Web: www.witpress.com Email witpress@witpress.com
Paper from: *Data Mining III*, A Zanasi, CA Brebbia, NFF Ebecken & P Melli (Editors).
ISBN 1-85312-925-9

Detecting sequential patterns for cross-selling fast moving consumer goods

G. Verstraeten, D. Van den Poel, A. Prinzie & P. Van Kenhove
Department of Marketing, Ghent University, Belgium.

Abstract

In the marketing domain, sequential patterns have been usefully deployed for predicting various aspects of customer purchase behavior. However, to date, the applications of the technique have mainly focused on improving algorithms for detecting sequentially related events, whereas the implications of the sequences, and their incorporation into a global structure of consecutive sequences have been treated to a lesser extent. In this paper, such a structure, that we will refer to as *sequential architecture*, will be empirically investigated for a specific case in a fast moving consumer goods setting. Hence, the goal of this paper was to introduce a new concept that might prove to be a relevant tool for marketing decision making rather than offering a sound solution within a clearly demarcated problem definition.

As opposed to the traditional sequence-analysis approaches, in this study, an array of binary logit analyses was applied for detecting significant sequences among category purchases. We use the output of the logit analyses to define the category that is most significantly influenced per newly purchased category, and we select these links for constructing the applicable sequential architecture. Finally, we provide empirical evidence that the methodology suggested is able to double the performance of predicting purchases in categories that were not purchased previously by the consumer, compared to a random model.

In summary, we have shown that (i) binary logit analysis provides a feasible alternative for detecting and selecting highly significant sequential relationships, (ii) a sequential architecture can be successfully compiled through the methodology offered in this paper, and (iii) the provided sequential architecture can be a useful tool in understanding and predicting customer behavior.

Future applications possibly lie ahead in the field of inter-category management, shelf-space allocation, store-layout decisions, retailer promotions, customer profiling and individual customer predictions.

Paper from: *Data Mining III*, A Zanasi, CA Brebbia, NFF Ebecken & P Melli (Editors).
ISBN 1-85312-925-9

1 Introduction

The analysis of sequential patterns is a widely researched topic in a vast amount of disciplines, including genetics, chemistry, archeology, history, psychology, sociology and economics. In the marketing domain, sequential patterns have been usefully deployed for estimating future purchase behavior, where they optimally serve as decision aid for implementing cross-selling strategies. However, while sequence analysis has generated a lot of attention from scientists, the application of the technique has mainly resulted in improved algorithms (see, e.g. [1]) for detecting sequentially related events (such as the purchasing of products), whereas the implications of the sequences, and their incorporation into a global structure of consecutive sequences have been treated to a lesser extent. In this paper, this structure, that we will refer to as *sequential architecture*, will be investigated for a fast-moving consumer goods setting.

As a basic starting point of sequential architecture, among other researchers, Kasulus et al. [2] analyzed through a Guttman scalogram analysis [3] that the order of acquiring financial assets is relatively stable across different age cohorts. In their analysis, they link the pattern to the existence of strong societal norms for acquiring e.g. a life insurance before buying bonds. The Guttman scalogram embraces the assumption that a cumulative, one-dimensional scale exists on which the categories offered can be ranked. Although this approach has been refined over time to accommodate latent class/latent trait methods that can hold different alternative rankings for different segments (see, e.g. [4], [5]), it is a priori unknown to what extent this covers the reality in a fast-moving consumer goods setting. Thus, while in non-fast moving consumer goods, a priority pattern of acquisition has been investigated, such a one-dimensional pattern might prove inappropriate to accommodate all fast moving consumer goods that are e.g. sold in a retail outlet. The main goal of this paper is to offer an approach that broadens the topic to a different angle. Hence, this paper is rather intended towards expanding the focus of the analysis than offering a sound solution within a clearly demarcated problem definition. Therefore, the focus of this analysis is more descriptive rather than predictive.

We will argue, however, that the applications that could result from extended refinements of the approach described here, might refuel a number of domains within marketing such as inter-category management, promotional strategy, customer profiling and individual customer predictions.

2 Methodology

Throughout this paper, we define a *period* as a single purchase occasion (i.e. a 'ticket'), containing often several purchased items (i.e. 'ticket lines'). A category *opening* is defined in this paper as a purchase in a category that was not purchased by the customer in a previous period.

As opposed to the traditional sequence-analysis approaches, in this paper, we apply an array of binary logit analyses for the same goal. Undoubtedly, logistic regressions are one of the most frequently used techniques in modeling purchase

Paper from: *Data Mining III*, A Zanasi, CA Brebbia, NFF Ebecken & P Melli (Editors).
ISBN 1-85312-925-9

behavior. While the binary logit is most often used in repeat-purchase modeling (i.e. will the consumer repurchase in a given period, see, e.g. [6]), multinomial logit is a widespread technique for predicting cross-selling behavior (to determine which category has the highest probability of being purchased next, see, e.g. [7]). In the application of both techniques for modeling purchase behavior, very often a rich array of behavioral variables is created based on the customers' individual purchase history to improve the predictive performance. Contrarily to these attempts, since the approach in this paper is descriptive rather than predictive (cf. supra), we limit ourselves to the restricted set of predictors that is relevant for determining sequence architectures.

In our approach, we will apply a different binary logistic regression - performed by a maximum likelihood estimation - for each product category separately. We define the (binary) dependent variable as whether the customer has opened this category (1 or 0) in a period where he or she has opened at least one product category. We define the independent variables as an array of binary variables 1 to n where n represents the total number of categories investigated. Each of these predictors measures whether the category in case has been opened by the same customer in the previous period.

As it is plausible that a category opening is significantly influenced by different category openings, and that this category significantly influences multiple category openings, in this first attempt, we will restrict the analysis in order to ensure a tractable, intuitively comprehensible architecture of sequences. The outcome of the models will serve to indicate the product category that is most significantly influenced by the category in case. In this way, we introduce the constraint that a product category can influence one and just one other category, provided that the link between the categories is significant. Hence, if we define a significance matrix as a matrix containing the p-values of the

Table 1: Significance matrix of the binary logit analyses.

	X_1	X_2	...	X_j	...	X_n
Y_1	.	s_{12}	...	s_{1j}	...	s_{1n}
Y_2	s_{21}	.	...	s_{2j}	...	s_{2n}
...	...	...		...		...
Y_i	s_{i1}	s_{i2}	...	s_{ij}	...	s_{in}
...	...	...		...		...
Y_n	s_{n1}	s_{n2}	...	s_{nj}	...	.

explanatory variables (as a result of the Wald test for individual parameters), being all possible preceding categories X per predicted category Y (i.e. one line per binary logit analysis), it is rather straightforward to find that category Y_i that is most significantly influenced by the product category X_j in case (see Table 1). For category j, it is the category i for which $s_{ij} = \min[s_{1j}, s_{2j},..., s_{ij},...,s_{nj}]$.

By definition, when each category is bound to influence strictly one other category, at least one loop must be existing in the data. A priori, a large variety

Paper from: *Data Mining III*, A Zanasi, CA Brebbia, NFF Ebecken & P Melli (Editors).
ISBN 1-85312-925-9

of sequential architectures might offer the result to this analysis. Without pretending to sketch an exhaustive overview, some alternatives (for an exemplary situation containing only 8 product categories, numbered 1 to 8) are graphed in the overview presented in Figure 1.

Note that the Guttmann scalogram [3], that was discussed previously in the introduction, can be represented as a loop containing all categories (see Figure 1, architecture (a)), in which at least one of the causal relationships is highly insignificant, while the other relationships are highly significant.

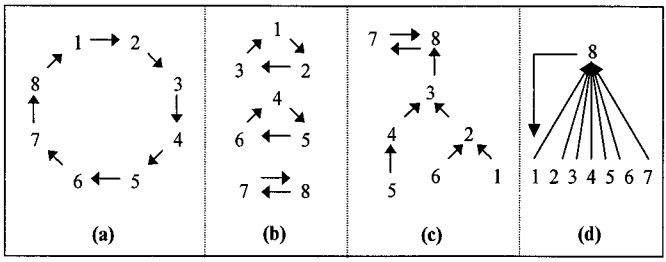

Figure 1: Examples of possible sequence architectures.

It should be clear that the outcome of the analysis will have a large impact on the applications that might result from this paper. For example, while architecture (b) might lead to a clustering of different product categories and a reengineering of meta-categories, architecture (c) might e.g. be used to indicate leading product categories as categories having a high amount of successors.

In a final effort in this paper, the resulting graph will be validated by applying it to suggest cross-selling actions to the customers included in the analysis. While it must be clear that this is not the ultimate goal of this research project, it will prove useful in assessing the validity of the approach offered, and might clear the path for a broader range of applications.

3 Design of the study

3.1 Description of the data

We conducted our research on a transaction database of a large Belgian retail chain. As the data available was spread over 10 months, starting from April 1st 2000 until January 31st 2001, the information was left censored (i.e. not all transactions of all customers were available from the beginning of their relationship with the company). Accordingly, it was also unclear when customers opened (i.e. started purchasing) a new category, because previous transactions were missing.

To overcome the problems according to left-censored data, all transaction occasions before a certain point in time were grouped per customer, to represent the history of regularly purchased items for the customer in case. Since the

Paper from: *Data Mining III*, A Zanasi, CA Brebbia, NFF Ebecken & P Melli (Editors).
ISBN 1-85312-925-9

database contained information about 4,143,841 purchase occasions of 242,605 customers, the mean number of tickets available per customer was approximately 17. Considering the fact that a certain amount of tickets was required for building the list of historical purchasing categories, we selected the 69,563 customers having an amount of purchase transactions larger than the mean number in the database. We used all data before period 11 for building the list of frequently purchased items on a customer basis. For periods after period 11 on which product category openings were performed by a certain customer, we predicted the opening of each category that was not opened at the time by the list of categories that were opened during the last period by the customer in case. In this effort, as described in the methodology, the information about the categories to be predicted was regrouped, so a separate model could be built per category.

Additionally, because some categories in the large purchase assortment were only bought occasionally, we have made a selection of the product categories used in this study. Hence, of the 140 available product categories, we selected the 80 categories with the highest purchase frequency to be included in the analysis. We refer to the Appendix of this paper for a complete list of the categories used in this study.

As a last step in the data preparation, we have assigned the customer list on a random basis to a training set and a test set, in order to validate the outcome of the logit analysis.

3.2 Results

3.2.1 Sequential architecture

The results of the model, in terms of the sequential architecture, were inspiring. Considering the threshold of a 0.05 significance limit, all models counted in between 4 and 20 significant predictive variables, with a mean value of 9.7. Alternatively, considering the same significance threshold, each independent variable was a significant predictor in between 2 and 20 models, again with a mean value of 9.7. Translated to our applications, this implies that all category openings are significantly influenced by previous category openings, and that all categories serve to predict other category openings. Hence, by applying our restriction that all categories are bound to influence just one other category, we were confident that all remaining influences were significant. This immediately has as a consequence that the Guttman scalogram [3] discussed previously did not present the optimal architecture, as it embraced the assumption that at least one of the remaining influences would be non-significant (cf. supra).

Following the methodology described above, we reached the sequential architecture presented in Figure 2.

Paper from: *Data Mining III*, A Zanasi, CA Brebbia, NFF Ebecken & P Melli (Editors).
ISBN 1-85312-925-9

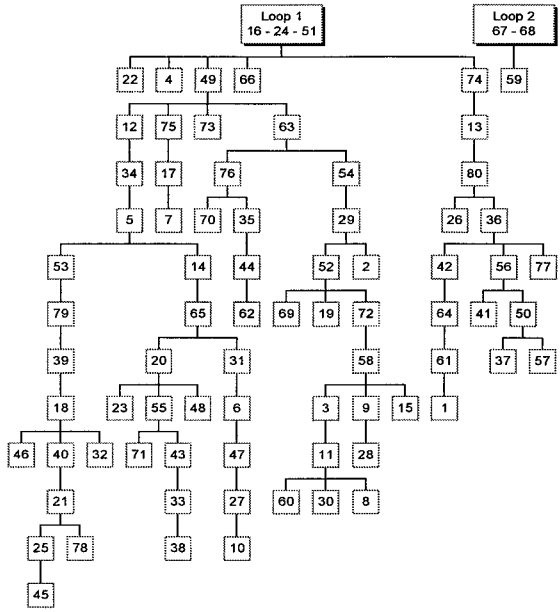

Figure 2: Sequential architecture applicable to the case study.

In Figure 2, the links between the categories imply that the lower category significantly influences the upper category. For example, starting on the lowest level, an opening of category 45 is predicted to lead to an opening of category 25, which in turn should lead to a future opening of category 21, etc. Roughly, the resulting architecture consists of 2 parts: one large tree of successive influences ending in a loop of three categories, and a separate 'tree' of only three categories, of which two form a loop at the end. As a small example, we extracted the structure of the second part (loop 2) of the sequential architecture in Figure 3.

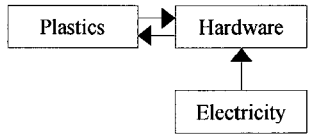

Figure 3: Detail of the sequential architecture.

Here, it is clearly shown that purchasing electricity-related products might result in purchasing hardware, which in turn influences the purchases of plastics and vice versa. It is, however important to notice that most sequential relationships cross the barriers of traditional meta-categories. For example, as opposed to this subpart, loop 1 consists of three mutually influencing categories, being pasta & rice, washing products and intimate hygiene products, which clearly belong to

Paper from: *Data Mining III*, A Zanasi, CA Brebbia, NFF Ebecken & P Melli (Editors).
ISBN 1-85312-925-9

different meta-categories. For the complete list of descriptions of the product category numbers in Figure 2, we refer again to the Appendix.

As a final conclusion about the proposed sequential architecture, we can state that the outcome of the array of binary logit analyses clearly favored architecture (c) in Figure 1, thus resulting in the appropriate implications (cf. infra).

3.2.2 Predictive performance

While we have already stated that predictive performance is not the major concern of this exploratory study, a basic amount of predictive capacity is required in order to validate the potential of the proposed sequential architecture. As a first insight, we assessed the predictive performance of the array of binary logit analyses by the benchmark of Morrison [8], who provided evidence that the correct benchmark (C_{pro}), from a marketing perspective, would consider the random probability of selecting an opening considering the given class distribution, as in eqn (1):

$$C_{pro} = \alpha^2 + (1 - \alpha)^2 \quad (1)$$

where

α is the proportion of cases belonging to class 1
$1 - \alpha$ is the proportion of cases belonging to class 2.

According to this formula, considering a class distribution of 96.02 % non-openings versus 3.98 % openings, the benchmark was defined as being 92.36 %. The actual accuracy of our model rose to 92.85 % on the test set, implying a slight but distinct improvement over the benchmark. Because these initial results only seem to imply a limited incremental advantage of our model above the benchmark, in the next paragraph, the results can be found for a validation of the findings in a more realistic setting, simulating a case where the sequential architecture would be used to predict the following category opening.

4 Validation of the findings

In this validation of the proposed sequential architecture, we will apply the pattern presented in Figure 2 for all customers in our sample, in order to predict future category openings at the individual level. As a first step in this exercise, we indicated the categories that have been purchased until period 11 in the database. This information was used to generate predictions for category openings in the following period during which at least one opening occurred. As all customers had invariably opened more than one category by that time, an array of possible future openings was presented per customer. For example, suppose customer Z had purchased categories 1, 41 and 60, according to the proposed sequential architecture (see Figure 2), the categories 61, 56 and 11 would be the product categories that have a high potential of being opened during the following period. In order to check the performance unambiguously, we selected one potential category out of the array of future categories that was relevant to the customer in case.

Paper from: *Data Mining III*, A Zanasi, CA Brebbia, NFF Ebecken & P Melli (Editors).
ISBN 1-85312-925-9

To this end, we computed the conditional probabilities of each link, representing the confidence of the links. This measure was formed by the percentage of times a category was opened in period $t+1$ given that its relevant preceding category was opened in period t. We thus selected the category that (i) belonged to the array of suggested openings considering all previous category openings of the customer in case and (ii) showed the largest confidence of being opened. We compared our prediction with the real openings, resulting in a binary variable that indicated whether the category proposed by the sequential architecture belonged to the array of product categories that was actually opened during the following period in which at least one opening occurred.

As a benchmark to this, we determined the random probability for selecting a product that would be opened during the next period. For this calculation, we used a ratio featuring (i) the number of remaining categories that were unopened by the customer during previous transactions, and (ii) the number of categories that were actually opened during the next period in which at least one opening occurred. To conclude our example, if the confidence of the link 1-60 was higher than the confidence of the links 41-56 and 60-11, the category 60 would have been selected as the category showing the largest potential for being opened by customer Z. Considering the remaining potential of the customer being 77 (i.e. the total number of categories in the analysis minus the number of previously opened categories by customer Z), and supposing that customer Z opened 3 product categories during the next period, the probability of selecting a correct product category by chance was defined as approximately 3.9 % (i.e. 3/77).

By performing this exercise for all customers, we were able to compare the relative frequency of selecting the correct category with the mean a-priori probability of selecting the correct category by a random procedure. In order to test the statistical difference of both measures, we use a bootstrap procedure [9]. To approximate the empirical distribution of our findings, we generated 100 bootstrap samples on a sample of 1500 randomly selected customers. We also calculated the difference between the random benchmark and the results

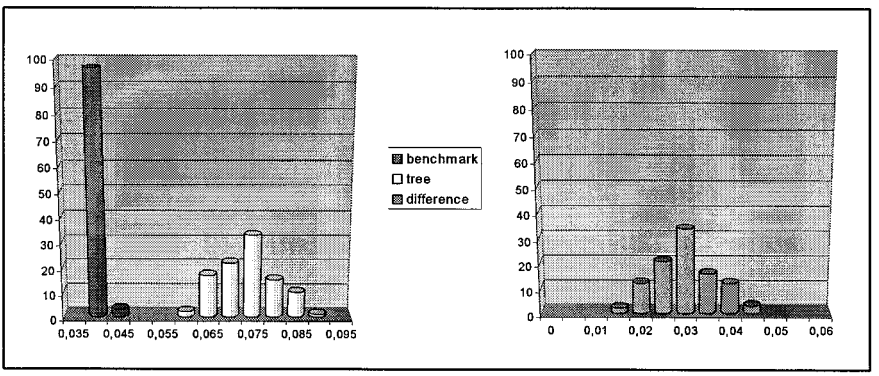

Figure 4: Bootstrapped performance of a random procedure versus a prediction based on the sequential architecture.

Paper from: *Data Mining III*, A Zanasi, CA Brebbia, NFF Ebecken & P Melli (Editors).
ISBN 1-85312-925-9

proposed by our model. The frequency distribution of this exercise is presented in Figure 4. Considering the highly significant difference between both measures ($p < 0.0001$), and the fact that the prediction using the computed sequential architecture outperforms a random estimation by a factor 2, it is clearly shown that the sequential architecture proposed in this paper improves the performance when predicting category openings compared to a random model. Hence, although we do not pretend to have offered an *optimal* solution to this topic, we have thus reached our goal to introduce a new concept that might prove to be a relevant tool for marketing decision making. These successive applications form the topic of the last paragraph in this paper.

5 Issues for further research

As a first improvement of the current findings, it is important to note that, in such an exploratory study, several decisions had to be taken without a proper benchmark being available in literature. The authors thus fully acknowledge the need for further research through extensive, yet time-consuming exploration of the impact of changing several parameters that were suggested throughout the methodology as well as the application. For example, opportunities arise in examining the extent to which the predictive modeling can be improved by gradually relaxing the restrictions on the causal influences, e.g. including the 2nd most significant link, … provided that these links are still significant. A second important improvement to the model might be accomplished by introducing a latent class approach, as it is unlikely that only one sequential architecture exists that describes the relevant structure of consecutive sequences for all customers. Thirdly, in order to assimilate the external validity of the findings, it would be necessary to reduplicate the study in different environments than this Belgian retail setting.

It is important to notice, however, that none of these alterations threaten the validity of this research study, because the main goal of this effort was to indicate and provide evidence for a potentially interesting gap in the literature rather than to offer a fully developed solution to the issue.

Potential utilization of the detected sequential architecture might refuel a wide range of applications within the marketing domain, including inter-category management, shelf space allocation, store layout decisions and retailer promotions, where the scheme can be deployed to explore information containing the product categories involved, customer profiling and individual customer predictions, after enriching the general scheme with individual customer behavioral information, and e.g. profiling of retail outlets based on the potential of its customers. However, it must be clear that, for each of these utilizations, a number of consecutive studies should be undertaken in order to determine the exact impact of the sequential architecture in the focused domain.

Paper from: *Data Mining III*, A Zanasi, CA Brebbia, NFF Ebecken & P Melli (Editors).
ISBN 1-85312-925-9

References

[1] Srikant, R. & Agrawal, R., Mining Sequential Patterns: Generalizations and Performance Improvements, *IBM Almaden Research Center paper*, 1996.
[2] Kasulus, J.J., Lush, R.F. & Stafford, E.F., Consumer Behavior in Accumulating Household Financial Assets. *Journal of Business Research*, 10, pp. 397-417, 1982.
[3] Guttman, L., The basis for scalogram analysis. *Measurement and prediction*, ed. S.A. Stouffer, Princeton University Press, NJ, pp. 60-90, 1950.
[4] Feick, L., Latent Class Models for the Analysis of Behavioral Hierarchies, *Journal of Marketing research*, 24, pp. 174-186, 1987.
[5] Kamakura, W., Ramaswami, S. & Srivastava, R., Applying latent trait analysis in the evaluation of Prospects for Cross-selling of Financial Services, *International Journal of Research in Marketing*, 8, pp. 329-349, 1991.
[6] Van den Poel, D., *Response Modeling for Database Marketing using Binary Classification*, PhD thesis, K.U. Leuven, 1999.
[7] Lilien, G.L., & Rangaswamy, A., Marketing Engineering, Computer-Assisted Marketing Analysis and Planning, Addison-Wesley: NY, 1998.
[8] Morrison, D.G., On the interpretation of discriminant analysis. *Journal of Marketing Research*, 6, pp. 156-163, 1969.
[9] Efron, B. & Tibshirani, R., *An Introduction to the Bootstrap*. Chapman and Hall: NY, 1993.

Appendix

1 Cheese	28 Bodycare	55 Dry fruit
2 Vegetables & potatoes	29 Juices	56 Prepared meals A
3 Yoghurt& soft white cheeses	30 Ice cream & desserts	57 Aperitives & cocktails
4 Fats	31 Confectionery	58 Sugar
5 Condiments & sauces	32 Canned fish & meat	59 Electricity
6 Biscuits	33 Wines	60 Petfood
7 Sliced cold meat	34 Haircare	61 Patisserie
8 Canned vegetables	35 Waters	62 Shavecare
9 The kitchen	36 Vegetables	63 Meat & fish
10 Milk UHT / fresh	37 Canned fruit	64 Prepared meals B
11 Bread	38 DSD Bakery	65 Strong alcohols
12 Ready to eat	39 Dental care	66 Insects & plants
13 Fruit	40 Facial care & make up	67 Plastics
14 Culinary aids	41 Bread preparation	68 Hardware
15 Aperitif snacks	42 Fries & potatoe products	69 Fire making products
16 Pasta & rice	43 Beers	70 Worldfood
17 Hot beverages	44 Shrimps	71 Tights
18 Soft drinks	45 Fish preparations	72 Greeting cards
19 The surfaces	46 Flakes breakfast	73 Cassettes & CD
20 Fresh juice	47 Dessert preparation	74 Syrups
21 Spreadables	48 Soups	75 Babycare
22 Ingredients	49 Stationary	76 Fresh poultry
23 Chocolate	50 Entrées and crustaceans	77 Sewing-thread & soles
24 Washing	51 Intimate hygiene	78 Champagnes & sparkling wines
25 Eggs	52 Parafarmacy	79 Maintenance / washing / ironing
26 Oils & vinegars	53 Tabledecoration	80 Cigarettes
27 The toilet	54 Candy bars	