ANNALS of the ORADEA UNIVERSITY.

Fascicle of Management and Technological Engineering, Volume VII (XVII), 2008

ASSOCIATION AND SEQUENCE ANALYSIS IN BUSINESS

Simona Aurelia Bodog
University of Oradea
Faculty of Electrical Engineering and Information Technology
sbodog@uoradea.ro
Alexandru Constăngioară, Mirela Bucurean
University of Oradea, Faculty of Economic Science
sandu_oradea@yahoo.com

Key words: association analysis, sequence analysis, business application.

Abstract: Forbes (Palmeri 1997) reported that a major retailer has determined that customers who buy Barbie dolls have a 60% likelihood of buying one of three types of candy bars. The confidence of the rule Barbie ⇒ candy is 60%. The retailer was unsure what to do with this nugget. The online newsletter Knowledge Discovery Nuggets invited suggestions (Piatesky-Shapiro 1998). This paper focuses on association and sequence analysis and their relevance for business world.

1.Introduction.

For each transaction, there is a list of items. Typically, a transaction is a single customer purchase and the items are the things that were bought. An association rule is a statement of the form (item set A) ⇒ (item set B).

The aim of the analysis is to determine the strength of all the association rules among a set of items. The strength of the association is measured by the support and confidence of the rule. The support for the rule A ⇒ B is the probability that the two item sets occur together. Note that support is reflexive. The confidence of an association rule A ⇒ B is the conditional probability of a transaction containing item set B given that it contains item set A. The lift of the rule A ⇒ B is the confidence of the rule divided by the expected confidence, assuming the item sets are independent. The lift can be interpreted as a general measure of association between the two item sets. Values greater than 1 indicate positive correlation; values equal to 1 indicate zero correlation; and values less than 1 indicate negative correlation. Note that lift is reflexive.

Of course, the issue is how we can interpret association. Does a strong rule imply causality? The answer is no. Given a strong rule, correlation is not mandatory as shown in the following example. Suppose there are two classes of customers, as in Table 1.

Table 1. Strong rule and negative correlation

	CHECKING ACCOUNT		
SAVING ACCOUNT		NO	YES
	NO	500	3500
	YES	1000	5000

Source: Palmeri, 1997

In the above example the rule Saving Account ⇒ Checking Account is strong having 50% support and 83 % confidence. However the two variables are negatively correlated since

not having a saving account results in a superior probability for having a checking account. One can see that the lift for our rule is 83% /85% so less the one.

2. Association analysis.

A bank wants to examine its customer base and understand which of its products individual customers own in combination with one another. It has chosen to conduct a market-basket analysis of a sample of its customer base. The bank has a data set that lists the banking products/services used by 8211 customers. The data set has over 32,000 rows. Each row of the data set represents a customer-service combination. Therefore, a single customer can have multiple rows in the data set, each row representing one of the products he or she owns. The median number of products per customer is three.

The support is the percentage of customers who have all the services involved in the rule. For example, approximately 54% of the 7,991 customers have a checking and savings account and approximately 25% have a checking account, savings account, and an ATM card. Similarly Confidence column gives the conditional probability of a customer having B given that it has A. For first observation wee see that a customer who has saving account has a probability of 87.5 % of having a checking account. Among those customers that have both a savings account and a credit card, over 97% have a checking account.

Lift is the third defining a rule. For first observation we see that a customer having saving account is 1.02 more likely to have a checking account than a random customer. The lift of the relationship CKCRD ==> CCRD is 3.19. Therefore, if you select a customer who has a check/debit card, the relative frequency of that customer having a credit card is more than 3 times higher than an individual chosen at random.

Table 2 Rules identified in the dataset

	Relations	Lift	Support(%)	Confidence(%)	Transaction Count	
1	2	1.02	54.17	87.56	4329.0	SVG ==> CKING
2	2	1.02	54.17	63.15	4329.0	CKING ==> SVG
3	2	1.10	36.19	42.19	2892.0	CKING ==> ATM
4	2	1.10	36.19	94.11	2892.0	ATM ==> CKING
5	2	1.08	25.69	41.53	2053.0	SVG ==> ATM
6	2	1.08	25.69	66.81	2053.0	ATM ==> SVG
7	2	1.17	16.47	100.00	1316.0	HMEQLC ==> CKING
8	2	1.17	16.47	19.20	1316.0	CKING ==> HMEQLC
9	2	1.04	15.72	25.40	1256.0	SVG ==> CD
10	2	1.04	15.72	64.08	1256.0	CD ==> SVG
11	2	1.04	15.58	89.31	1245.0	MMDA ==> CKING
12	2	1.04	15.58	18.16	1245.0	CKING ==> MMDA
13	2	1.12	14.85	17.32	1187.0	CKING ==> CCRD
14	2	1.12	14.85	95.96	1187.0	CCRD ==> CKING
15	2	1.17	11.30	13.17	903.00	CKING ==> CKCRD
16	2	1.17	11.30	100.00	903.00	CKCRD ==> CKING
17	2	1.09	11.15	18.02	891.00	SVG ==> HMEQLC
18	2	1.09	11.15	67.71	891.00	HMEQLC ==> SVG
19	2	1.07	10.22	16.53	817.00	SVG ==> CCRD
20	2	1.07	10.22	66.05	817.00	CCRD ==> SVG

Suppose instead that you are particularly interested in those associations that involve automobile loans. You would like to create a subset of the rules, to include only those rules with the product AUTO. Results are shown in Table 3.

Table3. Automobile loan

	Relations	Lift	Support(%)	Confidence(%)	Transaction Count	Rule
1	2	1.07	8.52	91.78	681.00	AUTO ==> CKING
2	2	1.07	6.14	66.17	491.00	AUTO ==> SVG
3	2	1.25	4.47	48.11	357.00	AUTO ==> ATM
4	3	1.19	5.97	64.29	477.00	AUTO ==> SVG & CKING
5	3	1.30	4.35	46.90	348.00	AUTO ==> CKING & ATM

There are 5 rules involving automobile loans. We see that support is week although confidence for some of them is high. Lift indicates positive correlation although week in most cases.

3. Sequence analysis.

Beside association in business is important to know the sequence of the transaction. In the following I have employed a sequence analysis using the same dataset. The transaction count is the total number of customers that have purchased products in this order. The percent support is the transaction count divided by the total number of customers, which would be the maximum transaction count. The percent confidence is the transaction count divided by the transaction count for the left side of the sequence. So, for example, of the customers that got an automobile loan, 82.21% got a second automobile loan.

Table 4 Sequence rules

	Chain Length	Support(%)	Confidence(%)	Transaction Count	Rule
1	2	54.17	63.15	4329	CKING ==> SVG
2	2	36.19	42.19	2892	CKING ==> ATM
3	2	25.69	41.53	2053	SVG ==> ATM
4	2	21.39	55.61	1709	ATM ==> ATM
5	2	20.99	24.46	1677	CKING ==> CD
6	2	16.47	19.20	1316	CKING ==> HMEQLC
7	2	15.72	25.40	1256	SVG ==> CD
8	2	15.58	18.16	1245	CKING ==> MMDA
9	2	14.85	17.32	1187	CKING ==> CCRD
10	2	13.58	21.95	1085	SVG ==> SVG
11	2	11.30	13.17	903	CKING ==> CKCRD
12	2	11.30	100.00	903	CKCRD ==> CKCRD
13	2	11.15	18.02	891	SVG ==> HMEQLC
14	2	10.22	16.53	817	SVG ==> CCRD
15	2	9.21	37.55	736	CD ==> CD
16	2	8.82	10.28	705	CKING ==> IRA
17	2	8.81	53.50	704	HMEQLC ==> HMEQLC
18	2	8.53	22.19	682	ATM ==> HMEQLC
19	2	8.52	9.93	681	CKING ==> AUTO
20	2	7.97	12.88	637	SVG ==> CKCRD

Suppose you are only interested in those sequences that involve automobile loans. Results are presented in Table 5.

Table 5 Sequence rules for automobile loans

	Chain Length	Support(%)	Confidence(%)	Transaction Count	Rule
1	2	8.52	9.93	681	CKING ==> AUTO
2	2	7.63	82.21	610	AUTO ==> AUTO
3	2	6.14	9.93	491	SVG ==> AUTO
4	2	4.47	11.62	357	ATM ==> AUTO
5	2	2.54	10.36	203	CD ==> AUTO
6	2	2.53	15.35	202	HMEQLC ==> AUTO
7	2	2.21	14.31	177	CCRD ==> AUTO
8	3	6.87	80.62	549	CKING ==> AUTO ==> AUTO
9	3	5.97	11.02	477	CKING ==> SVG ==> AUTO
10	3	4.87	79.23	389	SVG ==> AUTO ==> AUTO
11	3	4.47	100.00	357	ATM ==> AUTO ==> AUTO
12	3	4.35	12.03	348	CKING ==> ATM ==> AUTO
13	3	3.30	12.86	264	SVG ==> ATM ==> AUTO
14	3	2.53	15.35	202	CKING ==> HMEQLC ==> AUTO
15	3	2.40	11.45	192	CKING ==> CD ==> AUTO
16	3	2.18	14.66	174	CKING ==> CCRD ==> AUTO
17	3	2.10	82.76	168	CD ==> AUTO ==> AUTO

Thus one can easily identify the sequence of transactions involving automobile loans and the strength of the rules measured by all three dimensions.

References.

1. Piatetsky-Shapiro, G. (1991), Discovery, analysis, and presentation of strong rules, in G. Piatetsky-Shapiro & W. J. Frawley, eds, 'Knowledge Discovery in Databases', AAAI/MIT Press, Cambridge, MA;
2. R. Agrawal; T. Imielinski; A. Swami: Mining Association Rules Between Sets of Items in Large Databases", SIGMOD Conference 1993: 207-216;
3. Jochen Hipp, Ulrich Güntzer, and Gholamreza Nakhaeizadeh. Algorithms for association rule mining - A general survey and comparison. SIGKDD Explorations, 2(2):1-58, 2000;
4. Jian Pei, Jiawei Han, and Laks V.S. Lakshmanan. Mining frequent itemsets with convertible constraints. In Proceedings of the 17th International Conference on Data Engineering, April 2 - 6, 2001, Heidelberg, Germany, pages 433-442, 2001.