Market Basket Analysis Using Variable Cluster (VARCLUS) Method - An Alternative Approach to Association Rules

Karthikeyan Nallakamachi Asst. Manager, Analytics, GENPACT India, Bangalore

Introduction

The purpose of this paper is to break the traditional paradigm of doing market basket analysis (MBA) using association rules and to show that factorial algorithms like PROC VARCLUS can also be used to arrive at market baskets. This paper shows step by step procedure of doing MBA using VARCLUS method. At the same time, this paper also gives insight on the usage of PROC VARCLUS procedure in arriving at product groups that are bought together more frequently by the customers.

Background

Market Basket Analysis is a data mining tool used to identify the co-occurrence or co-existence of nominal or categorical observations. It is widely used to identify purchasing pattern of customers in a retail stores using transaction level data. The outcome of the analysis will give rules like "People who bought items on set X often also bought items on set Y".

The most widely used technique to conduct market basket analysis is *association rules technique*. Although it is most widely used technique, the following are the limitations[1] of the application of association rules to arrive at market baskets:

1. Both independent and dependent data is nominal or categorical type. Therefore, the frequency of pattern can be counted. Such as, market basket analysis using association rule is sometimes called as *mining frequent pattern*. If the data is quantitative, we need to categorize it into some interval (but the meaning is actually nominal) such as: age 0 to 1, age 1-5, age 5 to 12, age 12 to 19, etc.
2. In MBA, the quantity of each item that the customers bought is usually not considered. Whether a customer buys one kg of apple or 10 kg of apple would be considered as the same set of apple.
3. Not all recorded transactions are used. Only transactions of purchase of more than one item are considered as data. Transactions of single item are not used for the analysis.
4. The input data is assumed to be clean from error and noise.

In this paper we are trying to demonstrate, how to do MBA using Factorial algorithm i.e., PROC VARCLUS – SAS™ Procedure. Usage of VARCLUS also helps to overcome the above limitations. In addition to this, factor analysis technique helps in saving lot of time in terms of computing resources like space and server run time.

Data Preparation

Data Format

The input SAS™ data must contain three columns of variables. The first column required is customer identifier. This is typically unique to the Merchandise Division (MD) in which the customer made purchase. The second required column is MD identifier. This is typically a product or product grouping identifier. The third required column is purchase amount made on each of the MDs by the customer. The data set should contain one

[1] http://people.revoledu.com/kardi/tutorial/MarketBasket/Characteristics.htm

row of data for each unique combination of customer identifier and MD identifier with respect to the purchase amount. A visual example of the base data can be seen in Figure -1.

Figure-1

CUSTOMER IDENTIFIER	PRODUCT GROUP IDENTIFIER	PURCHASE AMOUNT
Customer 1	Product Group A	$ XXX
Customer 2	Product Group B	$ XXX
Customer 2	Product Group C	$ XXX
Customer 3	Product Group A	$ XXX
Customer 3	Product Group B	$ XXX
Customer 3	Product Group D	$ XXX

Once the base dataset is created using the transaction level data as explained above, the data needs to be prepared for running the PROC VARCLUS procedure. The following program would help us to do that:

```
PROC SUMMARY DATA=<Base Dataset> NWAY MISSING;
CLASS <<customer identifier product identifier>>;
VAR <Purchase Amount>;
OUTPUT OUT=< dataset-1> (drop = _type_ _freq_)SUM=;
RUN;

PROC TRANSPOSE DATA=<dataset-1> OUT=<dataset-2>;
ID <Product_Group_Identifier>;
VAR <Purchase Amount>;
BY <Customer Identifier>;
RUN;
```

The input dataset for the PROC VARCLUS procedure will transpose the data to have each product group identifier as one column. At this level, all the duplicate customer identifiers are eliminated from the dataset so that each row is unique to the customer identifier. A visual example of the base data is shown in Figure -2.

Figure-2

CUSTOMER IDENTIFIER	PRODUCT GROUP - A PURCHASE AMOUNT	PRODUCT GROUP - B PURCHASE AMOUNT	PRODUCT GROUP - C PURCHASE AMOUNT	PRODUCT GROUP - D PURCHASE AMOUNT
Customer 1	$ XXX	$ XXX	$ XXX	$ XXX
Customer 2	$ XXX	$ XXX	$ XXX	$ XXX
Customer 3	$ XXX	$ XXX	$ XXX	$ XXX

Once the base dataset is transformed using PROC transpose, the data is cleaned through missing value treatment and outlier removal / treatment. Missing value treatment implies replacing the null values into 0. This is because if a customer did not make any purchase in a particular MD, it can also be said as the customer made 0 purchases in that MD. Outlier analysis implies removing extraordinary purchase amount or replacing the extraordinary purchase amount by the average purchase amount in that particular MD by all the customers.

Once the data cleaning is done, we can run the PROC VARCLUS procedure. The following program would help doing that:

```
PROC VARCLUS DATA=final_data2  OUTTREE=tree;

VAR sales:;
RUN;
```

PROC VARCLUS to Arrive MD Groups

The VARCLUS procedure attempts to divide a set of variables into non-overlapping clusters in such a way that each cluster can be interpreted as essentially unidimensional. For each cluster, PROC VARCLUS computes a component that can be either the first principal component or the centroid component and tries to maximize the sum across clusters of the variation accounted for by the cluster components. A large set of variables can often be replaced by the set of cluster components with little loss of information.

PROC VARCLUS creates an output data set that can be used with the SCORE procedure to compute component scores for each cluster. A second output data set can be used by the TREE procedure to draw a tree diagram of hierarchical clusters[2].

For this exercise, PROC VARCLUS attempts to divide the set of MD variables into non-overlapping MD groups. This can be interpreted as the each unidimensional cluster contains MDs that are highly correlated among themselves but have least correlated with other unidimensional clusters. In other way, each MD variable cluster contains the MDs that are most likely bought together as compared to the MDs in other MD variable clusters.

Methodology Summary

In summary, the methodology of doing MBA analysis using factorial algorithm like PROC VARCLUS would look like this:

1. Customer level MD purchase history data for the customers who were purchase active for the last 12 month time period is prepared.
2. The data is transformed by product group to have MD purchase data as columns and customer level data as rows.
3. Data cleaning is done through missing value treatment and outlier analysis of the MD variables.
4. PROC VARCLUS is run with the MD level purchase history data.
5. Merchandise Division “Groups” are formed based on PROC VARCLUS factor loading and business logic of grouping Merchandise Divisions.
6. The stability of “Clusters” is validated by developing the MD Groups based on MD level purchase history data at quarterly level or out of time period validation is done.

Conclusion

In this paper, an attempt is made to break the industry practice of doing market basket analysis using association rules by suggesting the factorial algorithm such as PROC VARCLUS – SAS™ procedure, which is a principle component factor analysis method. In addition to that, it is also observed that using factor analysis technique helps in saving lot of time in terms of computing resources like space and server run time.

References

1. Kardi Teknomo, PhD. “Characteristics of Market Basket Analysis”, http://people.revoledu.com/kardi/tutorial/MarketBasket/Characteristics.htm
2. Chapter-68, The VARCLUS Procedure, SAS user guide, http://www.okstate.edu/sas/v8/saspdf/stat/chap68.pdf

[2] http://www.okstate.edu/sas/v8/saspdf/stat/chap68.pdf

ACKNOWLEDGMENTS

The author appreciates the review and suggestions provided by M. Muthu Mangai, PSS. Moorthy, Milind Kelkar and Pankaj Bagri, GENACT India, Bangalore. The author is very much thankful to the management of GENPACT India for giving the consent to publish this paper.

About the Author

Karthikeyan Nallakamachi is working as Assistant Manager, Analytics with GENPACT's Collections & Operations analytics practice. With 2 Years of economics research experience and 5+ years of experience in Marketing, Collections and Operations Analytics, he has built marketing, collections and operations strategies. An expert in data research, econometric model building and analysis, Karthikeyan holds, Masters in Economics Science from Madras School of Economics, Anna University, Chennai, India.

About Genpact

Genpact is a global leader in business process and technology management, offering a broad portfolio of enterprise and industry-specific services. The company manages over 3,000 processes for more than 400 clients worldwide. Putting process in the forefront, Genpact couples its deep process knowledge and insights with focused IT capabilities, targeted analytics and pragmatic reengineering to deliver comprehensive solutions for clients. Lean and Six Sigma are an integral part of Genpact's culture and Genpact views the management of business processes as a science.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Karthikeyan Nallakamachi,
GENPACT India, # 99, Surya park, Ring Road, Electronic City, Bangalore – 560 100
Web: www.genpact.com
Email: Karthikeyan.n1@genpact.com