6-2002

Modeling direct marketing response : bayesian networks with evolutionary programming

Geng CUI
gcui@ln.edu.hk

Man Leung WONG
mlwong@ln.edu.hk

MODELING DIRECT MARKETING RESPONSE: BAYESIAN NETWORKS WITH EVOLUTIONARY PROGRAMMING

ABSTRACT

Given the explosive growth of customer information, data mining can potentially discover new knowledge to improve decision making in marketing. This study proposes a data mining approach to modeling direct marketing response using Bayesian Networks (BN) and Evolutionary Programming (EC) and applies these methods to direct marketing data. The results suggest this approach generate superior results than the conventional method of logistic regression. Future research in this area should devote more attention to applying data mining methods to solving complex problems facing today's businesses.

INTRODUCTION

Conventional marketing research is a process in which researchers design a study, collect the data, and then manually analyze them to explore the relationships among various factors defined by the researcher. Such methods usually consider a limited number of variables and compare a few alternative models. Nowadays, many businesses generate and collect a huge amount of data in a relatively short period. Even with powerful computers and versatile statistical software, much of the useful marketing insights into customer characteristics and their purchase patterns are largely hidden and untapped (Peacock 1998). The explosive growth of data requires a more efficient way to extract useful knowledge. Thus, marketing is a major area for applying data mining that aims at discovering novel, interesting and useful knowledge from databases (Shaw et al, 2001). By definition, data mining can help automatically discovering complex relationships among various factors and extracting meaningful knowledge to improve the efficiency and quality of managerial decision making.

Recently, several researchers have applied methods of artificial intelligence including genetic algorithm (GA) as a data mining tool to solving marketing problems, for instance, direct marketing response modeling, and produced encouraging results (Bhattacharyva 2000). However, like other artificial intelligence methods such as artificial neural networks (ANN), data mining using GA for direct marketing face several significant challenges. First, although current GA methods can generate the empirical results, GA procedures are computationally intractable and remain a "black box" operation. Researchers cannot understand how and why a certain solution has been derived or what the nature of the discovered relationships among the variables. Secondly, current applications of GA in direct marketing can describe the discovered model through the parameter estimates based on the fitness function emulated by the algorithm. But it is similar to conventional statistical methods that compare only a limited number of alternative models or the variety of structures. Thus, it is possible that some alternative representations of model structures are not explored, leaving some knowledge hidden (Bhattacharyva 2000). Thirdly, current GA applications lacks a mechanism to evaluate the performance (i.e., fitness) of the alternative models and does not lend an opportunity to understand how a superior or optimal model has been derived.

Based on the recent works in artificial intelligence and machine learning, we propose

a data mining approach to modeling direct marketing response using Bayesian Networks (BN) and evolutionary programming (EP). First, we introduce the background literature on data mining and the research problems in direct marketing. Secondly, we describe the data mining approach to direct marketing response modeling and delineate the learning process using BN and EP. We elaborate the advantages of BN in representing the knowledge structure and EP as an efficient tool for learning BNs and for searching the space to arrive at the optimal solution among all eligible models using the evolutionary mechanism. Thirdly, we apply these methods to two data sets of direct marketing and find that these methods are superior than logistic regression in predicting consumer response to direct marketing. Finally, we explore the implications for data mining in marketing and directions for further research.

RECENT DEVELOPMENT

The increasing and widespread use of computers by businesses has led to an explosion of marketing information. Even with the best researchers and state-of -the-art methodologies, many hidden and potentially useful relationships may not be recognized by the analysts. The voluminous data can be best used if hidden knowledge can be uncovered, thus making data mining an important research topic for today's businesses. Within a broad definition, data mining encompasses "confirmation" or the testing of relationships in the discovery process. Narrowly defined, data mining is the automated discovery of "interesting" non-obvious patterns hidden in a database that have a high potential for contributing to the bottom line (Peacock 1998). Data mining is the core of the knowledge discovery in database (KDD) process. Thus, the two terms are often used interchangeably (Fayyad et al. 1996).

Data mining can be useful for many real-world problems, and especially for marketing organizations. With the computerization of marketing operations, huge amount of customer and transactional data are being collected, presenting significant challenges for marketing researchers. The problem of data growth is even more acute in the field of direct marketing. A catalog marketing company with multiple product lines can easily accumulate large amounts of transactional data, in additional to the consumer demographics and credit data that the company has collected or purchased from commercial vendors. Often, such databases include hundreds of thousand or even millions of customers and thousands of variables or attributes. As the amount of data grows exponentially, how marketers can take advantage of the vast resources to gain insight into consumer behavior and devise more

effective marketing strategies is among the top challenges for marketing professionals.

One of the areas that can benefit from data mining is direct marketing. Until recently, direct marketing and consumer response modeling have been neglected in the mainstream marketing literature (Bodenberg and Roberts 1990). However, the last few decades have witnessed phenomenal growth of direct marketing. According to the Direct Marketing Association (2001), direct marketing is a US$1.7 trillion industry in the U.S. alone in 2000. Besides the traditional direct marketers such as catalog companies and telemarketers, many large corporations have also added direct marketing as one of their strategies, making accurate predicting consumer response to direct marketing a top priority for many businesses to improve the cost-effectiveness of direct marketing. Even a small increment in predicting consumer response to direct marketing may lead to increased sales, tremendous cost-savings, and improved profitability. Due to its distinctive nature, direct marketing companies are among the first businesses to computerize their marketing operations. Consequently, direct marketing is often referred to as database marketing. Although data mining has been increasingly used by marketing organizations, research on data mining has appeared mostly in the information systems and other related areas and received scanty treatment in the marketing field.

Conventional research is based on an *a priori* approach to build predictive models. Researchers typically rely on theory development and hypothesis testing. For instance, in direct marketing response modeling, marketing practitioners have developed a theory of purchase behavior, which has come to be known as the RFM model (David Shepard Associates 2000). The model states that the likelihood of consumers responding to a promotion is predicted by the *recency* of the last purchase, the *frequency* of purchase over the past years, and the *money* value of a given customer's purchase history (Berger and Magliozzi 1992). While some researchers continue to improve the accuracy in predicting consumer response, for instance, by controlling for unobserved consumer heterogeneity (DeSarbo and Ramaswamy 1994; Gönül, Kim and Shi 2000), others have adopted a profit-maximization approach which aims at identifying high profit consumers as well as low profit or unprofitable consumers so that precious marketing resources can be saved to augment profitability (Bult and Wansbeek 1995; Bitran and Mondschein 1996).

The merits of the conventional methods not withstanding, the size of real problems may render computational solution of optimization impossible given the number of variables

and large dataset (Bitran and Mondschein 1996). Traditional research approaches, which can be very powerful in their own rights, can only handle a limited number of variables and effectively compare a few alternative solutions. Furthermore, many urgent problems facing businesses are practical issues that are poorly structured, falling out of the established theoretical frameworks. Given a large amount of data and increasing variety of customer data, marketing researchers need new innovative methods to efficiently discover knowledge from data and gain additional insight into consumer behavior. Thus, data mining is increasingly used by many companies to discover useful knowledge and to improve marketing efficiency and quality of decision-making (Shaw et al, 2001).

Data mining has many potential uses in marketing including customer acquisition, customer retention, customer abandonment and market basket analysis. In addition to query tools and descriptive statistics, data mining experts have developed various knowledge discovery systems based on both conventional and innovative research methods to extract knowledge from data including visualization tools, regression-type models, association rules, decision tree analysis, and case-based reasoning (Peacock 1998). Meanwhile, recent development in artificial intelligence and machine learning has presented more powerful data mining techniques and analytical methods to the marketing researchers' arsenal toolkit, including artificial neural networks (Thieme, Song and Calantone 2000) and genetic algorithm (Hurley, Moutinho and Stephens 1995)

Genetic algorithm (GA) is one of the promising methods of evolutionary computation (EC) for solving marketing problems. GA was originally developed in the field of computer science (Holland 1975; Goldberg 1989). Its principles and methods have been adopted by researchers in business management. Genetic algorithms operate through procedures modeled upon the evolutionary biological processes of selection, reproduction, mutation, and survival of the fittest to search for good solutions to prediction and classification problems (Holland 1975; Goldberg 1989; Peacock 1998). They are particularly effective for solving poorly understood and poorly structured problems because they attempt to find many solutions simultaneously, whereas a linear regression model, for example, focuses on presumably a single best solution. Another strength of GAs is that they can explicitly model any decision criterion in the "fitness function," an objective system used to assess a GA's performance (Hurley, Moutinho and Stephens 1995; Peacock 1998).

Data mining using GA has several advantages. First, by definition, data mining using GA searches the space for all possible alternative representations of the knowledge and then determines the best possible solution among all eligible candidates based on a fitness criterion. Comparing to other procedures, GAs can produce innovative solutions and discover relationships not defined by the researchers (Hurley and Moutinho and Stephens 1995). They may discover combinations of predictor variables that no one would have expected to predict beforehand (Peacock 1998). Secondly, unlike conventional research that emphasizes hypothesis testing based on *a priori* model with a limited number of variables selected by the researcher, data mining discovers the relationships and then presents the *posterior* results. Thirdly, although GA is dissimilar to conventional statistical methods, the empirical results of the data mining process allow for comparison with those generated by other methods based on common evaluation criteria so that they can assist managerial decision-making. Moreover, GA can also handle multiple criteria such as sales and profitability using the Pareto frontier analysis (Bhattacharyva 2000).

Such beneficial features can be helpful for knowledge discovery in the field of marketing. Recently, methods based on GAs have been applied to marketing problems such as product design (Balakrishnan and Jacob 1996), inventory control and product assortment management (Urban 1998), brand competition (Midgley, Marks and Cooper 1997), and marketing mix elasticities (Klemz 1999). Recent research in this area has focused on how to apply GAs to specific marketing problems and comparing its results to those generated by other conventional methods. One of the data mining applications in marketing is to identify responsive, loyal or profitable customers so that marketers can design more accurate targeted marketing to improve sales and profitability of direct marketing operations. Thus, several researchers in the field of artificial intelligence have applied GA methods to modeling direct marketing response (Bhattacharyva 2000; Eiben et al 2000; Levin and Zahavi 2001). Findings of these studies have found that optimization based on genetic algorithms can produce superior results in predicting consumer response to direct marketing.

Despite the recent progress, data mining using GA for direct marketing face several significant challenges. First, although current GA methods can generate the empirical results comparable to those of logistic regression such decile analysis used in direct marketing (explained later), GA procedures like other artificial intelligence methods including neural networks are computationally intractable. The "black box" operation does not lend an

opportunity for researchers to understand how and why a certain solution has been derived. Secondly, current applications of GA in direct marketing delineate the learned knowledge or to describe the discovered model by presenting the parameter estimates based on the fitness function that the algorithm emulates. Thus, it is similar to conventional statistical methods that compare only a limited number of alternative models or the variety of structures, including the tree analysis method. Thus, such methods explore limited search space in the optimization process. Consequently, it is possible that some alternative representations of model structures are not explored, leaving some knowledge hidden (Bhattacharyva 2000).

Therefore, to improve the applicability of data mining in marketing and the interpretability of results, several issues need to be addressed. First, GA relying on binary bits are still limited in representing alternative model structures. Thus, methods that are flexible enough to represent all alternative structures of knowledge of a specific domain and delineate the complex relationships among variables are more desirable for data mining purposes. So that the degree of freedom is maximized in the search space. Secondly, due to the black box nature of current GA methods, improved methods should allow direct interpretation and easy understanding of the discovered knowledge, i.e. the relationships among the variables of the model and the strengths of such relationship, based on the mathematical or statistical attributes. Thirdly, although data mining using GA seeks the best possible solution among all eligible candidates based on a fitness criterion, the fitness function only has to do with the computational algorithm to emulate the specific marketing problem. Researchers need a method to evaluate the fitness of alternative models and to understand why a superior or optimal model has emerged.

DATA MINING FOR DIRECT MARKETING

Recent development in machine learning and artificial intelligence has greatly improved the efficiency of the search process and can help address the above issues for data mining for direct marketing (Lam et al.; 1998; Wong, Lam and Leung 1999). Specifically, we propose a data mining approach to modeling consumer response to direct marketing using Bayesian Networks (BN) and evolutionary programming (EP). In the following sections, we delineate a data mining approach to modeling direct marketing response and describe the learning process using BN and EP.

In Figure 1, we describe the five steps in the process of the data mining using these methods (Ngan et al 1999). First, a selection is made to extract a relevant or a target data set from the database. Then, preprocessing is performed to remove noise and to handle missing data fields. Transformation is performed to reduce the number of variables under consideration. The third and fourth steps induce knowledge from the preprocessed data. A suitable data mining algorithm is applied to the prepared data. The causality and structure analysis learns the overall relationships among the variables. In the fifth step, the discovered knowledge is verified and evaluated by the domain expert. The domain experts may discover and correct mistakes in the discovered knowledge. The discovered knowledge can be used to refine the existing domain knowledge or incorporated into an expert system for decision making. If the discovered knowledge is not satisfactory, the five steps will be reiterated (Figure 1).

In this study, we focus on the third and fourth steps. For causality and structural analysis, we use Bayesian Networks (BN) to represent the knowledge structure and build models. To learn a plausible BN model, we adopt evolutionary programming (EP) as the data mining algorithm.

(insert Figure 1 here)

BAYESIAN NETWORKS (BN)

Bayesian network is a method for formal knowledge representation based on the well-developed Bayesian probability theory. Although the underlying theory of Bayesian probability has been around for nearly 250 years and Bayesian Networks were proposed in the 1960s (Earman 1990), building and executing realistic Bayesian Network models has only been made possible because of recent development of algorithms and software tools that implement them (Jensen, 1996; Pearl 1988). Bayesian networks have made tremendous progress and have been adopted by researchers in many fields. Several authors have given excellent introduction to Bayesian networks and detailed comparisons with other methods (D'Ambrosio 1999; Geiger and Heckerman 1996; Haddawy 1999; Heckerman and Wellman 1995).

The key feature of Bayesian networks is the fact they provide a method for

decomposing a probability distribution into a set of local distributions. The independence semantics associated with the network topology specifies how to combine these local distributions to obtain the complete joint-probability over all the random variables represented by the nodes in the network model, making it an effective tool for solving prediction and classification problems (Haddawy 1999). For example, the probability of hearing a dog barking is the joint probability of the local probabilities of other events including the dog is let out, due to bowel problem, a family member is out, and the light is turned on (Figure 2). Real-world applications of the Bayesian networks method have been successful in solving many problems including software engineering, space navigation, and medical diagnosis (Haddawy 1999).

(insert Figure 2 here)

The most common computation performed using BN is the determination of the posterior probability of some random variables in the network. Because of the symmetric nature of conditional probability, this computation can be used to perform both diagnosis and prediction (Haddawy 1999). In essence, a Bayesian network captures the conditional probabilities between variables and therefore, can be used to perform reasoning under uncertainty. In practice, a Bayesian network is a directed acyclic graph (DAG), such as the one in Figure 2. Each node represents a domain variable, and each edge represents a dependency between two nodes. An edge from node A to node B can represent causality, with A being the cause and B being the effect. Each node is associated with a set of parameters. Thus, let N_i denotes a node and Π_{N_i} denotes the set of parents of N_i. And the parameters of N_i are conditional probability distributions in the form of $P(N_i \mid \Pi_{N_i})$ with one distribution for each possible instance of Π_{N_i}.

The main task of learning BN from data is to automatically find directed edges between the nodes to identify a network model that can best describe the causalities. Once the network structure is constructed, the conditional probabilities are calculated based on the data. Bayesian network learning can also be implemented by imposing limitations and assumptions to guide the search process thus improving its compatibility with other methods. For instance, in addition to the more general algorithms (Heckerman, Geiger and Chickering 1995; Spirtes, Glymour and Scheines 2000), the algorithm of Rebane and Pearl (1989) can learn networks

with tree structures. The algorithms of Cooper and Herskovits (1992) require the variables to have a total ordering. More recently, Larrañaga et al. (1996) applied GA to learn Bayesian networks.

The success of Bayesian networks lies largely in the fact that the formalism introduces structure into probabilistic modeling and cleanly separates the qualitative structure of a model from the quantitative aspect (Haddawy 1999). Although the formal definition of a Bayesian network is based on conditional independence, in practice a Bayesian network typically is constructed using notions of cause and effect, making it powerful for identifying and analyzing the structural relationships among variables (Heckerman and Wellman, 1995). Like logistic regression, the Bayesian Networks approach is free from the normality assumption. The value of each variable can be discrete and continuous. There are several methods for incorporating continuous variables within Bayesian networks (Chang and Fung 1991; Lauritzen 1990; Olesen 1991). Unlike the regression-based methods and most learning algorithms that rely on a specific fitness function, Bayesian Networks can take any shape. Such "freedom of expression" allows exploring the complex interrelationships among the variables. In addition, the Bayesian networks method offers several other benefits for marketing research. Based on the generated model, Bayesian Networks method also calculates a probability score for each case, which is useful for predicting consumer responses to marketing activities. Bayesian Networks also tests for independence among the variables so that spurious relationships and redundant measures can be identified and avoided.

EVOLUTIONARY COMPUTATION

Evolutionary computation is a general term to describe computational methods that simulate the natural evolution based on the Darwinian principle of evolution to perform function optimization and machine learning (Bäck, Fogel and Michalwicz 2000). In theory, an evolutionary algorithm maintains a group of individuals, called the population, to explore the search space. A potential solution to the problem is encoded as an individual, or a Bayesian Network model in this study. A fitness function evaluates the performance of each individual to measure how close it is to the solution. The search space is explored by evolving new individuals. Based on the Darwinian principle of evolution through natural selection, the fitter individual has a higher chance of survival and tends to pass on its

favorable traits to its offspring. A 'good' parent is assumed to be able to produce "good" or even better offspring. Thus, an individual with a better score in the fitness function has a higher chance of undergoing evolution to produce new individuals. New individuals are generated by applying genetic operators that alter the underlying structure of individuals. EP is a general, domain independent method that does not require any domain-specific heuristic to guide the search.

Methods of evolutionary computation include genetic algorithms (GA), genetic programming (GP), evolutionary programming (EP), and evolution strategy (ES). They mainly differ in the evolution models assumed, the evolutionary operators employed, the selection methods, and the fitness functions used. Based on the genetics theory of chromosomes, GA uses a fixed-length binary bit string as an individual (Holland 1975; Goldberg 1989; Hurley, Moutinho and Stephens 1995). Three genetic operators are used to search for better individuals. Reproduction operator copies the unchanged individual. Crossover operator exchanges bits between two parents. Mutation operator randomly changes individual bits. While the GA methods treat the individuals as a binary string, GP extends GA to represent an individual using a tree structure (Koza 1992). ES emphasizes on the individual, i.e. the phenotype, as the object to be optimized. A genetic change in the individual is within a narrow band of the mutation step size, which has self-adaptations (Schewefel 1981). Since data mining can be considered as a search problem, which tries to find the most accurate knowledge from all possible hypotheses, and evolutionary algorithms are robust and parallel search algorithms, thus they can be used in data mining to find interesting knowledge in noisy environment.

EVOLUTIONARY PROGRAMMING (EP)

By comparison, Evolutionary Programming (EP), the method of choice in this study, has several distinctive advantages. First, unlike GA focusing on binary bits or GP using tree structures, EP does not require any specific genotype in the individual. Thus, there is no constraint on the representation or structure in EP. The individual can be a binary string, a tree structure, or any other shape, and evolve into new structures during the evolutionary process. In essence, the structural representation can simply follow from the problem. Secondly, while GA uses reproduction, cross-over and mutation operators, reproduction and mutation are the only genetic operators used for evolution in EP. While cross-over operation

may lead to invalid models such as a recursive model, mutation operator is in self more powerful. The mutation operator allows for simultaneous modification of all variables at the same time. Thus, cross-over operation designed to recombine building blocks is not utilized in the general forms of EP. Thirdly, mutations in EP attempt to preserve behavioral similarity between offspring and their parents, rather than seeking to emulate specific genetic operators as observed in nature (Fogel 1994). A "child" is generally similar to its parent at the behavioral level with slight variations. EP assumes that the distribution of potential offspring is under a normal distribution around the parents' behavior. The severity of mutations is according to a statistical distribution. Therefore, EP employs a model of evolution at a higher abstraction.

Due to these unique advantages, EP's flexibility and freedom from various constraints of make it an ideal tool as the search mechanism for data mining purposes (Fogel 1994; Fogel, Owens and Walsh 1966). Recently, Wong, Lam, and Leung (1999) performed a series of experiments to compare EP approach for learning Bayesian networks with the classical GA approach proposed by Larrañaga et al. (1996). Their findings suggest that the EP approach is superior in terms of both quality of solutions and computational time in most data sets they tested.

In a typical process of EP (Table 1), a set of individuals is randomly created to make up the initial population. Each individual is evaluated by the fitness function. Then each individual reproduces a child by mutation. There is a normal distribution of different types of mutation, ranging from minor to extreme. Minor modifications in the behavior of the offspring occur more frequently and substantial modifications occur less. The offspring are also evaluated by the fitness function. Then tournaments are performed to select the individuals for the next generation. For each individual, a number of rivals are selected among parents and their offspring. Tournament score of an individual is the number of rivals with lower fitness scores than itself. Individuals with higher tournament scores are selected as the population of next generation. There is no requirement that the population size is held constant. The process is iterated until the termination criterion is satisfied or when no more improvement.

(insert Table 1 here)

MODEL EVALUATION

In the proposed knowledge discovery process, structure analysis process induces a Bayesian network from the data. Although complex Bayesian networks may more accurately represent a model of the data, they have several disadvantages. In the worst case, it is intractable to compute posterior probabilities in complex Bayesian networks (Cooper 1990). Moreover, Bayesian networks with more connections between their nodes require more probability parameters and more computer storage to store these parameters. Hence, complex networks suffer from both the time and space complexity problems. Complex networks also have several conceptual disadvantages. The structure of a Bayesian network represents information about the underlying causal and probabilistic relationships in the domain. Bayesian networks with complex structures are difficult to understand and explain. Consequently, it is beneficial to discover simpler networks from databases if they are sufficiently accurate. The dilemma between model simplicity and accuracy is a significant challenges for model evaluation in new methods of artificial intelligence such as artificial neural networks.

Model evaluation is a significant and necessary step in the evolutionary process using EP to produce better models by each generation. To balance the trade-off between model simplicity and accuracy, the Minimum Description Length (MDL) metric is employed as the fitness criterion to evaluate Bayesian networks, and the EP procedure searches for the best network structure based on this metric. Rissanen (1978) first proposed the minimal description length principle, which are subsequently developed by Lam and Bacchus (Lam 1998; Lam and Bacchus 1994) for evaluating Bayesian networks. The metric is asymptotically equivalent to the Bayesian scoring function, thus asymptotically correct (Cooper and Herskovits 1992; Heckerman et al. 1995). In other words, when the number of samples increases, the learned model converges to the underlying true distribution of data with probability equal to one (Heckerman 1995; Geiger et al. 1996). The main advantage of the MDL metric is that it balances the accuracy and simplicity of Bayesian networks. It allows a less complex network to be learnt if that network is accurate enough. If no simpler network is sufficiently accurate, the metric still allows a complex network to be induced. Recently, several researchers successfully applied the MDL metric to learning Bayesian networks in an EP environment (Lam et al.; 1998; Wong, Lam and Leung 1999).

The MDL Metric

The MDL principle is motivated by information coding (Rissanen 1978). Assume that a collection *C* of data items is given and it is necessary to store this collection of data items in computer storage. In order to conserve storage, a compressed version of *C* is stored. We can find a suitable model for *C* so that the encoder can use it to produce a compressed version of *C*. Since we want to be able to recover *C*, the model used by the encoder to compress *C* must also be stored. The total description length is then defined as the sum of the length of the compressed version of *C* and the length of the description of the model used in the compress. The MDL dictates that the optimal model explaining a collection of data items is the one that minimizes the total description length.

For example, suppose that we have *n* data points, (x_1, y_1), (x_2, y_2), …., (x_n, y_n), and we want to compress these data points by using an order *k* polynomial where *k* </= *n*. We need to store *k*+1 numbers that specify the coefficients of the polynomial. The size of storage used to store these numbers is the description length of the polynomial (model). Obviously, the description length of an order 2 polynomial is smaller than that of an order 3 polynomial, because 3 and 4 parameters are respectively stored for these models. In other words, the model description length measures the simplicity of the model.

The compressed version of the data includes the values of $x_1, x_2, \ldots, x_n$. Moreover, one cannot guarantee that the order *k* polynomial precisely fits the data. In other words, there may be some error e_i between the y-value of the polynomial evaluated at x_i and the actual value of y_i. Thus, to completely compress the data points, it is also necessary to store these errors $e_i, e_2, \ldots e_n$ along with $x_1, x_2, \ldots, x_n$. The size of storage used to store these values is the description length of the compressed data. Since the size of storage used to store $x_1, x_2, \ldots, x_n$ is fixed for all models, if the data description length for one model is smaller than that of another model, it implies that the size of storage used to store the errors for the former model is smaller than that for the latter model. In other words, a model is more accurate if the corresponding data description length is smaller.

Therefore, the MDL metric measures the total description length $D_t(B)$ of a network structure *B*. A better network has a smaller value on this metric. Let $N = \{N_1, \ldots N_n\}$ denote the set of nodes in the network (and thus the set of variables, since each node represents a

variable), and Π_{Ni} denotes the set of parents of node *Ni*. The total description length of a network (D_t) is the sum of description lengths of each node:

$$D_t(B) = \sum_{N_i \in N} D_t\left(N_i, \Pi_{N_i}\right) \tag{1}$$

The total description length (D_t) is based on two components, the network description length (D_n) and the data description length (D_d):

$$D_t\left(N_i, \Pi_{N_i}\right) = D_n\left(N_i, \Pi_{N_i}\right) + D_d\left(N_i, \Pi_{N_i}\right) \tag{2}$$

The formula for the network description length is:

$$D_n\left(N_i, \Pi_{N_i}\right) = k_i \log_2(n) + d(s_i - 1) \prod_{j \in \Pi_{N_i}} s_j \tag{3}$$

where k_i is the number of parents of variable N_i, S_i is the number of values N_i can take on, S_j is the number of values a particular variable in Π_N can take on, and d is the number of bits required to store a numerical value. This is the description length for encoding the network structure. The first part in the addition is the length for encoding the parents, while the second part is the length for encoding the probability parameters. In this case, the network length measures the simplicity of a network.

The formula for the data description length (D_d) is:

$$D_d\left(N_i, \Pi_{N_i}\right) = \sum_{N_i \in \Pi_{N_i}} M\left(N_i, \Pi_{N_i}\right) \log_2 \frac{M\left(\Pi_{N_i}\right)}{M\left(N_i, \Pi_{N_i}\right)} \tag{4}$$

where $M(.)$ is the number of cases that match a particular instantiation in the database. This is the description length for encoding the data. A Huffman code is used to encode the data using the probability measure defined by the network. Thus, this data description length measures the accuracy of a network.

COMBINING BN AND EP

The learning algorithm for combining BN and EP is given in Table 2. Each individual represents a network structure model, which is a directed acyclic graph (DAG). Firstly, a set of individuals is randomly generated to make up the initial population. Each graph is evaluated by the MDL metric described above. Then, each individual produces a child by performing a number of mutations, which are described the in the next section. The child is also evaluated by the MDL metric. The next generation of population is selected among the parents and children by tournaments. Each DAG B is compared with other randomly selected DAGs. The tournament score of B equals to the number of rivals that B can win, that is, the number of DAGs among those selected that have higher MDL scores than B. In our setting, q = 5. One half of DAGs with the highest tournament scores are retained for the next generation. The process is repeated until the maximum number of generations is reached. The number of the maximum number of generations depends on the complexity of the network structure. If we expect a simple network, the maximum number of generations can be set to a lower value. The network with the lowest MDL score emerges as the final result, i.e., the optimal solution.

MUTATION OPERATORS

Mutation, an asexual operation, is the only genetic operator used in EP for evolution. Offspring in EP are produced using a specific number of mutations. The probabilities of using 1, 2, 3, 4, 5 or 6 mutations are set to 0.2, 0.2, 0.2, 0.2, 0.1 and 0.1 respectively. The mutation operators modify the edges of a DAG. If a cyclic graph is formed after the mutation, edges in the cycles are removed to keep it acyclic. Four mutation operators, with the same probabilities of being selected, are used for evolution:

1. Simple mutation randomly adds an edge between two nodes or randomly deletes an existing edge from the parent.

2. Reversion mutation randomly selects an existing edge and reverses its direction.

3. Move mutation randomly selects an existing edge. It moves the parent of the edge

to another node, or moves the child of the edge to another node.

4. Knowledge-guided mutation is similar to simple mutation. However, the MDL scores of the edges guide the selection of the edge to be added or removed. The MDL metric of all possible edges in the network is computed before the learning algorithm starts. This mutation operator stochastically adds an edge with a small MDL metric to the parental network or deletes an existing edge with a large MDL metric.

METHOD

To test the robustness and feasibility of the proposed methods for modeling direct marketing responses, we need to run the Bayesian Networks learning experiments with data from direct marketing operations. To facilitate the data mining process as well as model evaluation and comparison, the research team includes a data mining expert and a marketing domain expert. While the data mining expert performs the BN learning, the results of the learned BN models are evaluated by the marketing domain expert. We have applied the proposed methods to two direct marketing data sets to generate models of consumer response.

The first data set for this study comes from a direct mail promotion program from the credit card division of a major U.S. bank. The database contains the data of 308,857 people who received an "invitation to apply" direct mail promotion from the bank. The data include over 2,000 variables, including consumer demographics and financial information as well as response data of the consumers to the most recent credit card promotion from the bank. The number of responders to the promotion was 1,623, representing a response rate of 0.53%, which is close to the industry average.

First, we sampled 3,785 records or 1.2% from the database, including 100% of the responders (1,623) and 0.7% non-responders (2,162). Following the industry practice, over-sampling of the responders is performed to ensure nearly symmetric distribution of responders and non-responders for the logistic regression model (Scott and Wild 1986, 1997). Since Bayesian Networks also calculate the distribution of probabilities, the same concern is also relevant. Thus, Bayesian Networks learning uses the same sample for direct comparisons with the results of logistic regression.

The second dataset comes from a U.S.-based catalog direct marketing company that sells multiple product lines of general merchandise ranging from gifts, apparel to consumer electronics. This particular database from the Direct Marketing Education Foundation stores records of 106, 284 consumers' response to a recent promotion as well as their purchase information over a twelve year period, including the credit information and demographic information appended from the 1995 U.S. Census. Each record contains 361 variables. This study focuses on the most recent catalog promotion with a 5.4% response rate or 5, 740 responders.

In our experiments, logistic regression was used to for data reduction using the forward selection method to select the variables for model building. Logistic regression has been widely used by researchers in direct marketing to select potential respondents. Due to budget constraints and other consideration, typically only the names in the top two deciles (i.e., those with the highest probabilities to respond) will receive the promotion materials from a company. Thus, the logistic model is also used to as a benchmark baseline model for comparison with Bayesian network models.

Following the standard practice in direct marketing, we split the sample into two sets for logistic regression and data mining with Bayesian networks using EP. A training set is used for developing the model and a testing set is used for validating the model. We have also performed a 10-fold cross-validation for the catalog promotion dataset to make further comparison concerning the robustness of the response models (Mitchell 1997). For all experiments in learning Bayesian networks using EP, the population size is 50 and the maximum number of generations is 5,000.

RESULTS

The First Experiment - Credit Card Promotion

For the credit card promotion, we first randomly split the sample into two equal subsets and then developed a logistic regression with the training set and validated with the testing set. Total 13 variables were selected for model building because they were considered important for mail operations by the bank's research department. The 13 variables include the following information: response (CLASS), household income (INCCODE), education

(GE2YR), marital status (MARTL), number of adults (NUMADLT), owner occupied housing (OWNER), dwelling size (DWSIZE), number of vehicles (NUMCAR), vehicle value (VEHVAL), number of direct marketing mails received (TMAIL), number of pre-screened offers received in the last twelve months (TMAILPS), club membership (TYFLAG), and baby registry (BBFLAG).

The logistic regression model has a Cox and Snell R-square of 0.101 and correctly classifies 64.5% of the cases. In addition, the Hosmer and Lemeshow test has an insignificant chi-square of 15.41 (DF=8, sig.=0.052), suggesting that the results predicted by the model are not significantly different from those being observed. Thus, the logistic regression model has a good fit of the data. Then, we generated the empirical results -- decile analysis of cumulative lift in a so-called gains table -- a standard measure of the model fitness in direct marketing (Shepard 2000). The gains table indicates the first two deciles have cumulative lifts of 274 and 218 respectively, suggesting that by mailing to the top two deciles alone, logistic regression model generates over twice as many respondents as a random mailing without a model (Table 3). However, the lift in the fourth declines sharply to 78, which is lower than the next three deciles (94. 82, 81), suggesting instability in the model.

(insert Table 3 here)

Then, the Bayesian networks method using the same set of variables was performed, first with the training set and then validated with the same testing set. We generated the gains table for the model to compare with those of the logistic regression model (Table 4). Comparing to the cumulative lift of 274 in the top decile of the logistic regression model, the Bayesian network model has only a cumulative lift of 261 in the top decile, even though its lift of 167 in the second decile is slightly higher than that of 163 in the logistic regression. Overall, the results of the Bayesian network model fall slightly short of the logistic regression model. The Bayesian network model repeats the drop of lift in the third decile (91), again suggesting instability in the model (Table 4).

(insert Table 4 here)

Furthermore, we generated the DAG for the Bayesian network model using all 13 variables (Figure 3). The relationship structure among the variables discovered by the

Bayesian networks appears to be much more complex than that of the logistic regression model. Most of the relationships discovered by the Bayesian network learning are meaningful and easy to understand based on the interpretation by the marketing domain expert. For instance, income is related to owner occupied housing, which is related to both marital status and the number of adults in the household. Response to the promotion is predicted by income, marital status, the number of regular and pre-screened mails received, education, and club membership. In the logistic regression, they are simply treated as separate endogenous variables.

(insert Figure 3 here)

The Second Experiment - Catalog Promotion

For the catalog promotion data, we split the data set into two parts, one for training the response model and the other one for testing. The training set contains 2,870 respondents and 5,740 non-respondents. The testing set contains 2,870 respondents and 94,804 non-respondents. Nine variables were selected for model building: responder (CLASS) cash payment (CASH), total promotion orders (PRORD85), frequency of purchase in the last 36 months (FREQUENT), money used in the last 36 months (CRCPR85), use of house credit card (NCARD), lifetime orders (SALCAT), telephone order (TEL), and recency (RECMON). The logistic regression model has cumulative lifts of 350 and 259 in the top two deciles, which are not exceptionally high given a response rate 5.4% response rate. The logistic regression model has a Cox and Snell R-square of 0.141 and correctly classifies 72% of the cases. The results show a gradual decline of lifts from the top decile to the lower deciles (Table 5).

(insert Table 5 here)

The same training and testing data sets were also used for Bayesian network learning. The results in Table 6 show that the Bayesian network model has a cumulative lift of 396 in the top decile and 290 in the second decile, significantly higher than those of the logistic regression model. In fact, all cumulative lifts in the first seven deciles are higher than those of the logistic regression model. Specifically, the percentage and number of actives captured by BN is significantly higher than that of logistic regression (58.15% vs. 51,85%, and 1,669 vs.

1,448). Overall, the Bayesian network model performs significantly better than the logistic regression model in predicting consumer response to direct mail promotions. BN performs better with the catalog data than the credit card data. We attribute this difference to the fact that the catalog data set is much bigger and has a much higher response rate than the credit card data, thus making the Bayesian network learning process more plausible and efficient. The DAG for the model is presented in Figure 4.

(insert Table 6 and Figure 4 here)

To make a further comparison concerning the robustness of the response models, we have employed a 10-fold cross-validation for performance estimation with non-overlapping samples for the catalog data. Firstly, we partition the dataset into 10 disjoint subsets of equal size. We then train and test the logistic regression and the Bayesian network learning algorithm for 10 times, using each of the 10 subsets in turn as the testing set, and using all remaining data as the training set. Thus, the algorithms are tested on 10 independent testing sets. In Tables 7 and 8, the experimental results for the logistic regression and our approach are presented respectively. We collect the statistics on the predicted probabilities, percentage of active response, lift, and cumulative lift at each decile averaged over the ten experiments. The numbers beside the "±" sign are the standard deviations. The experimental results suggest that the Bayesian network model predicts consumer response more accurately than the logistic regression model. Moreover, it provides higher cumulative lift in the first few deciles.

(insert Table 7 and Table 8 here)

DISCUSSION

Conclusions

First, the results of the two experiments and the cross-validation suggest that Bayesian Networks and evolutionary programming are plausible methods for optimizing modeling to predict consumer response to direct marketing. Although the results of the Bayesian networks method fall slightly short of the logistic regression with a small data set, the Bayesian networks approach generates superior results with a larger sample, suggesting that the

Bayesian network model furnishes a significant better representation of the structure of data. Comparing the empirical results of the logistic regression model, the Bayesian network model captures a larger percentage of buyers in the top two deciles and therefore, can potentially help improve sales and profitability of direct marketing programs. Despite these problems, our study shows that the Bayesian networks approach with evolutionary programming can potentially become a powerful and efficient data mining tool for marketing professionals.

The contribution of the current study is many-fold. First, the conventional methods of modeling consumer response to direct marketing can only compare a limited number of alternative models. This study proposes and tests a data mining approach to direct marketing response modeling using Bayesian Networks and evolutionary programming, which automatically optimize the response model by comparing thousands of alternative models with many different structures. Comparing to GA used in several studies, EP method is free from certain constraints and operates at a higher level of abstract while focusing the behavioral linkages between the parents and their offspring. Secondly, unlike other methods such as logistic regression where model structures are constrained by a specific learning algorithm or fitness function, the Bayesian network method is flexible and can represent any model structures. In addition to the empirical results (i.e., decile analysis of consumer response probability), the Bayesian networks also generate a graphic model to represent the discovered knowledge. It helps putting a face on the data mining process and de-mystifying the methods of artificial intelligence such as machine learning and evolutionary computation. Thirdly, while other methods such as GA can model consumer response as some kind of fitness criterion designed for the problem, our approach also adopts a fitness function (i.e., the MDL metric) to evaluate the fitness of alternative models in the process of evolution and optimization. Fourthly, the proposed methods are also computationally efficient. A single experiment with one sample takes only a few minutes on a standard personal computer.

Implications

The explosive growth of data is one of the most significant challenges facing marketing managers in the information age. The power of computers to collect data about consumers and their behaviors has far outstripped the ability of analysts to process them into usable, value-added information. The methods proposed in this study, i.e., Bayesian network models

and evolutionary programming, present effective and efficient tools for marketing managers to discover useful insight from databases and to develop meaningful knowledge to assist decision making. The proposed methods provide two significant advantages. First, Bayesian network models offer superior representation of data structure over the traditional methods such as logistic regression. The BN method is flexible, assumption free, and more importantly, it considers the interrelationships among variables. Secondly, given the large amount of data, evolutionary programming presents a robust and efficient tool to search and discover the best possible BN model to represent the discovered knowledge. In essence, the combination of Bayesian Network models and evolutionary programming lead to a more powerful tool for data mining than if applied alone or separately.

To take advantage of the mounting data, marketing researchers and data mining experts have devised various methods to discover new knowledge to assist decision making. As the internet provides an additional channel for direct marketing, the need for intelligent decision support systems and tools will grow. Although research problems in social sciences are often affected by a whole array of factors, the conventional method to marketing research, like many social sciences studies, is often theory driven in that the research tests the hypotheses about the relationships among the interested variables while statistically controlling those uninterested factors (Malhotra, Peterson and Bardi 1999). Research without a theoretical model is often considered lacking intellectual merits and analytical rigor. The merits of theory-driven research notwithstanding, the current environment also demand problem-oriented research and feasible methods to explore the vast quantities of disaggregated data (Silk 1993). The explosive growth of marketing data require efficient data mining tools in order to assist managers uncover useful knowledge from data bases for decision making.

Suggestions for Future Research

Data mining and artificial intelligence are among the several new technologies that will clearly have a major impact across a wide range of industries (Baker and Baker 1998). However, the experiments also reveal several problems associated with the proposed data mining methods. To apply Bayesian Network using evolutionary programming to direct marketing response modeling, the following issues demand attention in future research. First, the BN approach with evolutionary programming appears to be sensitive to sample size.

Sample size and proportion of buyers in the sample apparently affect the performance of the method as they do with regression analysis (Berger and Magliozzi 1992). With a small sample size, evolutionary programming may not have ample opportunities to learn the data structure in order to extract more accurate representations. Thus, how such methods perform with different types of data and sample size needs further examination.

Secondly, given the need for information on a real time basis, methodological and technological advances should be undertaken to greatly reduce the marketing research cycle time and complete projects in a few hours rather than a few months (Malhotra, Peterson and Bardi 1999). The proposed method maximizes the search space and optimizes the modeling process by comparing many alternative models, including the invalid ones. Efficiency of the data mining process may suffer as invalid models may be explored in the process. Putting constraints on the learning algorithm based on existing domain knowledge may help guide the search progress and improves its overall efficiency by avoiding the invalid models. Despite the declining cost of computing power, model building and validation using evolutionary computation methods are still time-consuming for large data sets with a greater number of variables. For large databases, EP procedures are computationally demanding and may perform less efficiently than mathematical optimization techniques. Thus, more research is needed to improve the computing efficiency of the evolutionary algorithms so that computing time can be reduced.

Thirdly, related to the above issue of computational efficiency is the challenge of meaningful data reduction or variable (feature) selection. In data-rich marketing environments, managers face an ever-growing need to reduce the number of variables effectively (Naik, Hagerty and Tsai 2000). Although researchers may exercise their judgment in a trial-and-error selection process, the increasing variety and number of variables would make an automated or semi-automated process more desirable. In order to mine useful information from large databases, a more efficient method is needed to automate or semi-automate the process of selecting meaningful variables for subsequent analyses and model building. Emerging techniques such as the filter method, the wrapper method, and the Bayesian method of variable selection may help solving such problems.

Finally, comparing to regression models, EP solutions are usually difficult to interpret since they do not have standard interpretative statistical measures that enable the user to

understand why the procedure arrives at a particular solution. Thus, the generated model needs the input from the domain expert to evaluate the validity of the discovered knowledge. DAG can help interpretation of results. It becomes more difficult with models with a larger number of variables. Visualization tools can help improve the interpretation of complex model and knowledge structures. While evolutionary programming is a powerful tool for searching and optimizing decision problems, such methods need to be made user-friendlier to marketing professionals and more flexible to handle a greater variety of variables and marketing problems.

REFERENCES

Bäck, T., D. Fogel and Z. Michalwicz (eds.). 2000. *Evolutionary Computation 1: Basic Algorithms and Operators*. Institute of Physics Publishing.

Baker, Sunny and Kim Baker. 1998. "Mine over matter". *The Journal of Business Strategy*, 19(4): 22-26.

Balakrishnan, P. V. and Varghese S. Jacob. 1996. "Genetic algorithms for product design". *Management Science*, Vol. 42 (8): 1105-1118.

Berger, Paul and Thomas Magliozzi. 1992. "The Effect of Sample Size and Proportion of Buyers in the Sample on the Performance of List Segmentation Equations Generated by Regression Analysis". *Journal of Direct Marketing*, Vol. 6 (1): 13-22.

Bhattacharyva, Siddhartha. 2000. Evolutionary Algorithm in Data Mining: Multi-Objective Performance Modeling for Direct Marketing, KDD-2000: 465-473.

Bitran, G. and S. Mondschein. 1996. Mailing Decisions in the Catalog Sales Industry. *Management Science*, 42(9), 1362-1381.

Bodenberg, T.M. and M.L. Roberts. 1990. "Integrating marketing research into the direct-marketing testing process: The market research test". *Journal of Advertising Research*, 30 (5): 50-60.

Bult, J. R. and T. Wansbeek. 1995. "Optimal Selection for Direct Mail". *Marketing Science*, 14 (4): 378-394.

Chang, K. C. and R. M. Fung. 1991. "Symbolic Probabilistic Inference with Continuous Variables". *Proceedings of the Eleventh Annual Conference on Uncertainty in Artificial Intelligence*. San Francisco, CA: Morgan Kaufmann.

Chickering, D, D. Geiger, D. and D. Heckerman. 1995. "Learning Bayesian Networks: Search Methods and Experimental Results". *Proceedings of the Fifth Conference on Artificial Intelligence and Statistics*: 112-128.

Cooper, G. 1990. "The Computational Complexity of Probabilistic Inference using Bayesian Belief Networks". *Artificial Intelligence*, 42: 393-405.

Cooper, G. and E. A. Herskovits. 1992. "The Computational Complexity of Probabilistic Inference using Bayesian Belief Networks". *Artificial Intelligence* (9): 309-347.

D'Ambrosio, Bruce. 1999. "Inference in Bayesian Networks". *AI Magazine* (summer): 21-36.

David Shepard Associates. 1999. The New Direct Marketing, 3rd ed.. New York: McGraw-Hill.

DeSarbo, W. S. and V. Ramaswamy. 1994. "CRISP: Customer Response Based Iterative Segmentation Procedures for Response Modeling in Direct Marketing". *Journal of Direct Marketing*, 8(3): 7-20.

Earman, John. 1990. "Baye's Bayesianism". *Studies in History and Philosophy of Science*, 21(3): 351.

Eiben, A. E., T. Euverman, W. Kowalczyk, E. Peelen, F. Slisser and J. Wesseling. Comparing Adaptive and Traditional Techniques for Direct Marketing.

Fayyad, U. M., G. Piatetsky-Shapiro, and P. Smyth. 1996. "From Data Mining to Knowledge Discovery: An Overview". In U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy (eds.). *Advances in Knowledge Discovery in Data Mining*, 1-34. Menlo Park, CA: AAAI Press.

Fogel, D. B. 1994. "An introduction to simulated evolutionary optimization". *IEEE Transactions on Neural Network*, 5:3-14.

Fogel, L., A. Owens, and M. Walsh. 1966. *Artificial Intelligence through Simulated Evolution*. New York: John Wiley and Sons.

Geiger, Dan and David Heckerman. 1996. "Knowledge Representation and Inference in Similarity Networks and Bayesian Multinets". *Artificial Intelligence* 82: 45-74.

Geiger, Dan, David Heckerman, and C. Meek. 1996. “Asymptotic Model Selection for Directed Graphs with Hidden Variables”. *Proceedings of the Twelfth Conference on Uncertainty in Artificial Intelligence*. San Francisco, CA: Morgan Kaufmann: 283-290

Goldberg, D. E. 1989. “Genetic Algorithm in Search, Optimization and Machine Learning”. Reading , MA: Addison-Wesley.

Gönül, Füsun F., Byung-Do Kim and Mengze Shi. 2000. “Mailing Smarter to Catalog Customers”. *Journal of Interactive Marketing*, 14 (2): 2-16.

Haddawy, Peter. 1999. “An Overview of Some Recent Developments in Bayesian Problem-solving Techniques”. *AI Magazine* (Summer): 11-19.

Heckerman, David and Michael P. Wellman. 1995. “Bayesian Networks”. *Communications of the ACM*, 38 (3): 27-30.

Heckerman, D., D. Geiger and D. M. Chickering. 1995. “Learning Bayesian Networks: the Combination of Knowledge and Statistical Data”. *Machine Learning* 20 (3): 197-243.

Holland, J. 1975. *Adaptation in Natural and Artificial Systems*. Ann Arbor: The University of Michigan Press.

Hurley, S., L. Moutinho and N.M. Stephens. 1995. “Solving marketing optimization problems using genetic algorithms”. *European Journal of Marketing*, Vol. 29 (4): 39-56.

Jensen, F. V. 1996. An Introduction to Bayesian Networks. UCL Press.

Klemz, Bruce R. 1999. “Using genetic algorithms to assess the impact of pricing activity timing”. *Omega*, Vol. 27 (3): 363-372.

Koza, J. R. 1992. *Genetic Programming: on the Programming of Computers by Means of Natural Selection*. Cambridge, MA: MIT Press.

Lam, W. 1998. “Bayesian Network Refinement via Machine Learning Approach”. IEEE Transactions Pattern Anal Machine Intelligence.

Lam, W. and F. Bacchus. 1994. “Learning Bayesian Belief Networks -- An Approach based on the MDL Principle”. *Compu Intelligence*, 10 (3): 269-293.

Lam, W., M.L. Wong, M.S. Leung, and P. s. Ngan. 1998. “Discovering Probabilistic Knowledge from Databases using Evolutionary Computation and Minimum Description Length Principle”. Genetic Programming Proceedings of the Third Annual Conference.

Larrañaga, P., M. Poza, Y. Yurramendi, R. Murga, and C. Kuijpers. 1996. “Structure Learning of Bayesian Network by Genetic Algorithms: A Performance Analysis of Control Parameters”. *IEEE Trans Pattern Anal Machine Learning*, 18 (9): 9.

Lauritzen, S. L. 1990. “Propagation of Probabilities, Means, and Variances in Mixed Graphical Association Models”. *Research Report R90-18*. Institute for Electronic Systems, Aalborg University.

Levin, Nissan and Jacob Zahavi. 2001. “Predictive Modeling Using Segmentation”. *Journal of Interactive Marketing*, 15 (2): 2-22.

Ling, Charles X. and Chenghui Li. 1998. “Data Mining for Direct Marketing: Problems and Solutions”. KDD-98: 73-79.

Malhotra, Naresh K., Mark Peterson and Susan Bardi. 1999. “Marketing research: A state-of-the-art review and directions for the twenty-first century”. *Journal of the Academy of Marketing Science*, 27(2): 160-183.

Malthouse, Edwrad C. 2001. “Assessing the Performance of Directing Marketing Scoring Models”. *Journal of Interactive Marketing*, 15(1): 49-62.

Midgley, David F., Robert E. Marks and Lee G. Cooper. 1997. “Breeding Competitive Strategies”. *Management Science*, Vol. 43 (3): 257-275.

Mitchell, T. 1997. Machine Learning. New York: McGraw-Hill.

Naik, Prasad A., Michael R. Hagerty and Chih-Ling Tsai. 2000. “A new dimension reduction approach for data-rich marketing environments: Sliced inverse regression”. *Journal of Marketing Research*, 37(1): 88-101.

Ngan, Po Shun, Man Leung Wong, Wai Lam, Kwong Sak Leung, and Jack C. Y. Cheng. 1999. “Medical data mining using evolutionary computation”. *Artificial Intelligence in Medicine*, 16: 73-96.

Olesen, K. G. 1991. “Causal Probabilistic Networks with both Discrete and Continuous Variables”. *Research Report R91-29*. Institute for Electronic Systems, Aalborg University.

Peacock, Peter R. 1998. “Data mining in marketing: Part 1”. *Marketing Management*, Vol. 6 (4): 8-18.

Pearl, J. 1988. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. San Mateo: Morgan Kaufmann.

Rebane, G. and J. Pearl. 1989. “The Recovery of Causal Poly-trees from Statistical Data”. *Uncertainty in Artificial Intelligence 3*. Amsterdam: North-Holland: 175-182.

Rissanen, J. 1978. “Modeling by Shortest Data Description”. *Automatica*, 14: 465-471.

Rud, Olivia Parr. 2001. Data Mining Cookbook. John Wiley & Sons, Inc. New York.

Schewefel, H. P. 1981. *Numerical Optimization of Computer Models*. New York, NY: Wiley.

Scott, A. J. and Wild, C. J. 1986 “Fitting logistic regression models under case-control or choice based sampling”. *Journal of the Royal Statistical Society B*, 48: 170-182.

Scott, A. J. and Wild, C. J. 1997. “Fitting Regression Models to Case-Control Data by Maximum Likelihood” *Biometrika*, 84: 57-71.

Shaw, Michael J., Chandrasekar Subramaniam, Gek Woo Tan and Michael E. Welge. 2001. “Knowledge management and data mining for marketing”. *Decision Support Systems*, 31(1): 127-137.

Shepard, David. 1999. The New Direct Marketing. The David Shepard Associates.

Silk, Alvin J. 1993. “Marketing Science in a Changing Environment”. *Journal of Marketing Research*, 30 (4): 401-404.

Spirtes, P., Glymour, C. & Scheines, R. *Causation, Prediction and Search, Second Edition.* MIT Press, MA, 2000.

Thieme, R. Jeffrey; Michael Song and Roger J. Calantone. 2000. “Artificial neural network decision support systems for new product development project selection”. *Journal of Marketing Research*, 37(4): 499-507.

Urban, Timothy L. 1998. “An inventory-theoretic approach to product assortment and shelf-space allocation”. *Journal of Retailing*, Vol. 74 (1): 15-35.

Wong, Man Leung, W. Lam and K. S. Leung. 1999. “Using Evolutionary Computation and Minimum Description Length Principle for Data Mining of Probabilistic Knowledge”. IEEE Transactions: Pattern, Analysis, and Machine Intelligence. Engineering in Medicine and Biology, July/Aug., 45-55.

Figure 1. An Innovative Approach to Knowledge Discovery Using Bayesian Networks

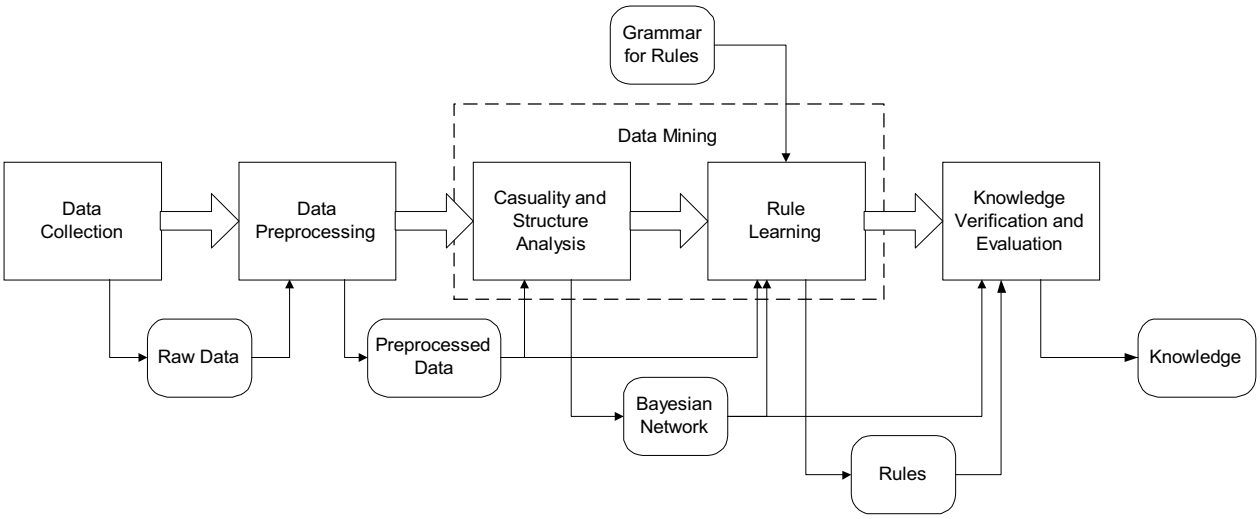

Figure 2. A Directed Acyclic Graph (DAG) using Bayesian Networks

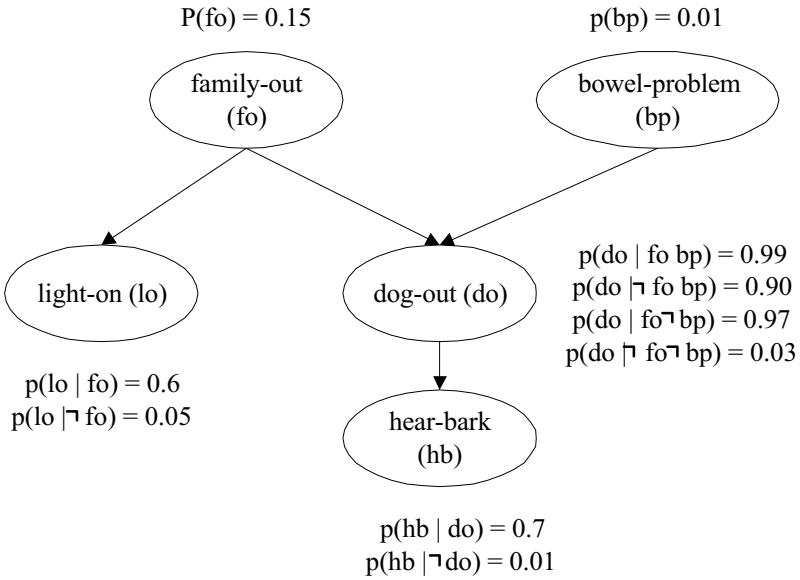

Figure 3. A Directed Acyclic Graphic Model for the Credit Card Promotion

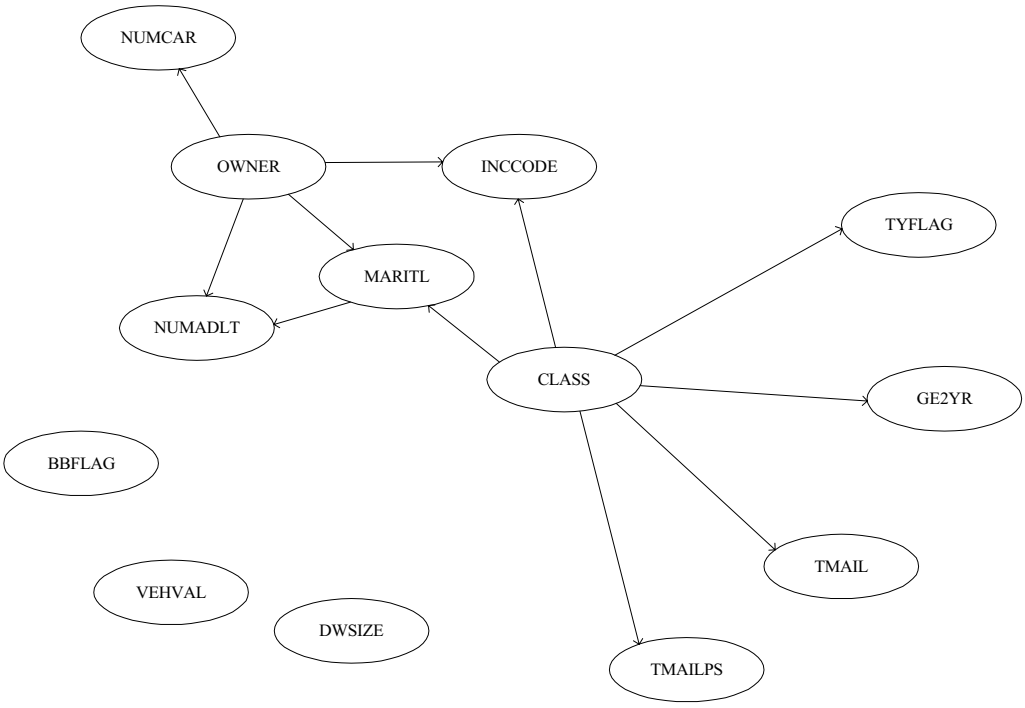

Figure 4. A Directed Acyclic Graphic Model for the Catalog Promotion Response

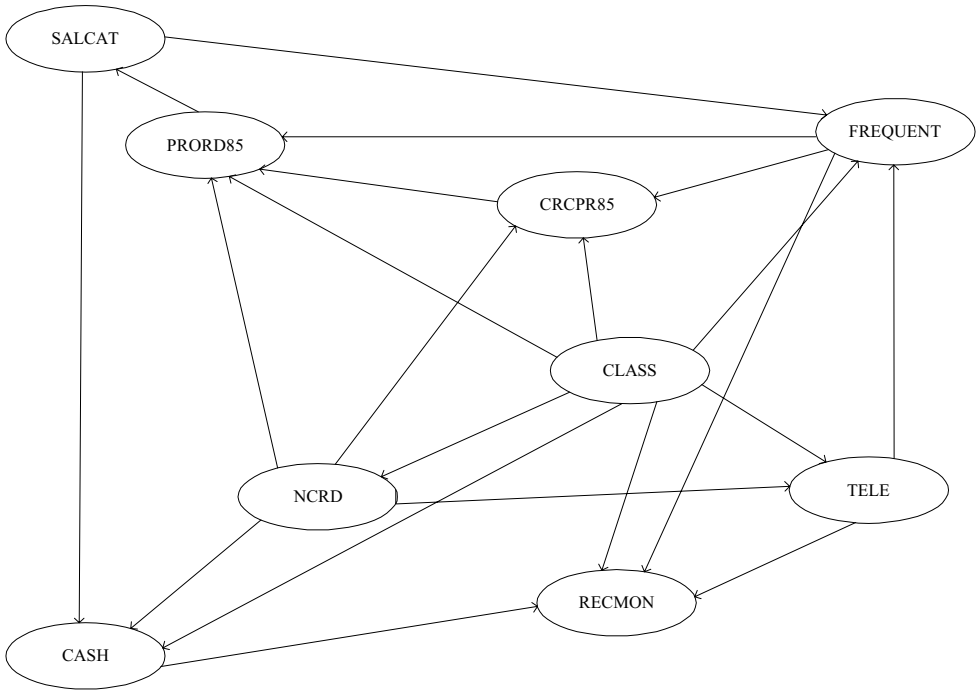

Table 1: The Algorithm of Evolutionary Programming

---

initialize the generation, t, to be 0.
Initialize a population of individual, Pop(t)
Evaluate the fitness of all individual in Pop(t)
While the termination criteria is not satisfied
    Produce one or more offspring from each individual by mutation
    Evaluate the fitness of each offspring
    Perform a tournament for each individual
    Put the individuals with high tournament scores into Pop(t+1)
    Increase the generation t by 1
Return the individual with the highest fitness value

Table 2. The Algorithm for Evolutionary Programming to Learn Bayesian Networks

---

- Set t to 0.
- Create an initial population, pop(t), of PS random Bayesian Networks.
- Each Bayesian Network in the population Pop(t) is evaluated by using the fitness function.
- While t is smaller than the maximum number of generation G
    - Each Bayesian Network in Pop(t) produces one offspring by performing a number of mutation operations. If the offspring has cycles, delete the edges of the offspring that invalidate the acyclic condition.
    - The Bayesian Networks in Pop(t) and all new offspring are stored in the intermediate population Pop(t'). The size of Pop(t') is 2*PS.
    - Conduct a number of pairwise competitions over all Bayesian Networks in Pop(t'). Let Bi be the Bayesian Network being conditioned upon, q opponents are selected randomly from Pop(t') with equal probability. Let Bij, 1 j q, be the randomly selected opponent Bayesian Networks. The Bi gets one more score if Dt(Bi) Dt(Bij), 1 j q.
    - Select PS Bayesian Networks with the highest scores from Pop(t') and store them in the new population Pop(t+1).
    - Increase t by 1.
- Return the Bayesian Network with lowest fitness value found in any generation of a run as the result of the algorithm

---

Table 3. Gains Table for Logistic Regression of Credit Card Promotion

Decile	Records	% of File	Prob. of Active	Percent Active	Cum. % Active	# of Actives	% of Total Actives	Cum. # of Actives	Cum. % of Tot Actives	Lift	Cum. Lift
0	30833	10%	0.64	1.44	1.44	445	27.42	445	27.41	274	274
1	30794	20%	0.54	0.85	1.15	264	16.27	709	43.68	163	218
2	30721	30%	0.48	0.62	0.97	191	11.77	900	55.45	118	185
3	30798	40%	0.45	0.40	0.83	126	7.76	1026	63.21	78	158
4	30825	50%	0.42	0.49	0.77	153	9.42	1179	72.64	94	145
5	30805	60%	0.39	0.43	0.71	133	8.19	1312	80.84	82	135
6	30803	70%	0.34	0.42	0.67	131	8.07	1443	88.91	81	127
7	30768	80%	0.29	0.31	0.62	96	5.91	1539	94.82	59	119
8	30725	90%	0.22	0.17	0.57	53	3.26	1592	98.09	33	109
9	30845	100%	0.11	0.10	0.53	31	1.91	1623	100.00	19	100
Total	307917					1623	100				

Table 4. Gains Table for Bayesian Network Model of Credit Card Promotion

Decile	Records	% of File	Prob. of Active	Percent Active	Cum. % Active	# of Actives	% of Total Actives	Cum. # of Actives	Cum. % of Tot Actives	Lift	Cum. Lift
0	30644	10%	0.64	1.37	1.37	420	25.88	420	25.88	261	261
1	30789	20%	0.55	0.88	1.12	271	16.70	691	42.58	167	214
2	30664	30%	0.50	0.48	0.91	146	9.00	837	51.58	91	173
3	30682	40%	0.47	0.53	0.82	164	10.11	1001	61.68	102	155
4	30680	50%	0.45	0.53	0.76	162	9.98	1163	71.67	100	144
5	30689	60%	0.41	0.56	0.72	171	10.54	1334	82.20	106	138
6	30622	70%	0.37	0.34	0.67	104	6.41	1438	88.61	65	127
7	30603	80%	0.32	0.28	0.62	85	5.24	1523	93.85	53	118
8	30867	90%	0.24	0.18	0.57	56	3.45	1579	97.30	35	109
9	32616	100%	0.12	0.13	0.53	44	2.71	1623	100.01	26	100
	307917					1623	100				

Table 5. Gains Table for Logistic Regression of Catalog Promotion

Decile	Records	Prob of Active	Percent Active	Cum. % Active	# of Actives	% of Total Actives	Cum. # of Actives	Cum. % of Tot Actives	Lift	Cum. Lift
0	9768	0.57	10.30	10.30	1006	35.05	1006	35.05	350	350
1	9768	0.50	4.93	7.62	482	16.79	1488	51.85	167	259
2	9768	0.47	4.39	6.54	429	14.95	1917	66.79	149	222
3	9768	0.43	2.50	5.53	244	8.50	2161	75.30	85	188
4	9768	0.38	1.98	4.82	193	6.72	2354	82.02	67	164
5	9768	0.32	1.55	4.27	151	5.26	2505	87.28	52	145
6	9768	0.26	1.26	3.84	123	4.29	2628	91.57	42	130
7	9768	0.19	0.94	3.48	92	3.21	2720	94.77	32	118
8	9768	0.14	0.84	3.19	82	2.86	2802	97.63	28	108
9	9762	0.08	0.70	2.94	68	2.37	2870	100.00	23	100
	97,674				2870	100				

Table 6. Gains Table for Bayesian Network Model of Catalog Promotion

Decile	Records	Prob of Active	Percent Active	Cum. % Active	# of Actives	% of Total Actives	Cum. # of Actives	Cum. % of Tot Actives	Lift	Cum. Lift
0	9768	0.98	11.65	11.65	1138	39.65	1138	39.65	396	396
1	9768	0.62	5.44	8.54	531	18.50	1669	58.15	185	290
2	9768	0.38	3.71	6.93	362	12.61	2031	70.77	126	235
3	9768	0.29	1.74	5.63	170	5.92	2201	76.69	59	191
4	9768	0.22	1.96	4.90	191	6.66	2392	83.34	66	166
5	9768	0.15	1.27	4.29	124	4.32	2516	87.67	43	146
6	9768	0.10	1.26	3.86	123	4.29	2639	91.95	42	131
7	9768	0.07	0.92	3.49	90	3.14	2729	95.09	31	118
8	9768	0.05	0.76	3.19	74	2.58	2803	97.67	25	108
9	9762	0.02	0.69	2.94	67	2.33	2870	100.00	23	100
	97,674				2870	100				

Table 7. Cross-validation Gains Table for Logistic Regression of Catalog Promotion

Decile	Prob. of Active	Percent Active	Lift	Cum. Lift
0	0.2489±0.0020	0.1849±0.0131	342.0±17.1	342.0±17.1
1	0.1559±0.0016	0.0864±0.0089	159.4±15.1	250.8±7.8
2	0.1267±0.0014	0.0714±0.0073	131.8±11.8	211.1±5.7
3	0.1045±0.0011	0.0612±0.0052	113.1±10.0	186.7±4.6
4	0.0850±0.0011	0.0374±0.0084	68.7±14.2	163.1±2.3
5	0.0663±0.0011	0.0302±0.0072	55.0±10.3	145.0±1.8
6	0.0489±0.0009	0.0238±0.0053	43.3±8.3	130.5±1.8
7	0.0345±0.0006	0.0182±0.0052	33.3±9.8	118.4±1.2
8	0.0236±0.0004	0.0150±0.0023	27.5±4.4	108.2±0.9
9	0.0138±0.0002	0.0114±0.0039	20.6±7.4	100.0±0.0

Table 8. Cross-validation Gains Table for Bayesian Network Model of Catalog Promotion

Decile	Prob. of Active	Percent Active	Lift	Cum. Lift
0	0.2789±0.0053	0.2143±0.0073	397.2±19.0	397.2±19.0
1	0.1058±0.0029	0.0897±0.0102	165.5±16.0	281.4±8.9
2	0.0619±0.0017	0.0584±0.0096	107.5±14.1	223.4±7.2
3	0.0452±0.0013	0.0500±0.0093	92.0±15.6	190.5±2.8
4	0.0344±0.0009	0.0330±0.0046	60.5±6.7	164.5±1.6
5	0.0247±0.0005	0.0298±0.0046	54.5±6.5	146.3±1.0
6	0.0176±0.0004	0.0208±0.0045	37.7±7.4	130.8±1.0
7	0.0126±0.0001	0.0161±0.0021	29.4±3.6	118.1±1.0
8	0.0085±0.0002	0.0130±0.0034	23.5±6.4	107.5±0.7
9	0.0034±0.0002	0.0149±0.0026	27.1±5.0	100.0±0.0