中國統計學報
第 42 卷 第 2 期
九十三年六月
209-221 頁

Effect of Simpson's Paradox on Market Basket Analysis

Heather Y. Ma and Dennis K.J. Lin

Department of Statistics and Department of

Supply Chain & Information Systems

The Pennsylvania State University, University Park, PA 16802 U.S.A.

ABSTRACT

Association rules discovery is an important database mining algorithm that finds interesting association or correlation relationships among a set of items. One of the well-studied problems in data mining is pruning for association rules in Market Basket Analysis. In this paper, we address the effect of Simpson's Paradox on the decision maker in Market Basket Analysis and proposed a method, called common improvement, to handle Simpson's Paradox in selecting association rules. The proposed method is suited for detecting association rules that are likely to be ignored from the existing methods.

Key words and phrases: Association Rule, confidence, improvement, Odds Ratio Test, support.

AMS 2000 subject classifications: Primary 62E05; secondary 60G55.

1. Introduction

We live in the Age of Information. The importance of data collection, which reflects scientific endeavors in an effort to achieve competitive advantage, is now widely recognized. Databases today can be several terabytes in size. Within these masses of data lies hidden information that can be of great strategic importance. The problem is how to extract the information we need from these masses of data.

One process to address this concern is data mining - which uses sophisticated statistical analysis and modeling techniques to uncover patterns and relationships hidden in organizational databases. Typically, a data-mining analyst is presented with a problem to solve and an existing database that is relevant to the problem. The analyst strives to discover patterns in the data that can be used to solve the problem. For example, a grocery store may be troubled by a low percentage of repeat customers in its customer base. By analyzing products customers bought, a data mining algorithm might uncover patterns that can be used to predict which products customers are likely to purchase in the future, such that these customers can be enticed to shopping there again.

Data mining is the process of discovering advantageous patterns and relationships in data. The patterns here refer to a parsimonious statement about a probability distribution (George, 1997). Pattern definition restricts a pattern to be a statement about a probability distribution. An important role of the analysts in the data mining process is to turn the resulting patterns into a plan of action, and to test this plan to be sure that is gives some advantage

In essence, data mining attempts to discover two main patterns: predictive and descriptive patterns. Predictive patterns are built to solve a specific problem of predicting one or more attributes in a database from the rest of the data with known results. For example, use the payment history of loan recipients to help identify people who are likely to default on loans. Predictive patterns are not always used to foretell the future-the important characteristic is that they make an educated guess about the value of an unknown attribute given the values of other known attributes. The most common tools for finding predictive patterns include regression analysis and classification analysis. On the contrary, descriptive patterns present interesting patterns in existing

data to guide the decision maker rather than to solve a specific problem. Descriptive patterns are harder to evaluate than predictive patterns, because their real value lies in whether they suggest any actions to the decision maker, and how effective those actions are. The most common tools for finding descriptive patterns include Clustering and Association Rules. It is association rule that we are particularly interested in this study.

This paper is organized in five sections. Followed the general introduction of data mining and association rules in this section, one of major association rules tools, Market Basket Analysis (MBA) is described in section 2. In section 3, we describe Simpson's Paradox and discuss its impact to MBA. It is shown that the results from MBA may be misleading when the Simpson Paradox exists. In Section 4, a solution to deal with the Simpson's paradox in MBA is proposed. In section 5, a summary and conclusion is given, and some potential future work is discussed. Examples are intertwined throughout the paper.

2. Market basket analysis

Market Basket Analysis is a modeling technique based upon the phenomenon that if clients buy a certain group of items, then the clients are more (or less) likely to buy another group of items. The typical example is {if a customer buys diapers then the customer will also buy beer.}. It is used to determine which products sell together and is one of the most common and useful types of data analysis for marketing. The name is derived from the study of items purchased by supermarket shoppers. The results can be useful to any company that sells products, whether in a store, a catalog, or directly to the customer.

The purpose of Market Basket Analysis is to determine what products customers purchase together and to improve the effectiveness of marketing and sales tactics using customer data already available to the company. Knowing what products people will purchase as a group can be very helpful to a retailer. A store could use this information to place products frequently sold together in the same area, while a catalog or World Wide Web merchant could use it to determine the layout of their catalog and order

from. Direct marketers could use Market Basket Analysis results to determine what new products to offer their repeat customers.

Market Basket Analysis (MBA) is a powerful method for association rule induction which aims at finding regularities in the shopping behavior of supermarket customers, mail-order companies, on-line shops and the like. From the presence of certain products in a shopping cart, one can infer a high probability that certain other products are present. These techniques enable analysts and researchers to uncover hidden purchasing patterns in large data sets. Such information, expressed in the form of association rules, can often be used to increase the number of items sold, for instance, by appropriately arranging the products in the shelves of a supermarket.

In general, MBA is a rule that implies certain association relationships among a set of objects in a database, such as "occur together", "one implies the other", and "if condition, then results". It is a rule of the format: A $\Rightarrow$ B, where A and B stand for two different sets of items. These are two sets of items and do not share common items. The rule can be read as {if A then B}. For the example above {if a customer buys diapers then the customer will also buy beer} can be read as {if diapers then beer}. The A is "diapers", and the B is "beer". A set of items is called an itemset. The common measurements used to select the usefulness of the association rules are *support, confidence,* and *improvement,* as described below.

The *support* of a rule is the percentage of transactions that contain the itemset in both A and B. This can be expressed mathematically:

$$\begin{aligned} Support(if\ A\ then\ B) &= \frac{Transactions\ with\ itemsets\ in\ both\ A\ and\ B}{Total\ Transactions} \\ &= P(A \bigcap B) \end{aligned}$$

An itemset with a support higher than a given minimum support is called a frequent itemset. Support considers only the combination, not the direction. That is, the support for if A then B is the same as the support of if B then A.

The *confidence* of a rule is defined as the ratio of the support for the combination of A and B divided by the support for the A. It is the conditional probability of B, given A.

$$confidence = P(B|A) = \frac{P(A \bigcap B)}{P(A)}$$

Confidence is directional i.e. Confidence for {if A then B} Confidence for {if B then A}.

The *Improvement* (also known as *Lift, Interest* or *Correlation*) of a rule is defined as the confidence of the combination of A and B divided by the support of the B.

$$\text{Improvement} = \frac{P(B|A)}{P(B)} = \frac{P(A \bigcap B)}{P(A) \times P(B)}$$

It explains how much more confident we can be that a transaction contains the B if we know the transaction contains the A. When improvement is greater than 1, then the resulting rule is better at predicting the result than random chance. The further the value of improvement is from 1, the more the dependence between A and B. The values above 1 indicate a positive dependence, while the values below 1 indicate negative dependence.

In general, a potential useful association rule shall have a high support, a high confidence, and a high improvement. The main difficulty is, of course, there are too many potential rules to be considered. We simply cannot evaluate every single of them. Efficient algorithms are needed to restrict the search space and check only a subset of all rules, without missing useful rules.

Table 1. An Illustrative Example of Simpson's Paradox

With consideration of C

C=0

	A = 0	A = 1
B = 0	49	18
B = 1	573	472

$OR_1 = 2.24$

Table 1(a)

C=1

	A = 0	A = 1
B = 0	89	297
B = 1	45	348

$OR_2 = 2.32$

Table 1(b)

Without C

	A = 0	A = 1
B = 0	138	315
B = 1	618	820

$OR_3 = 0.58$

Table 1(c)

3. Simpson's paradox

A marginal association of aggregated dataset can have a different direction from each conditional association of non-aggregated dataset. This is called the Simpson's Paradox (Simpson, 1951). Simpson's Paradox is caused by the lurking variable and two

unbalanced data size groups combined into a single data set. A lurking variable is the variable that is not included in the analysis and may obscure an association that exists within each stratum. It could substantially change the interpretation of the data and result in Simpson's Paradox.

In analysis of Simpson's Paradox, we usually use odds ratio to measures the association relationship. An odds ratio is a statistic that measures the direction and magnitude of the relationship between variables in each row and column in contingency tables. An easy way to calculate the odds ratio is to use the cross-product ratio, which is defined to be $OR = \text{oddsratio} = \dfrac{n_{11} \times n_{22}}{n_{21} \times n_{12}}$, where, n_{11}, n_{12}, n_{21} and n_{22} are the counts within each cell in the 2-way Contingency tables as displayed below

	Buy B	Not Buy B
Buy A	n_{11}	n_{12}
Not Buy A	n_{21}	n_{22}

The condition that the odds ratio equals to 1 implies the independence between the two variables-product A and product B. When the odds ratio > 1 , the association between buying Products A and B is positive, i.e., whenever consumers buy Product A, the consumers are likely to buy Product B as well. When the odds ratio < 1 , the association between buying Products A and B is negative, i.e., whenever consumers buy Product A, the consumers are less likely to buy Product B.

Consider the following example for illustration. Let A, B, and C be three binary variables and stand for three products. For simplicity of the presentation, the notation α (similarly for β and γ) represents the Product A as well as the binary variable (indicator) A, as defined below.

$$A = \begin{cases} 0: & \textit{Buy product } A \\ 1: & \textit{Do not buy product } A \end{cases}$$

$$B = \begin{cases} 0: & \textit{Buy product } B \\ 1: & \textit{Do not buy product } B \end{cases}$$

$$C = \begin{cases} 0: & \textit{Buy product } C \\ 1: & \textit{Do not buy product } C \end{cases}$$

The data is displayed in Table 1. Table 1(a) is for Case I—with the consideration of C (when $\gamma = 0$); Table 1(b) is for Case II—with consideration of C (when γ=1), and Table 1(c) is for Case III, the combined data- without consideration of C.

The odds ratios OR_1, OR_2 and OR_3 are displayed at the bottom of each table. In Table 1(a), the sample odds ratio of Products A and B, given buying Product C, is: $OR_1 = \frac{n_{11} \times n_{22}}{n_{21} \times n_{12}} = \frac{49 \times 472}{573 \times 18} = 2.24$. Namely, the odd of buying A and buying B is estimated to be 2.24 times the odd of buying A and not buying B—given buying C (γ =0). Similarly in Table 1(b), the odds ratio of A and B for consumers who do not buy C, is: $OR_2 = \frac{n_{11} \times n_{22}}{n_{21} \times n_{12}} = \frac{89 \times 348}{45 \times 297} = 2.32$. This implies that the odd of buying A and buying B is estimated to be 2.32 times the odd of not buying B—given not buying C ($\gamma = 1$). Thus, we conclude that A and B are positively correlated-meaning, the consumers are more likely to buy A for those who brought B, as compared to those who did not buy B, no matter γ is 0 or 1.

However, in Table 1(c), the odds ratio of A and B, without the consideration of C, is: $OR_3 = \frac{n_{11} \times n_{22}}{n_{21} \times n_{12}} = \frac{138 \times 820}{618 \times 315} = 0.58$. This implies that the odd of buying A and buying B is estimated to be only 0.58 times the odd of not buying B. We thus conclude that A and B are negatively correlated-meaning, the consumers are less likely to buy A for those who brought B, as compared to those who did not buy B. This conclusion contradicts conclusions from Table 1(a) and 1(b). This is Simpson's Paradox-marginal association has a different direction for each conditional association.

For Market Basket Analysis, Simpson's Paradox refers to the reversal of the direction of an association when non-aggregated data combined into one dataset. Therefore, the decision maker may miss a useful association rule. The above example is used again to illustrate how Simpson's Paradox affects decisions in Market Basket Analysis.

The probabilities of all events for the three products are summarized in Table 2. The support, confidence, and improvement then can be evaluated for all potential association rules. This is given in Table 3. As seen from Table 3, with the consideration of C—Cases I and II (in Tables 3(a) and 3(b)—the improvements for Rule one (if A then B) are larger than 1, indicating that Rule one may be a useful rule. On the contrary, without the consideration of C—Case III in Table 3(c)—an opposite conclusion was resulted. The improvement for Rule one is only 0.7619: a weak negative relationship

between A and B is indicated. Decision makers may thus decide to drop this rule.

In summary, the conclusions from Cases I and II are different from the conclusions of Case III. Thus, further analysis is necessary to determine which conclusion is likely to be the correct one. The Breslow-Day test is recommended to decide whether the result is reasonably attributable to a significant lurking variable effect or just to sampling error. This allows us to make a correct decision on whether to aggregate the data.

Table 2. Percentages of all events

Event	Case I Probability	Case II Probability	Case III Probability
A	55.94%	17.20%	39.98%
No A	44.06%	82.80%	60.02%
B	6.03%	49.55%	23.96%
No B	93.97%	50.45%	76.04%
A and B	4.41%	11.42%	7.30%
A and No B	51.53%	5.78%	32.68%
B and No A	1.62%	38.13%	16.66%
No B and No A	42.45%	44.67%	43.36%

4. Proposed Method

To resolve the inconsistent conclusion from Simpson's paradox, we first determine whether the effect of a lurking variable is significant. We need to check the homogeneity association. Homogeneous association means that the association between A and B is the same for each level of C: $\gamma = 0$ and $\gamma = 1$.

The method used here is the Breslow-Day Test for homogeneity of odds ratios. The Breslow-Day statistic tests the null hypothesis of a homogeneous odds ratio. It tests whether the odds ratio between A and B is the same for different levels of C. It is a test of homogeneous association based on asymptotic chi-square distribution.

The general approach is as follow. If the data are not homogenous, we need to analyze the data separately. If the data are homogenous and the effect of a lurking variable is not significant, then the data can be combined. When this is the case, statisticians usually replace the "combined" odds ratio by the Mantel-Haenszel estimate

of Common Odds Ratio (OR_{MH}), as will be discussed after the Breslow-Day test.

Let us first test whether the data within two conditional tables, Table 3(a) and Table 3(b), are homogenous. This is to test the null hypothesis that two odds ratios are equal versus the alternative hypothesis that two odds ratios are not equal, i.e.,

$$H_0 : OR_1 = OR_2 \qquad \text{versus} \qquad H_a : OR_1 \neq OR_2.$$

Under the null hypothesis is true, the Breslow-Day test statistic has an asymptotic chi-square distribution with 1 degree of freedom. For a chi-squared distribution with 1 degree of freedom, the p-value for this test statistic is 0.091. Thus we fail to reject the null hypothesis that $OR_1 = OR_2$ and conclude that the association between A and B is the same regardless of the value of C (under significant level of α=0.05, say). This implies that the data shall be combined. In other words, the joint distribution of probability among the levels of A and B is homogeneous across the levels of C (lurking variable) and the effect of a lurking variable is insignificant. Of course, if the null hypothesis is rejected, then the data should not be combined and must be analyzed separately.

Because the potential of Simpson's Paradox, when the aggregated dataset is analyzed, we cannot follow the regular way for calculating the "improvement." Here, we propose a "Common Improvement" method for evaluating "improvement." This method is based on conditional probability and Bayes rule. In our previous example, improvements for associate rules are summarized in the Table 4. Recall that

$$\text{Improvement} = \frac{P(A \bigcap B)}{P(A) \times P(B)} = \frac{P(A|B)}{P(A)}$$

And

$$\begin{aligned} P(A|B) &= P[(A|B) \bigcap (C = 0)] + P[(A|B) \bigcap (C = 1)] \\ &= P[(A|B)|C = 0] \times P(C = 0) + P[(A|B)|C = 1] \times P(C = 1) \end{aligned}$$

Note that C is a binary variable and only takes two potential values: C = 0 and C =1. Divide P(A) on both sides, assuming P(A)$\neq$ 0 implies the common improvement is

$$\begin{aligned} \frac{P(A|B)}{P(A)} &= \frac{P[(A|B)|C = 0]}{P(A)} \times P(C = 0) + \frac{P[(A|B)|C = 1)]}{P(A)} \times P(C = 1) \\ &= (\text{Improvement}|C = 0) \times P(C = 0) + (\text{Improvement}|C = 1) \times P(C = 1). \end{aligned}$$

Specifically, in the example, the Common Improvement is: Common Improvement

$$
\begin{aligned}
&= \frac{P(A|B)}{P(A)} = \frac{P[(A|B)|C=0]}{P(A)} \times P(C=0) + \frac{P[(A|B)|C=1]}{P(A)} \times P(C=1) \\
&= (Improvement|C=0) \times P(C=0) + (Improvement|C=1) \times P(C=1) \\
&= 1.31 \times \frac{1112}{1891} + 1.34 \times \frac{779}{1891} = 1.31 \times 0.59 + 1.34 \times 0.41 \\
&= 0.7729 + 0.5494 \\
&= \underline{1}.\mathbf{3223}
\end{aligned}
$$

The Common Improvement for rule one is 1.3223, which is larger than 1. Thus we should retain the rule {If A then B} as a potential useful association rule.

Table 3(a). Measurements for all association rules. Case 1 — with C=0

	Rule	Support	Confidence	Improvement
1	If A then B	0.04406	0.0787781	**1.30748188**
2	If B then A	0.04406	0.7313433	**1.30748188**
3	If A then no B	0.51529	0.9212219	0.98028585
4	If no B then A	0.51529	0.5483254	0.98028585
5	If no A then B	0.01619	0.0367347	0.60968626
6	If B then no A	0.01619	0.2686567	0.60968626
7	If no A then no B	0.42446	0.9632653	1.0250249
8	If no B then no A	0.42446	0.4516746	1.0250249

Table 3(b). Measurements for all association rules. Case II — with C=1

	Rule	Support	Confidence	Improvement
1	If A then B	0.1142	0.6641791	**1.340402908**
2	If B then A	0.1142	0.2305699	**1.340402908**
3	If A then no B	0.0578	0.3358209	0.665660248
4	If no B then A	0.0578	0.1145038	0.665660248
5	If no A then B	0.3813	0.4604651	0.929280636
6	If B then no A	0.3813	0.7694301	0.929280636
7	If no A then no B	0.4467	0.5395349	1.069459731
8	If no B then no A	0.4467	0.8854962	1.069459731

Table 3(c). Measurements for all association rules. Case III — without consideration of C

	Rule	Support	Confidence	Improvement
1	If A then B	0.073	0.1825397	**0.76199236**
2	If B then A	0.073	0.3046358	**0.76199236**
3	If A then no B	0.3268	0.8174603	1.07497737
4	If no B then A	0.3268	0.4297636	1.07497737
5	If no A then B	0.1666	0.277533	1.15853196
6	If B then no A	0.1666	0.6953642	1.15853196
7	If no A then no B	0.4336	0.722467	0.95005913
8	If no B then no A	0.4336	0.5702364	0.95005913

Table 4. Improvements of all association rules

	Rule	Case I Improvement	Case II Improvement	Case III Improvement
1	If A then B	**1.307481883**	**1.340402908**	**0.761992361**
2	If B then A	**1.307481883**	**1.340402908**	**0.761992361**
3	If A then no B	0.980285851	0.665660248	1.074977372
4	If no B then A	0.980285851	0.665660248	1.074977372
5	If no A then B	0.609686263	0.929280636	1.15853196
6	If B then no A	0.609686263	0.929280636	1.15853196
7	If no A then no B	1.0250249	1.069459731	0.950059125
8	If no B then no A	1.0250249	1.069459731	1.0250249

5. Conclusion

This paper proposed a method to discover potentially useful association rules that may be missed by marginal and conditional associations, rather than by the minimum support and confidence thresholds in association rules analysis. The existence of Simpson's Paradox in a dataset usually brings the opposite associations, i.e. the same associate rule has its improvement value larger than 1 under a conditional situation, but its improvement value less than 1 under a marginal situation, respectively. The

main idea proposed in this thesis is a method to select useful association rules when data involve the occurrence of the Simpson's paradox. If the effect of the paradox is significant, as tested by Breslow-Day, then we need to analyze the data based upon conditional association. If the effect is not statistically significant, we then can combine the data. However, in this case, we should not follow the usual way to calculate the improvement based upon marginal association. We propose to use "common improvement" to evaluate the association rules in this case. If the Common Improvement is larger than 1, it means that the resulting rule is better at predicting the result than random chance. This approach gives decision makers the possibility of selecting the association rules that may be discarded by the analysis based upon the datasets in which Simpson's paradox effect presents. Association rules analysis has received a great deal of attention and a wide class of applications, such as, genes expression datasets, biological networks, and cellular states. We hope the proposed procedure given in this paper helps future development in selecting useful association rules in these different application areas. Association rules can also be applied to identify patterns of interest in the data, i.e., the B set is likely to occur whenever the A set occurs. Using Common Improvement will help us to find advantageous patterns in these areas.

Acknowledgement: *Min-Te Chao has been a humble hero and a truly leader in the statistical community. I have fortunately got to know him in the early stage of my career. Soon he became my professional model, and in many senses, my mentor. He has significant impact in my entire career: the way I think, the way I work and the way I see the world. It is a great honor to contribute my favorite piece of work in data mining to this special issue in honor of his retirement, understanding that data mining is indeed one of his recent research interests.

References

Fabris, C. C. and Freitas A. A. (1999). Discovering surprising patterns by detecting

occurrences of Simpson's paradox. *Research and Development in Intelligent Systems XVI (Proc. ES99, The 19th SGES Int. Conf. on Knowledge-Based Systems and Applied Artificial Intelligence)* 148–160. Springer-Verlag.

Freitas, A. A. (1997). *Generic, Set-oriented Primitives to Support Data-parallel Knowledge Discovery in Relational Database Systems.* PhD thesis. University of Essex, England.

Freitas, A. A. (1998). On objective measures of rule Surprisingness. Principles of Data Mining & Knowledge Discovery. *Artificial Intelligence* **1510**, 1–9.

George, J. H. (1997). *Enhancements to the Data Mining Process.* PhD thesis. Computer Science Department, School of Engineering, Stanford University, March 1997.

Megaputer Intelligence Inc., Available from http://www.megaputer.com/dm/dm101.php3

Michalski , R. S., Kerschberg , L., Kaufman , K. A. and Ribeiro , J. S. (1992). Mining For Knowledge In Databases: The Inlen Architecture, Initial Implementation And First Results. *Journal of Intelligent Information Systems.*

Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *J. Roy. Statistic. Soc. Ser. B* **13**, 238–241.

[Received March 2004; accepted June 2004.]