An Intelligent Recommendation System for Customer Targeting

YongSeog Kim and W. Nick Street
Management Sciences Department University of Iowa
Iowa City, IA 52242 USA
{yong-s-kim,nick-street}@uiowa.edu

Abstract

One of the key problems in database marketing is the identification and profiling of households who are most likely to be interested in a particular product or service. In this paper, we propose a new recommendation system for database marketing that uses artificial neural networks (ANNs) guided by simple genetic algorithms (GAs) to target households. We show that our system not only maximizes the hit rate at fixed target point but also selects a "best" target point where expected profit from direct mailing is maximized. Further our system produces models that are easier to interpret by using a smaller number of predictive features.

1 Introduction

The ultimate goal of decision analysis is to provide managers information that is useful for understanding various managerial aspects of a problem and to choose a best solution among many alternatives. In this paper, we focus on a very specific decision analysis paradigm on behalf of market managers who want to develop and implement efficient marketing programs by fully utilizing a customer database. This is important because, due to the growing interest in micro marketing, many firms devote considerable resources to identifying households that may be open to targeted marketing messages. This becomes more critical through the easy availability of data warehouses combining demographic, psychographic and behavioral information.

Both the marketing (Bult and Wansbeek, 1995; Gönül and Shi, 1998; Piersma and Jonker, 2000) and data mining communities (Bhattacharyya, 1998; Piatetsky-Shapiro and Masand, 1999; Kim et al., 2001; Domingos and Richardson, 2001) have presented various database-based approaches for direct marketing. Traditionally, the optimal selection of mailing targets has been considered one of the most important factors for direct marketing to be successful. Thus many models aim to identify as many customers as possible who will respond to a specific solicitation campaign letter, based on the customer's estimated probability of responding to marketing program.

Another important but often neglected functionality of the models for customer targeting is the decision support functionality that helps market managers make strategic marketing plans. For example, market managers want to know how many customers should be targeted to maximize the expected net profit or increase market share while at least recovering the operational costs of a specific campaign. In order to attain this goal, market managers need a sensitivity analysis that shows how the value of the objective function (e.g., the expected net profit from the campaign) changes as campaign parameters vary (e.g., the campaign scope measured by the number of customers targeted).

This problem becomes even more complicated when the interpretability of the model is important. For example, in database marketing applications, it is often important for managers to understand the key drivers of consumer response. A predictive model that is

essentially a "black box" is not useful for developing comprehensive marketing strategies.

In this paper, we propose a new approach to building predictive models that satisfies these requirments efficiently and effectively. First, we show how to build predictive models that help market managers identify prospective households. We also demonstrate that our approach can be used to determine the scope of marketing campaign given marginal revenue per customer and marginal cost per campaign mail. At the same time, we enhance the interpretability of our model by reducing the dimensionality of data sets.

Data reduction is performed via feature selection in our approach. Feature selection is defined as the process of choosing a subset of the original predictive variables by eliminating features that are either redundant or possess little predictive information. If we extract as much information as possible from a given data set while using the smallest number of features, we can not only save a great amount of computing time and cost, but also build a model that generalizes better to households not in the test mailing. Feature selection can also significantly improve the comprehensibility of the resulting classifier models. Even a complicated model - such as a neural network - can be more easily understood if constructed from only a few variables.

Our methodology exploits the desirable characteristics of genetic algorithms (GAs) and artificial neural networks (ANNs) to achieve two principal goals of household targeting at a specific target point: model interpretability and predictive accuracy. A standard GA is used to search through the possible combinations of features. The input features selected by GA are used to train ANNs. The trained ANN is tested on an evaluation set, and a proposed model is evaluated in terms of two quality measurements – cumulative hit rate (which is maximized) and complexity (which is minimized). We define the cumulative hit rate as the ratio of the number of actual customers identified out of the total number of actual customers in a data set. This process is repeated many times as the algorithm searches for a desirable balance between predictive accuracy and model complexity. The result is a highly accurate predictive model that uses only a subset of the original features, thus simplifying the model and reducing the risk of overfitting. It also provides useful information on reducing future data collection costs.

In order to help market managers determine the campaign scope, we run the GA/ANN

model repeatedly over different target points to obtain *local* solutions. A local solution is a predictive feature subset with the highest fitness value at a specific target point. At a target point i where $0 \leq i \leq 100$, our GA/ANN model is searching for a model that is optimal when the best $i\%$ of customers in a new data set is targeted based on the estimated probability of responding to marketing campaign. Once we obtain local solutions, we combine them into an *Ensemble*, a global solution that is used to choose the best target point. Note that our Ensemble model is different from popular ensemble algorithms such as Bagging (Breiman, 1996) and Boosting (Freund and Schapire, 1996) that combine the predictions of multiple models by voting. Each local solution in our Ensemble model scores and selects prospects at a specific target point independently of other local solutions. Finally, in order to present the performance of local solutions and an Ensemble, we use a lift curve that shows the relationship between target points and corresponding cumulative hit rate.

This paper is organized as follows. In Section 2, we explain GA for feature selection in detail. In Section 3, we describe the structure of the GA/ANN model, and review the feature subset selection procedure. In Section 4, we present experimental results of both the GA/ANN model and a single ANN with the complete set of features. In particular, a global solution is constructed by incorporating the local solutions obtained over various target points. We show that such a model can be used to help market managers determine the best target point where the expected profit is maximized. In Section 5, we review related work for direct marketing from both the marketing and data mining communities. Section 6 concludes the paper and provides suggestions for future research directions.

2 Genetic Algorithms for Feature Selection

A genetic algorithm (GA) is a parallel search procedure that simulates the evolutionary process by applying genetic operators. We provide a simple introduction to a standard GA in this section. More extensive discussions on GAs can be found in (Goldberg, 1989).

Since (Goldberg, 1989), various types of GAs have been used for many different application. However, many variants still share common characteristics. Typically, a GA starts with and maintains a population of chromosomes that correspond to solutions to the problem. A

chromosome can be represented by a number of different schema including but not limited to strings of bits or vectors of real numbers with fixed or variable length. A chromosome is typically represented by a fixed-size string and consists of a number of genes called alleles. In our approach for feature selection, the representation of an agent consists of D bits, where D is the number of input features, with each of the bits indicating whether the corresponding feature is selected or not (1 if a feature is selected, 0 otherwise).

Since a GA cannot typically search the whole space of chromosomes, it gradually limits its focus to more highly fit regions of the search space. One or more fitness values must be assigned to each string in order to guide a GA toward more promising regions. By default, standard GAs are designed for maximization problems with only one objective. Each fitness value can be determined by a fitness function that we want to optimize. In our work, the fitness value for a string is determined by a neural network. Using information from households with an observed response, the ANN is able to learn the typical buying patterns of customers in the dataset. The trained ANN is tested on an evaluation set, and the proposed model is evaluated both on the hit rate (which is maximized) and the complexity (number of features, which is minimized) of the solution. Finally, these two quality measurements are combined into one and used to evaluate the quality of each string.

Note that combining multiple objectives into one is not recommended in general. However, the main goal of this paper is to propose an intelligent system for customer targeting that can support strategic marketing plans. In order to show the feasibility of such a model, we keep our model as simple as possible by using a standard GA that can consider only one objective. Further, combining multiple objectives is one of the most common practices (Yang and Honavar, 1998) and our framework can be easily modified to consider multiple objectives separately. For readers who are interested in a customer targeting problem using GAs that consider multiple objectives, we refer to previous work (Kim et al., 2001; Bhattacharyya, 2000).

Once fitness values are assigned to each chromosome in the current population, the GA proceeds to the next generation through three genetic operators: reproduction, crossover, and mutation. We describe three generic operators as follows:

Reproduction: This operator determines which strings in the current population survive

into the next generation. A fundamental rule is that a string with a higher fitness value has higher chance of surviving into the next generation. Roulette wheel selection is one frequently used method, in which each string survives in proportion to the ratio of its fitness to the overall population fitness. In our work, we use another well-known method, ranking-based reproduction. In this scheme, the strings are sorted by their fitness value, and a fixed number of the best elements from the current population are copied into the next generation.

Crossover: The crossover (or recombination) operator combines a part of one string with a part of another string and is controlled by a crossover probability Pr(*crossover*). Typically, it takes two strings called the parents as inputs, and returns two new ones, the offspring. In this way we hope to combine the good parts of one string with the good parts of another string, yielding an even better string after the operation. Many different kinds of crossover operators have been proposed including single-point, two-point, and uniform crossover (Mahfoud, 1995). Our crossover operator follows the commonality-based crossover framework assuming that commonly selected features in the parents are more likely to lead to the offspring with improved performance (Chen et al., 1999). In our implementation, it takes two agents, a parent a and a random mate, and scans through every bit of the two agents. If it locates a different bit, it flips a coin to determine the offspring's bit.

Mutation: This operator assigns a new value to a randomly chosen gene and is controlled by a mutation probability Pr(*mutation*). By introducing new genetic characteristics into the pool of chromosomes, it prevents the *gene depletion* that might lead a GA to find local optimum. The mutation operator in this paper always randomly selects one bit of each string and flips it.

3 GA/ANN Model for Customer Targeting

Our predictive model of household buying behavior is a hybrid of the GA and ANN procedures. In our approach, the GA identifies relevant consumer descriptors that are used by the

ANN to forecast consumer choice given a specific target point. Our final solution, Ensemble, consists of multiple local solutions each of which is an optimized solution at a specific target point. In this section, we present the structure of our GA/ANN model and the evaluation criteria used to select an appropriate predictive model.

3.1 Evaluation Metrics

We define two heuristic evaluation metrics, $F_{complexity}$ and $F_{accuracy}$, to evaluate selected feature subsets. These objectives are combined with equal weight to return one fitness value for candidate solutions.

$F_{complexity}$: This objective is aimed at finding parsimonious solutions by minimizing the number of selected features as follows:

$$F_{complexity} = 1 - \frac{d-1}{D-1} \tag{1}$$

where d and D represent the dimensionality of the selected feature set and of the full feature set, respectively. Note that at least one feature must be used. Other things being equal, we expect that lower complexity will lead to easier interpretability of solutions as well as better generalization.

$F_{accuracy}$: The purpose of this objective is to favor feature sets with higher discriminative power to discriminate buyers from non-buyers. In our application, $F_{accuracy}$ is same as cumulative hit rate, the ratio of the number of actual customers identified, AC, out of the total actual customers, TAC. Note that AC is dependent on how many customers are targeted. Cumulative hit rate can be represented in a mathematical form as follows:

$$F_{accuracy} = \frac{AC}{TAC}. \tag{2}$$

Often hit rate, the ratio of the number of actual customers identified out of the number of customers targeted, has been used as an alternative measurement. However, we also note that it is important to have the values of two evaluation metrics in the same range between 0 and 1. Since we can always attain this requirement by using cumulative hit

```
for each  target point i where i = 10, 20, ···, 90
   run GA/ANN, optimizing top i% of training sets
   select best_i based on the fitness value
   Ensemble(i) = best_i
endfor
```

Figure 1: Pseudo code for Ensemble construction

rate by targeting the smallest proportion of actual buyers (in case of data we used in this paper, about 6% of customers are actual buyers), we prefer cumulative hit rate to hit rate.

3.2 Algorithm outline

We show an abstract view of our algorithm in Figure 1. As a first step, our model searches for a set of local solutions optimized at a specific target point. A local solution $best_i$ is an optimal solution when the firm targets the best i% customers based on their estimated probability of responding to a solicitation letter. In this paper, we consider nine target points $(10\%, 20\%, \cdots, 90\%)$ but more refined target points can be analyzed without the need to modify the structure of our algorithm. The GA/ANN component in our algorithm is used for finding local solutions and discussed in detail in Section 3.3. Once all the local solutions are found, we combine them into the final model, Ensemble. Each local solution is a neural network model built using the feature subset specific to a target point, and our Ensemble is a collection of such models.

However, in order to estimate the performance of Ensemble on the evaluation data, we cannot simply combine the best estimates of local solutions over different target points. Rather, the estimate of Ensemble at target point i (i.e., when we target the best i% of customers) is the estimated performance of local solution $best_i$ optimized at target point i. Ideally, the performance of Ensemble should return the highest hit rate over all the target points compared to all the local solutions.

3.3 Structure of GA/ANN Model

The structure of GA/ANN model for finding local solutions is shown in Figure 2. First, the GA searches the exponential space of feature subsets and passes one subset of features to an

ANN. The ANN extracts predictive information from each subset and learns the patterns. Once an ANN learns the data patterns, the trained ANN is evaluated on a data set not used for training, and returns two evaluation metrics, $F_{accuracy}$ and $F_{complexity}$, to the GA. The two evaluation metrics are then combined with the equal weight into one fitness value. It is important to note that in both the learning and evaluation procedures, the ANN uses only the selected features. Based on the fitness value, the GA biases its search direction to

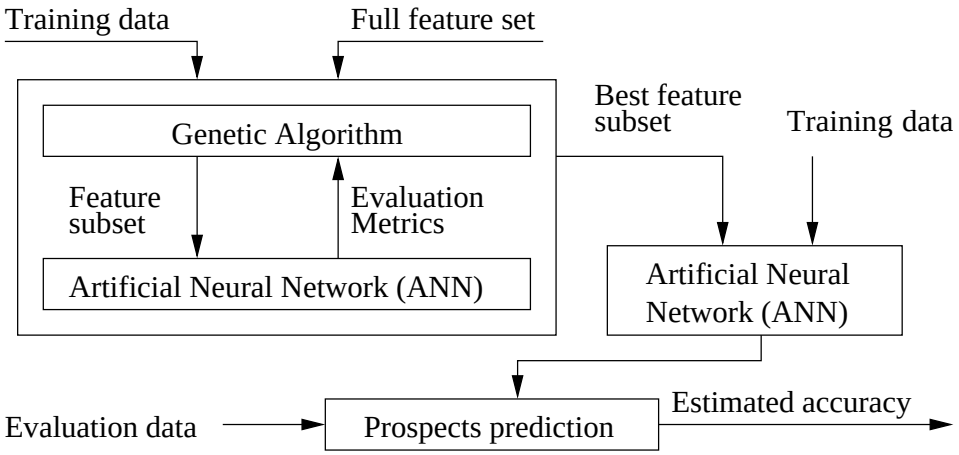

Figure 2: The structure of GA/ANN model. GA searches for a good subset of features and passes them to an ANN. The ANN calculates the "goodness" of each subset and returns two evaluation metrics to GA.

maximize the combined objective. This routine continues for a fixed number of generations. Among all the evolved models over the generations, we select the best model in terms of the fitness value. Once we choose the best model, we train the ANN using all the training points with the selected features only. The trained model is then used to rank the potential customers (the records in the evaluation set) in descending order by the probability of buying RV insurance (see Section 4, as predicted by the ANN. We finally select the top i% of the prospects in the evaluation set and evaluate the model's accuracy.

In order to provide a reliable estimate for all the candidates, we estimate their fitness with a rigorous estimation procedure, k-fold cross validation. In this procedure, the training data is divided into k non-overlapping groups. We train an ANN using the first $k-1$ groups of training data and test the trained ANN on the kth group. We repeat this procedure until each of the groups is used as a test set once. We then take the average of the performance measurements over the k folds. In our experiment, we set $k = 2$. This is a reasonable

compromise considering the computational complexity of systems like ours. Further, an estimate from 2-fold cross validation is likely to be more reliable than an estimate from a common practice using a single holdout set.

4 Application

The new GA/ANN methodology is applied to the prediction of households interested in purchasing an insurance policy for recreational vehicles. To benchmark the new procedure, we contrast the predictive performance of Ensemble to a single ANN with the complete set of features. We do not compare our approach to a standard logit regression model because a logit regression model is a special case of single ANN with one hidden node.

4.1 Data Description

The data are taken from a solicitation of 9,822 European households to buy insurance for a recreational vehicle. These data, taken from the CoIL 2000 forecasting competition (Kim and Street, 2000), provide an opportunity to assess the properties of the GA/ANN procedure in a customer prospecting application.[1] In our analysis, we use two separate datasets: a training set with 5822 households and an evaluation set with 4000 households. The training data is used to calibrate the model and to estimate the hit rate expected in the evaluation set. Of the 5822 prospects in the training dataset, 348 purchased RV insurance, resulting in a hit rate of $348/5822 = 5.97\%$. From the manager's perspective, this is the hit rate that would be obtained if solicitations were sent out randomly to consumers in the firm's database.

The evaluation data is used to validate the predictive models. Our Ensemble predictive model is designed to return the top $i\%$ of customers in the evaluation dataset judged to be most likely to buy RV insurance. The model's predictive accuracy is examined by computing the observed hit rate among the selected households. It is important to understand that only information in the training dataset is used in developing the model. Data in the evaluation

[1] We use a dataset on consumer responses to a solicitation for "caravan" insurance policies. A "caravan" is similar to a recreational vehicle in the United States. For more information about the CoIL competition and the CoIL datasets, refer to the Web site http://www.dcs.napier.ac.uk/coil/challenge/.

dataset is used exclusively for forecasting.

In addition to the observed RV insurance policy choices, each household's record also contains 93 additional variables, containing information on both socio-demographic characteristics (variables 1-51) and ownership of various types of insurance policies (variables 52-93). Details are provided in Table 1. The socio-demographic data are based upon postal code information. That is, all customers living in areas with the same postal code have the same socio-demographic attributes. The insurance firm in this study scales most socio-demographic variables on a 10-point ordinal scale (indicating the relative likelihood that the socio-demographic trait is found in a particular postal code area). This 10-point ordinal scaling includes variables denoted as "proportions" in Table 1. For the purposes of this study, all these variables were regarded as continuous. The psychographic segment assignments (attributes 4-13), however, are household-specific and are coded into ten binary variables.

In our subsequent discussion, the word feature refers to one of the 93 variables listed in Table 1. For example, the binary variable that determines whether or not a household falls into the "successful hedonist" segment is a single feature. Accordingly, in the feature selection step of the GA/ANN model, the algorithm can choose to use any possible subset of the 93 variables in developing the predictive model.

4.2 Experimental results

In our experiment, we first select local solutions over different target points and construct lift curves for each of them. A lift curve shows the percentage of all buyers identified in the group selected for a direct mail solicitation at the given target point out of all buyers in the database. We also construct the lift curve of Ensemble and compare it to those of the local models. Finally, we show how our Ensemble solution can be used to select the best target point where the expected profit is maximized under two different campaign scenarios.

We set the values for GA parameters as follows: $\Pr(mutation) = 1.0$, $\Pr(crossover) = 0.8$, $Population = 100$, and $Iteration = 200$. We use a three-layer ANN and train it with the standard backpropagation algorithm. We set the number of epochs = 10 and heuristically determine the number of hidden nodes using the formula $\min(3, \sqrt{node_{in}})$ where $node_{in}$ represents the number of input nodes.

Feature ID	Feature Description
1	Number of houses owned by residents
2	Average size of households
3	Average age of residents
4-13	Psychographic segment: successful hedonists, driven growers, average family, career loners, living well, cruising seniors, retired and religious, family with grown ups, conservative families, or farmers
14-17	Proportion of residents with Catholic, Protestant, others and no religion
18-21	Proportion of residents of married, living together, other relation, and singles
22-23	Proportion of households without children and with children
24-26	Proportion of residents with high, medium, and lower education level
27	Proportion of residents in high status
28-32	Proportion of residents who are entrepreneur, farmer, middle management, skilled laborers, and unskilled laborers
33-37	Proportion of residents in social class A, B1, B2, C, and D
38-39	Proportion of residents who rented home and owned home
40-42	Proportion of residents who have 1, 2, and no car
43-44	Proportion of residents with national and private health service
45-50	Proportion of residents whose income level is $<$ \$30,000, \$30,000-\$45,000, \$45,000-\$75,000, \$75,000-\$123,000, $>$\$123,000, and average
51	Proportion of residents in purchasing power class
52-72	Scaled contribution to various types of insurance policies such as private third party, third party firms, third party agriculture, car, van, motorcycle/scooter, truck, trailer, tractor, agricultural M/C, moped, life, private accident, family accidents, disability, fire, surfboard, boat, bicycle, property, social security
73-93	Scaled number of households holding insurance policies for the same categories as in scaled contribution attributes

Table 1: Household background characteristics

4.2.1 Analysis of lift curves

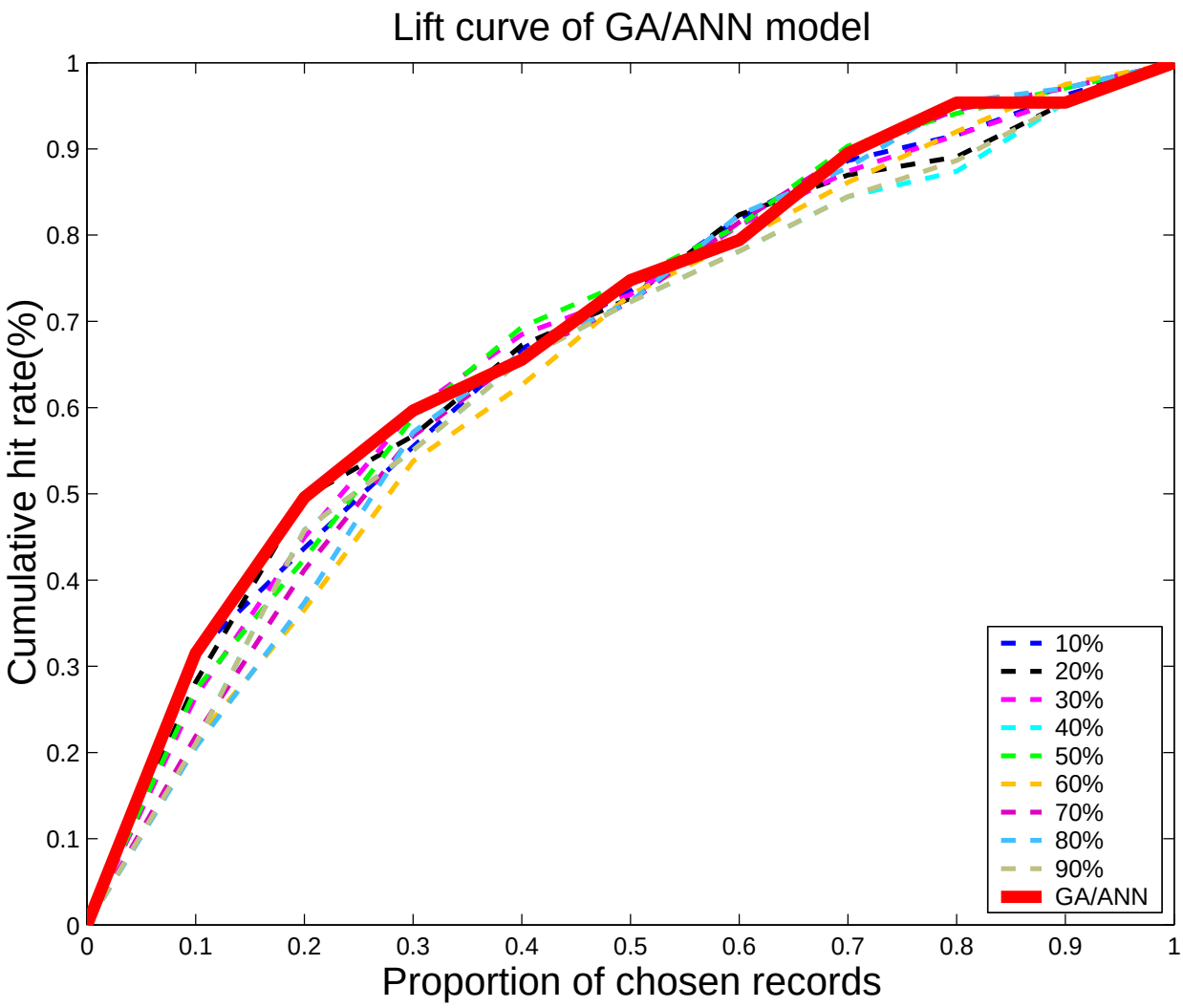

Figure 3: Lift curves of local and Ensemble solution. The lift curve of each local solution is represented as a thin dotted line. The lift curve of Ensemble is constructed by combining the lift curves of local solutions at various target points and shown as a thick solid line.

We first show the lift curves of local solutions (shown as dotted lines) and Ensemble (shown as the thick solid line) in Figure 3. The lift curve of a local solution $best_i$ is constructed as follows: We estimate the probability of buying new insurance for each prospect in the evaluation data with trained ANNs using the subset of features chosen by $best_i$. After sorting prospects in descending order of the estimated probability, we compute the values of $F_{accuracy}$ over various target points j where $j = 10, 20, \cdots, 100\%$. We define a cumulative hit rate of a model as the value of $F_{accuracy}$ at a given market point. Recall that $best_i$ is an optimal

solution at target point i. The lift curve shows the relationship between a set of market points and their corresponding cumulative hit rates.

The lift curve of Ensemble is constructed by combining a set of cumulative hit rates of $best_i$ at target point i where $i = 10, 20, \cdots, 100$.[2] Under random sampling, the lift curve is a 45-degree line starting at the origin of the graph. In an ideal case, the lift curve of Ensemble should be a convex hull of lift curves of local solutions because it is a combination of local optimal solutions. However, it is possible for GA to be stuck to local optimum. Among nine target points considered in our model, Ensemble shows the best performance at five target points (i = 10, 20, 30, 50, 80) and second best at one target point (i = 70). Compared to random sampling, our Ensemble returns 3.1 (2.5) times higher cumulative hit rate when we target the best 10% (20%) of customers.

For comparison purposes, we implemented an ANN with the complete set of input variables. We show the hit rates of Ensemble and the single ANN model on the evaluation data in Figure 4. Our Ensemble shows much better performance at target points $i = 10, 20$ but loses its advantages over middle target points $i = 30, \cdots, 60$. However, our model regains its superior performance at higher target points $i = 70, \cdots, 90$. We partially attribute to oversearching (Quinlan and Cameron-Jones, 1995; Murthy and Salzberg, 1995) the lower performance of our model over the middle target points. A wrapper model like ours can find *fluke* rules that fit the training data well but have low predictive accuracy on new data by overusing the estimate of accuracy of models. A recent dicussion on oversearching caused by multiple comparison procedures can be found in (Jensen and Cohen, 2000).

Though it is worthwhile to determine why the performance of Ensemble changes over different target points, we will leave this issue to future research in order to be consistent with the main purpose of this paper: proposing an intelligent recommendation system for customer targeting. Further, the single ANN requires all the input variables and provides more of a *black box* solution that market managers cannot utilize for making strategic marketing plans.

We also show the index of chosen features by local solutions in Table 2. Each local solution clearly highlights predictive features at the given target point. Note that one feature (55,

[2]In (Provost and Fawcett, 1998; Coetzee et al., 2001), a convex hull of multiple receiver operating characteristic (ROC) curves is taken as a final model to provide a more robust and accurate model.

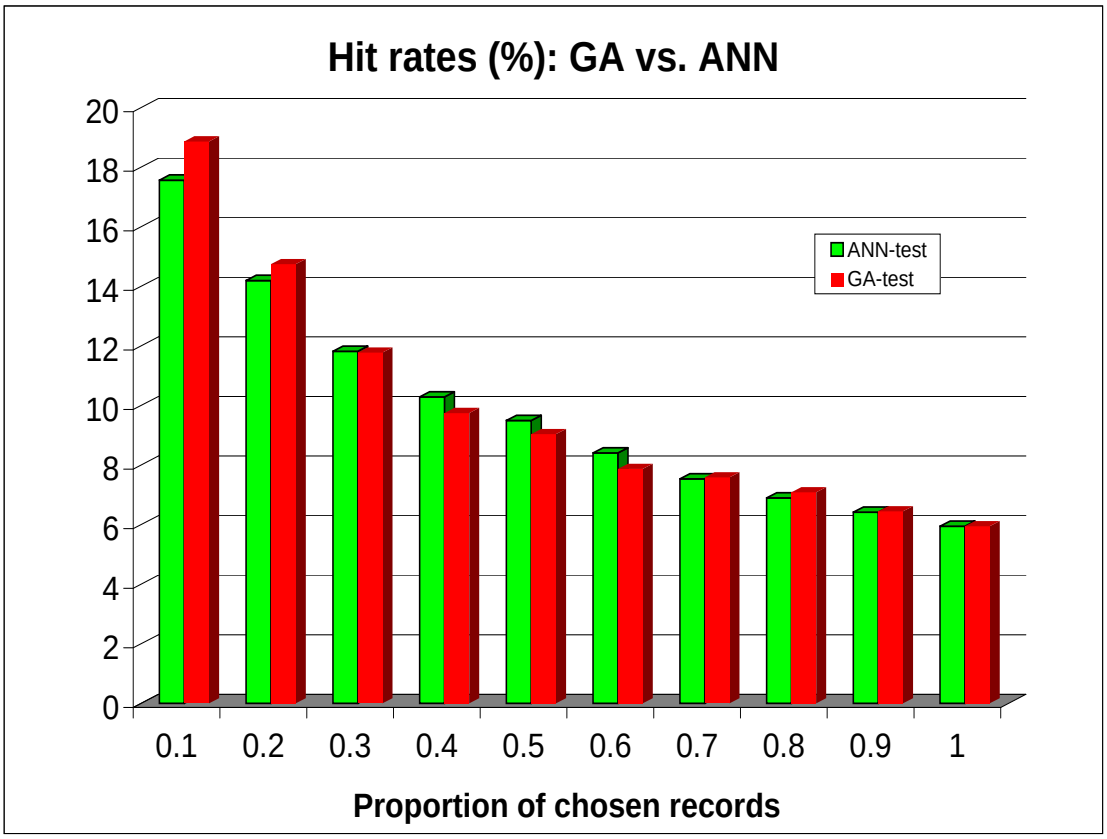

Figure 4: Hit rates of GA/ANN and single ANN model on the evaluation data

scaled contribution to car policy) is chosen by all local solutions. This makes considerable sense, given the fact that the firm is soliciting households to buy insurance for recreational vehicles. Other insurance related features (scaled contribution to and ownership of boat (69, 90) and fire policy (67, 88)) are also chosen by multiple solutions. Demographic variables such as income levels (45, 50), home ownership (39), and jobs (farmer, 29) are also selected by a few solutions. In general, the results are in line with marketing science work on customer segmentation, which shows that information about current purchase behavior is most predictive of future choices (Rossi et al., 1996). We refer to our previous work (Kim and Street, 2000) for readers who are interested in building a potentially valuable profile of likely

customers.

Local solution	Feature index	Local solution	Feature index
$best_{10}$	26, 39, 52, 55, 67, 90	$best_{50}$	29, 39, 55, 80, 90, others
$best_{20}$	54, 55, 67, 69, 90	$best_{60}$	8, 30, 55
$best_{30}$	40, 50, 55, 69, 88	$best_{70}$	29, 45, 55
$best_{40}$, $best_{90}$	55, 88	$best_{80}$	8, 45, 55

Table 2: Chosen features by local solutions

In terms of data reduction, all local solutions choose at most six features except one that chooses 26 features at target point $i = 50$. On average, local solutions choose 6 features, which reduces the data dimensionality by $(93 - 6)/93 \approx 93.5\%$. This implies that the firm could reduce data collection and storage costs considerably. This is possible through reduced storage requirements, and the reduced labor and data transmission costs.

4.2.2 Sensitivity analysis

In this section, we focus on the decision support functionality of Ensemble solution. In particular, we want to show how our Ensemble solution can be used to select the best target point where the expected profit is maximized. We consider two different marketing strategies: one case targeting customers with a nice-looking brochure of higher cost per mailing ($7) and the other case using a plain letter of lower cost per mailing ($0.71).[3] However, we make two common assumptions for both cases that the marginal revenue per buyer is $70 and the firm has a list of one million prospects to target. We show our experimental results in Figure 5.

In order to choose the best target point, we compute the expected profits over different target points using the estimated hit rates of Ensemble on the training data. When the cost per mailing is not expensive ($0.71), our Ensemble returns the maximized expected profit when the best 80% of customers in the train set are targeted. This makes sense because it is wise to target as many customers as possible up to a certain point. Once we determine the target point i, we can compute the expected profits of three different models when we

[3]Note that it is not our intention to relate brochure quality to campaign outcomes.

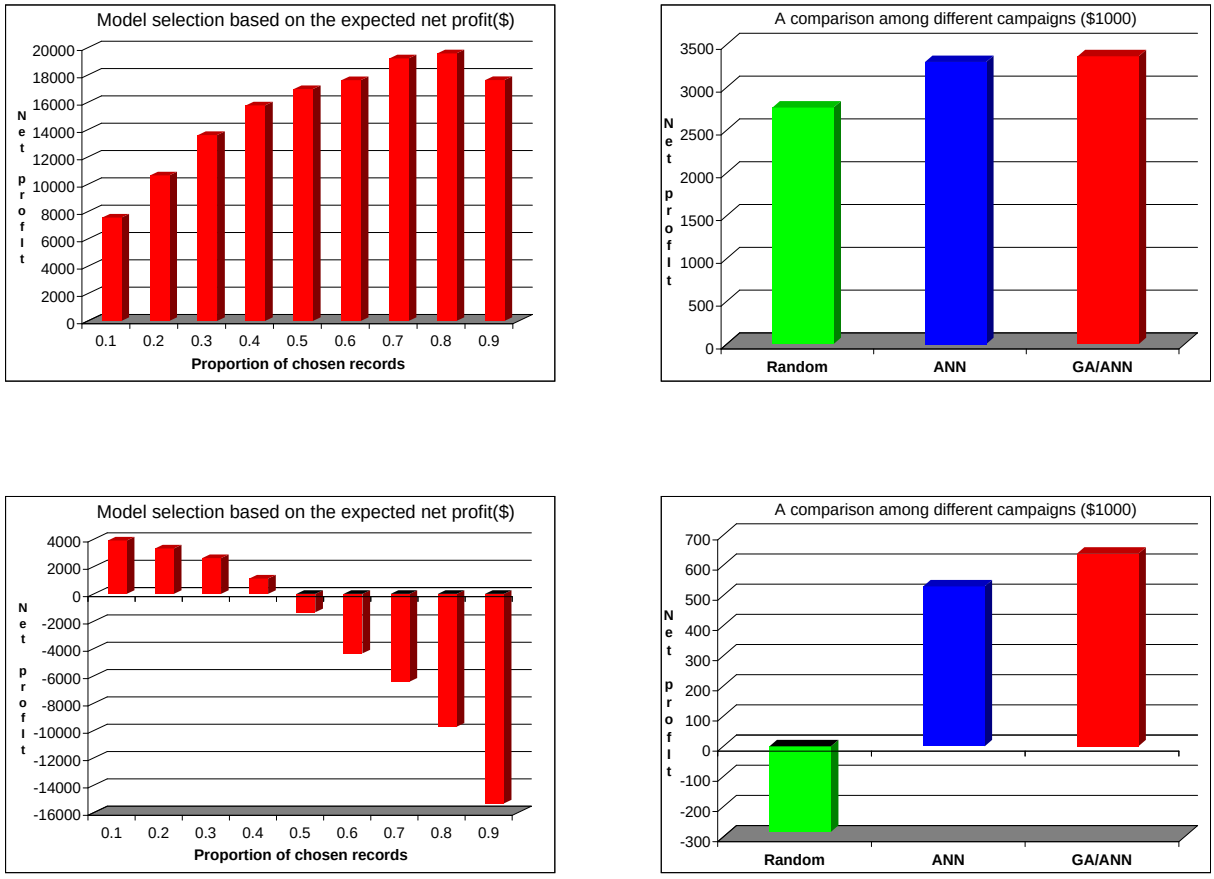

Figure 5: The selection of the best target point and the estimation of expected profit of market campaign using Ensemble. Two figures at the top panel show results when the cost per mailing is \$0.71. The other two figures at the bottom are results when the cost per mailing is \$7.

target the top i% of one million prospects. Though all models return positive profits, the Ensemble solution returns the maximized profit again.

When the cost per mailing is expensive (\$7), the best target point is $i = 10$ in terms of maximized profit. This time, random targeting returns a negative profit as we target the best 10% of customers because of the increased mailing cost. The Ensemble solution returns the highest expected profit followed by the single ANN solution. Note that a market manager can select the best target point at $i = 40$ when her main interest is not in maximizing the expected profit but in increasing the market share while recovering at least campaign costs. This is particularly useful when marginal revenue per customer and/or mailing cost

information is not known as in our case.

5 Related work

Various multivariate statistical and analytical techniques in marketing community have been applied to the database marketing problem. Routine mailings to existing customers are typically based upon the RFM (recency, frequency, monetary) approach that targets households using knowledge of the customer's purchase history (Schmid and Weber, 1998). However, the RFM model has major disadvantages including its limited applicability to current customers only and redundancy because of inter-dependency among RFM variables. Mailings to households with no prior relationship with the firm are based upon the analysis of the relationship between demographics and the response to a test mailing of a representative household sample. A brief summary of other traditional models such as Chi-square Automatic Interaction Detection (CH-AID), multiple regression methods, discriminant analysis, and gains chart analysis can be found in (Green and Green, 1988; Bult and Wansbeek, 1995).

A more elaborate model for identifying optimal targets was presented by Bult and Wansbeek (Bult and Wansbeek, 1995). Their profit maximization (PM) approach is mainly based on the gains chart model (Banslaben, 1992) in a single period campaign. In their model, only those prospects whose expected return is higher than the marginal cost of the mailing are selected for direct mailing. From a comparative experiment on a real data set, they reported improved performance compared to CH-AID and the gains chart model. Bitran and Mondschein (Bitran and Mondschein, 1996) presented a stochastic model to determine the optimal mailing frequency and assumed that multiple mailings would result in only one response.

Gönül and Shi (Gönül and Shi, 1998) presents a Markov decision model for the optimal mailing policy over an infinite period. In their model, two different objectives from both customers (maximizing utility) and direct mailers (maximizing profit) are optimized simultaneously. Their model utilized the estimated impact of recency and frequency on the customer's response probability, and showed significantly improved performance over a single-period model. However, their recency- and frequency-based model cannot be ap-

plied to a certain types of commodities (e.g. seasonal goods). Further their model can be computationally expensive once monetary value is added in the state definition.

Piersma and Jonker (Piersma and Jonker, 2000) studied a problem for optimizing the frequency for the direct mailing over a long-term but finite campaign horizon. Their model is simpler than (Gönül and Shi, 1998) because of finite campaign horizon considered, but more general than (Bitran and Mondschein, 1996) in the sense that they allow multiple responses to multiple mailings. Through comparative experiments on a data set from Dutch charitable organization, they showed that their model significantly increased the customer profitability and reduced wasteful mailings. Other models for direct marketing include a split-hazard model for estimating a physician's propensity of using new drug in (Kamakura and Kossar, 1998) and a latent trait and a latent class model for cross-selling of financial services in (Kamakura et al., 1991) and (Kamakura et al., 2000).

Researchers from machine learning and data mining community have developed models without making prior assumptions about data distribution. Problems for effectively profiling users have been studied in (Fawcett and Provost, 1996; Chan and Stolfo, 1998; Raghavan et al., 2000). Fawcett and Provost (Fawcett and Provost, 1996) presented a model for detecting fraudulent usage of cellular calls. Chan and Stolfo (Chan and Stolfo, 1998) presented a cost-sensitive model for detecting fraudulent usage of credit cards. Noting that the original distribution can be different from the desired distribution for optimal training, they divided a given data set into subsets with the appropriate class distribution through preliminary experiments. They obtained their final model by combining multiple classifiers trained on different subsets and reported some success. In (Raghavan et al., 2000), a predictive profiling model was presented for profiling customers of an internet service provider who are most likely to stop using internet service. In (Ling and Li, 1998), difficulties encountered in the process of applying data mining techniques to direct marketing were discussed.

A more relevant response model for direct mail campaigns can be found in (Bhattacharyya, 1998). In (Bhattacharyya, 1998), Bhattacharyya proposed a GA-based approach for developing optimal models at different target points. In his framework, each candidate solution was expressed as a linear combination of the input variables and was evaluated in terms of two evaluation criteria – response rate and robustness. He further analyzed trade-

offs of multiple objectives by varying weights assigned to multiple criteria at different target points. He reported significantly improved performance over the traditional logit regression model at the first few target points. Recent works (Bhattacharyya, 2000; Kim et al., 2001) studied the same problem in a *Pareto optimization* framework, where multiple criteria are not combined but considered independently in order to avoid a subjective weighting scheme.

In (Piatetsky-Shapiro and Masand, 1999) the profitability condition of a campaign was explicitly formulated as a function of the lift of the model, uniform campaign cost per mailing, and marginal revenue per identified positive record. A brief review of evaluation metrics for marketing campaigns can be found in (Rosset et al., 2001). They also provided heuristics to estimate the expected profit from targeting a subset of records without going through a lengthy data mining process. Chou et al. (Chou et al., 2000) devised an effective model for identifying prospective insurance buyers when buyer versus non-buyer information is not available. Gersten et al. (Gersten et al., 2000) presented a model to select prospects in the automotive industry where the buying decision takes a long time.

Domingos and Richardson (Domingos and Richardson, 2001) view a market as a social network where each customer has a different *network value*, influence on other customers' probability of buying a product. The network value of a customer in their model is computed as the expected profit from additional sales to customers whom she recursively influences to buy. Considering network effects can change optimal marketing strategy dramatically. For example, it is worth marketing to one whose intrinsic value is lower than the cost of marketing when her influence on others' purchasing decision is strong. From several experiments on a publicly available data (www.research.compaq.com/src/eachmovie/), they not only reported superior performance to traditional direct marketing strategy but also profiled good customers to target.

6 Conclusion

In this paper, we presented a novel approach for customer targeting in database marketing. We used a genetic algorithm to search for possible combinations of features and an artificial neural network to score customers. One of the clear strengths of the GA/ANN approach is

its ability to construct predictive models that reflect the direct marketer's decision process. In particular, with information of campaign costs and profit per additional actual customer, we show that our system not only maximizes the hit rate at fixed target point but also selects a "best" target point where expected profit from direct mailing is maximized. Further, our system is easier to interpret by using smaller number of features.

In future work we will look into a customer targeting model that assumes the heterogeneous structure of marginal revenue per each prospect. The data set we analyzed in this paper does not include critical information such as monetary values that each customer spent to purchase a caravan insurance policy. It is also reasonable to assume that there is relatively small difference in terms of monetary values in insurance options available to customers. However, assume the case that a non-profit organization sends solicitation letters to possible donors for charity. For this organization, maximizing the total amount of donated money is more important than identifying as many donors as possible. This is because, in an extreme case, single donor can donate more money than all the other donors. In this case, value-based customer targeting becomes critical. By considering raised monetary value by targeted customers as one objective, our GA/ANN model will be able to find an optimal solution with maximized monetary value.

Related with this direction of research, it is interesting to see if a bi-level model that has two separate procedures, one for estimating donation probability and the other for estimating donation amounts, can do better. It was claimed that any single classifier model that learns two parameters is likely to make more errors in learning decision rules than a bi-level model (Zadrozny and Elkan, 2001). However, their claim was not supported by a statistical significance test, which warrants further investigation.

Another research direction is to investigate whether or not the chosen feature subsets are related with target points. Our experimental results in Section 4.2.1 show that different feature subsets are chosen at different target points except for a few common features. We expected a good predictive subset of features to appear in most of the local solutions. We suspect that a certain subset of features can discriminate well buyers from non-buyers at a target point, but not as well as other features at different points. It could also happen because of strong correlation among insurance-related features. However, it warrants further

investigation to support our speculation.

Acknowledgments

The authors wish to thank Peter van der Putten and Maarten van Someren for making the CoIL data available for this paper. This work is partially supported by NSF grant IIS-99-96044.

References

Banslaben, J. (1992). Predictive modeling. In Nash, E. L., editor, *The Direct Marketing Handbook*. McGraw-Hill, New York, NY.

Bhattacharyya, S. (1998). Direct marketing response models using genetic algorithms. In *Proc. of 4th Int'l Conf. on Knowledge Discovery & Data Mining (KDD-98)*, pages 144–148.

Bhattacharyya, S. (2000). Evolutionary algorithms in data mining: Multi-objective performance modeling for direct marketing. In *Proc. of 6th ACM SIGKDD Int'l Conf. on Knowledge Discovery & Data Mining (KDD-00)*, pages 465–473.

Bitran, G. R. and Mondschein, S. V. (1996). Mailing decisions in the catalog sales industry. *Management Science*, 42:1364–1381.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24(2):123–140.

Bult, J. R. and Wansbeek, T. (1995). Optimal selection for direct mail. *Marketing Science*, 14(4):378–394.

Chan, P. K. and Stolfo, S. J. (1998). Toward scalable learning with non-uniform class and cost distributions: A case study in credit card fraud detection. In *Proc. of 4th ACM SIGKDD Intl. Conf. on Knowledge Discovery & Data Mining (KDD-98)*, pages 164–168.

Chen, S., Guerra-Salcedo, C., and Smith, S. (1999). Non-standard crossover for a standard representation – commonality-based feature subset selection. In *GECCO-99: Proc. of the Genetic and Evolutionary Computation Conference*, pages 129–134. Morgan Kaufmann.

Chou, P. B., Grossman, E., Gunopulos, D., and Kamesam, P. (2000). Identifying prospective customers. In *Proc. of 6th Int'l Conf. on Knowledge Discovery & Data Mining*, pages 447–456.

Coetzee, F., Glover, E., Lawrence, S., and Giles, C. L. (2001). Feature selection in Web applications using ROC inflections. In *Symposium on Applications and the Internet, SAINT*, San Diego, CA.

Domingos, P. and Richardson, M. (2001). Mining the network value of customers. In *Proc. of 7th ACM SIGKDD Int'l Conf. on Knowledge Discovery & Data Mining (KDD-01)*, pages 57–66.

Fawcett, T. and Provost, F. (1996). Combining data mining and machine learning for effective user profiling. In *Proc. of 2nd ACM SIGKDD Int'l Conf. on Knowledge Discovery & Data Mining (KDD-96)*, pages 8–13.

Freund, Y. and Schapire, R. (1996). Experiments with a new boosting algorithm. In *Proc. of 13th Int'l Conf. on Machine Learning*, pages 148–156, Bari, Italy.

Gersten, W., Wirth, R., and Arndt, D. (2000). Predictive modeling in automotive direct marketing: Tools, experiences and open issues. In *Proc. of 6th Int'l Conf. on Knowledge Discovery & Data Mining*, pages 398–406.

Goldberg, D. E. (1989). *Genetic Algorithms in Search, Optimization and Machine Learning.* Addison-Wesley, New York, MA.

Gönül, F. and Shi, M. Z. (1998). Optimal mailing of catalogs: A new methodology using estimable structural dynamic programming models. *Management Science*, 44(9):1249–1262.

Green, P. E. and Green, D. S. (1988). *Research for Marketing Decisions*, pages 507–552. Prentice-Hall, Inc, Englewood Cliffs, New Jersey, fifth edition.

Jensen, D. D. and Cohen, P. R. (2000). Multiple comparisons in induction algorithms. *Machine Learning*, 38(3):309–338.

Kamakura, W. A., de Rosa, F., Wedel, M., and Mazzon, J. A. (2000). Cross-selling financial services with database marketing. Unpublished working paper.

Kamakura, W. A. and Kossar, B. S. (1998). A factor analytic split-hazard model for database marketing. Technical Report 98-12-009, Department of Marketing, University of Iowa, Iowa City, IA.

Kamakura, W. A., Ramaswami, S. N., and Srivastava, R. K. (1991). Applying latent trait analysis in the evaluation of prospects for cross-selling of financial services. *International Journal of Research in Marketing*, 8:329–349.

Kim, Y. and Street, W. N. (2000). CoIL challenge 2000: Choosing and explaining likely caravan insurance customers. Technical Report 2000–09, Sentient Machine Research and Leiden Institute of Advanced Computer Science. http://www.wi.leidenuniv.nl/~putten/library/cc2000/.

Kim, Y., Street, W. N., Russell, G. J., and Menczer, F. (2001). Customer targeting: A neural network approach guided by genetic algorithms. *Management Science*. Submitted.

Ling, C. X. and Li, C. (1998). Data mining for direct marketing: Problems and solutions. In *Proc. of 4th ACM SIGKDD Int'l Conf. on Knowledge Discovery & Data Mining (KDD-98)*, pages 73–79.

Mahfoud, S. M. (1995). *Niching Methods for Genetic Algorithms*. PhD thesis, Department of general engineering, University of Illinois at Urbana-Champaign, Champaign, IL.

Murthy, S. and Salzberg, S. (1995). Lookahead and pathology in decision tree induction. In Mellish, C. S., editor, *Proc. of 14th Int'l Joint Conf. on Artificial Intelligence*, pages 1025–1031. Morgan Kaufmann.

Piatetsky-Shapiro, G. and Masand, B. (1999). Estimating campaign benefits and modeling lift. In *Proc. of 5th ACM SIGKDD Int'l Conf. on Knowledge Discovery & Data Mining (KDD-99)*, pages 185–193.

Piersma, N. and Jonker, J. (2000). Determining the direct mailing frequency with dynamic stochastic programming. Technical Report EI2000-34A, Econometric Institute.

Provost, F. and Fawcett, T. (1998). Robust classification systems for imprecise environments. In *Proc. of 15th National Conf. on Artificial Intelligence (AAAI-98)*, pages 706–713.

Quinlan, J. R. and Cameron-Jones, R. M. (1995). Oversearching and layered search in empirical learning. In *Proc. of 14th Int'l Joint Conf. on Artificial Intelligence*, pages 1019–1024. Morgan Kaufmann.

Raghavan, N., Bell, R. M., and Schonlau, M. (2000). Defection detection. In *Proc. of 6th Int'l Conf. on Knowledge Discovery & Data Mining*, pages 447–456.

Rosset, S., Neumann, E., Eick, U., Vatnik, N., and Idan, I. (2001). Evaluation of prediction models for marketing campaigns. In *Proc. of 7th ACM SIGKDD Int'l Conf. on Knowledge Discovery & Data Mining (KDD-01)*, pages 456–461.

Rossi, P. E., McCulloch, R., and Allenby, G. (1996). The value of household information in target marketing. *Marketing Science*, 15(3):321–340.

Schmid, J. and Weber, A. (1998). *Desktop Database Marketing.* NTC Business Books.

Yang, J. and Honavar, V. (1998). Feature subset selection using a genetic algorithm. *IEEE Intelligent Systems and their Applciations*, 13(2):44–49.

Zadrozny, B. and Elkan, C. (2001). Learning and making decisions when costs and probabilities are both unknown. In *Proc. of 7th ACM SIGKDD Int'l Conf. on Knowledge Discovery & Data Mining (KDD-01)*, pages 204–213.