

Paper 132-2007

Mining Transaction/Order Data Using SAS® Enterprise Miner™ Association Node

Xinli Bao, Ph.D., Visa U.S.A. Inc.

ABSTRACT

Among all the products SAS provides, Enterprise Miner (EM) is one that not only makes it easy for an SAS neophyte to build models, but also provides a rich set of tools flexible enough for advanced SAS users.

This paper focuses on the EM association node in particular. This node provides useful analysis procedures for merchants, such as retailers or publishers who have large amounts of transaction or order data available. Mining from them can be a daunting task. Yet it can give valuable insights into which products are most likely to be purchased/ordered together by customers. When transaction data from multiple merchants are available, the association node can look at the customers of a particular merchant and discover which other merchants they also frequently shop at.

This paper will first explain the EM association node briefly and demonstrate its power in analyzing transaction/order data. Then it will present methods one can utilize to customize existing procedures, so as to fit one's special analytical needs. Finally, this paper will explain how to compute a statistic that shows the significance of lift values generated by the association node. This statistic is necessary for determining how much one can trust the lift when the quantity in the category is very small.

This paper requires readers to have some knowledge of SAS Enterprise Miner.

INTRODUCTION

The EM association node is designed to perform association discovery and sequence discovery. Association discovery is the identification of things that happen together in a given event. For example, it identifies what products are purchased together by a customer in a transaction. This technique—identification of products which are purchased together—is also known as market basket analysis.

Sequence discovery is the identification of events that happen in a sequence. In other words, sequence discovery discovers the relationship between the occurrence of event A and the occurrence of event B next time.

The association node requires the input data set to have at least two variables: one has an ID role and the other one has a target role for association discovery. For sequence discovery, a third variable, which identifies the sequence of events, is necessary.

Table 1: SAS Data Set "ASSOC"



CUSTOMER	TIME	PRODUCT
1.0	5.0	olives
1.0	6.0	corned_b
2.0	0.0	avocado
2.0	1.0	cracker
2.0	2.0	artichok
2.0	3.0	heineken
2.0	4.0	ham
2.0	5.0	turkey
2.0	6.0	sardines
3.0	0.0	olives
3.0	1.0	bourbon
3.0	2.0	coke
3.0	3.0	turkey
3.0	4.0	ice_crea

Table 1 shows a portion of a SAS sample data set "ASSOCS" that comes with Enterprise Miner. Any data set having similar structures can be analyzed by the association node.

EM provides multi-way association analysis, meaning multiple items can be associated together with multiple items. The association node outputs the number of rules specified by the user. Table 2 gives the output of a 3-way analysis. The first record is a rule indicating that 44.93% of those who purchased steaks also purchased corned_b and apples. The Column 6 transaction count gives the number of customers who purchased the three items together, for example, 102 customers purchased steak, corned_b and apples together. The level of support column gives the percentage of customers, out of all customers, who purchased these three items together. Using the first record as an example, 10.19% of all customers purchased steaks, corned_b and apples together.

Column 3 Confidence in table 2 measures the percentage of an item/event occurs given the condition that another item/event occurs. For example, out of all those who have purchased steaks, 44.93% of them also purchased corned_b and apples. The expected confidence is the percentage of people who have purchased corned_b and apples out of all customers. Since 15.08% of all customers have purchased corned_b and apples, we would expect that about 15.08% of those who have purchased steaks will also purchase corned_b and apples. In fact, there are 44.93% of them purchased corned_b and apples. Thus the lift is the confidence divided by the expected confidence, that is, $44.93\%/15.08\%=2.98$. The lift measures the strength of association.

Sequence discovery is similar to association discovery. The result of sequence discovery is a set of association rules that can reveal sequential patterns. Like a multi-way association discovery, EM also does multi-chain sequence discovery. EM requires a third variable with a role of "Sequence" in order to perform the sequence discovery. Set the "Time" variable in the sample data set to a role of sequence. Table 3 shows a portion of a 3-chain sequence report. The interpretation of the numbers is similar to the association report except that it reflects the sequence of event occurrences. For example, for the first record, 69.06% of those who purchased cracker have purchased Heineken at a subsequent visit. Out of all customers, 33.67% made cracker => Heineken purchase and the transaction count is 337.

Table 2: Output of 3-way Association Discovery

Association Report						
Relations	Expected Confidence (%)	Confidence (%)	Support (%)	Lift	Transaction Count	Rule
3	15.08	44.93	10.19	2.98	102.00	steak ==> corned_b & apples
3	22.68	67.55	10.19	2.98	102.00	corned_b & apples ==> steak
3	29.57	87.86	12.29	2.97	123.00	ice_crea & chicken ==> coke
3	13.99	41.55	12.29	2.97	123.00	coke ==> ice_crea & chicken
3	29.57	87.42	13.19	2.96	132.00	sardines & ice_crea ==> coke
3	15.08	44.59	13.19	2.96	132.00	coke ==> sardines & ice_crea
3	29.57	86.67	11.69	2.93	117.00	sardines & chicken ==> coke
3	13.49	39.53	11.69	2.93	117.00	coke ==> sardines & chicken
3	31.27	89.80	13.19	2.87	132.00	sardines & coke ==> ice_crea

Table 3: Output of 3-chain Sequence Discovery

Sequence Report								
Chain Length	Transaction Count	Support (%)	Confidence (%)	Rule	Chain Item 1	Chain Item 2	Chain Item 3	
2	337	33.67	69.06	cracker ==> heineken	cracker	heineken		
2	235	23.48	48.35	hering ==> heineken	hering	heineken		
2	233	23.28	49.26	olives ==> bourbon	olives	bourbon		
2	229	22.88	47.12	hering ==> corned_b	hering	corned_b		
2	226	22.58	46.50	hering ==> olives	hering	olives		
2	225	22.48	57.40	baguette ==> heineken	baguette	heineken		
2	220	21.98	69.18	soda ==> cracker	soda	cracker		
2	220	21.98	56.12	baguette ==> hering	baguette	hering		
2	220	21.98	46.51	olives ==> turkey	olives	turkey		
2	218	21.78	68.55	soda ==> heineken	soda	heineken		
3	218	21.78	99.09	soda ==> cracker ==> heineken	soda	cracker	heineken	
2	217	21.68	73.31	coke ==> ice_crea	coke	ice_crea		

CUSTOMIZATION

The analysis that the association node provides is excellent. However, sometimes it might be of interest to know the association rules for one particular item/product with all the rest. For example, one might be only interested in the association rules related to coke. That is, the relationships between all other products, such as those between steak and apples are not subjects of interest. The association node still needs to carry out the calculations of all association rules before one can single out the coke rules that are of interest. It does not provide a condition in its property sheet to specify or to restrict the rules to calculate. This might not be a problem if the transaction data set is small.

However, it is very common for a transaction data set to have billions of records. Fortunately, SAS provides a SAS code utility node that allows users to write SAS code and customize the analysis procedures and results. EM also outputs some very useful data sets in the project directory for analysts to utilize.

How can one make EM only calculate the association rules between an item (e.g. coke) and the others? It can be implemented using the sequence discovery process.

First, create a new data set or alter the original data set to make it contain at least the following three variables:

1. A unique ID variable, e.g. Customer; Set its role as "ID".
2. A target variable, e.g. Product; Set its role as "Target".
3. An artificial variable and set its value to 1 if the item is coke and 2 for all others; Set its role as "Sequence".

Then perform a sequence discovery analysis. Use the maximum item sub-property under the association property to specify the number of items that you would like to include in the chain.

The result of a 2-chain sequence with maximum item of 1 is shown in table 4. It gives the association rules between coke and all other individual items. An artificial variable called "VISIT" is added to the original data set and set to the role of "Sequence". A portion of the revised data set is shown in table 5. Notice that the frequency an item is purchased or purchased together with other items is not changed in the new data set. All items purchased by a customer are still treated as being together, except that each item retains the timestamp indicating when it was purchased. The sequence discovery process essentially follows the same procedures for finding groups of items that appear together as the association discovery process does.

Table 4: Output of 2-chain Sequence Discovery with 1 Maximum Item

Sequence Report						
Transaction Count	Support (%)	Confidence (%)	Rule	Chain Item 1	Chain Item 2	
220	21.98	74.32	coke ==> ice_crea	coke	ice_crea	
174	17.38	58.78	coke ==> heineken	coke	heineken	
148	14.79	50.00	coke ==> olives	coke	olives	
147	14.69	49.66	coke ==> sardines	coke	sardines	
140	13.99	47.30	coke ==> bourbon	coke	bourbon	
139	13.89	46.96	coke ==> chicken	coke	chicken	
119	11.89	40.20	coke ==> turkey	coke	turkey	
75	7.49	25.34	coke ==> cracker	coke	cracker	
71	7.09	23.99	coke ==> apples	coke	apples	
68	6.79	22.97	coke ==> hering	coke	hering	
66	6.59	22.30	coke ==> baguette	coke	baguette	
63	6.29	21.28	coke ==> ham	coke	ham	
62	6.19	20.95	coke ==> corned_b	coke	corned_b	
55	5.49	18.58	coke ==> peppers	coke	peppers	
53	5.29	17.91	coke ==> artichok	coke	artichok	
52	5.19	17.57	coke ==> soda	coke	soda	
46	4.60	15.54	coke ==> avocado	coke	avocado	
42	4.20	14.19	coke ==> steak	coke	steak	
23	2.30	7.77	coke ==> bordeaux	coke	bordeaux	

Table 5: The Data Set with the Added Variable

CUSTOMER	TIME	PRODUCT	VISIT
1.0	5.0	olives	2.0
1.0	6.0	corned_b	2.0
2.0	0.0	avocado	2.0
2.0	1.0	cracker	2.0
2.0	2.0	artichok	2.0
2.0	3.0	heineken	2.0
2.0	4.0	ham	2.0
2.0	5.0	turkey	2.0
2.0	6.0	sardines	2.0
3.0	0.0	olives	2.0
3.0	1.0	bourbon	2.0
3.0	2.0	coke	1.0
3.0	3.0	turkey	2.0
3.0	4.0	ice_crea	2.0

Table 6: Sample Code to Calculate Expected Confidence, Lift and Z-Score

```

proc sort data=EMWS1.ASSOC_LINKS out=sbaodata.coke2;
by TO;
run;
data _NULL_;
set EMWS1.ASSOC_ASSOC;
if SET_SIZE=0 then call symput ('EM_NUM_ID', COUNT);
if ITEM1='coke' then call symput ('ANALYFREQ',COUNT);
run;
data sbaodata.cokeassoc;
merge EMWS1.ASSOC_ASSOC sbaodata.coke2 (rename=(TO=ITEM COUNT=CO_OCCURCT));
by ITEM;
where ITEM ^= 'coke';
EXPCONF=COUNT/&EM_NUM_ID*100;
LIFT=CONF/EXPCONF;
pstar=(CO_OCCURCT+&ANALYFREQ) / (COUNT+&EM_NUM_ID);
qstar=1-pstar;
Z_SCORE=(CONF-
EXPCONF)/sqrt(pstar*qstar*(COUNT+&EM_NUM_ID)/(COUNT*&EM_NUM_ID))/100;
drop pstar qstar LINKID;
run;
title 'Association of Coke with Other Products';
proc print data=sbaodata.cokeassoc;
run;

```

However, the output of the sequence discovery does not have the expected confidence and lift statistics as the association discovery does. To obtain these statistics, we have to compute them by ourselves. EM outputs several tables that contain the intermediate results. One of them is called `assoc_assoc`. It contains a count of all distinct IDs and counts of distinct IDs by Target. Another table is called `assoc_links`, which contains all columns printed out in the sequence report. Utilizing the information in these two tables can produce the expected confidence and lift. It is also possible to calculate other interesting statistics, such as a Z_Score. A Z-Score indicates whether the confidence is significantly better or worse than the expected confidence. Roughly speaking, a score greater than 2 means significantly better and a score less than -2 means significantly worse. A sample code to calculate the expected confidence, lift and z-scores is provided in table 6.

We can type the code into a SAS code node and connect it with the association node as shown in figure 1. The result of running the SAS code in figure 1 is shown in table 7.

Figure 1: EM Diagram



Table 7: Output of Sequence Discovery for Product Coke

Association of Coke with Other Products									
Obs	ITEM	COUNT	FROM	CO_OCCURCT	SUPPORT	CONF	EXPCONF	LIFT	Z_SCORE
1	apples	314	coke	71	7.09	23.99	31.3686	0.76466	-2.5444
2	artichok	305	coke	53	5.29	17.91	30.4695	0.58765	-4.3411
3	avocado	363	coke	46	4.60	15.54	36.2637	0.42854	-7.8036
4	baguette	392	coke	66	6.59	22.30	39.1608	0.56938	-6.4536
5	bordeaux	74	coke	23	2.30	7.77	7.3926	1.05109	0.0686
6	bourbon	403	coke	140	13.99	47.30	40.2597	1.17480	2.5781
7	chicken	315	coke	139	13.89	46.96	31.4685	1.49227	5.0974
8	corned_b	391	coke	62	6.19	20.95	39.0609	0.53624	-6.9496
9	cracker	488	coke	75	7.49	25.34	48.7512	0.51974	-9.8046
10	ham	305	coke	63	6.29	21.28	30.4695	0.69853	-3.1458
11	heineken	600	coke	174	17.38	58.78	59.9401	0.98071	-0.4918
12	hering	486	coke	68	6.79	22.97	48.5514	0.47317	-10.7603
13	ice_crea	313	coke	220	21.98	74.32	31.2687	2.37695	13.6141
14	olives	473	coke	148	14.79	50.00	47.2527	1.05814	1.0732
15	peppers	296	coke	55	5.49	18.58	29.5704	0.62837	-3.7386
16	sardines	296	coke	147	14.69	49.66	29.5704	1.67945	6.4035
17	soda	318	coke	52	5.19	17.57	31.7682	0.55299	-5.0057
18	steak	227	coke	42	4.20	14.19	22.6773	0.62570	-2.5852
19	turkey	283	coke	119	11.89	40.20	28.2717	1.42201	3.7891

CONCLUSION

Mining your transaction/order data can provide lots of benefits for making business decisions. For example, strongly-associated products can be recommended to customers who purchased a particular product; customers of one merchant can be contacted for special shopping offers by other strongly-associated merchants. SAS Enterprise Miner makes the discovery of the association rules between products or services easy and quick. Methods are also available for customizations of the existing processes.

In summary, we hope this paper can benefit other SAS users and analysts who need to mine very large transaction data sets.

CONTACT INFORMATION

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

Xinli Bao
 Visa U.S.A. Inc.
 P.O. Box 8999
 San Francisco, CA 94128-8999
 Email: sbao@visa.com or baoxinli@gmail.com

REFERENCES

1. Matthew Redlon. "A SAS Market Basket Analysis Macro: The 'Poor Man's Recommendation Engine'". *Proceedings of the twenty-eighth Annual SAS Users Group International Conference*. April 2003.
 2. SAS Enterprise Miner™ Help Menu
- SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries, ® indicates USA registration.
 Other brand and product names are trademarks of their respective companies.