Proceedings of the 6th INFORMS Workshop on Data Mining and Health Informatics (DM-HI 2011)
P. Qian, Y. Zhou, C. Rudin, eds.

A Dynamic Bayesian Network Approach to Market Basket Analysis

Bumsoo Kim
Department of Decision Sciences
George Washington University
Washington, DC
bk4498@gwu.edu

Srinivas Prasad
Department of Decision Sciences
George Washington University
Washington, DC
prasad@gwu.edu

Refik Soyer
Department of Decision Sciences
George Washington University
Washington, DC
Soyer@gwu.edu

Abstract

Researchers in multiple fields have been intrigued by the problem of market basket analysis, where different methods and tools are used to find meaningful associations in consumer purchase data. Although there are many different methods used in the literature in tackling this classic data mining problem, the fundamental goal in market basket analysis involves searching and analysing association rules. These associations could be some type of correlation relationships such as the sales of product A and B are positively related or in some cases may be causal relationships such as promotion efforts on product A increases sales on product A. In this study, we investigate the Bayesian Network approach to the market basket analysis problem and focus specifically on the marketing application of the method. We demonstrate how Bayesian networks can be used to identify and incorporate direct and indirect effects of promotion in marketing datasets.

Keywords: Bayes Networks, Promotion, Data Mining, Consumer Behavior

Introduction

Market basket analysis is a classic problem in data mining and has been studied extensively in the literature, as it is a clear case where data mining methods can be applied to aid business management in planning marketing and operation strategies. Although MBA is not a new idea, it is still an extremely relevant and interesting topic. In spite of the ease of accessibility to grocery stores and even with the recent expansion in online grocery shopping, consumers still plan and purchase multiple items over multiple categories of products in a single shopping trip. Furthermore, with the free distribution of various store cards (albeit with minimum commitment of first order personal information) coupled with improvements in information technology, retailers can now follow individual customers and gather transaction level data. This enables the retailer to analyze micro level consumer behavior as well as macro level product association patterns and devise an operational strategy to maximize store profits. In this paper, we highlight the computational difficulties faced by researchers in dealing with MBA datasets and the Bayesian network approach which was proposed

to overcome this problem[1]. In section 3, we briefly describe the cross-category effects in marketing literature and extend previous study by integrating the promotion variable to the MBA model. In section 4, we describe the data and show our findings. Finally, we conclude in section 5.

Bayesian network approach to Market Basket Analysis

The single most difficult computational challenge faced by researchers in dealing with MBA problems is the volume and sparseness of typical MBA datasets. Suppose one obtains a MBA dataset with 50 product categories over a 2 year period (104 weeks) then one has to deal with a 104×50 data matrix. Also, in order to identify association rules between products, one must deal with a contingency table with 2^{50} levels, which may also contain many empty cells. This characteristic of the MBA dataset makes it difficult, and in many cases infeasible, for the researchers to analyze the full model directly. Another important problem faced by researchers is the problem of model validity. Typically there is little theoretical knowledge available on different associations structures which may be identified from the analysis, and it is difficult to score or rank multiple structures.

Bayesian network approach to the MBA problem has two distinctive advantages. First of all, as the Bayesian approach is founded upon the computation of probabilities it is natural to score the possible networks by the *posterior probability*. In other words the posterior probability of a possible model M_i given a dataset D can be computed by (1) where (2) is the marginal likelihood.

$$P(M_i|D) \propto P(D|M_i)P(M_i) \tag{1}$$

$$P(D|M_i) = \int P(D|M_i, \theta)P(\theta|M_i)d\theta \tag{2}$$

Hence, by specifying prior distributions one can calculate the posterior probability of the model in question, and this probability can be interpreted easily when comparing multiple model candidates. Other than the technical advantages, Bayesian network approach is intuitively better equipped in solving market basket problems. One advantage is that by DAGs, Bayesian network can study causal links between variables, and this is useful when we are interested in applications of MBA such as finding the causal link between promotion variables and product variables. Bayesian network also benefits from the usual Bayesian paradigm of *learning* and *expertise*, meaning that expert knowledge in the field can be utilized through the prior distribution and *learning* from data can be achieved through the Bayesian updating process. These are key advantages one can get from using Bayesian networks as it allows business managers to learn and continuously update beliefs on consumer behavior from market basket data and apply them in forming business strategies.

There are challenges in using the Bayesian network approach as well. Specification of a suitable prior distribution is always a non-trivial task in performing Bayesian analysis. Furthermore, when the set of possible candidate models is large as is the case in MBA datasets, it is practically infeasible to compute the exact posterior probabilities and compare them among all possible models simultaneously. Previous studies[1,2] have shown that these problems can be solved by choosing the Dirichlet distribution as the prior distribution and utilizing Markov Chain Monte Carlo methods. It is also shown that especially MC^3 algorithm proposed by Madigan and York[3] works well in model selection. In more detail, Giudici and Passerone[2] show that in analyzing a dataset with 26 variables, model selection using Bayesian network methods can be performed directly on the joint 2^{26} contingency table of the products whereas, in the classical case this was infeasible.

Integrating Promotion to Market Basket Analysis

One of many possible operational strategies which could be used by retailers is that of product promotion. Blattberg, Briesch and Edward's study[4] on how promotions work give a thorough review of the promotion literature and there they define promotion activities as not only price promotions (*temporary* price discounts to the customer) but to also include distribution of coupons, display and feature advertising activity direct to the consumer. Specifically, the study highlights that promotion activities can either change the magnitude consumers purchases and/or enhance store traffic. Another important study regarding promotion strategy is done by Mulhern and Leone[5], where they emphasize the

presence of demand interrelationships among multiple products within a store. As we have discussed the underlying problem is a multi-product multi-category problem, and single product promotion strategy and implications of simple buyer-seller models have limited value in devising promotion strategies. Therefore, we will look at the promotion problem in a multi-product setting identifying and analyzing interdependent structure between products.

Through out the marketing literature, researchers agree that promotion activities in one category have impacts on the sale of products within that category, but they differ in modelling cross-category promotion effects depending on the nature of the underlying relationship between categories. All in all, there are three modelling paradigms related to cross-category promotion effects. Direct promotion effect proposed by Manchanda[6] can be described as promotion activities on category A directly affecting the purchasing behavior not only on category A but also category B. Second paradigm, albeit indirect promotion effect described by Russell[7], is where promotion activities in a category A only affects purchasing decisions in this category but the changes in purchasing patterns in category A affects the purchasing behavior in category B. Therefore promotion activities in category A indirectly affects purchasing decisions in category B. Finally there is the mixed promotion model where both direct and indirect promotion effects are present across product categories (Figure 1).

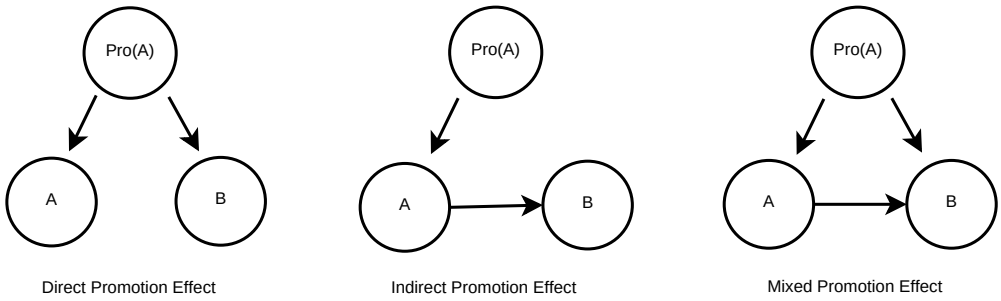

Figure 1: Bayesian network representation of cross-category promotion effects

How one models cross-category promotion is entirely up to the researcher. If there exists some prior knowledge on product dependency structure then one can specify these structures a priori, where they could be direct, indirect or mixed promotion models. If there is no past data then one can also follow a pure data mining approach and try to find the best cross-category dependence structure from scratch, just as it was done in previous studies[1,2]. In any case, we propose Bayesian network in approaching this promotion integrated MBA problem, as (i) it can handle all three promotion models described above, (ii) it is easy to represent and model the dependence structure between categories and (iii) it can be used to learn the structure when there is no prior structural knowledge available.

To be more specific, there are two preliminary goals in our modelling of the promotion integrated MBA problem. First, we want to identify the global Bayesian network of product clusters with the promotion variable included(Figure 2). This is a direct extension of the work which was done previously[1,2] which mainly looked to identify the product clusters. Once these clusters are identified, it can provide the retail managers with valuable information on planning marketing and operational strategies. Also, we evaluate the strength and direction of pairwise cross-category relationships, in terms of finding the posterior probability of edges, given the presence of promotion variables. This is more in line with traditional marketing research and it can give retail managers an idea on the products to focus their marketing efforts on.

Findings

The MBA dataset we used include sales data from multiple grocery stores in the Chicago Metropolitan area, where individuals were followed for a 2 year period. The data also included marketing data from the grocery stores and the consumer demographic data. In order to facilitate the analysis for this preliminary research, we chose twelve product categories (detergent, softener, towel, tissue, yogurt, cereal, soap, cleanser, hotdog, egg, cookie and cracker) some of

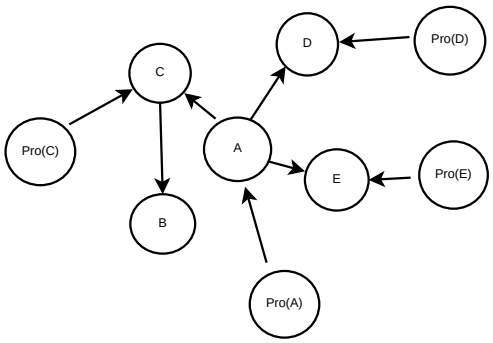

Figure 2: A possible Bayesian network identified from data

which have different functions and some which are closely related. Also, in defining the *promotion* variable, we combine the three types of marketing activities (price discount, display and feature) available in the data into one binary promotion variable. Hence, if any of the three marketing activity is present then the promotion variable will have the value of 1 and zero otherwise. For the purpose of this research, we assume that the impact of these three marketing activity on sales of the product is the same. We also limit the learning process, by not allowing any parents to the promotion variable. For space purposes, we do not illustrate the MC^3 algorithm that we employed, however details of the MC^3 algorithm can be found in [3].

The primary result we find, as seen from the figures below, is that the prevalent promotion effect between product categories is that of indirect promotion. More specifically, promotion activities on product A directly impacts the sales of product A. Now, the sales or no-sales decision of the product A will in turn affect the sales of its complement (or substitue) product B. When using a cutoff value of 0.9, there were three multi-category clusters identified, namely (towel, tissue, soap, hotdog), (softner, detergent) and (cereal, cracker, cookie). Using a cutoff value of 0.6, the three clusters are merged into two clusters represented in figure 3. Hence, the results show that Bayesian network obtained from the data with minimal restrictions can capture the multi-category relationship and give us insight on what may be the underlying promotion mechanism.

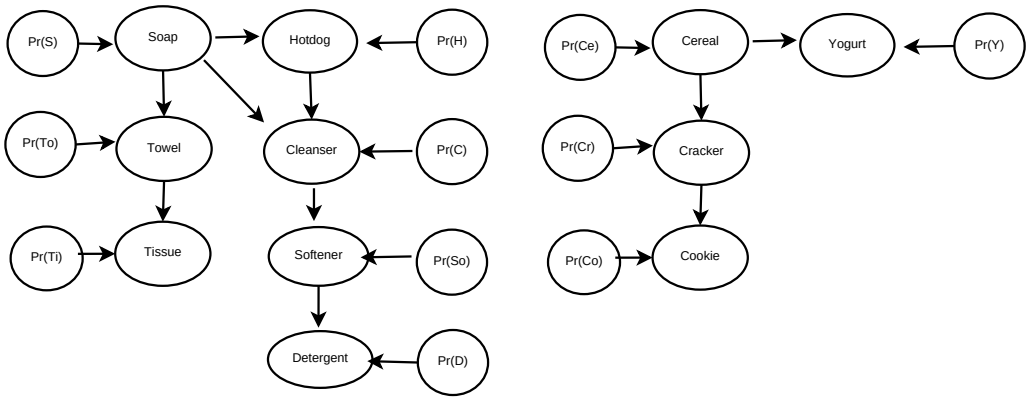

Figure 3: Two product clusters identified with cutoff value 0.6

We now turn our attention to the pairwise cross category relationship and compare the strength and direction of association by looking at the corresponding edge probabilities. This allows us to identify the primary and secondary categories in a pair of complimentary products. After analyzing pairwise associations, we identified three pairs of use compliment products as (Detergent, Softner), (Soap, Cleanser) and (Towel, Tissue). We now then isolated the strong compliment products (Detergent, Softner)and computed the posterior probabilities for the three networks illustrated in

figure 4. The posterior probabilities were 0.98,0.01 and 0.01 respectively. This suggests that the softener is the primary product in this pair of complimentary products and suggests that the retail manager may want to focus on marketing softeners.

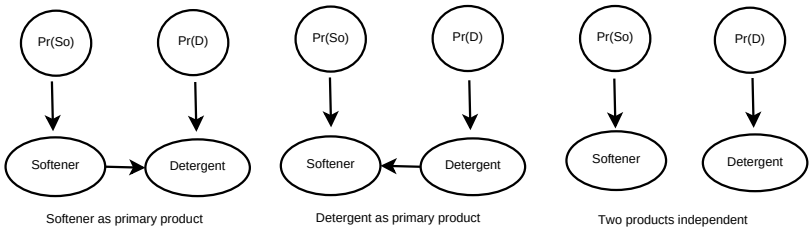

Figure 4: Three possible networks for the complimentary product pair (Detergent, Softner)

Concuding remarks

In this paper, we concentrated on modelling association structures in market basket analysis by using Bayesian networks, with applications in formulating marketing strategies. We first, described the advantages of Bayesian network in identifying association structures in market basket data by highlighting previous work[1,2] in the literature which deals with this problem. We then identified different types and effects of promotion as discussed in the marketing literature and also shown how they can be integrated to the previous Bayesian MBA methods. The results show that the most prevalent form of promotion found from the analysis is that of indirect promotion. Coupled with this result, we have also found that when dealing with a pair of close compliment products one can identify the primary product which may drive the relationship.

These results may have strong impacts on retail managers making good decisions regarding marketing activities as it is crucial to identify and understand the unobservable relationships between consumer purchase decisions. For example, the retail manager may interpret our results as to identify the most effective product to promote, and therefore maximize the effects of indirect promotion and at the same time minimize the number of products to promote. Also, the retail manager should only promote the primary product when faced with a pair of complimentary products. The retailer can also decide on when to promote the products by identifying the causal links in the Bayesian network and run a sequential promotion strategy (i.e. first on product A, then product B and so on). This type of sequential strategy might be useful if there is some type of diminishing effects of promotion present, and hence running all four promotion activities simultaneously may not be optimal.

There are certain limitations to this study. One of the limitations is that we have merged different types of promotion in creating the promotion variable. These promotion activities will certainly vary in their effectiveness and we may have over simplified in creating our model. Also, we have not incorporated the available demographic data. This added element will certainly complicate the problem in the computational sense, however it would give us a more complete picture of the association structure and could give insights on consumer purchasing behavior. Finally, in the future we would like to extend the problem and explore dynamic Bayes networks in modelling the promotion integrate MBA problem in a dynamic setting. The data we use here spans over a 2 year period and it is fair to assume that consumer's purchasing decisions on a given week has an impact on purchasing decisions of the following week. As mentioned before, with the improvements in computer technology, retailers can now issue promotion coupons as customers go through check out almost simultaneously. If the dynamic model can be successfully implemented then the retailer can input products in the customer's current shopping basket and print *correct* coupons which would have the biggest impact on customer's purchasing decisions in his/her next visit to the store.

References

[1] Giudici, P., and Castelo, R., 2002, "Data mining of association structures to model consumer behavior," Computational Statistics and Data Analysis, 38, 533-541.

[2] Giudici, P., and Castelo, R., 2003, "Improving Markov Chain Monte Carlo Model Search for Data Mining," Machine Learning, 50, 127-158.

[3] Madigan, D., and York, F., 1995, "Bayesian graphical models for discrete data," International Statistical Review, 63, 215-232.

[4] Blattberg, B., Briesch, R., and Edward, F., 1995, "How promotions work," Marketing Science, 14(3), 122-132.

[5] Mulhern, F., and Leone, R., 1991, "Implicit price bundling of retail products: a Multiproduct to maximizing store profitability," Journal of Marketing, 55, 63-76.

[6] Manchanda, P., Asim, A., and Sunil, G., 1999, "The Shopping Basket: a Model of multicategory purchase incidence decisions," Marketing Science, 18(2), 95-114.

[7] Russell, G., Ratneshwar, S., Shocker, A., David, B., and Bodapati, A., 1999, "Multiple-Category Decision Making: Review and Synthesis," Marketing Letters, 10(3), 319-332.