Selection of Consistent Rules with Market Basket Analysis and Overall Variability Association Rules Method (OVARM)

Fajrianza Adi Nugrahanto, Hari Wijayanto, Farit Mochamad Afendi

Abstract—Market Basket Analysis (MBA) is one of the most popular data mining techniques used in retail environment. To monitor rules performance over time a variation indicator is needed, which OVARM can fulfill. This study aimed to apply MBA and OVARM on transaction data from a retailer with several supermarket stores to find the most consistent item pair bought by consumers. Application of MBA with minimum support 0.1% and minimum confidence 50% found eleven rules with high support, confidence, and lift values. Applying OVARM on those eleven rules resulted in a product association "If consumers buy Blueberries, then Bananas also tend to be bought" as the most consistent rule with 12% value in OVAR (Overall Variability Association Rules).The selected rule was a trivial one, therefore several modifications on OVARM usage were proposed in this study to acquire the most consistent non-trivial rule.

.

Index Terms—association rules, market basket analysis, OVARM, retail

1 Introduction

Market Basket Analysis (MBA) is one of data mining techniques that used to reveal products tend to be bought together by consumers [1]. This technique will produce rules, which are "if-then" statements, as its result. Rules can give insight into consumer buying pattern and affect retail decision making processes, such as product placement or product marketing campaign.

One main consequent that need to be looked from this technique is it will produce a large number of rules when the number of products increases. To find the most interesting and informative rules could be a tedious task as not all rules give meaningful insight. Goodness indicators, such as support, confidence, and lift, are often used to sort and select rules. Larger indicators value will indicate better rules.

Often times goodness indicators alone are not enough to select interesting rules as consumers buying pattern changes throughout some period. A variation indicator is needed to measure how well each rules perform over certain period. Rules with low variation can be considered as the consistent one, thus they are valuable for long term decision making. In their study [2], Papavasileiou and Tsadiras proposed a variation indicator called Overall Variability Association Rules Method (OVARM).This indicator will measure the fluctuation of goodness indicators value for each rule over the time.

In this study, OVARM will be applied to transaction data from a supermarket chain to find the most consistent rule. The selected rule will be then compared to the best rules by Market Basket Analysis. In addition, performance of OVARM on supermarket transaction data will be evaluated.

2 Research Method

2.1 Data

The data used in this study was transaction data from a retail that can be acquired from https://www.dunnhumby.com/sourcefiles with title The Complete Journey. There were 275,000 baskets and 92,000 distinct products which had been accumulated over two years. Two different data formats were used. The first one was data which contain all transactions for two years period and the second one was transaction data per month (there were 24 datasets with this format). Details on transaction data per month can be seen on Table 1.

- *Fajrianza Adi Nugrahantois currently pursuing masters degree program in applied statistics in Bogor Agricultural University, Indonesia, PH +6281385884228E-mail: fajrianza.adin@gmail.com*
- *Hari Wijayanto is a Lecturer in Departement of Statistics, Bogor Agricultural University, Bogor, Indonesia*
- *Farit Mochamad Afendi is a Lecturer in Departement of Statistics, Bogor Agricultural University, Bogor,Indonesia*

TABLE 1
DETAILS ON DATA PER MONTH

Month	Week	# Basket	Month	Week	# Basket
1	1-4	1493	13	49-52	11797
2	5-8	3703	14	53-56	11768
3	9-12	6514	15	57-60	12041
4	13-16	10219	16	61-64	12363
5	17-20	12192	17	65-68	12394
6	21-24	12150	18	69-72	12237
7	25-28	11613	19	73-76	12005
8	29-32	11621	20	77-80	11778
9	33-36	11635	21	81-84	11862
10	37-40	11529	22	85-88	12064
11	41-44	11751	23	89-92	12103
12	45-48	11739	24	93-96	11254

2.2Methodsof Data Analysis

There were three main analysis steps in this study:

1. Generating rules from two year period data with Market Basket Analysis. Rules were generated with apriori algorithm [3] with minimum support 0.1% and minimum confidence 50%.
2. Calculating goodness indicators value (support, confidence, and lift) for every generated rules in point 1 from data per month. All goodness indicators value then were summarized into their average and standard deviation per rule.
3. Performing Overall Variability Association Rules Method (OVARM) by following steps:
   a. Calculating Coefficient of Variation (CV) for confidence and lift for each rule. CV can be acquired by dividing average value with standard deviation.
   b. Calculating Overall Variation Association Rules (OVAR) by averaging CV confidence and CV lift for each rule.
   c. Calculating Overall Variation Product (OVP) for every product from all rules. OVP can be acquired by averaging OVAR value from rules that contain certain product.
   d. Selecting rule that contains product with the lowest OVP value in the left hand side (precedent) part.
   e. Selecting rule from point (d) that contains product with the lowest OVP value in the right hand side (antecedent) part.
   f. Rule selected from point (e) is the most consistent rule according to OVARM.

3 Result and Discussion

3.1 Rules Generated with Market Basket Analysis

Rules that were generated from two year period transactions data with Market Basket Analysis by using apriori algorithm (with minimum support 0.1% and minimum confidence 50%) can be seen on Table 2 whereas their goodness indicators values are on Table 3.

TABLE 2
LIST OF RULES

Rules	Rules with Product ID
R01	{968215} ➔ {1082185}
R02	{1098248} ➔ {1082185}
R03	{880427} ➔ {1082185}
R04	{7024990} ➔ {1082185}
R05	{879528} ➔ {1082185}
R06	{900491} ➔ {1053763}
R07	{901062} ➔ {1082185}
R08	{901666} ➔ {1082185}
R09	{1050131} ➔ {1053763}
R10	{1127831,866211} ➔ {1082185}
R11	{1029743,1127831} ➔ {1082185}

TABLE 3
GOODNESS INDICATORS

Rules	Support (%)	Confidence (%)	Lift
R01	0.16	53.2	4.97
R02	0.17	55.7	5.21
R03	0.11	50.6	4.73
R04	0.11	50.3	4.70
R05	0.15	51.6	4.83
R06	0.11	51.0	150.76
R07	0.40	51.3	4.79
R08	0.13	55.9	5.23
R09	0.11	52.8	156.13
R10	0.11	61.5	5.75
R11	0.12	54.3	5.07

From Table 3 it can be seen that rule R07 was rule with the highest support (0.40%), rule R10 was rule with the highest confidence (61.5%), and rule R09 was rule with the highest lift (156.13). Those three rules will then be considered as the best rules generated with MBA and will be compared with the most consistent rule selected by OVARM.

As it has been mentioned on Section 2.2, every rule generated from two year period data transaction then would be generated again from transaction data per month. Ideally, every rule would have goodness indicators value for 24 months, but there were months where some rules cannot be generated, thus leaving some missing data. Summarized goodness indicators value can be seen on Table 4 for average and Table 5 for standard deviation.

TABLE 4
GOODNESS INDICATORS AVERAGE VALUE

Rules	Average		
	Support (%)	Confidence (%)	Lift
R01	0.16	55.6	5.12
R02	0.21	58.7	5.48
R03	0.12	52.3	4.85
R04	0.35	52.0	5.01
R05	0.15	54.1	4.98
R06	0.10	48.5	145.49
R07	0.39	51.4	4.76
R08	0.13	59.8	5.60
R09	0.11	52.1	156.50
R10	0.11	64.9	6.09
R11	0.13	54.7	5.07

TABLE 5
GOODNESS INDICATORS STANDARD DEVIATION

Rules	Standard Deviation		
	Support (%)	Confidence (%)	Lift
R01	0.03	12.0	0.87
R02	0.18	18.0	1.59
R03	0.06	13.7	1.30
R04	0.22	6.2	0.61
R05	0.04	11.7	0.85
R06	0.04	13.0	43.38
R07	0.16	9.5	0.92
R08	0.09	18.6	1.97
R09	0.04	8.1	35.08
R10	0.07	16.0	1.58
R11	0.07	8.8	0.93

3.2 Overall Variation Association Rules Method

To find the most consistent rule, variation indicator, which is OVAR, need to be calculated. OVAR value for all rules can be seen on Table 6. Rule R04 had the lowest OVAR value (12%) and will be indicated as the most consistent rule, whereas rule R08 will be indicated as the most incosistent one with OVAR value (33.2%). The last step of OVARM is to find OVP value for every product contained in every rule. From Table 7 it can be seen that product 7024990 had the lowest OVP value (12.0%). There was only one rule that contains product 7024990 on its precedent, which was rule R04. Because it was the only rule that had product 7024990, rule R04 will then be considered as the most consistent rule by OVARM.

TABLE 6
CV AND OVAR VALUE

Rules	CV Confidence (%)	CV Lift (%)	OVAR (%)
R01	21.6	16.9	19.3
R02	30.7	29.0	29.9
R03	26.2	26.8	26.5
R04	11.8	12.1	12.0
R05	21.7	17.0	19.3
R06	26.9	29.8	28.3
R07	18.4	19.2	18.8
R08	31.2	35.2	33.2
R09	15.6	22.4	19.0
R10	24.7	26.0	25.3
R11	16.0	18.3	17.2

TABLE 7
OVP VALUE

Product ID	OVP (%)
7024990	12.0
1029743	17.2
901062	18.8
1050131	19.0
968215	19.3
879528	19.3
1127831	21.2
1082185	22.4
1053763	23.7
866211	25.3
880427	26.5
900491	28.3
1098248	29.9
901666	33.2

3.3 Comparison

From Table 8 it can be concluded that OVARM succesfully selected the most consistent rule, which was rule R04, from two year period observation when compared to Market Basket Analysis result (R07, R10, and R09). Another main thing that need to be taken into consideration from this result is that there was trade-off between goodness indicators and variation indicator for rules. Rules with highest goodness values didn’t automatically had the lowest variance and vice-versa.

TABLE 8
COMPARISON BETWEEN MBA AND OVARM RESULT

Rules	Support (%)	Confidence (%)	Lift	OVAR (%)
R07	0.40	51.3	4.79	18.8
R10	0.11	61.5	5.75	25.3
R09	0.11	52.8	156.13	19.0
R04	0.11	50.3	4.70	12

3.4 Notes on OVARM Result

Rule R04 was the most consistent rule selected by OVARM. Referencing to its products' details, this rule could be interpreted as "If consumers buy Blueberries, then Banana also tend to be bought.". Blueberries and Banana are two fruit products which can be considered as daily product. Consumers have tendency to consistenly buy daily product than other non-daily product such as cakes, sweet products, etc. As such, rule R04 was a trivial rule, a rule that give no new information to retail. Similar result can also be found on [2] with the most consistent rules selected by OVARM was {Vegetables, Dry Toast ➔ Fruits}. Those three products were also daily products. From this two results (especially with supermarket transaction data) it could be seen that daily products will dominate the selected most consistent rule by OVARM. To avoid trivial rule to be selected there were two solutions proposed in this study:

1. OVARM is applied on transaction data based on product categories, mainly non-daily products. This approach will avoid daily products to be included in resulted rule. For example, if OVARM is used only on products in electronic category, then the selected rule will be consisted of electronic products.
2. OVARM is only applied on certain time period. For example, OVARM is only used on near-Thanksgiving-day transaction data or on weekday transaction data. This approach will reduce the chance of daily products to appear as they are bought consistently over time.

In this study, solution number one was applied to available transaction data in order to measure the effectiveness of proposed solutions. The result can be seen on Table 9.

TABLE 9
NON-TRIVIAL RULES RESULTED FROM FIRST SOLUTION

Rules	OVAR (%)
Category : Electronics { Inside Frost Bulbs ➔ Children's Books }	58.69
Category : Ice Cream { Traditional ➔ Soft Drink 2 Liter Carbonated }	42.15
Category : Seafood { Fresh Catfish ➔ Misc. Candy }	52.08
{ Fresh Catfish ➔ Snack Crackers }	54.90
Category : Spices { Spices & Seasonings ➔ Peppers Green Bell }	54.95

Solution number one was used on eight different product categories, which were, batteries, cakes, condiments, cookies, candies, electronics, ice creams, seafood, and spices. From those eight categories only four categories that gave non-trivial rule, for only the precedent part that was forced to contain product from chosen categories. All non-trivial rules on Table 9 need further examination as they concealed previously hidden consumer buying pattern.

4 Conclusion

OVARM succesfully selected the most consistent rule. One main consideration when using OVARM is this method will choose trivial rules as the most consistent one. To minimize this occurrence, two different solutions were proposed. One of the solutions was applied in this study and gave satisfactory result as it revealed non-trivial rules.

Reference

[1] RC Blattberg, BD Kim, SA Neslin, *Database Marketing: Analyzing and Managing Customers*. New York: Springer, 2008.

[2] V Papavasileiou, A Tsadiras, " Evaluating Time Variations to Identify Valuable Association Rules in Market Basket Analysis". *Intelligent Decision Technologies*, vol. 7, pp. 81-90, 2013.

[3] R Agrawal, T Imeilinski, A Swami, "Mining Association Rules between Sets of Items in Large Databases",*Proceedings of the 1993 ACM SIGMOD Conference*, Washington DC, USA, May 1993.