Two Essays on " Mining Market Basket Data: Models and Applications in Marketing"

by

Xiaojun Li

B.E. Computer Engineering, Shandong University of Technology, 1997
M.E. Computer Engineering, Chinese Academy of Sciences, 2000

A Dissertation submitted to

The Faculty of
School of Business
of The George Washington University
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy

May 18, 2008

Dissertation directed by

Srinivas Prasad
Associate Professor of Decision Sciences
Pradeep Rau
Professor of Marketing

The School of Business of The George Washington University certifies that Xiaojun Li has passed the Final Examination for the degree of Doctor of Philosophy as of March 24, 2008. This is the final and approved form of the dissertation.

Two Essays on "Mining Market Basket Data: Models and Applications in Marketing"

Xiaojun Li

Dissertation Research Committee:

Srinivas Prasad, Associate Professor of Decision Sciences, Dissertation Director

Pradeep Rau, Professor of Marketing, Co-Director

Refik Soyer, Professor of Decision Sciences, Committee Member

Acknowledgments

I would like to thank my advisor Dr. Srinivas Prasad for his guidance and mentoring during my Ph.D. study. He has been so patient with me. I would also like to thank my co-advisor, Dr. Pradeep Rau, for his guidance on this research. I am also grateful to Dr. Sanal Mazvancheryl. He provided numerous suggestions to this research. I would also like to thank Dr. Refik Soyer for his advice on statistics. Dr. David Bell kindly provided us with the data set.

I would like to thank my daughter Shirley and my son Arthur. They bring so much joy in my life. Finally and most importantly, I would like to thank my wife Yi. Without her love and support, this accomplishment would be impossible.

Abstract of Dissertation

Two Essays on " Mining Market Basket Data: Models and Applications in Marketing"

Abstract

A retailer typically makes decisions on what products to promote, how, and when. These decisions, often referred to as marketing mix activities, are especially important given that pricing and promotion decisions in one category can not only influence sales in the promoted category, but also in other categories. For example, a price reduction in cake mix may not only boost its own sale but also sales of cake frosting. It would thus be in the retailer's interest to coordinate the marketing activities across products in multiple categories so as to maximize profit. One such model for maximizing store profit could be stated as follows. Suppose there are n products. Let the decision variables that relate to the pricing and promotion decisions for each product in week t be represented by $\mathbf{p}_t = \{p_{1t}, ...p_{nt}\}$ and $\mathbf{Promo}_t = \{Promo_{1t}, ..., Promo_{nt}\}$, respectively. Clearly, given the presence of cross-category effects, the sales of product i during week t could be expected to be a function of $\mathbf{Promo}_t$ and $\mathbf{p}_t$, resulting in the following expression for the revenue R_t during week t,

$$R_t = \sum_{i=1}^{n} s_{it}(\mathbf{Promo}_t, \mathbf{p}_t) \times p_{it}, \tag{1}$$

where s_{it} is the sales of product i at week t and p_{it} is the price of product i at week t. The decision problem then is to determine values for the $\mathbf{Promo}_t$ and $\mathbf{p}_t$ variables such that the revenue in equation 1 is maximized.

Before one can address such a goal, two pieces of critical information are needed. First, given the large number of products carried in a store, one needs to know what products are associated. Second, one needs to understand how promotions work

across categories. Since the 1990s, advances in information technology have made the collection and storage of consumers' purchase history and shopping basket content technically and economically feasible. Such market basket data contain valuable information about product association and promotion effects, and make possible the analysis of coordinated marketing mix activities. This dissertation consists of two essays in data mining models of market basket data.

Essay 1 is titled "Market basket analysis using Bayesian Networks". This essay addresses the question of how promotions work across categories. Promotions in one product category can affect sales of products in another category either directly or indirectly. Given a set of product categories and market basket data, we analyze the presence of cross category impacts using Bayesian Networks. We model the occurrence of a product category, and not the number of units (of a product category) in a basket. The data set we employ is an IRI market basket data set that contains transactions including 22 categories over 2 years for 500 panelists. Bayesian Networks are learned from this data and are used to identify the underlying dependencies across product categories. Specifically, we study how the associations across categories vary based on marketing mix activities, and also based on demographics. The results from such an analysis can help in 1) identifying clusters of categories wherein associations exist primarily between categories within a cluster and not across clusters, and 2) in making predictions on basket choices given a set of specific marketing mix activities. The ability of Bayesian networks to learn based on new evidence also makes such an approach possible in an online context when customers' choices can be observed, and marketing activities can be dynamically customized.

Essay 2 is titled "Localized rule discovery in market basket data" and builds on traditional association rule mining algorithms to identify pairs of products that are associated. Due to consumer heterogeneity, the association between products may vary across consumer segments. Two products can be globally associated or locally

associated. For the latter, associated products are matched with consumer segments within which the localized associations are strong. We illustrate the algorithm on the same shopping market basket data set used in Essay 1. The results from this analysis should help the retailer in identifying customer segments in which specific pairs of products are strongly associated, and also help determine the marketing mix effects on cross product associations.

Table of Contents

List of Figures

Essay One

Essay Two

List of Tables

Essay One

Essay Two

Market Basket Analysis using Bayesian Networks

Xiaojun Li

School of Business, the George Washington University

Washington DC

December, 2007

Abstract

This essay addresses the question of how promotions work across categories. Promotions in one product category can affect sales of products in another category either directly or indirectly. Given a set of product categories and market basket data, we analyze the presence of cross category impacts using Bayesian Networks. We model the occurrence of a product category, and not the number of units (of a product category) in a basket. The data set we employ is an IRI market basket data set that contains transactions including 22 categories over 2 years for 500 panelists. Bayesian networks are learned from this data and are used to identify the underlying dependencies across product categories. Specifically, we study how the associations across categories vary based on marketing mix activities, and also based on demographics. The results from such an analysis can help in 1) identifying clusters of categories wherein associations exist primarily between categories within a cluster and not across clusters, and 2) in making predictions on basket choices given a set of specific marketing mix activities. The ability of Bayesian networks to learn based on new evidence also makes such an approach possible in an online context when customers' choices can be observed, and marketing activities can be dynamically customized.

1 Cross Category Promotion Modeling

Consumers typically purchase multiple products from multiple categories in one shopping trip. Advances in information technology have made the collection and storage of basket level transaction data economically and technically feasible. It is in the retailers' interest to leverage information hidden in these data so as to increase store profit. One useful piece of information is how promotions work, and in particular how they work across categories.

1.1 Motivation

A broader definition of promotion includes not only price discount and couponing, but also display and feature advertising activities, etc. Marketing researchers have found that promotions can either change the magnitude of consumers' purchases and/or enhance store traffic. A thorough review of how promotions work can be found in Blattberg, Briesch, and Edward (1995). One important finding from past research is that promotion in one category affects sales in complementary categories and substitute categories. This would suggest that retail pricing strategy should incorporate demand interdependencies such as complementary and substitute to maximize store profitability (Mulhern and Leone 1991).

Two product categories can be use complements such as cake mix and frosting, substitutes such as butter and margarine, or just independent. Since it is hard to observe use complements by a retailer, we use the idea of purchase complements. A pair of categories are defined to be purchase complements "if marketing actions (price and promotion) in one category influence the purchase decision in the other category"(Manchanda et al. 1999). A more detailed explanation of complements and substitutes is given in (Manchanda et al. 1999, Russell et al. 1999).

According to consumers' purchasing decisions, Seetharaman et al. (2005) classify cross-category models into incidence models, brand choice models, and quantity models. Incidence models can be further classified into "whether to buy" models, "when

to buy" models, and "bundle choice" models. This essay extends the "whether to buy" models to explicitly identify category clusters.

1.2 Cross-Category Models

Both statistical and data mining models have been proposed to understand promotions among multiple categories. Approaches to measure cross category effects in the marketing literature have been theory driven, and typically based on utility theory models. These utility theory based models usually decompose the utility of a category into its own effects and cross-category effects. Some models go further to isolate cross-category marketing mix effects from cross-category co-incidence effect, examples being multinomial logistic models (Russell and Peterson, 2000) and multivariate probit models (Manchanda et al. 1999, Chib et al. 2002). These researchers all agree that promotions in one category have impacts on its own sales. However they differ in how to model the cross category promotion effects. There are three paradigms in modeling cross category promotion effects. One is marketing actions effects choices of other categories directly (Manchanda et al. 1999). In this type of model, the consumers' decision process is viewed as a black box. The final choices in a basket are modeled as a function of marketing mix variables. A second paradigm is that promotions in a category only effect its own sales, but presence of this category affects decisions in other categories (Russell et al. 1999, Hruschka et al. 1999). Therefore, marketing mix activities impact indirectly across categories. The last paradigm attempts to model category choices using both promotions and presence of other categories.

The quality of these statistical approaches depends on the practitioner's choice of modeling paradigm. In this essay, we propose a data mining paradigm which learns the cross category promotion effects based on data. We use Bayesian networks to learn the dependencies among variables suggested by the data. Advances in Bayesian network make it possible to learn the multivariate relationship from data. Model uncertainty is also considered using Bayesian networks. Bayesian networks were first introduced as an expert system tool. Because they have both causal and probabilistic

semantics, Bayesian networks can represent causal relationships in a problem domain. They can also be used to predict the consequences of intervention. Bayesian networks have several other advantages as a research tool(Heckerman 1996). For example, they can handle missing data readily and avoid over-fitting of data.

Bayesian networks have been used to model association among products in past research(Giudici and Passerone 2002, Giudici and Castelo 2003). Note however that our research differs in that we model not only co-occurrence but also how marketing mix works across multiple categories using transaction level market basket data as opposed to aggregated basket data used in earlier research (Giudici and Passerone 2002).

This essay is organized as follows. Section 2 is a brief introduction on Bayesian networks. Section 3 discusses the details of learning a Bayesian network of discrete data. Section 4 applies the Bayesian network to retailing data and summarizes the research.

2 Bayesian networks

Intuitively, a Bayesian network is a graphical representation of conditional independence and dependence among variables regardless numerical or functional details. Also known as Directed Acyclic Graph (DAG) or belief network, a Bayesian network is a type of graphical model that combines the science of statistics and graph theory. As shown in Figure 1, a Bayesian network consists of:

- A directed graph with nodes representing variables and edges representing dependencies.
- A set of probability distributions associated with the edges.

In this figure, A is called the parent of B, and B is parent of C. A and C are said to be conditionally independent given B since $p(C|A \cap B) = p(C|B)$.

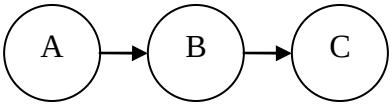

Figure 1: An example of Bayesian network

More rigorously, let $X = x_1, ..., x_n$ represent a set of variables and S represents a network structure. S needs to encode a set of conditional independence statements about X. P represents the set of local probability distribution of each variable in X, $p(x_i|Pa_i, \xi)$. Pa_i denotes the parents of node x_i in structure S and the corresponding variable in X. An important property of Bayesian network is the chain rule, which means the joint probability distribution for X is

$$p(X) = \prod_{i=1}^{n} p(X_i|Pa_i, \xi) \tag{1}$$

As we see, a Bayesian network can be viewed as a collection of local probabilistic/regression models. As we observe the state of some nodes in the network, we can update the probability of other nodes' states.

2.1 Graphical Representation of Cross-Category Effects

To represent cross category marketing effects, we need to define two types of nodes in the network. They are:

1. X_i : Marketing mix variable for product i. It is a binary variable with $X_i = 1$ meaning there is a promotion for the product. We do not need to estimate $p(X_i)$, which is up to the retailer.

2. Y_i : Purchase decision of product i. It is a binary variable with $Y_i = 1$ meaning product purchase.

In this study, we have partial knowledge about the underlying network structure.

- A category node's parents can be either marketing variables or other category nodes.

- A marketing variable does not have any parent node.

Thus, the three cross-category effects modeling paradigms mentioned earlier can be represented as in

1. Figure 2: Direct Paradigm. Sales of category B is directly dependent on category A's promotions.
2. Figure 3: Indirect Paradigm. Sales of category B is dependent on the purchase decision of category A.
3. Figure 4: Mixed Paradigm. Sales of category B is dependent on category A's both promotions and purchase decision.

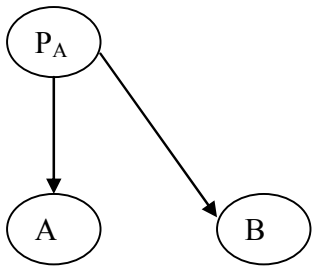

Figure 2: Direct Paradigm

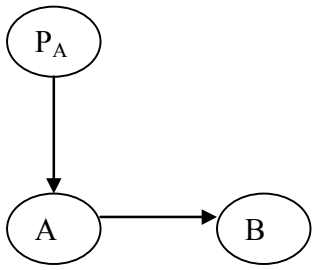

Figure 3: Indirect Paradigm

2.2 Learning Bayesian Networks

A Bayesian network can be constructed based on an expert's knowledge. In such a context, both statistical parameters and network structure are specified by domain

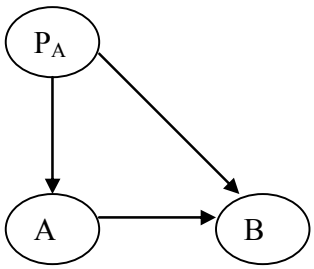

Figure 4: Mixed Paradigm

experts. Once actual values of some variables are observed, the Bayesian network updates the posterior probabilities of other variables by an inference process. For example, one can use Bayes' rules to reverse the arcs step by step in the network until the requested probabilities are answered (Shachter 1988). However, either exact inference or approximate inference is NP-hard (Cooper 1990,Dagum and Luby 1993). For Bayesian networks of discrete variables, the most commonly used algorithm is the junction tree algorithm(Lauritzen and Spiegelhalter 1988, Jensen and Lauritzen 1990, Dawid 1992).Probabilistic inference is performed using several mathematical properties of the junction tree.

So far we have assumed the network structure is known. In domains where we have little knowledge, machine learning techniques are enlisted to help learn structures from data. The most straightforward approach is to compare all possible networks based on some measure. The network that optimize this measure will be selected as the best Bayesian network. The challenge in this approach is the number of possible networks given n nodes explodes as n increases.Table 1 gives examples of candidate network counts. There is no closed form formula known for the number of structures. Robinson (1977) gives the following recursive formula

$$f(n) = \sum_{i=1}^{n} (-1)^i \frac{n!}{(n-i)!i!} 2^{i(n-i)} f(n-i)$$

Given a number of competing networks, we need both a measure and a search strategy to find the most promising network. Measures used include maximum likelihood, predictive assessment, and posterior probabilities. As Chickering (1996) shows, it is

number of variables	*number of possible DAGs*
2	3
3	25
4	543
5	$29,000$
10	4.2×10^{18}

Table 1: Number of Network Structures

NP-hard to learn the structure of a Bayesian network. A variety of heuristic search algorithms has been introduced, such as greedy search, greedy search with restarts, best-first search, and Monte-Carlo methods (Heckerman 1996). The most straightforward search and scoring approach is greedy search. Here is a brief introduction of the greedy search algorithm. Let E represents all eligible changes to a Bayesian network and use log of relative posterior probability as the network score.

$$\log P(D, S^h) = \log P(S^h) + \log P(D|S^h)$$

Let $\delta(e)$ represent the changes of the network score caused by change e.

- Choose an initial network.
- Change one edge in the network at a time and evaluate the change. Pick the one with maximum $\delta(e)$.
- Stop the search when no e can make a positive contribution.

This approach may hit a local maximum. To escape from local maxima, we need to restart the search process randomly with a new initial network. Another way to find a global maxima is to use approximation approach such as Markov Chain Monte Carlo Model Composition, or MC^3 (Madigan and York 1995).

If one has some partial knowledge of causal relations among variables, the search space can be reduced by ruling out unlikely models or by placing a restriction that a node can have at most u parents ($u < n-1$) (Cooper and Herskovits 1992). It is very

likely there is no single dominant model learned. In this case, instead of selecting a single true model, model uncertainty is accounted by averaging all the models.

Let Δ be the quantity of interest. Its posterior distribution conditional on data D is

,

$$pr(\Delta|D) = \sum_{k=1}^{K} pr\left(\Delta|M_k, D\right) pr(M_k|D) \tag{2}$$

To find the M_k and their posterior probability $pr(M_k|D)$, an algorithm called Markov Chain Monte Carlo Model Composition (MC^3) (Madigan and York 1995) can be used. Markov Chain Monte Carlo (MCMC) is a simulation method that generates samples from complex and nonstandard distributions. Developed by Metropolis et al. (1953), and generalized by Hasting, the Metropolis-Hastings algorithm is an implementation of the Markov Chain Monte Carlo method.

A brief introduction of the Metropolis-Hastings algorithm follows - for a more detailed introduction, see Chib and Greenberg (1995). A MCMC algorithm draws samples of a target probability density $\pi(x)$ by constructing a Markov chain, which converges to the target probability distribution. Define the candidate generating density as $q(x, y)$, from which a value y is generated when the process is in state x. When $\pi(x)q(x, y) > \pi(y)q(y, x)$, the process moves from x to y more often than from y to x. Thus we need to specify a probability of move $\alpha(x, y)$ to meet the requirement of reversibility,

$$\pi(x)q(x, y)\alpha(x, y) = \pi(y)q(y, x)\alpha(y, x)$$

Given an initial state $X^{(0)}$,

- Repeat for $j = 1, ..., N$
- Generate y from $q(x^{(j)}, \cdot)$ and u from $U(0, 1)$
- IF $u \leq \alpha(x^{(j)}, y)$, set $x^{(j+1)} = y$
- ELSE set $x^{(j+1)} = x^{(j)}$
- return the values $\left\{x^{(1)}, x^{(2)}, \ldots, x^{(N)}\right\}$

MC^3 is a Metropolis-Hastings based method, which build a Markov chain of graphs that will converge to the distribution of the model that generates the data. MC^3 has been successfully used for linear regression models (Raftery et al. 1997).

3 Bayesian Network of Discrete Data

This section introduces the technical details of learning Bayesian networks of discrete data using the MC^3 algorithm. We assume all variables in the network are discrete and follow Dirichlet distribution (Heckerman 1996). More discussion of the Markov Chain Monte Carlo Model Composition algorithm can be found in Madigan and York (1995), and Giudici and Castelo (2003).

3.1 Network Specifications

Now let g be a Bayesian network of variables $\mathbf{X}$. Each variable $X_i \in \mathbf{X}$ is discrete, $X_i = x_i^1, ...x_i^{r_i}$. Denote X_i's parents as Pa_i

$$p(x_i^k | Pa_i^j, \theta_i, g) = \theta_{ijk} > 0 \tag{3}$$

Assume data completeness and parameter independence, the joint probability of the parameters is

$$p(\theta_g | D, g) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} p(\theta_{ij} | D, g) \tag{4}$$

For discrete variables, we will assume Dirichlet prior distribution:

$$p(\theta_{ij} | g) = Dir(\theta_{ij} | \alpha_{ij1}, ..., \alpha_{ijr_i}) \tag{5}$$

where r_i is the number of realizations under such a configuration. Assume a uninformative assignment with equivalent sample size, $\alpha_{ijk} = 1/(r_i \times q_i)$, q_i being the

number of configurations of parents. The Posterior distribution is:

$$p(\theta_{ij}|D,g) = Dir(\theta_{ij}|\alpha_{ij1} + N_{ij1}, ..., \alpha_{ijr_i} + N_{ijr_i}) \tag{6}$$

where N_{ijk} is the number of observations in dataset D with $X_i = x_i^k$ and $Pa_i = Pa_i^j$ Following the four assumptions given in Cooper and Herskovits (1992):

1. All the variables are discrete.
2. Given the Bayesian network model, the observed cases are independent.
3. There is no missing value.
4. Before observing data set D, we are indifferent regarding the numerical probabilities to assign to a given network structure.

Based on the above, Coopers and Herskovits (1992) show that the marginal likelihood of a discrete DAG model is given by

$$L(g) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(N_{ij} + \alpha_{ij})} \prod_{i=1}^{r_i} \frac{\Gamma(N_{ijk} + \alpha_{ijk})}{\Gamma(\alpha_{ijk})} \tag{7}$$

where i is the index of nodes in the network, j is the index of its parents' configurations.. $\alpha_{ij} = \sum \alpha_{ijk}.$ and $N_{ij} = \sum_{k=1}^{r_i} N_{ijk}.$

3.2 Learning network structure using MC^3

Learning the network structure is accomplished by constructing a Markov chain of DAG's. The irreducibility of a DAG is guaranteed in the case of a DAG. However,acyclicity needs to checked for each graph. Define the neighborhood of a given graph g as nbd(g), which is the collection of graphs that can be reached from g in one step by adding or deleting one edge, including itself. There are three types of moves, addition, deletion, and reversal. Reversal of an arc can be viewed as to perform a

removal move first followed by an addition move.A move can not generate directed cycles. One way to guarantee no directed cycle is to traverse the whole network.

When the chain moves from g to g', $g' \in nbd(g)$, the acceptance probability of the move is

$$min\{1, \frac{\#(nbd(g))p(g'|D)}{\#(nbd(g'))p(g|D)}\}$$

where $\#(nbd(g)$ represents the cardinality of graph g's neighborhood. If the move is not accepted, the chain stays in state g.

Since g and g' are neighbors, it is reasonable to assume that

$$\frac{\#\left(nbd\left(g\right)\right)}{\#\left(nbd\left(g'\right)\right)} \approx 1$$

Since $p(g|D) \propto p(D|g)p(g)$, the acceptance ratio relates only to the Bayes factor $\frac{p(D|g')}{p(D|g)}$. The Bayes factor in case of addition and deletion is

$$\frac{L(g', i)}{L(g, i)} \tag{8}$$

where i is the variable whose parent set is different in g and g'. Bayes factor in case of reversal is

$$\frac{L(g', i)L(g'', j)}{L(g, i)L(g, j)} \tag{9}$$

Recall that

$$L\left(g, i\right) = \prod_{j=1}^{q_i} \frac{\Gamma\left(\alpha'_{ij}\right)}{\Gamma\left(N_{ij} + \alpha'_{ij}\right)} \prod_{k=1}^{r_i} \frac{\Gamma\left(N_{ijk} + \alpha'_{ijk}\right)}{\Gamma\left(\alpha'_{ijk}\right)} \tag{10}$$

3.3 Average Network

As the number of nodes increase, the result of MC^3 learning are many DAGs with small likelihood. We can apply Bayesian model averaging to present the relation among the nodes in one overall network (as in Giudici and Castelo, 2003). Let e be an edge in a graph. $P(e|D)$ measures the probability of its presence given data D. Only those edges whose $P(e|D) > \mu$ will be drawn in an aggregate graph.

$$P(e|D) = \sum_i I(e|g_i) * P(g_i|D)$$

where

$$I(e|g_i) = \begin{cases} 1 & if\ e \in g \\ 0 & if\ e \notin g \end{cases}$$

These edges can be plotted to create an average network. The average network can be used to identify clusters of variables. Note that the average network constructed in this manner may not be a valid DAG.

4 Application in Marketing

There are two goals in this application. First, we want to identify product clusters using Bayesian network given presence of other products, promotions, or customer demographics. We believe once these clusters are identified, they can provide a retailer with useful information on planning marketing activities or segmenting customers. Second, we specifically want to evaluate pairwise cross-category relationship given presence of other products, promotions, or customer demographics. This is more in line with traditional marketing research. We build three Bayesian network models. They are:

- Model 1: Using Bayesian network to map the relationship of the products only. This is an equivalent model of (Giudici and Passerone 2002, Giudici and Castelo 2003).
- Model 2: Using Bayesian network to model product relationship with marketing mix variables included.
- Model 3: Using Bayesian network to model product relationship with customer demographics variables included.

The data set include sales data from multiple grocery stores in a Metropolitan area. Also included are marketing data and consumer demographic data. Twelve categories are chosen for the analysis. They are detergent, softener, towel, tissue, yogurt, cereal, soap, cleanser, hotdog, egg, cookie, and cracker. These categories include products which have different functions or closely related.

Section 4.1 introduces implementation and validation the MC^3 algorithm. Section 4.2 and 4.3 discusses details of product clustering and pairwise assessment. Section 4.4 summarizes findings and discusses managerial implications of Bayesian network in marketing research.

4.1 Algorithms Implementation and Validation

The analytic software used is developed on the basis of Bayesian Network Inference with Java Objects or BANJO, which is an open source software originally developed by Duke University researchers. Here is a brief description of the implementation of the MC^3 algorithm.

1. Given an initial network.

2. Randomly propose a move. It can be addition, deletion, or reversal of an edge.

3. Check whether the proposed move leads to cycles. If it does, repeat from step 2.

4. Evaluate the move according to the criteria given in last chapter.

5. Decide if the move is accepted so that Markov chain gets into a new state. Otherwise it stays in the same state. Repeat from step 2 until the chain converges.

6. Maintain a database of Bayesian networks. Two key pieces of information, structure of the network, and its frequency, are kept for calculation of the average network.

As the number of nodes increases, the number of legal networks increases exponentially. The Markov chain will converge very slowly. To speed up, instead of starting the Markov Chain from a random generated network, we start it from a network structure learned using greedy search algorithm. We find that this speeds up the convergence by more than a factor of two.

In this research, we use two random samples to validate the MC^3 algorithm. Two measures are used to validate the algorithm. Let e_{ij} be the edge from node i to node j. Given two datasets D_1 and D_2, E_1 represents all edges learned in D_1 and E_2 represents all edges learned in D_2.

The first measure is on the difference of each edge's posterior probability. Let $P(e_{ij}|D)$ represents the probability of edge e_{ij} being present in data set D. The difference of the two datasets is captured in $\delta_{ij} = P(e_{ij}|D_2) - P(e_{ij}|D_1)$. Using sales data of the twelve categories, there are 41 edges learned from the two networks. Figure 5 is the histogram of the differences, from which we can see there is no major difference between the two edge sets.

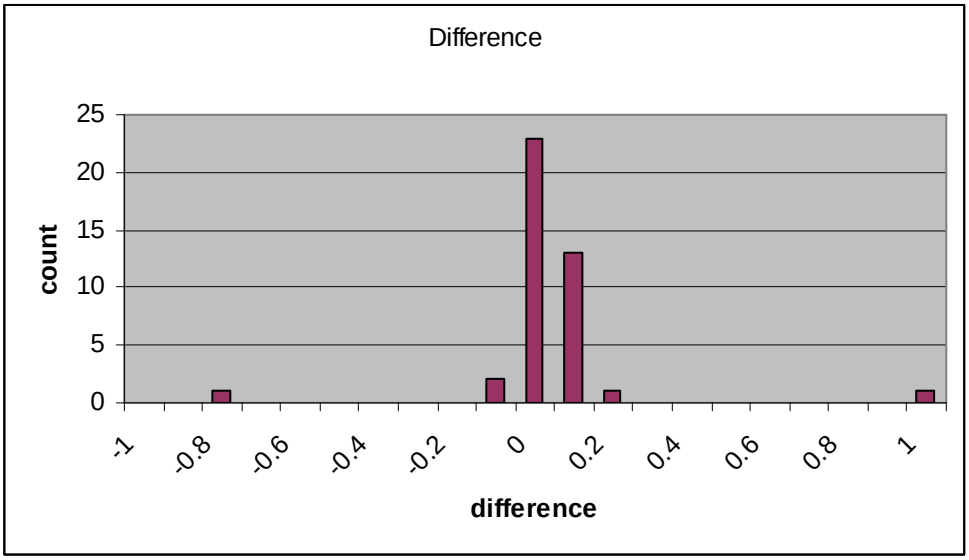

Figure 5: Difference between edge sets

We now present some validation statistics based on the similarity of edges generated in the network structures learned from the two random data sets. At a given cutoff point, suppose the average network from data set D_1 has a set of edges E_1. The size of E_1, which is the number of edges in it, is m_1. Average network from dataset D_2 has a set of edges E_2 with a size m_2. Let m_{12} represent the size of $E_1 \bigcap E_2$. Define validation rate $r = \frac{m_{12}}{m_1}$. r is a number between 0 and 1. Set cutoff point at $0.9, 0.7$, and 0.5, the validation rate is 0.9, 0.9, and 0.91. Based on the above, we can reasonably conclude that the MC^3 learning algorithm generates valid network structures.

4.2 Product Clusters Identification

Results of Model 1

We start with 12 category sales data only. Based on the average network with a cutoff value 0.9, there are one large network with most categories involved, and several one-category clusters. They are (tissue, towel, cleanser, detergent, softener, cereal, cookie, hotdog, cracker), (egg), (yogurt), and (soap), see Figure 6. This result is different from Giudici and Castelo (2003). In Giudici and Castelo (2003), there are two-category, three-category, and five-category clusters. The difference might be due to two facts. First, their sales data are weekly aggregate data. Second, they use 1 to indicate the weekly sale of some category is greater than median, otherwise 0. The possible explanation for the absence of clusters in our findings can be that most product categories are related to each other at the transaction level.

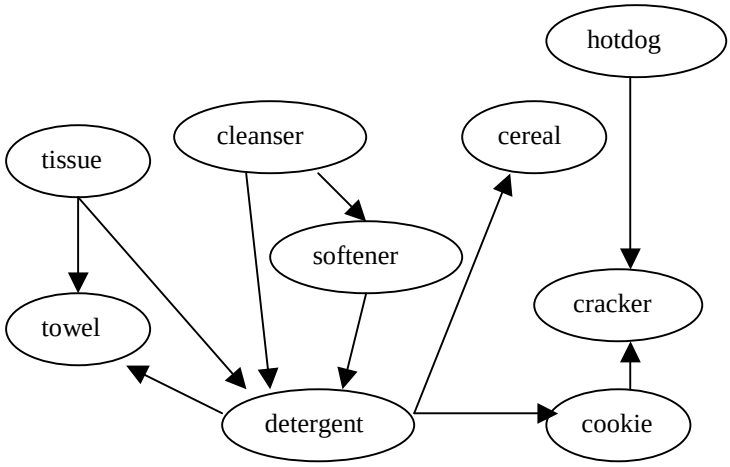

Figure 6: Product cluster

Results of Model 2

We now add promotion information in the learning process. There are three types of marketing activities available in the IRI data. They are price discount, display, and feature. In this application, these three variables are combined into one binary promotion variable to indicate if there is at least one marketing activity for the

category. We set a limit on the MC^3 learning process in this model. That is, there should be no parents for any promotion variable. At cutoff point 0.9, we identify three clusters, which are shown in Figure 7, 8, and 9.

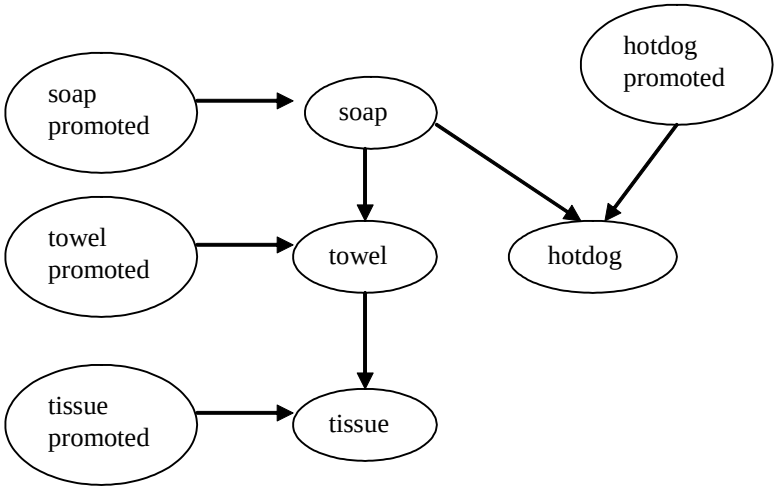

Figure 7: Product cluster one with promotions included, cutoff at 0.9

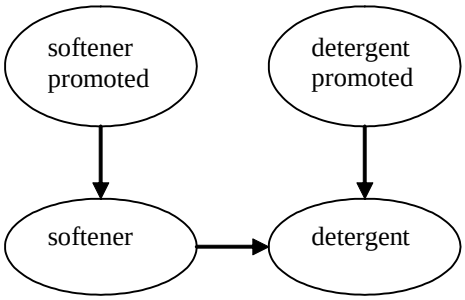

Figure 8: Product cluster two with promotions included, cutoff at 0.9

We find that cross category promotions work through indirect effects primarily. That means, promotions in category A impacts sales of A. In turn the sale or no-sale decision of category A will influence sales of its complement (or substitute) category B. Thus promotions in category A indirectly impact sales of category B. Using a cutoff value 0.9, there are three multi-category clusters. They are (towel, tissue, soap, hotdog), (softener, detergent), and (cereal, cracker, cookie). Using a cutoff value 0.6, the three clusters are merged into two cluster,which are shown in Figure 10, and Figure 11.

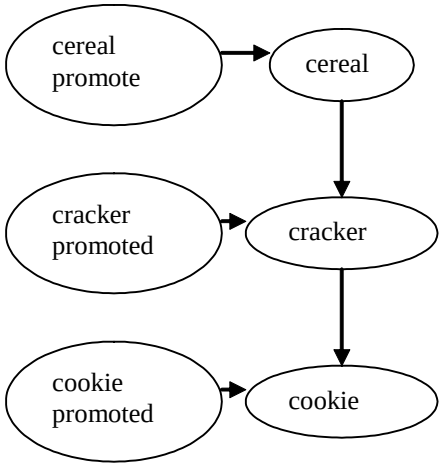

Figure 9: Product cluster three with promotions included, cutoff at 0.9

Results of Model 3

The two types of demographic information used are family size and income. Here is the definition of the two variables. If there are more than two members in a family, the "family size" variables equals to one, otherwise zero. If the family income is over 35 thousand dollars, the "income" variable equals to one, otherwise zero.

Figure 12 illustrates the average network of the model. Family size is related to more categories than income. The possible explanation is that the categories in our data set are most staples. Despite income, most families needs to buy products from these categories. Like model 1, there is one large multi-category cluster. However, the direction of some edges differs, indicating customer segments whose shopping behavior is different from the mass.

These models show Bayesian network learned from data can capture the multi-product relationship. It also shows that with promotion data, Bayesian network can capture a model that is closer to the underlying mechanism.

4.3 Pairwise Assessment

We can see that with more information, Bayesian networks are better able to learn the presence of multi-category relationships. We can identify different typology of

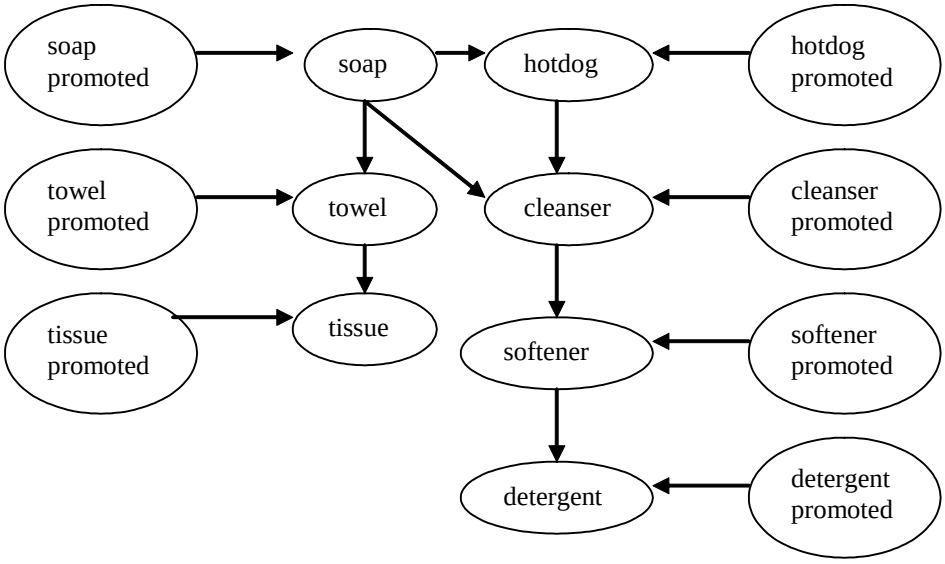

Figure 10: Product cluster one with promotions included, cutoff at 0.6

complements using Bayesian network combined with localized rule discovery. Table 2 gives examples of category pairs and contrasts each pair's strength of association with the corresponding edge probabilities in the three models.

Pair	Assoc. Rule lift	Prob. in Model 1	Prob. in Model 2	Prob. in Model 3
(Soap, Cleanser)	3.98	0.83	0.63	0.89
(Detergent, Softener)	3.55	0.99	0.99	0.99
(Towel, Tissue)	1.86	0.99	0.99	0.99
(Cracker, Cookie)	1.63	0.99	0.96	0.80
(Cereal, Yogurt)	1.68	0.84	0.62	0.99

Table 2: Comparison across different approaches

Pair	Strong BN Presence	Moderate BN Presence
High lift	(Detergent, Softener)	(Soap, Cleanser)
Moderate lift	(Towel, Tissue)	

Table 3: Comparing Lift to probability of presence in BN

These results are summarized in Table 3, and show that true use complements always have a strong relationship. For example, detergent and softener have both high lift and high presence probability. Spurious complements such as soap and cleanser will disappear when promotion effects are considered. Findings such as towel and tis-

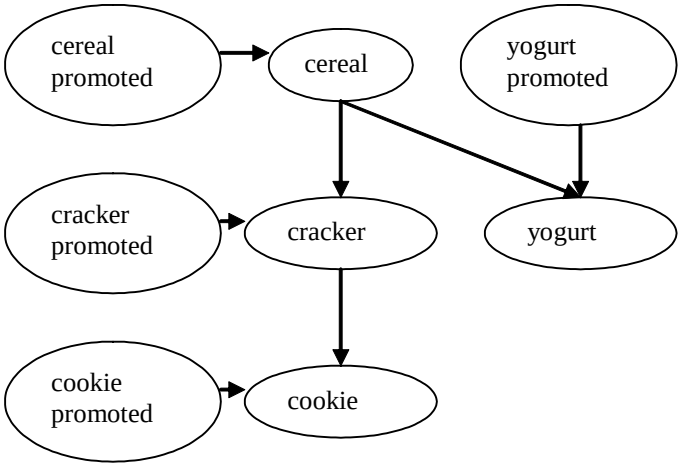

Figure 11: Product cluster two with promotions included, cutoff at 0.6

sue, and cracker and cookie can be utilized to cross sell. Another benefit of pairwise assessment is identification of primary and secondary categories in a pair of use complements. For strong complements detergent and softener , three Bayesian networks were learned with their sales data and their promotion data. Figure 13 has a probability of 98% and Figure 14 has a probability of 1%. There is less than 1% probability that the two categories are independent, which is not shown here. We may conclude that softener is the primary item in the Detergent-Softener relationship.

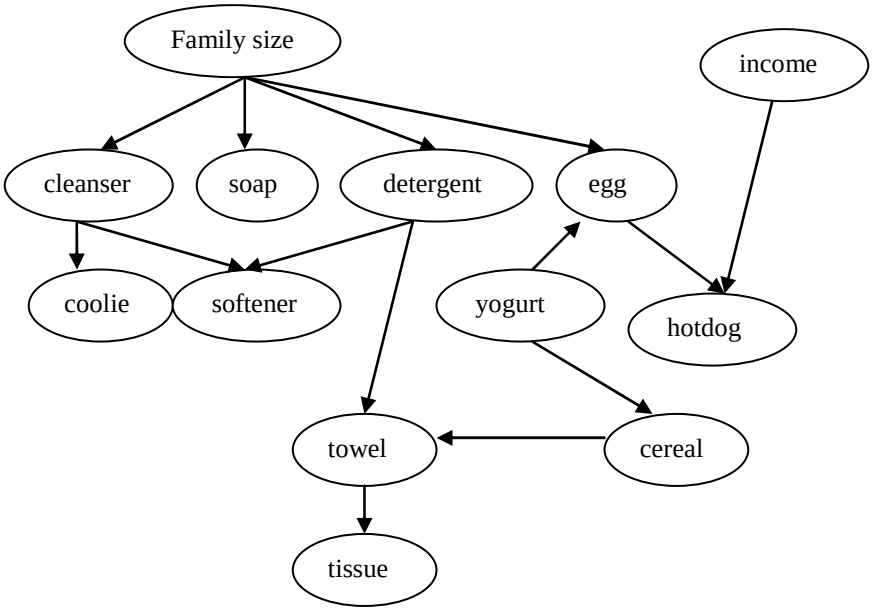

Figure 12: Product cluster with demographics included

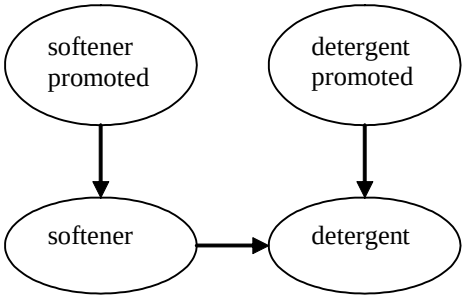

Figure 13: Softener as primary category

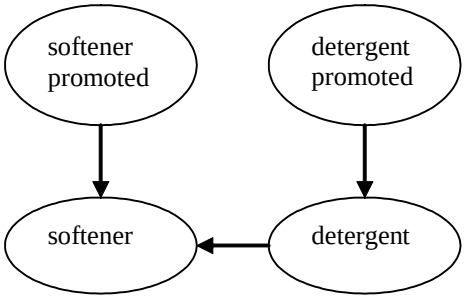

Figure 14: Detergent as primary category

;

4.4 Summary

There are three interesting findings in this application. With grocery data, promotion variable is more useful in identifying product clusters than sales data only. In grocery store, family size is more related to product relationships than income. Complements such as (detergent, softener) and (towel, tissue)can be identified when we compare the all the three networks. With findings like these, a retailer will be able to coordinate marketing activities and target specific customers accordingly. This research, coupled with essay two, can be applied to study the cross category effects of marketing mix activities as shown in Figure 15. Or they can be used to identify customer segmentations as shown in Figure 16. We would further extend this research to a multi-period model. In such a model, customers' purchase decisions in current week impact their purchase decisions of next week. Thus the model will be more closer to reality.

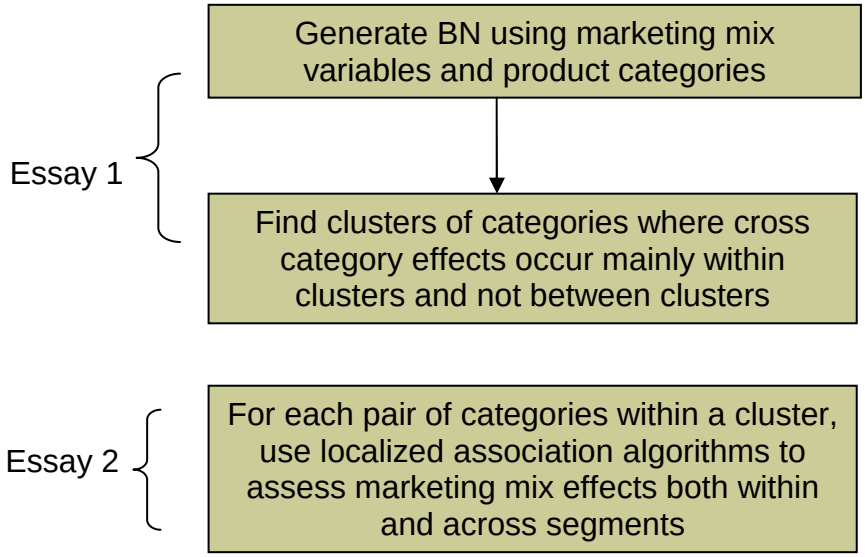

Figure 15: Cross category effects of marketing mix activities

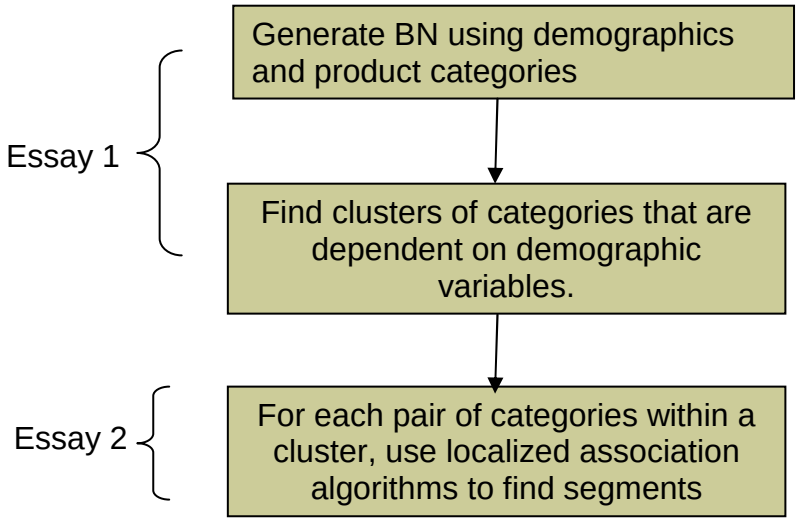

Figure 16: Customer segmentation using cross category associations

References

Baesens, B., S. Viaene. 2002. Bayesian neural networks learning for repeat purchase modeling in direct marketing. *European Journal of Operational Research* **138(1)** 191–211.

Blattberg, R., R. Briesch, F. Edward. 1995. How Promotions Work. *Marketing Science* **14(3)** 122–132.

Chib, S., E. Greenberg. 1995. Understanding the Metropolis-Hastings Algorithm. *The American Statistician* **49(4)** 327–335.

Chib, S., P. Seetharaman, A. Strijnev. 2002. Analysis of Multi-Category Purchase Incidence Decisions Using IRI Market Basket Data. *Econometric Models in Marketing* **16** 65–90.

Chickering, D. M. 1996. Learning Bayesian networks is NP-complete. *LECTURE NOTES IN STATISTICS* **114** 121–130.

Cooper, G. 1990. Computational complexity of probabilistic inference using Bayesian belief networks. *Artificial Intelligence* **42(2/3)** 393–405.

Cooper, G., E. Herskovits. 1992. A Bayesian Method for Induction of Probabilistic Networks from Data. *Machine Learning* **9(4)** 309–347.

Cooper, L. 2000. Strategic marketing Planning for Radically New Products. *Journal of Marketing* **64(1)** 1–16.

Cui, G., M. Wong, 2006. Machine Learning for Direct Marketing Response Models: Bayesian Networks with Evolutionary Programming. *Management Science* **52(4)** 597–612.

Dagum, P., M. Luby. 1993. Approximating probabilitic inference in Bayesian belif networks is NP-hard. *Artificial Intelligence* **60** 141–153.

Dawid, P. 1992. Applications of a general propagation algorithm for probabilistic expert systems *Statistics and Computing* **2** 25–36.

Giudici, P., R. Castelo. 2001. Association Models for Web Mining. *Data Mining and*

Knowledge Discovery **5** 183–196.

Giudici, P., G. Passerone. 2002. Data Mining of Association Structures to Model Consumer Behavior. *Computational Statistics and Data Analysis* **38** 533–541.

Giudici, P., R. Castelo. 2003. Improving Markov Chain Monte Carlo Model Search for Data Mining. *Machine Learning* **50** 127–158.

Hastings, W.K. 1970. Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika* **57(1)** 97–109.

Heckerman, D. 1996. A tutorial on learning with Bayesian networks. Microsoft Research.

Hruschka, H., M. Lukanowicz, C. Buchata. 1999. Cross category sales promotion effects. *Journal of Retailing and Consumer Services* **6** 99–105.

Jensen, F., S. Lauritzen. 1990. Bayesian updating in recursive graphical models by local computations. *Computational Statisticals Quarterly* **4** 269–282.

Lauritzen, S., D. Spiegelhalter. 1988. Local computations with probabilities on graphical structures and their application to expert systems. *Journal of Royal Statistical Society B*. **50** 157–224.

Madigan, D., J. York. 1995. Bayesian graphical models for discrete data. *International statistical Review* **63(2)** 215–232.

Manchanda, P., A. Asim, G. Sunil. 1999. the "Shopping Basket": a Model of Multicategory Purchase Incidence Decisions . *Marketing Science*. **18(2)** 95–114.

Metropolis, N., A. Rosenblth, M. Rosenbluth, A. Teller, E. Teller. 1953. Equations of State Calculations by Fast Computing Machines. *Journal of Chemical Physics* **21** 1087–1092.

Mulhern, F., R. Leone. 1991. Implicit Price Bundling of Retail Products: a Multiproduct to Maximizing Store Profitability. *Journal of Marketing* **55** 63–76.

Raftery, A., D. Madigan, J. Hoeting. 1997. Bayesian Model Averaging for Linear Regression Models. *Journal of the American Statistical Association* **92(437)** 179–191.

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

Russell, G., A. Petersen. 2000. Analysis of Cross Category Dependence in Market Basket Selection. *Journal of Retailing* **76(3)** 367–392.

Robinson, R. W. 1977. Counting unlabelled acyclic digraphs. *Lecture Notes in Mathematics: Combinatorial Mathematics V* Springer-Verlag.

Seetharaman, P.B., S. Chib, A. Ainslie, P. Boatwright, T. Chan, S. Gupta, N. Mehta, V. Rao, A. Strijnev. 2005. Models of Multi-Category Choice Behavior. *Marketing Letters* **16(3/4)** 239–254.

Shachter, R. 1988. Probabilistic inference and influence diagrams. *Operations Research* **36** 589-604.

Localized Rule Discovery in Market Basket Data

Xiaojun Li

School of Business, the George Washington University

Washington DC

December, 2007

Abstract

Association rule mining approaches have been proposed in the data mining literature to analyze market basket data. The outputs of such approaches are rules that identify pairs of associated products that imply the co-occurrence of particular products in a basket. As established in the marketing research literature, several factors can influence the co-occurrence of products. First, marketing mix activities such as pricing and promotions in one product category may influence a consumer's purchase decision of products in other categories. Second, due to consumer heterogeneity, the association of a set of products may vary across customer segments. Third, a basket of products may just be purchased together coincidentally since the consumer wants to spread the shopping cost of one trip. Finally, aggregate associations (or correlations) may differ from localized associations. In the extreme case, we may see Simpson's paradox, which means that a pair of products is positively correlated in every consumer segment but shows negative correlation in the aggregate data or vice versa. This happens when market baskets from multiple sources are pooled into one aggregate database. Thus product associations may be distorted in two ways - change of magnitude, or change of direction. As a result, association rules discovered by existing association rule mining may be spurious. We develop an exploratory rule mining algorithm based on transaction attributes such as consumer demographics or marketing mix variables that identifies segments of the data (a subset of the baskets) which exhibit strong associations between pairs of products that are not seen in the aggregate data set. Results are presented using an IRI market basket data set that contains transactions including 22 categories over 2 years for 500 panelists.

1 Introduction

Consumers typically purchase multiple products from multiple categories in one shopping trip. Marketing research has attempted to analyze the multi-category purchase information to plan marketing activities accordingly so as to maximize store profit. For example, a retailer can reduce the price of cake mix to sell more cake mix and frostings (Mulhern and Leon 1991), as a result of which the overall profit may improve. Earlier research attempted to capture these cross-category effects using data aggregated at the weekly or monthly level instead of the transaction (or market basket) level. In the 1990s, advances in information technology have made the collection and storage of basket level transaction data economically and technically feasible. Subsequently, multiple data analysis techniques have been developed so as to find associated products through analysis of market basket data.

Early research using market basket data focused on identifying what products are purchased together frequently. Given the number of products carried in a typical retailing store, this trivial question can be computationally challenging. For example, consider a store carrying a thousand products. To identify what two products have been purchased together frequently, one needs to check the frequency of roughly half a million pairs of products. This analysis needs to be performed in a reasonably short time since the retailer needs to adjust promotion and pricing decisions accordingly. To achieve this goal, association rule mining algorithms which are exploratory data analysis techniques, have been developed (Agrawal et al. 1993, Agrawal and Ramakrishnan 1994). An association rule is a statement such as "If a customer buys item A, then he will buy item B with probability c%". Association rule mining is the task of finding these rules in very large databases. A retailer can then decide as to which products to promote and at what levels based on these rules. Early research in association rule mining was based on frequency measures, but more recent research has started examining associations using other measures such as Chi-square, lift, etc. A more detailed discussion on the various measures of association is given in Section 2.2.

Approaches to measure cross category effects in the marketing literature have been theory driven, and typically based on utility theory models. Marketing researchers have long realized that products or categories can be complements or substitutes. Promotions in one category can improve not only its own sales but also sales of its complements, while suppressing sales of its substitutes. Sometimes, products may also appear in a basket together coincidentally without necessarily being complements, since a consumer may decide to purchase several products in one trip just in order to minimize the shopping cost. During the late 1990s, multivariate models have been developed to address the question of cross-category effects. These utility theory based models usually decompose the utility of a category into its own effects and cross-category effects. Some models go further to isolate cross-category marketing mix effects from cross-category co-incidence effect. An understanding of these effects can help retailers improve their profits. For example, if the co-occurrence is due to complementarity rather than just co-incidence, a retailer will be able to co-ordinate the marketing mix in order to maximize profits.

Compared to the data mining approach, these multivariate models provide more precise measures of cross-category effects, albeit at a high computational cost since they typically involve Markov Chain Monte Carlo (MCMC)models .(Chib et al. 2002, Manchanda et al. 1999, Russell and Peterson 2000). The number of products or categories in a real life database makes it infeasible to use these approaches to model all product associations simultaneously without selecting specific products a priori by experts. Therefore even as marketing models continue to develop, the sheer amount of data makes association rule mining indispensable in large data sets. As a data-driven approach association rule mining does not need prior knowledge of products across which associations may exist. One more drawback is that these statistical models usually assume linear relationships. They usually don't consider the interactions among predictors, which is more often the case in the real world.

Association rule mining, although used widely as an exploratory tool in market basket data analysis, is not without problems. It typically mines rules on the aggregate

dataset, which is an agglomeration of data subsets. Statisticians have observed that when subsets of data are pooled together, associations between two variables may change in magnitude or even in direction. This effect is called association reversal in the statistics literature.

In marketing, these subsets can be sales data from multiple locations over years, from different marketing plans, or from different consumer segments. For those weakly associated products in the aggregated dataset, there might be customer segments in which they are associated more strongly. As such, decisions made according to these aggregate level data can be misleading.

Research and Contribution

This paper proposes a new association mining approach that considers the influence of aggregation on association discovery in market basket data. The results from this analysis should help the retailer in identifying customer segments in which specific pairs of products are strongly associated, and also help determine the marketing mix effects on cross product associations.

Aggregate data will be partitioned into subsets according to such attributes as consumer demographics, marketing mix, etc. In each subset, the association will be evaluated, and since this association is calculated locally for the subset, it is called a local association. If the local associations are not strong (or interesting) enough, the subsets will be further partitioned until they meet the preset minimum requirements. The results of this process are localized association rules, which are defined as $(\mathbf{X} = \mathbf{x} \rightarrow (Y_i, Y_j))$, where $\mathbf{X}$ is a set of attributes based on either customer demographics, or marketing mix variables, and Y_i, Y_j are a pair of products. Following the association rule mining tradition, the strength of the association is assessed using the lift measure, which is described later in section 2. We also measure the homogeneity of associations across subsets of data in two different ways. First, if there is a change in the direction of the association, and secondly if there is a change in the magnitude of the association. We examine the data subsets (segments) to look for association

reversal, and also check if the association is homogeneous in magnitude within the sub segments.

Given a large number of transaction attributes over which the segments may be defined, there is a need to have a fast process that selects a parsimonious set of attributes that result in homogeneous segmentation. A similar process is used in classification tree algorithms such as CHAID. In this research, aggregate data is segmented in a tree structure, wherein branches in the tree correspond to transactions segmented according to attribute variables. The root node represents the aggregate data, and the leaf nodes correspond to a particular subset (perhaps indicating a customer segment).

The contributions this research makes are summarized as follows:

1. It extends traditional association rule mining by introducing localized associations. The ideas of association reversal and association homogeneity have not been discussed extensively in the data mining literature.

2. It proposes a heuristic approach for selecting attributes that impact the association of two binary variables mostly.

3. The output of the process is rules in the format of $(\mathbf{X} = \mathbf{x} \rightarrow (Y_i, Y_j))$. These rules are more informative and actionable than rules derived from aggregate data. The association of a pair of products is based on attributes such as consumer demographics, marketing mix, etc, and retailers can take actions accordingly based on such rules (Russell et al. 1999).

This paper is organized as follows. Section 2 reviews market basket analysis from both data mining and marketing research areas. Section 3 discusses a framework of association rule mining in segmented data, and provides an algorithmic implementation of the new framework. Section 4 presents results from the application of these algorithms in a retailing context.

2 Literature Review

Models of market basket data have been developed by both marketing and data mining researchers. Marketing models are theory guided. Data mining models are usually data driven. Section 2.1 briefly introduces choice models from marketing literature. Section 2.2 introduces association rule mining, Section 2.3 discusses the association reversal effect from statistics literature.

2.1 Marketing Research Models of Market Basket Data

How do consumers make purchase decisions across multiple product categories? Russell et al. (1999) point out three factors that may influence their purchase decisions:

- Cross-category consideration: The consumer needs to make one choice out of the alternatives that satisfy the same consumption purpose. These alternatives are often referred to as substitutes. For example, one may decide to choose one beverage from juice or soft drinks.
- Cross-category learning: a consumer's prior experience in one category impacts later choice in a different category.
- Product bundling: Products from multiple categories need to be combined together for consumption which results in them being purchased together. These products are usually consumption complements such as cake mix and frosting, etc.

According to Manchanda et al. (1999), the co-occurrence of two or more products purchased from different categories can result from co-incidence as well as consumption complementarity. For example, several unrelated products are bought together since the consumer decides to spread the cost of the trip over many items. Since consumption complementarity is usually not observable to the retailer, marketing researchers usually model purchase complementarity. Purchase complements are those

products wherein marketing actions in one category influence the purchase decisions in the other categories.

Several utility theory based choice models have been developed to model the co-occurrence of categories in market baskets. The utility of a category is a function of its own marketing mix and related categories' market mix. Purchase complementarity is captured by the coefficients associated with price or promotion variables. The co-incidence is usually modeled as the residual that follows a multivariate normal distribution. Customer heterogeneity also plays a part in the relationship among intrinsic utilities and price responses. Manchanda et al. (1999) build a multivariate probit model, which will be used to illustrate this modeling approach. Given J categories, the latent utility of category n for household h on trip t can be modeled as

$$u_{hnt} = \beta_{hn0} + \beta_{hn1} OwnEffects + \beta_{hn2} CrossEffects + \epsilon_{hnt}$$

The link between the utility of a category and a consumer's choice in the category is represented as follows

$$y_{hnt} = \begin{cases} 1 & if\ u_{hnt} > 0 \\ 0 & if\ u_{hnt} \leq 0 \end{cases}$$

The utility for household h on trip t for all the J categories can be expressed as

$$\mathbf{u_{ht}} = \mathbf{X_{ht}} \beta_{\mathbf{ht}} + \epsilon_{\mathbf{ht}}$$

where the jth row of the matrix $\mathbf{X_{ht}}$ are the causal (marketing mix) variables influencing the utility for the jth category. Unobserved influences on this trip are captured by $\epsilon_{\mathbf{ht}} \sim \mathbf{MVN}(\mathbf{0}, \mathbf{\Sigma})$, where $\mathbf{\Sigma}$ is a $J \times J$ covariance matrix. This correlated structure of errors represents the co-incidence among categories. For example, if $cov(\epsilon_{hit}, \epsilon_{hjt}) > 0$, then an increase of utility in category i will result in an increase of utility in category j.

Since the consumers are heterogeneous, the household level parameters can be modeled as linear function of demographic variables,

$\beta_{\mathbf{h}} = \mathbf{D_h}\mu + \lambda_{\mathbf{h}}$

where $\beta_{\mathbf{h}} = \{\beta_{\mathbf{h0}}, \beta_{\mathbf{h1}}, \beta_{\mathbf{h2}}\}$, where $\lambda_{\mathbf{h}} = \mathbf{MVN}[\mathbf{0}, \mathbf{\Lambda}]$, and $\mathbf{D_h}$ is a matrix containing the demographics and other household specific variables. μ captures the impacts of demographic variables. All other unobserved heterogeneity is captured in the vector $\lambda_{\mathbf{h}}$.

Manchanda et al. (1999) fit their model on two product categories at a time. Chib et al. (2002) fit a multivariate probit model on twelve categories simultaneously. One empirical finding is that a model with fewer categories underestimates the magnitude of cross-category correlations and overestimates the effectiveness of the marketing mix such as price, promotions. This finding implies the importance of modeling a large amount of products. Another empirical finding of Chib et al. (2002) is that ignoring unobserved heterogeneity will overestimate cross-category correlations.

2.2 Association Rule Mining

We now discuss approaches from the data mining standpoint, and in particular the Association Rule Mining literature. Association Rule Mining is the process of finding rules in the market basket data (Agrawal et al. 1993), where a rule $Y_1 \rightarrow Y_2$ consists of a left hand condition and a right hand consequence. Without considering physical storage format, a market basket data set is a matrix of 0s and 1s. Each column represents a variable, and each row represents a record. In data mining, the variables are called items, a set of items are called an itemset, and the rows are called transactions. After preprocessing the raw data, the two items or itemsets of a rule can be summarized in a contingency table as in Table 1.

The frequency of an itemset is usually measured by *support*, the fraction of transactions that contain the itemset.

	$Y_2 = 1$	$Y_2 = 0$	
$Y_1 = 1$	n_{11}	n_{10}	n_{1+}
$Y_1 = 0$	n_{01}	n_{00}	n_{0+}
	n_{+1}	n_{+0}	N

Table 1: A 2×2 contingency table

$$s(Y_1 Y_2) = P(Y_1 = 1, Y_2 = 1) = \frac{n_{11}}{N}$$

The confidence of a rule is a directional measure of association. It is a conditional probability, and can be written as,

$$c(Y_1 = 1 \Rightarrow Y_2 = 1) = P(Y_2 = 1 | Y_1 = 1) = \frac{n_{11}}{n_{+1}}$$

or

$$c(Y_2 = 1 \Rightarrow Y_1 = 1) = P(Y_1 = 1 | Y_2 = 1) = \frac{n_{11}}{n_{1+}}$$

Since a typical market basket dataset may have terabytes of data, the typical research focus in association rule mining is algorithm efficiency. The most straightforward way of solving the problem is to enumerate all possible combinations of items. However, this task is infeasible given the number of products carried in a typical market basket database. Many algorithms have been proposed to mine this type of data. They include a priori method, dynamic item counting, parallel and distributed computing, etc. A review of these algorithms can be found in Hipp et al. (2000).

To assess the quality of a rule, many association measures have been proposed (Tan and Kumar 2000). Since the support and confidence cannot measure the dependency between the two itemsets, other measures of association have been used in the literature. Examples are Chi-square and lift (Brin et al. 1997), Pearson's correlation coefficient (Xiong et al. 2004), mutual information (Smyth and Goodman 1992), among others. A thorough review can be found in (Tan and Kumar 2000, Tan et al. 2002). The next section introduces a few measures that are going to be used in this

research.

Association Rule Mining Measures

Based on Table 1, here are a few definitions of association rule mining measures.

1. Chi-Square

   Chi-square is used to determine whether two itemsets are associated.

$$\chi^2 = \frac{N(n_{11}n_{00} - n_{01}n_{10})^2}{n_{1+}n_{0+}n_{+1}n_{+0}}$$

2. Lift (or Interest)

   The lift for a pair of itemsets, also called the interest, is a measure of the strength of the association:

$$lift(Y_1Y_2) = \frac{Nn_{11}}{n_{+1}n_{1+}}$$

   When lift is greater than one, it indicates a positive association, when lift is equal to one it indicates independency, and when lift is less than one it indicates negative association.

3. Odds ratio

   The odds ratio is defined as:

$$\theta = \frac{n_{11}/n_{10}}{n_{01}/n_{00}} = \frac{n_{11}n_{00}}{n_{01}n_{10}}$$

   $\theta = 1$ represents independence between the respective itemsets. The farther away theta is from 1, stronger the association. If associations are the same but in opposite directions, the odds ration of one is the inverse of the other.

4. Pearson's Phi Coefficient,

$$\phi = \frac{n_{10}n_{01} - n_{11}n_{00}}{\sqrt{(n_{11} + n_{01})(n_{10} + n_{00})(n_{11} + n10)(n_{01} + n_{00})}}$$

assuming that

$$(n_{11} + n_{01})(n_{10} + n_{00})(n_{11} + n10)(n_{01} + n_{00}) > 0$$

5. Mutual Information

The relationship between two variables can be measured with mutual information,

$$I(Y_1; Y_2) = \sum_i \sum_j P(y_i, y_j) \log_2 \frac{P(y_i y_j)}{P(y_i)P(y_j)}$$

Mutual information is always positive or zero. It is zero when Y_1 and Y_2 are independent.

For a review of the strength and limitations of these measures, see Tan and Kumar (2000).

Association Based Rule Learning

Given a set of transactions, the most straightforward way of learning rules is to enumerate all combinations of items and count their frequencies. If they meet the constraint of a pre-defined minimum support, they will be identified as large itemsets. Once a large itemset is ready, candidate association rules will be generated and evaluated from it. One research focus of association rule mining is to discover all large itemsets fast.

The most widely used algorithm is the Apriori algorithm (Agrawal and Ramakrishnan 1994). Apriori algorithm uses a property that any subset of a large itemset must be a

large itemset. For example, if $s(Y_1Y_2) > minsup$, then $s(Y_1) > minsup$ and $s(Y_2) > minsup$. Therefore, if $s(Y_1) < minsup$ or $s(Y_2) < minsup$, then $s(Y_1Y_2) < minsup$. Apriori algorithm will assess all single items first, for those whose supports are lower than the minimum support, there is no need for further analysis.

In addition to support-based algorithms, several association measures based algorithms have also been proposed (Brin et al. 1997). One such correlation based algorithm called TAMPER (Xiong *et al.* 2004) is briefly discussed next. Using support values, we can re-write the formula of Pearson's Phi as

$$\phi = \frac{s(Y_1, Y_2) - s(Y_1)s(Y_2)}{\sqrt{s(Y_1)s(Y_2)(1 - s(Y_1))(1 - s(Y_2))}}$$

There is an upper bound of Pearson's Phi , which is

$$upper(\phi(Y_1, Y_2)) = \sqrt{\frac{s(Y_2)}{s(Y_1)}}\sqrt{\frac{1 - s(Y_1)}{1 - s(Y_2)}}$$

The upper bound of the Phi correlation coefficient has a monotone property. If we index the variables according to their support so that $s(Y_i) > s(Y_{i+1}) > ... > s(Y_{i+k})$, then we will have $upper(\phi(Y_i, Y_{i+1})) > ... > upper(\phi(Y_i, Y_{i+k}))$. If $upper(\phi(Y_1, Y_{i+1}))$ is less than the preset threshold, there is no need to calculate $\phi(Y_i, Y_{i+2})$, ..., $\phi(Y_i, Y_{i+k})$. Thus the algorithm reduces the search effort.

Information Theory Based Rule Learning

Classification rules can be considered as a special form of association rules in that they have one target variable. Mutual information has been used as the primary measure in learning classification rules. An example of decision tree is the algorithm ID3 (Quinlan 1986). At each node of the tree, the algorithm chooses an independent variable X that will maximize mutual information $I(Y; X)$. Fleuret (2004) aims to select independent variables that maximize conditional mutual information.

$$I(Y;Z|X) = \sum_{y,z,x} P(y,z,x) \log \frac{P(y,z|x)}{P(z|x)P(y|x)}$$

2.3 Association Reversion and Homogeneity of Associations

A common practice in the statistics literature is to test for the homogeneity of associations in subsets of the aggregate data. One example of association reversion is described as Simpson's paradox (Simpson 1951), wherein the direction of an association reverses when several groups of data are combined into one single group. Although this has been discussed in the statistics literature, it has not received as much attention by data mining researchers, perhaps due to the fact that the task of association rule mining has typically been defined in an aggregate data set.

Simpson's paradox is the phenomenon that $P(A|B) < P(A|\overline{B})$ even though $P(A|B) \geq P(A|\overline{B})$ under the additional conditions of both C and $\overline{C}$ (Simpson 1951). One real life Simpson's paradox is observed in Cohen and Nagel (1934). In 1910, the overall tuberculosis mortality rate was higher in New York city than Richmond, Virginia. However, in both white and non-white racial segments, the mortality rates were actually lower than Richmond, Virginia.

Two other real life examples are reported in Wagner (1982). In early 1979, the publishers of American History Illustrated were pleased to notice the overall renewal rate increased from 51.2% to 64.1%. The paradox was actual renewal rate declined in every category. Between 1974 and 1978, the tax rate decreased in each income category, yet the overall tax rate increased from 14.1% to 15.2%. Blyth (1972) points out that "...Simpson's paradox is the simplest form of the false correlation paradox in which the domain of x is divided into short intervals, on each of which y is a linear function of x with large negative slope, but these short line segments get progressively higher to the right, so that over the whole domain of x, the variable y is practically a linear function of x with large positive slope. "

It has been noticed that Simpson's paradox is an extreme example of confounding

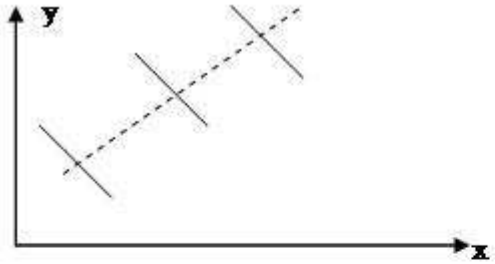

Figure 1: Simpson's Paradox

effects. Mittal (1991) and Samuels (1993) thoroughly discuss three types of paradoxes. Let us use α as a general measure of association, and let a 2×2 contingency table be represented by $A = (n_{11}, n_{10}; n_{01}, n_{00})$. Also, let the sub-tables be denoted as $A^i = (n^i_{11}, n^i_{10}; n^i_{01}, n^i_{00})$, wherein

$$\sum_i n^i_{11} = n_{11} \qquad \sum_i n^i_{10} = n_{10}$$

$$\sum_i n^i_{01} = n_{01} \qquad \sum_i n^i_{00} = n_{00}$$

Yule's Association Paradox (YAP): Two variables are independent in subsets, but associated in the aggregate, $\alpha(A^1) = \alpha(A^2) = 0$ but $\alpha(A) \neq 0$.

Yule's Reversal Paradox (YRP), or Simpson's paradox as it is known more popularly. It means the direction of association changes when pooling subsets into one set. It is defined as $n^i_{11}n^i_{00} - n^i_{10}n^i_{01} \geq 0$ (≤ 0) , for $i = 0, 1, 2, ...$ But $n_{11}n_{00} - n_{10}n_{01} \geq 0 (\leq 0)$. Fabris (1999) discusses the discovery of the Simpson's paradox in data mining. Given a goal variable G and two sets of attribute variables L_1, L_2, the algorithm exhaustively test whether an attribute variable A_2 from L_2 will change the relationship between G and A_1, where A_1 is an attribute variable from L_1.

Amalgamation Paradox (AMP): The aggregate association is either smaller than the smallest subset association, or greater than the greatest subset association $\alpha(A) < min_i(A^i)$ or $\alpha(A) > max_i(A^i)$. This is what we see in our data.

Homogeneity of Associations

The homogeneity of associations can be qualitative, which means the sign of the association are consistent in subpopulations . Mittal (1991) further classifies this type of homogeneity into row and column homogeneous. Meanwhile, there are several procedures to test the quantitative homogeneity of some association measures. For example, one can use Chi-square test to check the homogeneity of Pearson's Rho values:

$$\chi^2 = \sum (n_k - 3) z_{rk}^2 - \frac{[\sum (n_k - 3) z_{rk}]^2}{\sum (n_k - 3)}$$

$$df = k - 1, k > 1$$

where $z_r = 0.5 \times \log_e(\frac{1+r}{1-r})$, which is Fisher's transformation of correlation coefficient. Paul and Donner (1989, 1992) compare the performance of multiple procedures for testing the homogeneity of odds ratios in K 2×2 tables. A typical test procedure of odds ratios has been proposed in Breslow and Day (1980).

3 Methodology

This section will introduce localized rule discovery as a new market basket data analysis method. Section 3.1 will formally define the research problem. Section 3.2 describes the overall analysis process and compares it to the classical association rule mining method. Section 3.3 is the detailed implementation of the localized rule discovery process.

3.1 Problem Statement

The formal description of the research is:

Given a set of items $\mathbf{Y} = (Y_1, Y_2, ...Y_n)$, a set of attributes $\mathbf{X} = (X_1, X_2, ..., X_m)$, and a market basket dataset D, learn rules $(X^1 = x^1, X^p = x^p \rightarrow (Y_i Y_j))$ that meet preset

constraints. Note that $(X^1, ..., X^p) \subseteq \mathbf{X}$, $(Y_i Y_j) \subseteq \mathbf{Y}$. Rule $(X^1 = x^1, ..., X^p = x^p \rightarrow (Y_i Y_j))$ can be expressed as a conditional association, $(Y_i Y_j | X^1 = x^1, ..., X^p = x^p)$.

General preset constraints include:

1. Minimum support, which is the minimum fraction of transactions that contains $(Y_i Y_j)$.
2. In the dataset defined by $X^1 = x^1, ..., X^p = x^p$, $(Y_i Y_j)$ is homogeneous. Homogeneity can be
   - No association reversal phenomena. And/or
   - $(Y_i Y_j)$ is quantitatively homogeneous in the subsets.

The most straightforward way of fulfilling this task is to search through all subsets of the aggregate data for every pair of products. However given the large number of products and transaction attributes, exhaustive search would be a formidable task. There are two effective strategies to reduce search space:

1. Reduce the number of product pairs. Since a pair of use complementary products is very likely to display strong association in every subset, there is no need to further test these pairs. This research assumes retailers are not interested in those pairs with high support and strong lift in the aggregate data.

2. Reduce the number of subsets. Given m attributes $X = (X_1, X_2, ..., X_m)$, there are 2^m attribute-sets such as $\{X_1\}, ..., \{X_1 X_2\}, ..., \{X_1 X_2 ... X_m\}$. This can be done using heuristic algorithms. This research will use a tree-structured algorithm to divide data into segments in a piece-wise manner and assess the associations in each node.

The General Algorithm

Input: a set of items $\mathbf{Y}$; a set of attributes $\mathbf{X}$

1) Identify product pairs (Y_i, Y_j), that are candidate itemsets for localized rule discovery.

2) FOR each pair

3) construct a root node D_0 ;

4) build-a-tree (D_0, Y_i, Y_j, X);

5) NEXT

After the tree is complete, report rules found. A path from the root node to a leaf node will form a rule. Note in step 3 and step 4, all measures of associations such as support, chi-square and lift are calculated in the subset. There is a fundamental difference between this practice and the classical association rule mining. This difference will be shown in section 3.2. More details of step 3 and step 4 will be given in section 3.3. This algorithm is similar to classification tree with one difference: the target variable is the association measure, not the class variable.

3.2 Localized Rule Discovery

Localized Rule

Let Y_i be a binary variable, representing an item as usual. A transaction t is still a set of items. In addition to the items it contains, a transaction is also labeled by X_j , which are categorical variables such as consumer demographics or marketing mix variables, etc. An extended form of transaction t is a set of items and transaction attributes, $t = (y_1, ..., y_n, x_1, ..., x_m)$.

In this extended database, we study the problem of whether item1 and item2 are correlated given a condition. The condition is a set of the attribute variables $\mathbf{X} = \mathbf{x}$, which could refer to a particular group of customers or a specific marketing program. A localize rule is a three-part representation (condition, item1, item2), or $(\mathbf{X} = \mathbf{x} \rightarrow (Y_i, Y_j))$. In classical association rule mining, the condition and item1 will be combined together and expressed as (condition and item1, item2), or as $(\mathbf{X} = \mathbf{x} \wedge (Y_i) \rightarrow Y_j)$. In a localized definition of association rule $(\mathbf{X} = \mathbf{x} \rightarrow (Y_i, Y_j))$, relation of(Y_i, Y_j) is defined on a particular dataset $\mathbf{X} = \mathbf{x}$. If $\mathbf{X} = \Phi$, meaning no particular attributes on the dataset, the association rule mining will be performed

on the aggregate dataset. Thus the tradition definition of association rule becomes a particular case of localized rule mining.

	Y_j	$\overline{Y_j}$	
XY_i	n_{11}	n_{10}	n_{1+}
$\overline{XY_i}$	n_{01}	n_{00}	n_{0+}
	n_{+1}	n_{+0}	N

Table 2: Contingency table according to classical two-part expression

	XY_j	$X\overline{Y_j}$	
XY_i	n_{11}	n_{10}	n_{1+}
$X\overline{Y_i}$	n'_{01}	n'_{00}	n'_{0+}
	n'_{+1}	n'_{+0}	N

Table 3: Contingency table according to three-part local expression

In Table 2 and Table 3 , the contingency table based on aggregate data compares to the one based on subset data. The main difference is that $n'_{01} < n"_{01}$ and $n'_{00} < n"_{00}$. This leads to $N' < N$. Based on the Table 3, the localized definition of association measures would be

support: $s = \frac{n_{1+}}{N}$ or .

confidence: $c = \frac{n_{11}}{n_{1+}}$

lift: $l = \frac{n_{11}N'}{n_{+1}n_{1+}}$

Chi-square: $\chi^2 = \frac{N(n_{11}n_{00} - n'_{01}n'_{10})^2}{n_{1+}n'_{0+}n'_{+1}n'_{+0}}$

The impact on the measures is different. Local support becomes greater while confidence doesn't change. Chi-square and lift will change. The direction and magnitude of the change vary across subsets.

This paradigm helps solve a prevailing problem in real life association rule mining, that is, discovery of low-frequency association rules. In the context of local rule mining, the absolute frequency of an itemset may be low, but it is still considered frequent if it meets the minimum support constraint in a particular segment. For example, a marketing practitioner may decide that an itemset worth exploring as long as they are frequent enough in a particular customer group. Now a user can set one high and universal minimum support.

Partitioning Aggregate Data

This section will show that when aggregate data is partitioned, lift value may change and association reversion may happen. A database query according to $(X = x)$ will form a dataset, which will be represented using D. From now on $(X = x)$ and D will be used interchangeably in the general expression of a local rule. If D is split into mutually exclusive subsets $\{D_1, D_2, ..., D_j\}$, then

(1) There is at least one subset D' in which Y's support is not lower than the aggregate support. $s(Y_iY_j|D') \geq s(Y_iY_j|D)$. (2) There is at least one subset D" in which Y's support is not greater than the aggregate support$s(Y_iY_j|D') \leq s(Y_iY_j|D)$.

Proof by contradiction: Let $s(Y_iY_j|D')$ be the largest subset support. Let us assume that $s(Y_iY_j|D') < s(Y_iY_j|D')$. Let $n(Y_iY_j|D$ be the number transactions that contains both products and $n(D)$ be the total number transactions in the dataset.

$n(Y_iY_j|D) = \sum s(Y_iY_j|D_i) \times n(D_i)$

$\leq \sum s(Y_iY_j|D') \times n(D_i)$

$= s(Y_iY_j|D') \times \sum n(D_i)$

$= s(Y_iY_j|D') \times n(D)$

$< s(Y_iY_j|D) \times n(D)$

$= n(Y_iY_j|D)$

The result is contradictory, $n(Y_iY_j|D) < n(Y_iY_j|D)$. Clearly, the assumption is incorrect. Similarly, we can prove the second equation.

The bounds of lift is $s(Y_iY_j) \leq lift(Y_iY_j) \leq \frac{1}{s(Y_iY_j)}$ Proof:

$$1 \geq s(Y_i) \geq s(Y_iY_j)$$

$$1 \geq s(Y_j) \geq s(Y_iY_j)$$

$$s(Y_iY_j) \leq lift = \frac{s(Y_iY_j)}{s(Y_i)s(Y_j)} \leq \frac{s(Y_iY_j)}{s(Y_iYj)s(Y_iY_j)} = \frac{1}{s(Y_iY_j)}$$

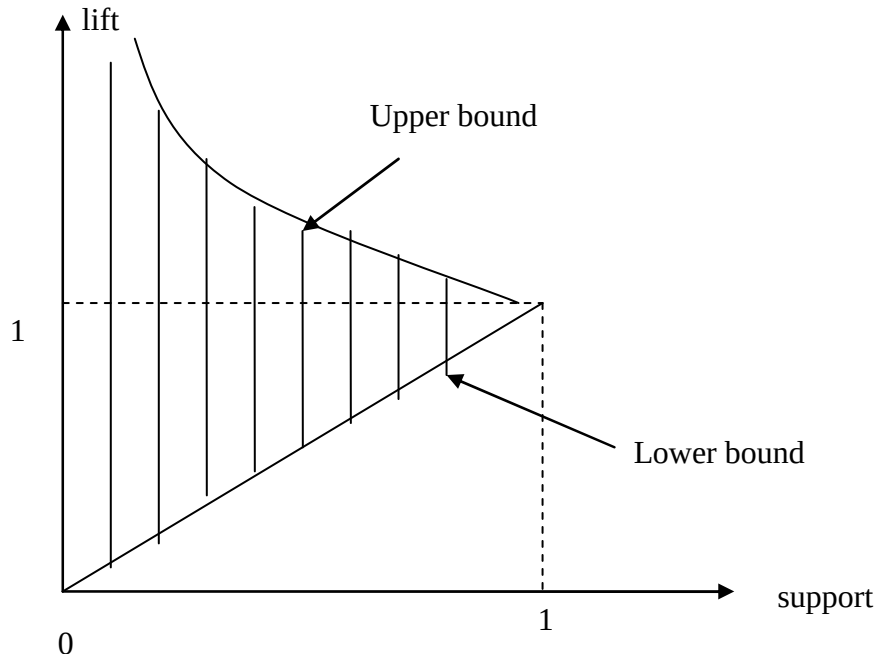

Figure 2: Bounds of lift value

Clearly, any split of current dataset into lower level subsets will lead to the change of support, which in turn will change the lift bounds. We need to test the homogeneity of an association in a segment so as to determine whether it needs to be split into smaller segments.

Homogeneity Test in Localized Rule Discovery

If for any attribute X_j , which partitions dataset D into multiple subsets $\{D_1, D_2, ..., D_j\}$, such that

$$lift(Y_1, Y_2|D) = lift(Y_1, Y_2|D_1 = lift(Y_1, Y_2|D_2 = ... = lift(Y_1, Y_2|D_J) \qquad (1)$$

(Y_1, Y_2) is homogeneous in dataset D . Of course, this is very likely not true when one starts to partition the dataset using combinations of attributes. In this research, the concept of homogeneity will be confined to the one-attribute partition. However even equation (1) is unlikely to be observed in data mining.

If D is split into mutually exclusive subsets $\{D_1, D_2, ..., D_j\}$, let $l(Y_1, Y_2)$ be the

current lift, $l_{max}(Y_1, Y_2)$ be the maximum subset lift, and $l_{min}(Y_1, Y_2)$ be the minimum subset lift. An even weaker definition of homogeneity in this research is

1. All subset lifts have same direction as the current lift; and

2. $l_{min}(Y_1, Y_2) < l(Y_1, Y_2) < l_{max}(Y_1, Y_2)$; and

3. The distance between $l_{max}(Y_1, Y_2)$ and $l_{min}(Y_1, Y_2)$ doesn't exceed a threshold η .

3.3 Implementations

The construction of the root node can be performed on the aggregate dataset. However, when the user wants to set a constraint of minimum support, the tree will stop at the first split if the database is sparse. To avoid this, we need to locate a large subset to start the process. This process is detailed in section 3.3.1. There are two ways to select the attributes. We can either select the ones that optimize the constraints directly, or select the ones that optimize the information gain and check whether it satisfies our constraints. These two approaches are discussed separately in section 3.3.2 and section 3.3.3.

Selection of Pairs

One strategy is to set a minimum dataset size, the least number of transactions a dataset needs to have. This constraint has practical meaning. For example, it may mean the minimum market size of a customer segment in retailing industry. Given a transaction database, in which most items have a relatively low frequency. One remedy is to set the minimum support really low, as low as 1% (Zheng et al. 2001). However this approach will lead to an explosion of rules. Another remedy is to set multiple minimum supports for different items (Liu et al. 1999, Wang et al. 2000).

To reduce the number of segments needed to be searched, we'll take advantage of an important fact about sparse database. Let D_i be a k-dimension transaction dataset, in which $\frac{n(Y|D_i)}{n(D_i)} << minsup$. Let D_j be a k+1 dimension subset of D_i . We know that $n(Y|D_j) < n(Y|D_i)$. If $\frac{n(Y|D_i)}{n(D_i)} < minsup$, there is no need to count $n(Y|D_j)$. We

can continue to test a smaller data set. Otherwise if $\frac{n(Y|D_i)}{n(D_i)} > minsup$, we will count $n(Y|D_j)$ to calculate support. If $\frac{n(Y|D_i)}{n(D_i)} > minsup$, we find one large segment. If not, we can repeat the same procedure in dataset D_j) . The procedure will work better when the transaction database is a sparse database. Usually, $n(Y|D_i) < n(Y|D_j)$. For example, if the transactions are split into two subsets by the gender of the customers, each subset will have approximately half of the total transactions, which is still a large number compared to most items' frequencies.

The Algorithm

Input: minimum support minsup, minimum data size minsiz

A 1-item list $S_0 = \{Y_1, ..., Y_I\}$, $n(Y_i|D_0)$

D=large subset sorted from high to low according to their size

for each Y_i in S_0

1) Initialize the candidate segment list, which is the full list of step 2.

2) The initial max frequency is $n(Y_i|D_0)$

3) Starting from the first segment in the segment list,

4) If $\frac{maxD_j}{n(D_j)} < minsup$, delete D_j from the candidate list;

5) If $\frac{maxD_j}{n(D_j)} \geq minsup$, Count $n(Y_i|D_j)$;

6) If $\frac{n(Y_i|D_j)}{n(D_j} < minsup$,

7) Compare $n(Y_i|D_j)$ to the maximums of D_j 's each subset, if less, then replace the maximum.

8) Delete D_j from the candidate list;

9) Go back to step b.

10) If $\frac{n(Y_i|D_j)}{n(D_j)} > minsup$,

11) Find one large segment, output (D_j, Y_i)

12) Delete all subsets of D_j from the candidate list.

13) Go back to step 4.

After we have (D_j, Y_i) pairs for all data segments and items, one can proceed to mine association rules in the segment D_j . If we treat D_j as an aggregate dataset, any traditional rule mining algorithms works.

Line 7 is necessary since a dataset has more than one superset. Its max is the minimum of the superset. The algorithm may output several large segments for item Y_i . It leaves to the user to judge what segment is worth further exploration.

Automatic Association Detection

The purpose of the algorithm is to select partitioning attributes according to their impact on the subset associations. It prefers attribute that lead to subsets in all of which the association between changes direction. Direction means positively associated, negatively associated, or independent. If there is no such an attribute, it prefers attribute that lead to subsets in some of which the association between changes direction. Otherwise, it prefers attribute that maximally differentiate subsets from its siblings in magnitude of lift values.

Automatic Association Detection (AAD) Algorithm

Input: root node D_0 , (Y_1, Y_2) , and $(X_1, X_2, ..., X_n)$

1) FOR node D_i ,

2) FOR each attributes X_j

3) Partition node D_i into subsets D_i' and D_i" according to

4) $L = lift(Y_i, Y_j|D_i)$

5) $L_{min} = min\{L = lift(Y_i, Y_j|D_i'), L = lift(Y_i, Y_j|D_i")\}$

6) $L_{max} = max\{L = lift(Y_i, Y_j|D_i'), L = lift(Y_i, Y_j|D_i")\}$

7) IF (L_{min} changes direction) and (L_{max} changes direction) THEN

8) set X_j as selected partition attribute

9) ELSEIF (one of L_{min} L_{max} changes direction) THEN

10) set X_j as the selected partition attribute if it is NULL

11) ELSE

12) Select the attribute that $max\underline{x}\frac{L_{max}}{L_{min}}$

set X_j as the selected partition attribute if it is NULL

13) END

14) Add two new nodes D_{2i+1} and D_{2i+2} to the tree

15) D_{2i+1} is the set of transactions where $X_j = x_j$

16) D_{2i+2} is the set of transactions where $X_j = x_j$

17) NEXT

After the tree are fully grown, the algorithm picks leafs whose lift value is 2 or 1/2 times of the root node. Then the algorithm will test if these leafs are statistically different from the remaining part of the data set using Chi-square test.

Multiple Mutual Information Based Algorithm

Multiple Mutual Information Algorithm

Input: D_0 , (Y_1, Y_2) , $(X_1, X_2, ..., X_n)$ and a threshold on information gain.

1) FOR node D_i, calculate the mutual information

2) $I(Y_1; Y_2) = \sum P(y_1, y_2) log_2 \frac{P(y_1,y_2)}{P(y_1)P(y_2)}$

3) Select X_j from the attributes which maximize the information gain

$maxI(Y_1; Y_2; X_j) =$

$I(Y_1; Y_2|X) = \sum P(y_1, y_2, x) log_2 \frac{P(y_1,y_2|x)}{P(y_1|x)P(y_2|x)}$

4) Add two new nodes D_{2i+1} and D_{2i+2} to the tree

D_{2i+1} is the set of transactions where $X_j = x_j$

D_{2i+2} is the set of transactions where $X_j = x_j$

7) NEXT

4 Market Basket Analysis using Localized Associations

For retailers, it is critical to understand how consumers respond to marketing activities of multiple products. This can be re-framed as two more specific questions:

1. What are the effects of marketing mix variables such as promotions on cross purchase associations in market basket data?

2. How do consumer demographics or psychographics impact such cross purchase associations?

The IRI scanner panel data used in this research includes over 500 households' purchases in a U.S. metropolitan market in a period of two years. It has twenty-two product categories: BBQ, butter, cat food, cereal, cleansers, coffee, cookie, crackers, detergent, hot dogs, eggs, ice creams, pill, nuts, pizza, snacks, soap, softener, soft drink, tissues, towel, and yogurt. Customer attributes include demographics such as age, income, family size, and psychographics such as cherry picking and store loyalty. Marketing mix includes price, display, and feature.

To validate the analysis algorithm, a validation sample of transactions will be hold out.

Section 4.1 will introduce the structure and descriptive statistics of the data set. Sect 4.2 will report the analysis result of the data set using traditional aggregate level association rule mining. Section 4.3 will report the analysis result of the data set using localized rule discovery.

4.1 Data Description

This part gives the definition of variables.

Since marketing mix variables are on the brand level, we need to construct category marketing mix variables. Category price is constructed as a weighted average of

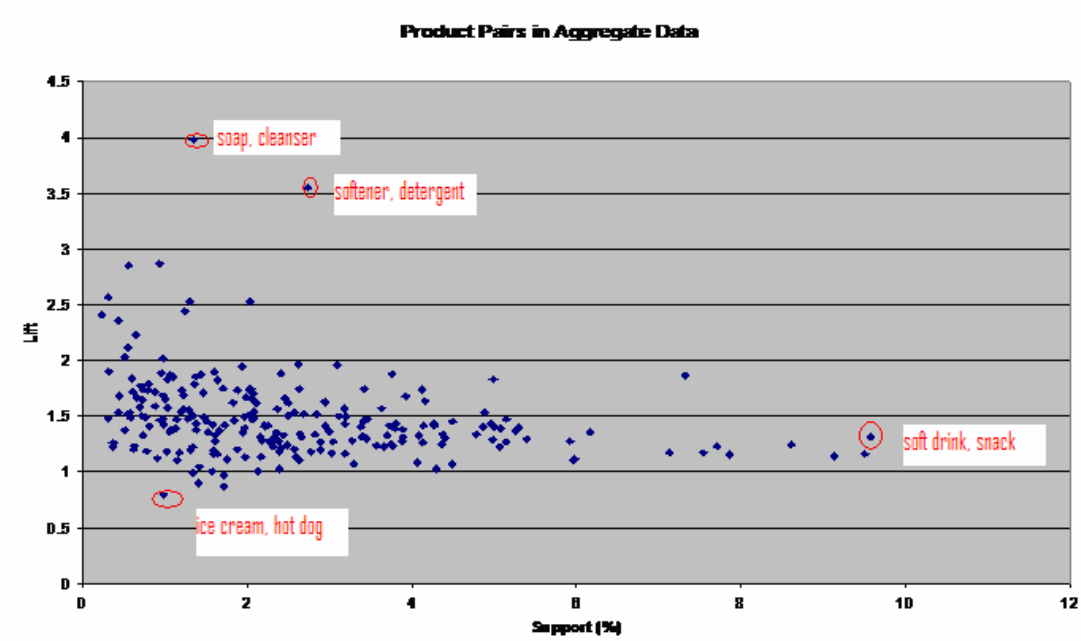

Figure 3: Overall Support-Lift chart of IRI data

brands' price where the weight is the share of each brand bought by each household (Manchanda et al. 1999). Category promotion is constructed as the weighted average of brands' promotion, where the weight is the share of each brand bought by each household. A brand is on promotion if it was either featured or displayed. These weighted averages will be transformed into discrete variables.

4.2 Analysis of Aggregate Data

This part will illustrate results from aggregate level association rule mining. Relative frequency of purchases at category level is shown at Table 4. Table 5 lists 20 pairs of categories based on high lift values. Table 6 lists 20 pairs of categories based on low lift values.

Category	Percentage	Category	Percentage
Softdrink	0.37	Hotdog	0.12
Tissue	0.21	Yogurt	0.11
Cereal	0.21	Cat food	0.11
Snack	0.19	Detergents	0.11
Towel	0.18	Coffee	0.10
Cookie	0.18	Pizza	0.08
Butter	0.17	Soap	0.07
BBQ	0.16	Softener	0.07
Eggs	0.16	Cleansers	0.05
Ice cream	0.14	Nuts	0.04
Crackers	0.14	Pill	0.03

Table 4: Frequency of category purchases

Category 1	Category 2	Lift
soap	cleansers	3.98
softener	Detergents	3.55
softener	cleansers	2.87
soap	pill	2.85
pill	cleansers	2.57
softener	soap	2.53
soap	detergents	2.53
detergents	cleansers	2.44
pill	nuts	2.41
softener	pill	2.35
pill	detergents	2.23
pill	coffee	2.12
softener	nuts	2.03
towel	pill	2.02
towel	soap	1.97
tissue	soap	1.96
tissue	cleansers	1.94
nuts	cleansers	1.90
towel	cleansers	1.90
nuts	crackers	1.89

Table 5: 20 pairs of categories based on high lift value

Category 1	Category 2	Lift
Yogurt	Cat Food	0.80
cookie	catfood	0.87
crackers	catfood	0.90
eggs	catfood	0.97
hotdogs	catfood	0.99
icecream	catfood	1.01
snack	catfood	1.01
softdrinks	catfood	1.03
cereal	catfood	1.03
yogurt	hotdogs	1.04
yogurt	softdrinks	1.07
tissue	icecream	1.07
softdrinks	coffee	1.08
soap	icecream	1.10
softdrinks	icecream	1.11
tissue	catfood	1.11
yogurt	icecream	1.11
softdrinks	crackers	1.12
soap	catfood	1.12
icecream	hotdogs	1.13

Table 6: 20 pairs of categories based on low lift value

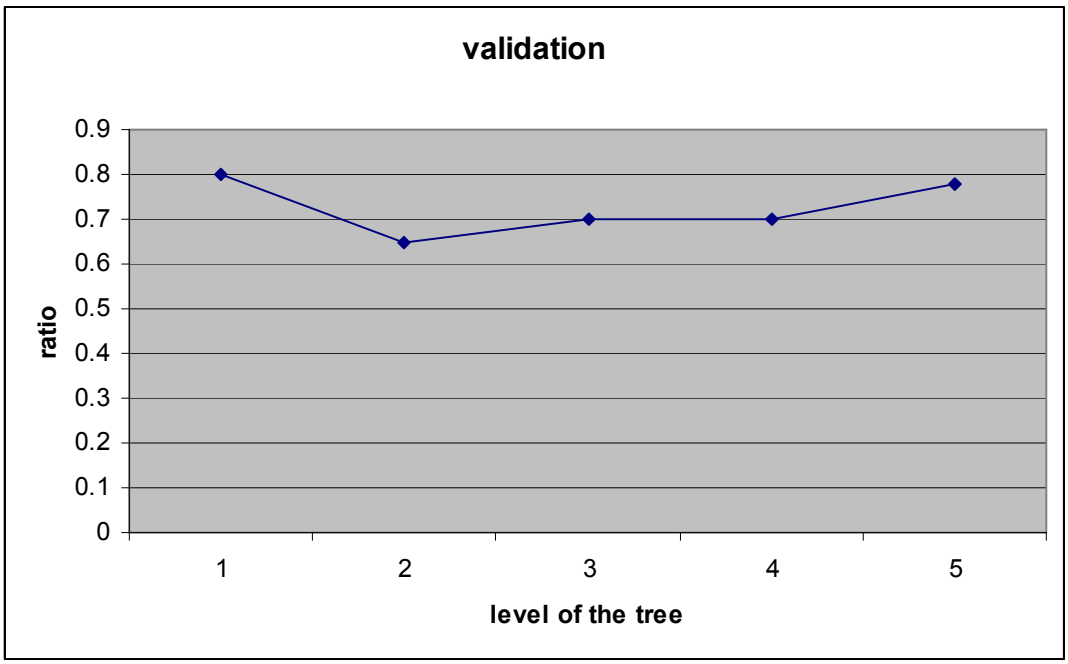

Figure 4: Validation Rate of Localized Rules

4.3 Analysis of Localized Data

Five pairs of categories have been picked. They are (Yogurt, Cereal), (Hotdog, BBQ), (Tissue, Towel), (Hotdog, Detergent), and (Detergent, Softener). They are selected since some of them are use complements such as detergent and softener. Some of them are closely related such as tissue and towel. Some of them are broadly linked such as yogurt, cereal, hotdog, and bbq.

A holdout set of data is used to validate rules found in the training dataset. A rule is validated using Chi-square goodness-of-fit test. As shown in Figure 4, most localized rules, from 70% to 80% can be verified.

For a balanced tree with depth D, The complexity of the algorithm is $O(N * M * D)$ where N is the number of transactions and M is the number of attributes.

Tree structure is used to illustrate the consecutive partition process.

Impact of Consumer Behavior: The Role of Cherry Picking and Store Loyalty

Cherry picking: if a consumer purchases in more than one stores on the same day, these transactions are defined as cherry picking transactions.

Store loyalty: if more than 4 transactions are made in the same store, these transactions are defined as store loyal transactions.

As expected Cherry Pickers baskets have lower lift values than the aggregate baskets. However, the effect is moderated by the complementarity of the pairs. There is no change in lift values in store loyal baskets.

Pair	Aggregate	CP=0	CP=1	CP=0 SL=0	CP=0 SL=1
Hotdogs-BBQ	1.60	1.60	1.27	1.52	1.59
Detergent-Softener	3.60	3.60	3.49	3.68	3.50
Hotdog-Detergent	1.53	1.55	1.14	1.32	1.56
Tissue-Towel	1.92	1.92	1.55	2.00	1.85
Yogurt-Cereal	1.69	1.68	1.21	1.45	1.69

Table 7: Consumer Behavior

Impact of Consumer Demographics: Family Size and Income

Aggregate cross category purchase effects can "hide" significant localized variations across consumer segments. In Table 8, "I" means income and "F" means family size. Income is defined as low (0) if family income is no more than $\$35,000$. Family size is defined as small(=0) if it is less than 3 people.

Pair	Aggregate	I=0 F=0	I=0 F=1	I=1 F=0	I=1 F=1
Hotdogs-BBQ	1.60	1.57	1.78	1.40	1.37
Detergent-Softener	3.60	4.14	3.79	3.10	2.46
Hotdog-Detergent	1.51	1.32	1.35	1.69	1.27
Tissue-Towel	1.92	1.88	1.90	1.90	1.82
Yogurt-Cereal	1.69	1.49	1.49	1.60	1.79

Table 8: Demographics

Impact of Marketing actions: The Role of Promotions (Feature and Display)

A category is considered on display in a week if the weighted weekly category-display is greater than the overall category-display. Same definition is made for category-feature. The results are shown in Figure 5 to Figure 9.

We find promotions have a negative impact on the co-purchase of complementary products. Category promotions lower the lift value of complementary product pairs. we further find that this effect is symmetric across complementary pairs.

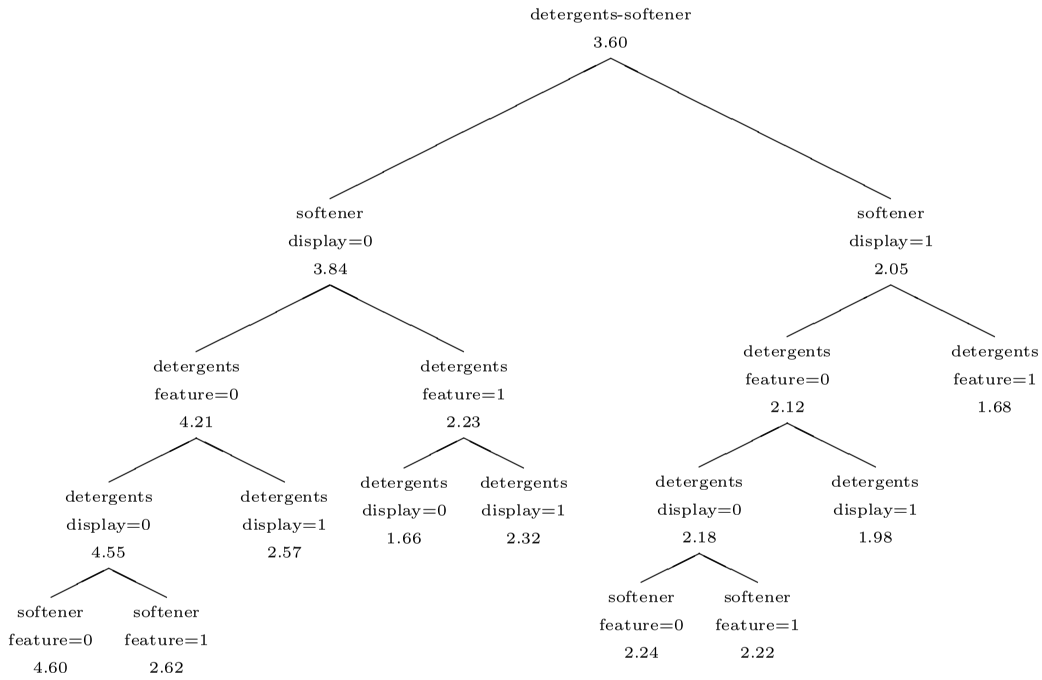

Figure 5: Role of Promotions on Detergent-Softener

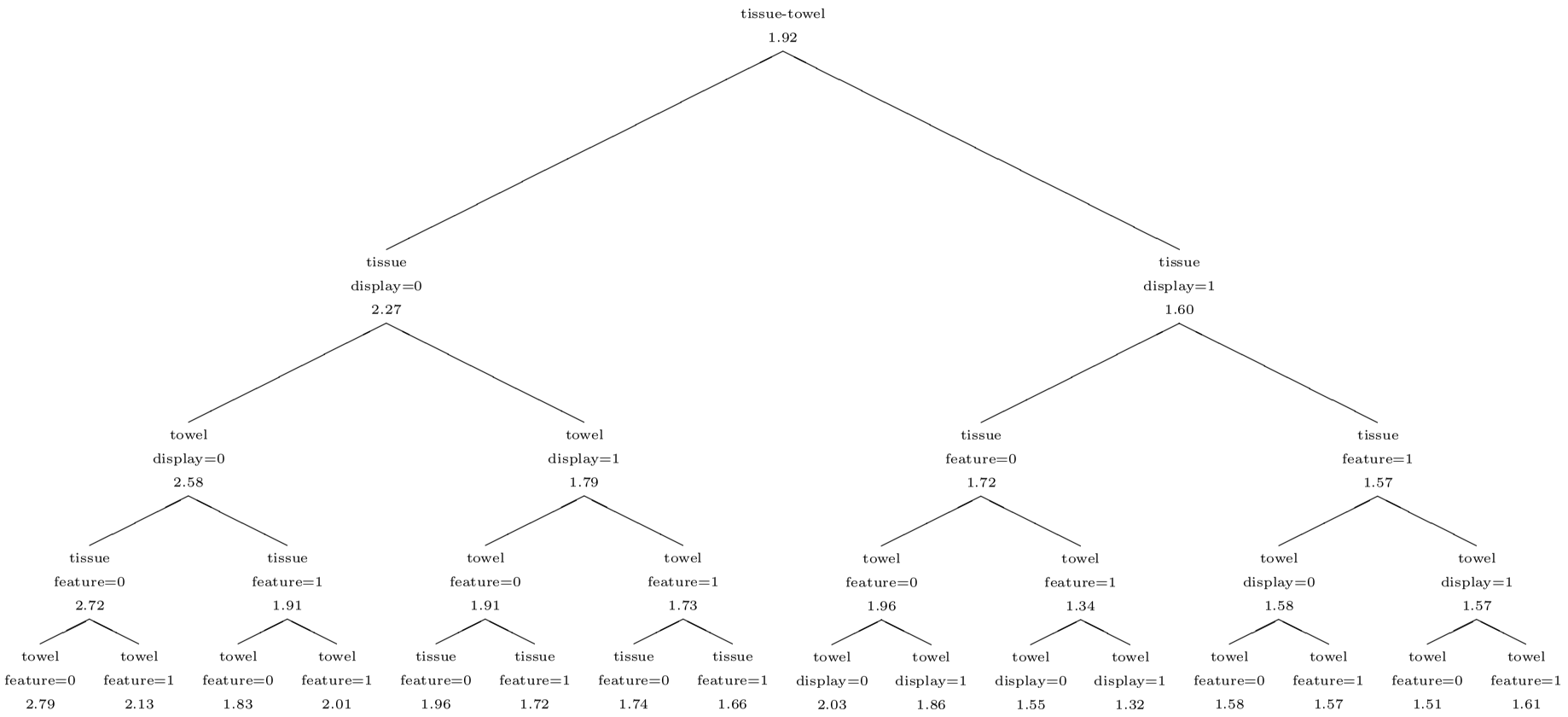

Figure 6: Role of Promotions on Tissue-Towel

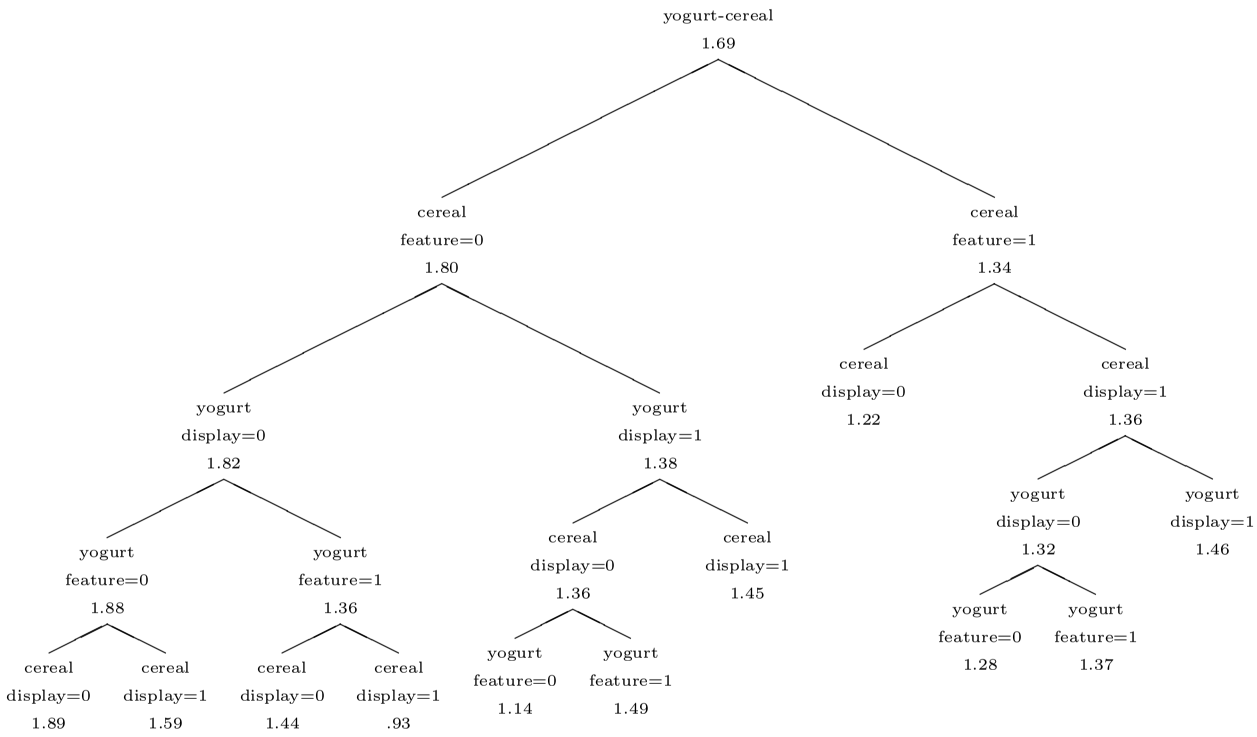

Figure 7: Role of Promotions on Yogurt-Cereal

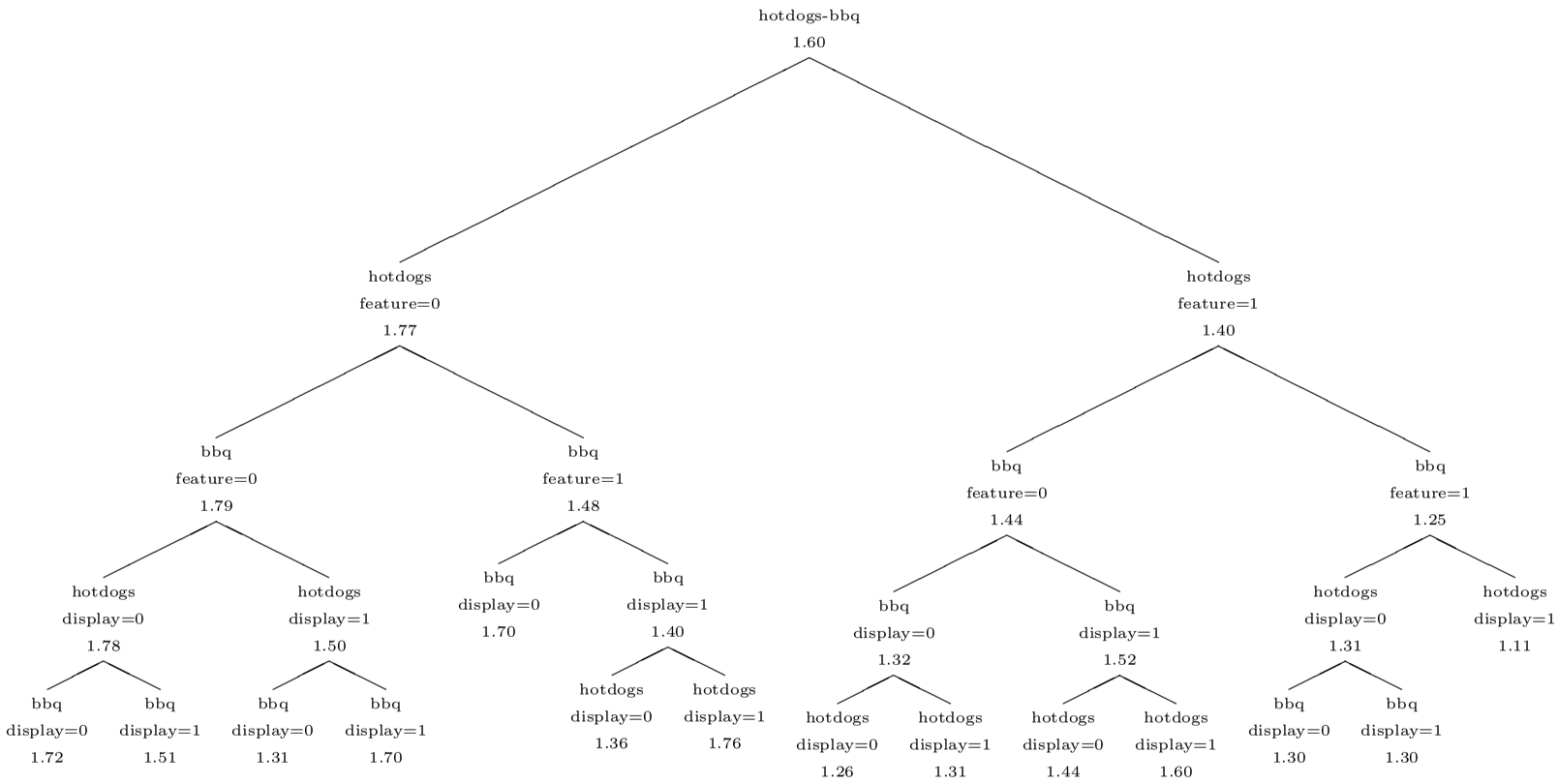

Figure 8: Role of Promotions on Hotdog-BBQ

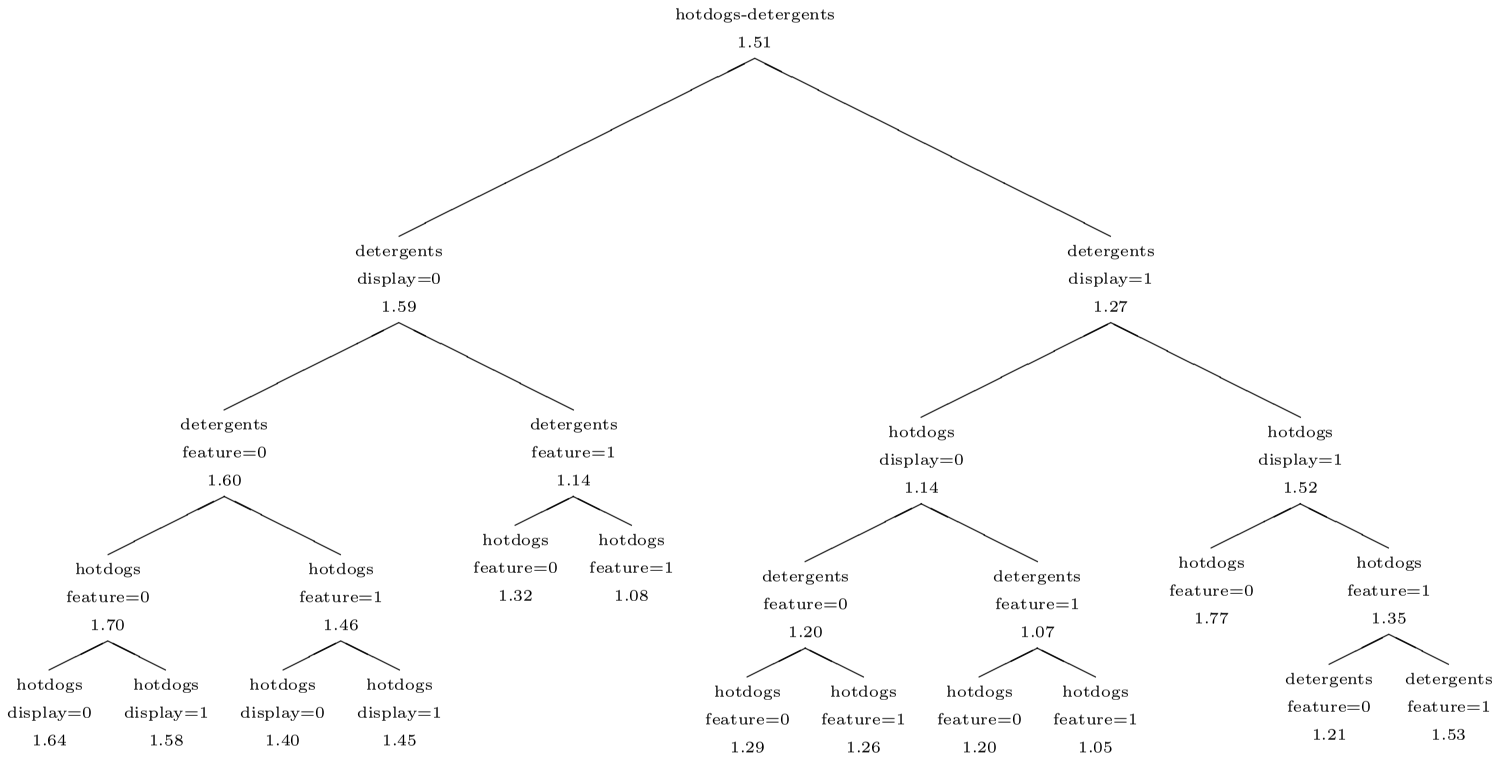

Figure 9: Role of Promotions on Hotdog-Detergent

Summary

This paper illustrates a new approach to analyze localized associations in market basket data. It applies the approach to study cross category choice in retailing data. It analyzed the retailing data given consumers' shopping behavior, demographics, and promotions. This research, coupled with essay one, can be applied to study the cross category effects of marketing mix activities as shown in Figure 10. Or they can be used to identify customer segmentations as shown in Figure 11. We would further extend pairwise category research to a multi-category tree model.

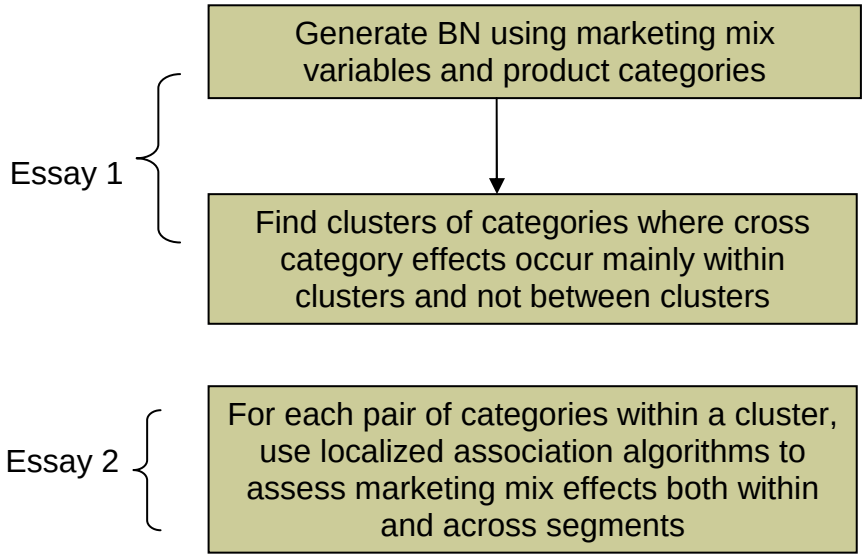

Figure 10: Cross category effects of marketing mix activities

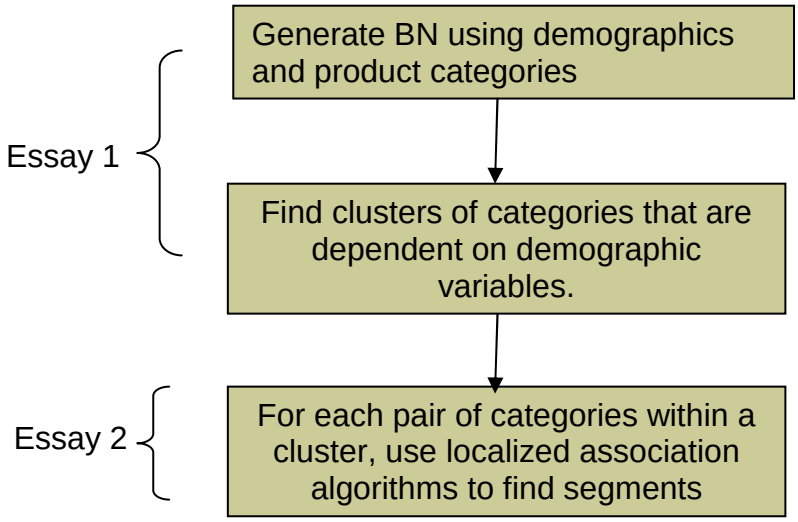

Figure 11: Customer segmentation using cross category associations

Reference

Agrawal, R., T. Iminlinski, A. Swami. 1993. Mining Association Rules between Sets of Items in Large Databases. 1993. *ACM SIGMOD* Washington DC.

Agrawal, R., S. Ramakrishnan. Fast Algorithms for Mining Association Rules. 1994. 20th VLDB Conference. Santiago, Chile. 487–499.

Blyth, R. 1972. On Simpson's Paradox and the Sure Thing Principle. *Journal of the American Statistical Association* **67** 364–366.

Breslow, N.E., N.E. Day. 1980. Statistical Methods in Cancer Research Volume I - the Analysis of Case-Control Studies. Lyon, the International Agency for Research on Cancer.

Brin, S., R. Motvani, C. Silverstein. 1997. Beyond Market Baskets: Generalizing Association Rules to Associations. *ACM SIGMOD International Conference on Management of Data*.

Chib, S., P. Seetharaman, A. Strijnev. 2002. Analysis of Multi-Category Purchase Incidence Decisions Using IRI Market Basket Data. *Econometric Models in Marketing* **16** 65–90.

Cohen, M., E. Nagel. 1934. An Introduction to Logic and Scientific Method. Harcourt, Brace.

Fabris, F. 1999. Discovering surprising patterns by detecting occurrences of Simpson's paradox. *Research and Development in Intelligent Systems XVI* 148–160. Springer-Verlag.

Fleuret, F. 2004. Fast binary feature selection with conditional mutual information. *Journal of Machine Learning Research* **5** 1531–1555.

Hipp, J., U. Guntzer, G. Nakhaeizadeh. 2000. Algorithms of Association Rule Mining: a General Survey and Comparison. *SIGKDD Explorations* **2(1)** 58–64.

Liu B., W. Hsu, Y. Ma. 1999. Mining association rules with multiple minimum supports. *KDD '99: Proceedings of the fifth ACM SIGKDD international conference*

on Knowledge discovery and data mining 337–341.

Manchanda, P., A. Asim, G. Sunil. 1999. the "Shopping Basket": a Model of Multicategory Purchase Incidence Decisions . *Marketing Science*. **18(2)** 95–114.

Mittal, Y. 1991. Homogeneity of subpopulations and Simpson's paradox. *Journal of American Statistical Association* **86** 167–172.

Mulhern, F., R. Leone. 1991. Implicit Price Bundling of Retail Products: a Multiproduct to Maximizing Store Profitability. *Journal of Marketing* **55** 63–76.

Paul, R., A. Donner. 1989. A Comparison of Tests of Homogeneity of Odds Ratios in K 2×2 tables *Statistics in Medicine* **8** 1455–1468.

Paul, R., A. Donner. 1992. Small Sample Performance of tests of Homogeneity of Odds Ratios in K 2×2 tables. *Statistics in Medicine* **11** 159–165.

Quinlan, J. 1986. Induction of Decision Trees. *Machine Learning* **1** 81–106.

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

Russell, G., A. Petersen. 2000. Analysis of Cross Category Dependence in Market Basket Selection. *Journal of Retailing* **76(3)** 367–392.

Samuels, M. 1993. Simpson's Paradox and Related Phenomena. *Journal of the American Statistical Association* **88(421)** 81–88.

Simpson, H. 1951. the Interpretation of Interaction in Contingency Tables. *Journal of the Royal Statistical Society, Ser.B.* **13** 238–241.

Smyth, P. R. Goodman. An information theoretic approach to rule inductionfrom databases. 1992. *IEEE Transactions on Knowledge and Data Engineering* **4(4)** 301–316.

Tan, P., V. Kumar. 2000. Interestingness Measures of Association Patterns: a Perspective. *KDD'00 Workshop on Postprocessing of Machine Learning and Data Mining* Boston, MA.

Tan, P., V. Kumar, J. Srivastava. 2002. Selecting the right interestingness measure for association patterns. SIGKDD'02. Edmonton, Alberta, Canada.

Wagner, H. 1982. Simpson's paradox in real life. *the American Statistician* **36** 46–48.

Wang K., Y. He, J. Han. 2000. Mining Frequent Itemsets Using Support Constraints. VLDB 2000, Proceedings of 26th International Conference on Very Large Data Bases, Cairo.

Xiong, H., S. Shekhar, P. Tan, V. Kumar. 2004. Exploiting a support-based upper bound of Pearson's correlation coefficient for efficiently identifying strongly correlated pairs. *KDD'04* , Seattle,Washington.

Zheng Z., R. Kohavi, L. Mason. 2001. Real world performance of association rule algorithms. *KDD '01: Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining* 401–406.