Paper 204-2012

Easily Add Significance Testing to your Market Basket Analysis in SAS® Enterprise Miner™

Michael Faron and Goutam Chakraborty, Oklahoma State University, Spears School of Business, Stillwater, OK, U.S.

ABSTRACT

Market Basket Analysis is a popular data mining tool that can be used to search through data to find patterns of co-occurrence among objects. It is an algorithmic process that generates business rules and several metrics for each business rule such as support, confidence and lift that help researchers identify "interesting" patterns. Although useful, these popular metrics do not incorporate traditional significance testing. This paper describes how to easily add a well-known statistical significance test, the Pearson's Chi Squared statistic, to the existing output generated by SAS® Enterprise Miner's Association Node. The addition of this significance test enhances the ability of data analysts to make better decisions about which business rules are likely to be more useful.

INTRODUCTION

Market Basket Analysis (MBA) or association discovery is accomplished in SAS Enterprise Miner via the Association Node which is located under the explore tab. This node's implementation follows industry and academic standards by outputting well-known metrics described below. These metrics are used to evaluate the quality and impact of discovered patterns that are often used as business rules. However, these metrics do not include traditional statistical testing.

After a brief introduction to core concepts of MBA, this paper will use example data from Enterprise Miner to provide details on configuring and using the Association Node. Then, the paper will introduce the Chi-Squared statistic and describe how to use a SAS Utility node to add the Chi-Squared statistic and its p-value to the output generated by the Association Node.

MARKET BASKET ANALYSIS

Although Market Basket Analysis (MBA) has been extended into areas such as loss-leader analysis and fraud detection, and into industries such as telecommunications and healthcare [5, 6], MBA has been and continues to be a popular practice in retail settings where researchers are interested in discovering patterns of customer purchasing behavior.

A typical MBA scenario involves items for sale at a grocery store. For example, by considering the purchases of shoppers, or the contents of their market baskets, researchers may be able to discover patterns of particular items occurring together much more frequently than expected. If we are interested in soda and popcorn, and in particular if the presence of soda increases the likelihood that popcorn is also present, we might discover that 10% of all transactions include Soda and Popcorn. We may also learn that of all the transactions that include Soda, Popcorn is present 30% of the time. Finally, if we only expect Popcorn to be present in 20% of all transactions, then the association rule "Soda => Popcorn" appears to be an interesting rule. The presence of Soda is associated with an increased likelihood that Popcorn is also present.

SAS Enterprise Miner (EM) includes the *Sampsio* library that in turn contains a dataset named *Assocs*. In this fictitious scenario there are 1,001 customers who each purchased 7 items out of a possible 20 items. In association analysis, only the presence of the item type is needed, not the quantity. Purchasing one avocado or five is equivalent in the sense that each results in a single "avocado" record. Table 1 shows a selection of customers and their purchases, or items.

Items purchased by two example customers

Customer	Items Purchased
742	herring, corned beef, apples, olives, steak, sardines, cracker
743	baguette, soda, herring, cracker, heineken, olives, apples

Table 1

ASSOCIATION NODE

To use the SAS EM Association Node, the data must be organized in a "transactional" structure. If we consider the purchases in Table 1, every customer/item combination becomes a "transaction" row in the dataset to be analyzed. If we consider the purchases of the two customers in Table 1, Figure 1 shows the purchasing data transformed into a transaction structure. Note that a TIME variable is present in the Sampsio.Assocs data. Time information is needed in Sequence Analysis, but time data is ignored in association analysis. If it is present, the time variable role needs to be set to rejected for performing MBA.

SAMPSIO.ASSOCS

Obs #	CUSTOMER	TIME	PRODUCT
5193	741	5	coke
5194	741	6	coke
5195	742	0	hering
5196	742	1	corned_b
5197	742	2	apples
5198	742	3	olives
5199	742	4	steak
5200	742	5	sardines
5201	742	6	cracker
5202	743	0	baguette
5203	743	1	soda
5204	743	2	hering
5205	743	3	cracker
5206	743	4	heineken
5207	743	5	olives
5208	743	6	apples
5209	744	0	avocado
5210	744	1	cracker
5211	744	2	artichok

Figure 1. "Basket" contents for customers 742 and 743 from the Assocs transaction table.

Note that if your source data is not in a transactional structure, PROC TRANSPOSE may be used to convert "wide" data to "narrow" data.

To use the Association Node for association analysis, the data must have two variables set to specific roles:

1. An ID variable such as customer ID must have the role of ID
2. An item or product variable must have the role of TARGET

Additionally, the dataset role must be set to TRANSACTION in the data creation step. Figure 2 shows a typical layout flow in EM for performing MBA.

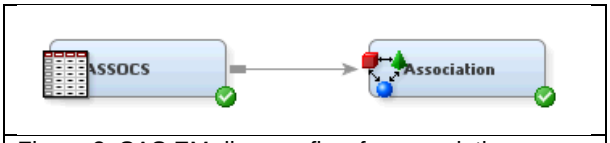

Figure 2. SAS EM diagram flow for association analysis.

MBA generates rules in the form of A ==> B, where A is called the antecedent or left hand side, and B is called the consequent or right hand side. Typically these rules are generated by counting combinations of instances of "A and B" that are present in the data. Examples of rules are Olives ==> Herring (2 items), Coke & Chicken & Avocados ==> Ice Cream (4 items), and Peppers & Avocado ==> Sardines & Baguette (4 items). Before running the Association Node, several properties of the node can be configured to customize the rule search and to manage the rule results.

The *Maximum Items* property specifies the number of items to consider for rule generation. With large datasets and/or higher Maximum Items, the search space for rules is exceedingly and sometimes prohibitively large. For example, suppose we are considering a collection of 10,000 items and looking only for rules containing two items in the left-hand-side and 1 item in the right-hand-side. Even under such restrictive rule conditions, there are approximately 1,000,000,000,000 such rules [2].

Minimum Support and *Minimum Confidence* (defined below) are used to constrain the potential number of rules generated by screening out those that don't meet minimum benchmarks. Similarly, *Rules to Keep* allows you to directly limit the number of rules returned. Finally, *Sort Criterion* allows you to specify which metric is used to order the resulting rules. Figure 3 shows the property settings of the Association node used in this paper's example.

Property	Value
Association	
Maximum Items	2
Minimum Confidence Level	10
Support Type	Percent
Support Count	1
Support Percentage	5.0
Sequence	
Rules	
Number to Keep	200
Sort Criterion	Lift
Number to Transpose	200
Export Rule by ID	No

Figure 3. Association Node properties used in this paper's example output.

MBA METRICS

MBA analysis often produces very large numbers of rules, especially when many items are involved. We've seen how setting constraints can help manage useful rules generation, but even then there can be hundreds or thousands of rules that meet user specified screening criteria. Once the rules are generated and an analyst is faced with gleaning important and actionable information, there is no single-best way to decide which rules are the "interesting" rules. Many approaches have been suggested and investigated [3, 8, 9], including a chi-square test for independence, but SAS EM outputs some of the most widely reported MBA metrics: Support, Confidence, Expected Confidence, and Lift.

Consider the rule A ==> B. Support for the rule is the joint probability that both items are in a basket. It answers the question what percent of baskets contain A and B? Confidence is the conditional probability that B is in the basket given that A is present. Expected Confidence is simply the probability that B is in a basket. Lift is the confidence divided by the expected confidence. It's the ratio of the likelihood of B being in a basket with A to the likelihood of B being in any basket. Figure 4 shows sample output from the Association Node.

Association Report

Relations	Expected Confidence (%)	Confidence (%)	Support (%)	Lift	Transaction Count	Rule
2	29.57	70.29	21.98	2.38	220.00	ice_crea ==> coke
2	31.27	74.32	21.98	2.38	220.00	coke ==> ice_crea
2	30.47	58.13	21.08	1.91	211.00	avocado ==> artichok
2	36.26	69.18	21.08	1.91	211.00	artichok ==> avocado
2	29.57	49.66	14.69	1.68	147.00	sardines ==> coke

Figure 4. Association Node output showing common MBA metrics for the first 5 of 200 rules.

Which are the best rules? Traditionally, support and confidence have been used to identify important rules, but "the most popular objective measure of interestingness is *lift*" [9]. When lift is greater than 1, A is said to "lift" the presence of B above what we would expect to see. But can we be sure that the rules, even the rules with lift > 1, are statistically significant? One experimental study showed even with minimum support and confidence constraints in place, up to 30% of the generated rules were statistically insignificant [4]. Adding a significance test for each rule would provide an additional criterion for judging rule importance. A test can also be used to prune rules with lift > 1 by setting aside those that are not statistically significant.

CHI SQUARE STATISTIC

Karl Pearson developed the Chi-squared test in 1900 [7]. For the rule A ==> B, the chi-squared test for independence evaluates the null hypothesis that the presence or absence of B is not related to the presence or absence of A. The test is based upon a 2 x 2 contingency table comprised of the cross-tab frequencies of A and B. Although SAS EM does not generate these counts for us to use, [1] derived the chi-squared statistic from the core MBA metrics, which SAS EM does provide (as seen in Figure 4). Figure 5 shows the equation.

$$\chi^2 = n\ (lift - 1)^2 \frac{supp\ conf}{(conf - supp)\ (lift - conf)}$$

Figure 5. Formula for chi-squared statistic derived from association analysis metrics [1].

By adding a SAS Code utility node to the existing diagram (Figure 6) and using the code listed in the appendix, you can easily append the statistic to the Association Output node.

Figure 6. SAS EM diagram flow for association analysis.

In this example based on the *Sampsio* library and *Assocs* data, 27 rules with lift > 1 are not statistically significant (alpha = .05). (The results are stored in a temporary work table, but the provided code in the appendix contains comments indicating how to save the table to your project library/hard drive.) A partial output of the SAS Code results is shown in Figure 7.

Obs	EXP_CONF	CONF	SUPPORT	LIFT	RULE	index	CHISQ	PVALUE
110	59.94	66.58	26.07	1.11	baguette ==> heineken	110	11.836	0.00058
111	47.25	51.89	16.48	1.10	soda ==> olives	111	4.015	0.04509
112	31.77	34.88	16.48	1.10	olives ==> soda	112	4.015	0.04509
113	30.47	33.40	15.78	1.10	olives ==> ham	113	3.644	0.05626
114	47.25	51.80	15.78	1.10	ham ==> olives	114	3.644	0.05626

Figure 7. Partial results of the SAS Code node. Note that rules 113 and 114 have lift > 1, but are not statistically significant at alpha = .05.

Although this example has the maximum items property set to 2, these results work for maximum items greater than 2 because multiple items on the left-hand or right-hand side are treated as a single item in the calculations. (Multi-item results have been tested and confirmed.) Note that chi-square results are unreliable when cell counts are less than 5. Setting an appropriate minimum support can help mitigate this occurrence.

CONCLUSIONS

Market Basket Analysis (MBA) is one of the most well-known analysis tools in the data mining toolkit. SAS Enterprise Miner easily allows analysts to make use of MBA via the Association Node. With data in the right structure and a few property settings, the Association Node quickly generates association rules accompanied by industry standard metrics such as support, confidence, and lift, which help determine which rules are the most important. However, even though some rules may appear interesting, they may not be statistically significant. We understand that data miners are typically less concerned with statistical significance of input variables in predictive models because of the large data sets used in data mining. However, in the case of association rules, the significance numbers provide another metric that helps data miners to quickly narrow down a large number of rules to a smaller set. As we have shown, with a few lines of Base SAS via a SAS Code utility node, a chi-square statistic and its p-value can be appended to the standard output to indicate which rules are statistically significant, thus providing another metric to help evaluate the importance of association rules.

REFERENCES

[1] Alvarez, S. (2003). Chi-squared computation for association rules: Preliminary results. *Technical Report BCCS-03-01, Computer Science Department, Boston College*.

[2] Association rule learning. (2011, October 5). In *Wikipedia, The Free Encyclopedia*. Retrieved 17:54, November 4, 2011, from http://en.wikipedia.org/w/index.php?title=Association_rule_learning&oldid=454134746

[3] Brijs, K., Vanhoof, K., & Wets, G. (2003). Defining interestingness for association rules. *International Journal Information Theories & Applications, 10* (4), 370-375.

[4] Dorn, M., Hou, W., Che, D., & Jiang, Z. (2008). An empirical study of qualities of association rules from a statistical view point. *Journal of Information Processing Systems, 4* (1), 27-31.

[5] Kotsiantis, S., & Kanelopoulos, D. (2006) Association rules mining: A recent overview. *GESTS International Transactions on Computer Science and Engineering, 32* (1), 71-82.

[6] Megiddo, N., & Srikant, R. (1998). Discovering predictive association rules. In *Proceedings of the fourth international conference on knowledge discovery and data mining (KDD-98),* (27-78). Menlo Park: AAAI.

[7] Pearson's chi-squared test. (2011, November 2). In *Wikipedia, The Free Encyclopedia*. Retrieved 06:08, November 3, 2011, from http://en.wikipedia.org/w/index.php?title=Pearson%27s_chi-squared_test&oldid=455325610

[8] Shaharanee, I., Dillon, T., & Hadzic, F. (2009). Ascertaining data mining rules using statistical approaches. In P. Sandhu (Ed.), *International Symposium on Computing, Communication and Control (ISCCC 2009)* (pp. 180-188). Singapore: International Association of Computer Science and Information Technology (IACSIT).

[9] Webb, G. (2007). Discovering Significant Patterns. *Machine Learning, 68* (1), 1-33. Netherlands: Springer.

CONTACT INFORMATION

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

Name: Michael Faron
E-mail: michaeljfaron@yahoo.com

Michael Faron is a recent graduate of the Graduate Data Mining Certificate program at Oklahoma State University. He is a Certified Predictive Modeler Using SAS® Enterprise Miner 6.1 and has over ten years of professional experience in Web and Database design and development. He is currently working as a consultant in the Healthcare IT Industry.

Name: Dr. Goutam Chakraborty
Enterprise: Oklahoma State University, Stillwater OK
E-mail: goutam.chakraborty@okstate.edu

Goutam Chakraborty is a professor of marketing and founder of SAS and OSU data mining certificate and SAS and OSU business analytics certificate at Oklahoma State University. He has published in many journals such as *Journal of Interactive Marketing, Journal of Advertising Research, Journal of Advertising, Journal of Business Research*, etc. He has chaired the national conference for direct marketing educators for 2004 and 2005 and co-chaired M2007 data mining conference. He is also a Business Knowledge Series instructor for SAS.

APPENDIX

SAS Code Utility Node

Appends a Chi-squared statistic to the Association Node's core output. SGNFCNT = 1 indicates statistical significance at the level specified in the code.

```
*\\ Set macro variable transN to total number of market baskets (or
transactions);
proc sql noprint;
       select count(distinct(%EM_ID))
       into :transN
       from &EM_IMPORT_TRANSACTION
       ;
quit;

*\\ Create a dataset that will add Chi Square statistic to existing Rules;
data work.ChiSquare;
       *\\ Start with Association Node Rules output table;
       set &EM_IMPORT_RULES;

       *\\ add Chi Square statistic to each rule;
       CHISQ = .;  *missing in case of divide by zero scenario;
       PVALUE = .;
       if NOT (CONF=SUPPORT or LIFT=CONF) then do;
              CHISQ = (&transN * (LIFT-1)**2)*((SUPPORT/100) * (CONF/100))
                           /
                           ((CONF/100 - SUPPORT/100) * (LIFT - CONF/100));
              PVALUE = 1-Probchi(CHISQ,1);
       end;

       *\\ keep desired columns;
       keep index rule exp_conf conf support lift chisq pvalue;

run;

proc sort data=work.ChiSquare;
  by descending lift;
run;

proc print;
run;

*\\ The Chi Square table is currently in the temporary Work library;
*\\ The following code saves the table to YOUR PROJECT LIBRARY;
* data <yourPrjLibName>.ChiSquare;
*        set work.ChiSquare;
* run;
```