ICEB 2001 Proceedings

International Conference on Electronic Business (ICEB)

Winter 12-19-2001

Data Mining for Decision Making in Direct Marketing: A Bayesian Network Approach with Evolutionary Programming

Geng Cui

Man Leung Wong

DATA MINING FOR DECISION MAKING IN DIRECT MARKETING: A BAYESIAN NETWORKS APPROACH WITH EVOLUTIONARY PROGRAMMING

Geng Cui
Dept. of Marketing & International Business
and
Man Leung Wong
Dept. of Information Systems
Lingnan University
Tuen Mun, New Territories, Hong Kong
Tel: (852) 2616-8245
Fax: (852) 2467-3049
E-mail: gcui@ln.edu.hk

ABSTRACT

Given the explosive growth of customer and transactional information, data mining can potentially discover new knowledge to improve managerial decision making in marketing. This study proposes an innovative approach to data mining using Bayesian Networks and evolutionary programming and applies the methods to direct marketing data. The results suggest that this approach to knowledge discovery can generate superior results than the conventional method of logistic regression. Future research in this area should devote more attention to applying this and other data mining methods to solving complex problems facing today's businesses.

INTRODUCTION

Conventional marketing research is a process in which data are analyzed manually to explore the relationships among various factors defined by the researcher. Even with powerful computers and versatile statistical software, many hidden and potentially useful relationships may not be recognized by the analyst. Nowadays, such problems are more acute as many businesses are capable of generating and collecting a huge amount of data in a relatively short period. 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. Through data mining, marketing researchers can discover complex relationships among various factors and extract meaningful knowledge to improve the efficiency and quality of managerial decision making.

In order for data mining to work for marketing managers, several issues have to be addressed. First, the process needs to adopt a method and produce results that can represent the structure of knowledge of the specific domain and specify the relationships among the variables. Secondly, the process should search the space for the best solution among all eligible candidates. Thirdly, the results of the data mining process should allow for comparison with existing methods using some common evaluation criteria to assist managerial decision making. Given these problems, we propose an innovative approach to knowledge discovery in marketing using Bayesian Networks and evolutionary programming. First, we introduce the background literature on data mining and the research problems. Secondly, we delineate the Bayesian Network learning process and evolutionary programming for data mining purposes. Thirdly, we apply these methods to two datasets of direct marketing and compare the results with those of logistic regression models. Finally, we explore the implications for data mining in marketing and directions for further research.

AN INNOVATIVE APPROACH TO KNOWLEDGE DISCOVERY

The increasing use of computers results in an explosion of information for businesses. Data can be best used if the hidden knowledge can be uncovered, thus making data mining an important research topic. 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 [19]. Within the broad-scope definition, data mining encompasses "confirmation" or the testing of relationships revealed through the discovery process. Data mining is the core of the knowledge discovery in database (KDD) process. Thus, the two terms are often used interchangeably [26]. Research in this area can be useful for many real-world problems

With computerization of marketing operations, a huge amount of customer and transactional data can be collected. Thus, there is a need for a way to automatically discover knowledge from data [26]. Data mining is increasingly used by many companies to improve marketing efficiency. Data mining has many potential uses in marketing, including customer acquisition, customer retention, customer abandonment and market basket analysis. In addition to query tools, descriptive statistics, visualization tools, regression-type models, association rules, decision tree analysis, and case-based reasoning, recent development in artificial intelligence and machine learning has presented more powerful data mining techniques and analytical tools, such as artificial neural networks (ANN) and evolutionary computation methods such as genetic algorithms [19].

Despite the promises of data mining, practical analytical tools that can assist managerial decision making need to be developed. One of the promising methods of evolutionary computation for solving marketing problems is genetic algorithms (GA). GA was originally developed in the field of computer science. Management researchers have adopted its principles and methods to solve business problems. 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 [19]. They are particularly effective for solving poorly understood, poorly structured problems because they attempt to find many solutions simultaneously, whereas a linear regression model, for example, focuses on a single best solution. Another strength of GA is that they can explicitly model any decision criterion in the "fitness function," an objective system used to assess a GA's performance [9] [19].

Recently, methods based on the evolutionary theory such as genetic algorithms have been applied to marketing problems such as product design [1], inventory control and product assortment management [24], brand competition [17], and marketing mix elasticity [11], direct marketing response modeling [2] [16]. For instance, to solve the problem of optimal product design using conjoint analysis, Balakrishnan and Jacob [1] used Genetic Algorithms (GA) as an alliterative procedure for generating "good" solutions for product design. Midgley, Marks and Cooper [17] adopted genetic algorithms to study how strategies may evolve in oligopolistic markets characterized by asymmetric competition. Subsequent simulations of repeated interactions using scanner data of brand actions show that the artificial agents bred in this environment outperform the historical actions of brand managers in the real market.

Recent research and studies in marketing focus on how to apply GA techniques to specific marketing problems and how the results compare to other conventional methods. Other major applications of GA include rule finding, pattern matching, and optimization. However, a major benefit of GA relative to other procedures is knowledge discovery in that they can produce novel solutions and discover relationships not defined by researchers. They may discover combinations of predictor variables that no one would have expected to be predictive beforehand [19]. Such beneficial features can be helpful for knowledge discovery in marketing and need to be explored.

As in other fields, data mining for marketing faces several significant challenges. First, conventional research emphasizes hypothesis testing based on *a priori* model with a limited number of variables selected by the researcher. Data mining, however, discovers the relationships and presents a *posterior* structure. Thus, the process needs to adopt a method and produce results that can represent the structure of knowledge of the specific domain and specify the relationships among the variables. Secondly, in the same vein, unlike conventional research that focuses on confirming an *a prior* model, data mining by definition should search the space for all possible alternative representations of the knowledge and then determine the best possible solution among all eligible candidates based on a fitness criterion. Thirdly, since data mining often adopts a method that is dissimilar to conventional statistical methods, the results of the data mining process should allow for comparison with those generated by other methods based on some common evaluation criteria so that they can assist managerial decision making. Against this backdrop, we propose an innovative approach to data mining in marketing.

The Knowledge Discovery Process

Data mining experts have developed various knowledge discovery systems to extract knowledge from databases. To apply data mining to marketing problems and to address the above issues, we propose an innovative approach to knowledge discovery in marketing using Bayesian Network (BN) models and evolutionary programming. In the following section, we delineate the novel approach to data mining and describe the learning process using the Bayesian Networks approach and evolutionary programming (EP).

First, we adapt the data mining process developed by Ngan et al. [18] and briefly describe its five steps in the process. Initially, 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 experts, who 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, these five steps will be reiterated [26].

In this study, we focus on the third and fourth steps. For causality and structural analysis, we use Bayesian Network models to represent the knowledge structure. To learn a plausible Bayesian Network model, we adopt evolutionary programming (EP) for the learning process. In the following sections, we describe the Bayesian Network models and evolutionary programming including the criteria for model evaluation and the learning process.

Bayesian Network Learning

Although the underlying theory of Bayesian probability has been around for a long time, building and executing realistic Bayesian Network models has only been made possible because of recent algorithms and software tools that implement them [10] [20]. Bayesian network is a method for formal knowledge representation based on the well-developed Bayesian probability theory. Bayesian networks have made tremendous progress and have been widely adopted by researchers in many fields. Several authors have given excellent introductions of Bayesian Networks and detailed comparisons with other methods [4] [6] [7] [8] [15].

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 [7]. The Bayesian network method has been successfully applied to solve many real-world problems including software engineering, space navigation, and medical diagnosis.

The most common computation performed using Bayesian Networks is 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 [7]. In essence, a Bayesian network captures the conditional probabilities between variables and can be used to perform reasoning under uncertainty. In practice, a Bayesian network is a directed acyclic graph (DAG). 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 a causality, with A being the cause and B being the effect. The value of each variable should be discrete. Each node is associated with a set of parameters. Thus, let N_i denotes a node and Đ_{N_i} denotes the set of parents of N_i. And the parameters of N_i are conditional probability distributions in the form of $P(N_i \mid \text{Đ}_{N_i})$ with one distribution for each possible instance of Đ_{N_i}.

The main task of learning Bayesian networks from data is to automatically find directed edges between the nodes so that the network can best describe the causalities. Once the network structure is constructed, the conditional probabilities are calculated based on the data. The problem of Bayesian network learning is computationally intractable. However, Bayesian network learning can be implemented by imposing limitations and assumptions. For instance, the algorithms of Rebane and Pearl [21] can learn networks with tree structures, while the algorithms of Cooper and Herskovits [3] require the variables to have a total ordering. More general algorithms include those by Heckerman, Geiger and Chickering [8] and Spirtes, Glymour and Scheines [23]. More recently, Larranaga et al [15] proposed algorithms for learning Bayesian networks using GA.

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 its quantitative aspect [7]. 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 [8]. In addition, the Bayesian networks method offers several other benefits for marketing research. Like logistic regression, the Bayesian networks approach is free from the normality assumption thus it can handle all types of data, binary, ordinal and continuous. Bayesian networks also test for independence among variables so that spurious relationships can be identified and avoided. 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.

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. The algorithms maintain a group of individuals to explore the search space. A potential solution to the problem is encoded as an individual. An evolutionary algorithm maintains a group of individuals, called the population, to explore the search space. A fitness function evaluates the performance of each individual, a Bayesian network model in this case, 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 able to produce "good" or even better offspring. Thus, an individual with a higher score in the fitness function has a higher chance of undergoing evolution. Evolution is performed by changing the existing individuals. New individuals are generated by applying genetic operators that alter the underlying structure of individuals. It is a general, domain independent method that does not require any domain-specific heuristic to guide the search.

Examples of algorithms in evolutionary computation include genetic algorithms (GA), genetic programming (GP), evolutionary programming (EP), and evolution strategy. They mainly differ in the evolution models assumed, the evolutionary operators employed, the selection methods, and the fitness functions used. GA uses a fixed-length binary bit string as an individual. 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. Meanwhile, GP extends GA by using a tree structure as the individual. But EP emphasizes on the behavioral linkage between parents and their offspring. Mutation is the only genetic operator in EP. There is no constraint on the representation in EP. In contrast, ES focuses on the individual, i.e. the phenotype, to be 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. Since evolutionary computation is a robust and parallel search algorithm, it can be used in data mining to find interesting knowledge in noisy environment. Data mining can be considered as a search problem, which tries to find the most accurate knowledge from all possible hypotheses.

Evolutionary Programming

Again, Evolutionary Programming (EP) emphasizes the behavioral linkage between parents and their offspring, rather than seeking to emulate specific genetic operators as observed in nature [5]. Different from GA, EP does not require any specific genotype in the individual. Thus, EP employs a model of evolution at a higher abstraction. Mutation is the only operator used for evolution. In a typical process of EP, a set of individuals is randomly created to make up the initial population. Each individual is evaluated by the fitness function. Then each individual produces an offspring by mutation. There is a 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 is 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 the parents and offspring. The tournament score of the individual is the number of rivals with lower fitness scores than itself. Then, 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.

EP has two distinctive advantages. First, there are no constraints on the representation. Mutation operator does not demand a particular genotype. The representation can follow from the problem. Second, mutations in EP attempt to preserve behavioral similarity between offspring and their parents. An offspring 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 parent' s behavior. Thus, the severity of mutations is according to a statistical distribution. The flexibility and freedom from constraints of EP make it an ideal tool as the search mechanism for data mining.

Structure analysis

In the proposed knowledge discovery process, structure analysis induces a Bayesian network from the data. The learning approach is based on the works of Lam [12] and Lam and Bacchus [13] to evaluate a Bayesian network by applying the Minimum Description Length (MDL) principle, which minimizes error terms while improving the accuracy of the model. EP is employed to optimize this metric in order to search for the best network structure.

The MDL Metric

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 $N = \{N_1, ...N_n\}$ denotes the set of nodes in the network (and thus the set of variables, since each node represents a variable), and Đ_{N_i} denotes the set of parents of node N_i. The total description length of a network is the sum of description lengths of each node:

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

This length 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 Đ_{N_i} 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. This length measures the simplicity of the network.

The formula for the data description length is:

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

As for the description length for encoding the data, a Huffman code is used to encode the data using the probability measure defined by the network. This length measures the accuracy of the network.

Combining MDL and EP

As suggested by Lam et al. [14] and Wong, Lam and Leung [25], we combine the MDL metric and EP for Bayesian network learning. Each individual represents a network structure, which is a directed acyclic graph (DAG). A set of individuals is randomly created to make up the initial population. Each graph is evaluated by the MDL metric described above. Then, each individual produces an offspring by performing a number of mutations. The offspring is also evaluated by the MDL metric. The next generation of population is selected among the parents and their offspring by tournaments. Each DAG B is compared with q 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 is output as the result.

Genetic operators

Mutation, the only genetic operator used in EP, is an asexual operation. An offspring in EP is produced by 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 the DAG. If a cyclic graph is formed after the mutation, edges in the cycles are removed to keep it acyclic. Our approach uses four mutation operators, with the same probabilities of being used:

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

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 in an "invitation to apply" direct mail promotion program 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 credit card promotions from a recent twelve-month period. 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 in the training set and testing set for the logistic regression model. Since the Bayesian network also calculate the distribution of probabilities, the same concerns are also relevant. Thus, Bayesian network learning uses the same samples so that the results can be compared with those of logistic regression.

The second dataset comes from a U.S. based catalog direct marketing company. The particular database stores records of 106,284 consumers' purchase information from 12 catalog promotions over a twelve year period, including demographic information appended from the 1995 Census data and credit information from a commercial vendor. Each case contains over 300 variables. In this study, we focus on a specific catalog promotion with a 5.4% response rate. To facilitate the data mining process as well as model evaluation and comparison, the research team includes a marketing domain expert and a data mining expert.

RESULTS

For both datasets, we split the sample into two sets, a training set and a testing set. For the first data set on credit card promotion, we developed a logistic regression using forward selection with the training set and validated with the testing set. Total 12 variables, considered important for mail operations by the bank's research department, were selected for model building, including response to the promotion, household income, marital status, number of people, number of children, owner occupied housing, number of vehicles, vehicle value, number of bank cards, number of direct marketing mails received, and number of pre-screened offers received in the last twelve months.

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 is not significantly different the one that is observed. Thus, the logistic regression model has a good fit of the data. Then, we generated the empirical results -- decile analysis of cumulative lift -- a standard measure by the direct marketing industry (Table 1). 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. The logistic regression model is used as the baseline model for comparison with the Bayesian network models. 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.

Table 1. 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				

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 so that the results could be compared with those of the logistic regression model (Table 2). 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) that appeared in the logistic regression, again suggesting instability in the model (Table 2).

Furthermore, we generated the DAG for the Bayesian network learning using all 12 variables. 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, dwelling size and marital status are directly related. The number of children and the number of adults are also related, which in turn determine the number of people in the household. In the logistic regression, they would simply be treated as separate endogenous variables.

For the catalog promotion data set, 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: cash payment, total promotion orders, frequency of purchase in the last 36 months, money used in the last 36 months, use of house credit card, lifetime number of orders, average order size, telephone order, and recency (number of months since the last order). The logistic regression model has cumulative lifts of 350 and 259 in the top two deciles, which are not exceptionally high given a 5.4% response rate. The results show a gradual decline of lifts from the top deciles to the lower deciles (Table 3).

Table 2. 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 3. 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 4. 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				

The same training and testing datasets were also used for Bayesian network learning. The results in Table 4 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. 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. Overall, the Bayesian network model performs significantly better than the logistic regression model in terms of predicting consumer response to direct mail promotions.

To make a further comparison concerning the robustness of the response models using these two methods, we have employed a 10-fold cross-validation for performance estimation. From the experimental results, the Bayesian network model predicts more accurately than the logistic regression model. Moreover, it provides higher cumulative lifts in the first few deciles.

DISCUSSION

Conclusions

Logistic regression has been widely adopted by researchers in direct marketing to select potential respondents. Most direct mail promotions only target the top two deciles. 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 can potentially help improve sales and profitability of direct marketing programs. Although the results of the Bayesian network method fall slightly short of the logistic regression with a small dataset, the Bayesian network approach generates superior results with a larger sample, suggesting that the Bayesian network model furnishes a significant better representation of the structure of data. Meanwhile, the proposed data mining methods also have several pending problems. First, the Bayesian network approach with evolutionary programming appears to be sensitive to sample size. With a small sample size, evolutionary programming may not have ample opportunities to learn the structure of data in order to extract more accurate representations. Secondly, results generated by Bayesian networks may be difficult to interpret and need the input from the domain expert to evaluate the validity of the discovered knowledge. Despite these problems, our study shows that the Bayesian network approach with evolutionary programming can potentially become a powerful and efficient data mining tool for marketing professionals.

Implications

The explosive growth of data is one of the most significant challenges facing marketing managers in the information age. The methods proposed in this study, i.e., Bayesian network models and evolutionary programming, provide efficient tools for marketing managers to mine useful knowledge from data warehouses to assist their decision making. The proposed methods have two significant advantages. First, Bayesian network models can offer superior representation of the structure of data over the traditional methods such as logistic regression. The Bayesian network method is flexible, assumption free, and more importantly, it considers the interrelationships among various factors. Secondly, given the large amount of data, evolutionary programming presents a robust and efficient tool to search and discover the best possible Bayesian network model. In essence, the combination of Bayesian network models and evolutionary programming lends a more powerful tool for data mining than if either method is applied alone.

In light of explosive growth of data, marketing researchers and database experts have devised various methods of data mining to discover new knowledge to assist management decision making. The conventional method in marketing research, like many social sciences studies, is often theory driven in that the researcher tests the hypotheses about the relationships among the interested variables. The current environment demands more problem-oriented research and efficient methods to explore the vast quantities of disaggregated data [22]. The explosive growth of marketing data requires efficient data mining tools in order to help managers uncover useful knowledge for decision making and improve sales and profitability.

Suggestions for Future Research

Wider applications of Bayesian networks and evolutionary programming to direct marketing response modeling face several significant challenges. First, EP procedures are computationally demanding and perform more slowly than mathematical optimization techniques. 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. More research is needed to improve the computing efficiency of the evolutionary algorithms so that computing time can be dramatically reduced. Secondly, a more efficient method is needed to automate or semi-automate the process of selecting meaningful variables for subsequent analyses and model building. Although researchers can always 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. Thirdly, in comparison 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. Sample size and proportion of buyers in the sample affect the performance of the method as they do with regression analysis. Finally, while evolutionary programming is a powerful tool for searching and optimizing decision problems, such methods need to be made user-friendlier to marketing researchers and more flexible to handle a greater variety of variables and marketing problems.

REFERENCES

[1] Balakrishnan, P.V. & Jacob, V.S. "Genetic algorithms for product design," *Management Science*, 1996, 42(8), 1105-1118.

[2] Bhattacharyva, S. "Evolutionary algorithm in data mining: multi-objective performance modeling for direct marketing," In *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2000, 465-473.

[3] Cooper, G. & Herskovits, E.A. "A Bayesian Method for the Induction of Probabilistic Networks from Data," *Machine Learning*, 1992, 9, 309-347.

[4] D'Ambrosio, B. "Inference in Bayesian Networks," *AI Magazine*, 1999, 20(2), 21-36.

[5] Fogel, D.B. "An introduction to simulated evolutionary optimization." *IEEE Transactions on Neural Network*, 1994, 5(1), 3-14.

[6] Geiger, D. & Heckerman, D. "Knowledge representation and inference in similarity networks and Bayesian multinets," *Artificial Intelligence*, 1996, 82(1-2), 45-74.

[7] Haddawy, P. "An overview of some recent developments in Bayesian problem-solving techniques," *AI Magazine*, 1999, 20(2), 11-19.

[8] Heckerman, D. & Wellman, M.P. "Bayesian Networks," *Communications of the ACM*, 1995, 38(3), 27-30.

[9] Hurley, S., Moutinho, L. & Stephens, N.M. "Solving marketing optimization problems using genetic algorithms," *European Journal of Marketing*, 1995, 29(4), 39-56.

[10] Jensen, F.V. *An Introduction to Bayesian Networks*, UCL Press, 1996.

[11] Klemz, B.R. "Using genetic algorithms to assess the impact of pricing activity timing," *Omega*, 1999, 27(3), 363-372.

[12] Lam, W. "Bayesian network refinement via machine learning approach," *IEEE Transactions on Pattern anal Machine Intelligence*, 1998, 20(3), 240-252 .

[13] Lam, W. & Bacchus, F. (1994), Learning Bayesian Belief Networks -- an approach based on the MDL principle," *Computational Intelligence*, 1994, 10(3), 269-293.

[14] Lam, W., Wong, M.L., Leung, M.S. & Ngan, P.S., "Discovering probabilistic knowledge from databases using evolutionary computation and Minimum Description Length principle," In *Genetic Programming: Proceedings of the Third Annual Conference*, 1998, 786-794.

[15] Larranaga, P., Poza, M., Yurramendi, Y., Murga, R. & Kuijpers, C. (1996), "Structure learning of Bayesian Network by Genetic Algorithms: A performance analysis of control parameters," *IEEE Transactions on Pattern anal Machine Intelligence*, 1996, 18(9), 912-926.

[16] Levin, N. & Zahavi, J. "Predictive modeling using segmentation," *Journal of Interactive Marketing*, 2001, 15(2), 2-22.

[17] Midgley, D.F., Marks, R.E. & Cooper, L.G., "Breeding competitive strategies," *Management Science*, 1997, 43(3), 257-275.

[18] Ngan, P.S., Wong, M.L., Lam, W., Leung, K.S. & Cheng, J.C.Y. (1999), "Medical data mining using evolutionary computation," *Artificial Intelligence in Medicine*, 16(1), 73-96.

[19] Peacock, P.R. "Data mining in marketing: Part 1," *Marketing Management*, 1998, 6(4), 8-18.

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

[21] Rebane, G. & Pearl, J. "The recovery of causal poly-trees from statistical data," In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, 1987, 222-228.

[22] Silk, A.J. "Marketing science in a changing environment," *Journal of Marketing Research*, 1993, 30(4), 401-404.

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

[24] Urban, T. L. "An inventory-theoretic approach to product assortment and shelf-space allocation," *Journal of Retailing*, 1998, 74(1), 15-35.

[25] Wong, M.L., Lam, W. & Leung, K.S. "Using Evolutionary Computation and Minimum Description Length principle for data mining of probabilistic knowledge," *IEEE Transactions: Pattern, Analysis, and Machine Intelligence*, 1999, 21(2), 174-178.

[26] Wong, M.L., Lam, W., Leung, K.S., Ngan, P.S. & Cheng, J.C.Y. "Discovering knowledge from medical databases using evolutionary algorithms," *IEEE Engineering in Medicine and Biology*, 2000, 19(4), 45-55.