at least $c\%$ also contain B (**Confidence**).

$$Support = \frac{\text{transactions that contain every item in } A \text{ and } B}{\text{all transactions}}$$

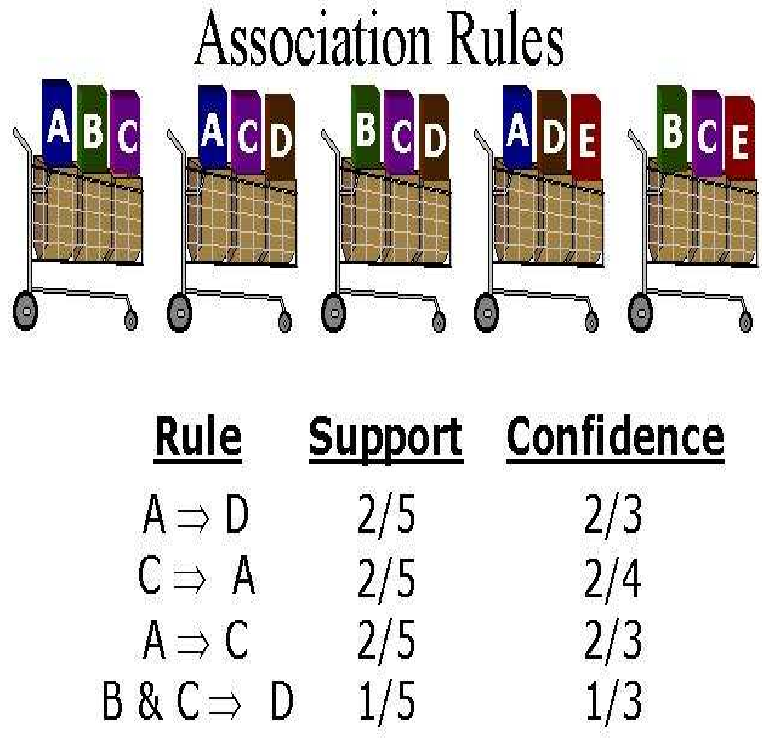

Figure 1: Association Rules

$$Confidence = \frac{\text{transactions that contain every item in } A \text{ and } B}{\text{transactions that contain the item in } A}$$

The support is a reflexive rule, contrary to the confidence that depends on the order of the variables. There are some drawbacks of Association Rules: they do not support negative implications and they don't work well when correlation is the appropriate measure.

Dependence and Correlation

Definition

1. Two events x and y are **independent** if $P[x \wedge y] = P[x]P[y]$.

2. Two variables A and B are **independent** (also called **uncorrelated**) if $P[A = v_a \wedge B = v_b] = P[A = v_a] \cdot P[B = v_b]$ for all possible values $v_a, v_b \in \{\texttt{TRUE,FALSE}\}$.

3. Events that are not independent are **dependent**.

4. Variables that are not independent (uncorrelated) are **dependent** (**correlated**).

If two variables A and B are independent (uncorrelated) then two events a and b (as well as $\bar{a}$ and b, etc.) are independent too, but the converse is not true in general.

Theorem

If a set of variables I is dependent (correlated), so is every superset of I.

Proof:

If a set of variables is dependent (correlated), then some set of events associated with the variables must also be dependent. Suppose, without loss of generality, that the variables are A and B and the events a and b are dependent. Assume, then, that some superset ABC is independent. Then all events associated with ABC must be independent. In particular, we must have $p(abc) = p(a)p(b)p(c)$ and $p(ab\bar{c}) = p(a)p(b)p(\bar{c})$. Then $p(ab) = p(abc) + p(ab\bar{c}) = p(a)p(b)p(c) + p(a)p(b)p(\bar{c}) = p(a)p(b)$. Thus, a and b are independent after all, contrary to our hypothesis, so ABC must actually be dependent (correlated).

Definition

A property ρ is said to be **upward-closed** if for every set with property ρ, all its supersets also have property ρ. Similarly, property ρ is said to be **downward-closed** if for every set with property ρ, all its subsets also have property ρ.

It is clear that the upward- and downward-closure are two faces of the same coin. In particular, if a property ρ is upward closed, then not having the property is downward-closed.

Definition

If an itemset I is dependent but no subset of I is dependent, we say I is minimally dependent.

The χ^2 (Chi-squared) Test for Independence

The χ^2 Test shall be used for testing the hypothesis that different items or attributes in frequency tables are independent.
The hypothesis tested is:

H_0=The variables are independent
H_1=Otherwise

The chi-squared statistic attempts to measure the degree of independence between different items. It is comparing the observed patterns of occurrence with the expected pattern of occurrence under the assumption of complete independence and a normal distribution of each attribute.
Let X be a Bernoulli random variable that denotes the number of successes in N independent trials where the probability of success in any given given trial is p. The expected number of successes is Np and the variance is $Np(1-p)$.

$$\chi = \frac{X - Np}{\sqrt{Np(1-p)}}$$

The above χ value is a standard normal variable. If we choose U variables (multinomial case) the distribution is chi-squared with $|U|$-1 degrees of freedom:

$$\chi^2 = \sum_{r \in U} \frac{(X_r - E(X_r))^2}{E(X_r)}$$

Let $R = \{i_1, \overline{i_1}\} \times \cdots \times \{i_k, \overline{i_k}\}$ be the Cartesian of the event sets corresponding to the presence or absence of items in a basket. An element $r = r1 \cdots r_k \in R$ is a single basket value, or an instantiation of all the variables. Each value of r denotes a cell (in a k-dimensional table - **Contingency Table**). Let $O(r)$ denote the number of baskets falling into cell r. To test whether a given a given cell is dependent, we must determine if the actual count in cell r differs sufficiently from the expectation.
In the chi-squared test, the expected count of an event is calculated under the assumption of independence. For a single event, we use the maximum likelihood estimators $E(i_j) = O_n(i_j)$ and $E(\overline{i_j}) = n - O_n(i_j)$. For

sets of events, we use the independence assumption to calculate $E(r) = n \times E(r_1)/n \times \cdots \times E(r_k)/n$. Then the chi-squared statistic is defined as follows:

$$\chi^2 = \sum_{r \in R} \frac{(O(r) - E(r))^2}{E(r)}$$

$$\chi^2 = \sum_{i=1}^{R} \sum_{j=1}^{K} \frac{(Ob_i j - Ex_i j)^2}{Ex_i j} \sim \chi^2_{(R-1)(K-1)}$$

R **Number of Rows**
K **Number of Columns**

The χ^2 statistic as defined will specify whether all k items are k-way independent. In order to determine whether some subset of items is dependent, for instance i_1,i_2, and i_7, we merely restrict the range of r to $(i_1, \overline{i_1}) \times (i_2, \overline{i_2}) \times (i_7, \overline{i_7})$.
No matter how r is restricted, the χ^2 test works as follows: Calculate the value of the chi-squared statistic. Corresponding to this value and a degrees of freedom count (always 1, for the Boolean variables) is a p value. This value, between 0 and 1, indicates the probability of witnessing the observed counts were the variables really independent. If this value is low (say, less than 0.05), we reject the hypothesis that the variables are independent. We say a set of items is dependent at **significance level** α if the p value of the set is at most $1 - \alpha$.

To put χ^2 values in perspective, for a p value of 0.05 with one degree of freedom, the χ^2 cutoff value is 3.84. Thus, any set of items with χ^2 value of 3.84 or more is significant at the 1-0.05=95% confidence level.
Example:

	i_8	$\overline{i_9}$	row-sum
i_9	1	2	3
$\overline{i_9}$	4	2	6
col-sum	5	4	9

The chi-squared value is:

$$\frac{(1 - 3 \times 5/9)^2}{3 \times 5/9} + \frac{(2 - 3 \times 4/9)^2}{3 \times 4/9} + \frac{(4 - 6 \times 5/9)^2}{6 \times 5/9} + \frac{(2 - 6 \times 4/9)^2}{6 \times 4/9}$$

=0.267+0.333+0.133+0.167=0.9
Since 0.9 is less than 3.84, we do not reject the independence hypothesis at the 95% confidence level.

Algorithm for Dependence Rules

Lets extend the definition of support:

Definition
A set of items S has **contingency table support (CT-support)** s at the $p\%$ level if at least $p\%$ of the cells in the contingency table for S have value s.

Theorem
The property of having CT-support s at the $p\%$ level is a downward-closed property.

Proof:
Suppose a set of items S has CT-support s at the $p\%$ level. Consider, without loss of generality, a subset $T = S\backslash\{i\}$. Then each cell of T has a value equal to the sum of two cells in S. In particular, cell I of T is equal to the sum of cells $I \cup \{i\}$ and $I \cup \{\bar{i}\}$ in S. Since s is an absolute number, if either of the two cells in S has support s, so will the cell in T. In terms of counting supported cells, the worst case is if both $I \cup \{i\}$ and $I \cup \{\bar{i}\}$ are supported. In this case, there are two supported cells in S corresponding to a single supported cell in T, causing the number of supported cells in T to be half that of S. But the number of total cells in T is half that of S, so the percent of supported cells cannot decrease.

Definition
We say that an itemset is significant if it is CT-supported and minimally dependent (correlated).

Algorithm

Input: A chi-squared significance level $\alpha = 95\%$, support s, support fraction

$p > 0.25$. Basket data B.
Output: A collection of minimal dependent (correlated) items, from B - `SIG`.

1. Initialize `CAND` $\leftarrow \emptyset$, `SIG` $\leftarrow \emptyset$, `NOTSIG` $\leftarrow \emptyset$.
2. **For** each item $i \in I$, **do** count $O(i)$. We can use these values to calculate any necessary expected values.
3. **For** each pair of items $i_a, i_b \in I$ such that $O(i_a) > s$ and $O(i_b) > s$, **do** add i_a, i_b to `CAND`.
4. `NOTSIG` $\leftarrow \emptyset$.
5. **If** `CAND` is empty, **then** return `SIG` and terminate.
6. **For** each itemset in `CAND`, **do** construct the contingency table for the itemset. **If** less than $p\%$ of the cells have count s, **then goto** Step 8.
7. **If** the χ^2 value for the contingency table is at least χ^2_a(constant) $\chi^2 \geq \chi^2_a$, **then** add the itemset to `SIG`, **else** add the itemset to `NOTSIG`.
8. **Continue** with the next itemset in `CAND`. **If** there are no more itemsets in `CAND`, **then** set `CAND` to be the collection of all sets S such that every subset of size $|S| - 1$ of S is in `NOTSIG`. **Goto** Step 4.

Theorem

An itemset at level $i + 1$ can be significant only if all its subsets at level i have CT-support and none of its subsets at level i are dependent (correlated).

Proof:

If some subset of I fails to have CT-support, then I also must fail to have CT-support, since CT-support is downward closed. If some subset of I is dependent, then I cannot be minimally dependent by definition.

Algorithm Description

For level $i + 1$, all we need is a list of the CT-supported but independent (uncorrelated) itemsets from level i. This list is held in `NOTSIG`. The list `SIG`, which holds the CT-supported and dependent itemsets, is the output set of interest.

`CAND` list build candidate itemsets for level $i+1$ from `NOTSIG` list at level i. Let S be a set of size $i+1$ for which every subset of size i is in `NOTSIg`. Then S is not ruled out by either CT-support pruning or significance pruning and is added to `CAND`. Once `CAND` has been constructed, we are done processing itemsets at level i. To start level $i+1$, we examine each set $S \in$ `CAND` to see if it actually does have the necessary CT-support. If so, we add it to either `SIG` or `NOTSIG` for level $i+1$, depending on its χ^2 value.

Implementation analysis

The running time of the Algorithm for level i is:

$$O(n|\mathtt{CAND}|min\{n, 2^i\} + i|\mathtt{NOTSIG}|^2).$$

In worst case the algorithm perform $2^{|n|}$ steps (all possible itemsets).

Possible suggestions to improve performance:

1. "Random Walker" is good in case when there are few independent (uncorrelated) itemsets.
2. Looking for maximally independent itemset (go from top of the graph to bottom).

References:

C. Silverstein, S. Brin and R. Motwani, **Beyond Market Baskets: Generalizing Association Rules**, Department of Computer Science, Stanford University CA. ,December 1997.