

Experimental Design for Solicitation Campaigns

Uwe F. Mayer

Fair, Isaac and Company, Inc.
5935 Cornerstone Court West
San Diego, CA 92121
1-858-799-8000

uwemayer@fairisaac.com

Armand Sarkissian

Fair, Isaac and Company, Inc.
5935 Cornerstone Court West
San Diego, CA 92121
1-858-799-8000

armandsarkissian@fairisaac.com

ABSTRACT

Data mining techniques are routinely used by fundraisers to select those prospects from a large pool of candidates who are most likely to make a financial contribution. These techniques often rely on statistical models based on trial performance data. This trial performance data is typically obtained by soliciting a smaller sample of the possible prospect pool. Collecting this trial data involves a cost; therefore the fundraiser is interested in keeping the trial size small while still collecting enough data to build a reliable statistical model that will be used to evaluate the remainder of the prospects.

We describe an experimental design approach to optimally choose the trial prospects from an existing large pool of prospects. Prospects are clustered to render the problem practically tractable. We modify the standard D-optimality algorithm to prevent repeated selection of the same prospect cluster, since each prospect can only be solicited at most once.

We assess the benefits of this approach on the KDD-98 data set by comparing the performance of the model based on the optimal trial data set with that of a model based on a randomly selected trial data set of equal size.

Categories and Subject Descriptors

G.3 [Probability and Statistics]: Experimental Design.

General Terms

Design, Experimentation, Performance.

Keywords

Experimental Design, solicitation campaign, data collection.

1. INTRODUCTION

Experimental design is a paradigm to systematically find out about the impact of individual characteristics on a system, where

the *system* could be anything ranging from a semi-conductor manufacturing machine to a solicitation campaign, and the *characteristics* could be anything influencing the system, such as temperature for a manufacturing process, or household income for a solicitation campaign. Usually the collection of data from which one may determine the relationship between characteristics and system behavior involves a cost, and hence there is a trade-off between more data and thus better description (lower variance) of the system and associated cost. For an excellent introduction to the subject, see [1].

In this paper, we focus on the task of deriving a statistical model for a solicitation campaign. The model shall be used to select those prospects from a relatively large pool of candidates who are most likely to make a financial contribution. We assume that, although no outcome data is available to us at the start of our modeling effort, we do have some prior business knowledge of the problem from which we can derive some insight. The task at hand is thus to gauge the available characteristics of the prospecting universe (i.e. the list of possible donors) and, based on a judicious choice of the relevant characteristics, to make a selection of whom to solicit in a first round. The aim of this first round of solicitations is of course to obtain sufficient data to fit the model with which the remainder of the prospecting universe will be evaluated.

The question of how to select the best learning exemplars for the model (i.e., whom to solicit in the first round) is at the heart of this paper. We wish to build a linear model that predicts the probability of a given prospect to donate, and we also wish to build a second model that predicts how much any given prospect will donate, under the assumption that a donation will be made. This problem setting falls squarely into the domain of Experimental Design, where one wants to choose the points at which to collect data in order to minimize, for example, the combined variation of the unknown coefficients of the linear model one postulates.

A key feature of Experimental Design is that the variance at a given data collection point can be reduced by measuring the outcome at that point more than once. While this feature is certainly applicable to a physical experiment, it is not practical in a solicitation campaign setting, where each prospect is usually solicited at most once. Hence standard Experimental Design algorithms need to be modified to exclude duplicate point selection.

Furthermore, in the context of a physical experiment, for which Experimental Design was initially developed, any combinations of the various relevant characteristics can be investigated. In a solicitation campaign, the possible combinations of characteristics are limited by the pool of existing prospects.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

SIGKDD '03, August 24-27, 2003, Washington, DC, USA.

Copyright 2003 ACM 1-58113-737-0/03/0008...\$5.00.

This restriction leads naturally to the consideration of optimal designs that rely on a given (fixed) list of candidate combinations. One such design is a D-optimal design. In essence, the D-optimality criterion minimizes the combined variance of the unknown coefficients of a postulated linear model. We used a Java implementation of the Fedorov exchange algorithm [2], which we modified so that points in the design can be selected at most once.

The application of Experimental Design to build a response model has the following complicating factor. In a solicitation campaign the outcome is binary, and we are interested in the probability of response, not just whether a prospect responded or not. To this aim, we cluster similar prospects (using a K-means algorithm with Euclidian distance) and then estimate the probability of response by computing the response rate for the cluster. If the Experimental Design dictates that a certain combination of characteristics be tested, it in effect selects a group of similar prospects which will all be solicited so that the response can be evaluated. This clustering approach also has the important side-benefit of making the problem size tractable with respect to the D-optimality search.

A related consideration is that we have modeled this probability of response with a maximum likelihood logistic regression. However, our Experimental Design criterion assumes a linear regression, which implies a least square error estimate of the coefficients. In practice, the data collection process of the Experimental Design benefits the logistic regression as well.

To evaluate our approach, we relied on a solicitation campaign data set made publicly available by a national veteran’s organization in the context of the KDD-Cup 98. For KDD-Cup 98, half of the prospecting universe was randomly selected to build a model, which was to be used to determine the best prospects in the remaining half. Soliciting half of all prospects involves a significant upfront cost and with it, a considerable financial risk exposure, which one might want to reduce. One could of course simply select less than half of the data for learning, and by selecting the right balance between learning set size and model performance; one can optimize the total net profit achieved. As we will show in the remainder of this paper, using Experimental Design, one can limit upfront exposure and still obtain comparable or superior profits to what can be obtained with a random selection of learning sets of larger size, and thus with corresponding larger initial upfront exposure.

2. POSTULATED MODEL

2.1 Model type

As previously mentioned, we modeled the probability of response of a prospect by using a logistic regression (response model), and, assuming response, we modeled the forecasted amount donated using a linear regression model (donation amount model). For a given prospect x , with an estimated probability of response $p(x)$ and estimated donation amount $a(x)$, the revenue to be expected from this prospect is thus $r(x) = p(x) \cdot a(x)$, with an associated cost $c(x)$. One would therefore choose to solicit a prospect x if the expected revenue $r(x)$ exceeds the cost $c(x)$. For the KDD Cup 98, the cost was assumed to be a fixed \$0.68 per solicitation.

2.2 Selection of prospect characteristics

This is an important step in the application of Experimental Design, and arguably one of the more difficult ones (the other one being the selection of the learning sample size). In general, one

can distinguish between two main approaches to the selection of the most relevant characteristics of the model. The first is purely data-driven and relies purely on statistical methods such as clustering, principal component analysis, etc. The second is to use prior knowledge of the problem obtained from experienced fundraisers or from previous campaigns. The latter carries a larger risk as important characteristics could easily be missed, resulting in a poor model. Since we assumed that no data were available to us, we have relied exclusively on the second approach to select the model characteristics. We chose 12 characteristics, listed in Table 1. They were chosen for both the response and donation amount model. Therefore the selection consists of variables that are believed to be predictive either of probability of response or of donation amount, or both.

Table 1. Characteristics in the KDD Cup 98 data set

Name	Description
INCOME	Household income
PEPSTRFL	PEP star donator flag
NUMPROM	Number promotions/solicitations send to prospect
RAMNTALL	Total received amount from all solicitations
NGIFTALL	Number of gifts received from all solicitations
LASTGIFT	Amount of last gift
LASTDATE	Date of last gift (transformed into how many months ago)
FISTDATE	Date of first gift (transformed into how many months ago)
AVGGIFT	Average gift received
PGIFT	Percentage of solicitations that resulted in a gift
RFA_2A	Amount of giving in most recent period
RFA_2F	Frequency of giving in most recent period

3. Clustering

Clustering is a basic component of our approach, not only because it reduces the problem size and complexity, but also because we are measuring the binary response instead of a continuous outcome. By grouping similar prospects, it becomes possible to estimate the probability of response by computing the response rate for the cluster. The size of the clusters should be constant to ensure uniform variance across clusters and commensurate with the response rate to ensure a minimum number of respondents in average per clusters.

All input characteristics were standardized to a mean of 0 and a standard deviation of 1 to ensure a uniform scale across characteristics. This approach is easily justified for continuous characteristics by assuming that they follow a Gaussian distribution. In the case of the binary characteristic PEPSTRFL, this scaling corresponds to a linear scale transformation that maps its two possible values to a scale comparable to that of the other characteristics. This step is necessary for the Experimental Design algorithm as well as for the K-means algorithm because it prevents large-valued variables from exerting an undue influence on the model.

The K-means algorithm does not naturally result in clusters of equal size. Since a probability of response is estimated from each cluster, a minimum cluster size needs to be imposed to maintain reasonable accuracy. We chose 50 prospects as this minimum, which seems reasonable since we expect a response rate for the KDD Cup 98 data set of a few percent, and thus each cluster should on average contain at least one respondent. To guaranty this minimum size, cluster seeds with a corresponding size lower than 50 were removed during the K-means iterations. At the end, larger clusters were randomly split into clusters of 50, and the remaining prospects discarded from the Experimental Design. To illustrate, a cluster with 277 prospects would give 5 clusters of 50 prospects each, and 27 prospects would be discarded. We kept the same centers for all of the sub-clusters.

4. EXPERIMENTAL DESIGN

4.1 D-optimal designs

An optimal design is a design that maximizes (or minimizes) some optimality criterion, where one commonly restricts the candidates for the design to a fixed list and chooses the design points with the help of a computer. The theory of computer-generated designs traces back to the work by Kiefer and Wolfowitz in the late 1950s [9, 10]. We have chosen one of the most well known optimality criterion known as the D-optimality criterion, which measures the volume of the combined confidence region of the unknown linear regression coefficients. Note that we are not yet concerned with the accuracy of the fit. At this stage, our objective is to minimize the uncertainty of the regression coefficients themselves. The Fedorov exchange algorithm [2] usually finds a near-optimal design by attempting to minimize the D-optimality criterion through selective swapping of points in and out of the design. We modified this algorithm so that any point can be incorporated at most once into the design.

It is important to keep in mind that a D-optimal design is optimal only with respect to a pre-determined parametric linear model. In other words, an optimal design always requires an associated choice of a *model class*, which also includes the pre-selection of all characteristics in the model. Note that the model is assumed to be linear in the regression coefficient but there is no such restriction on the characteristics themselves. For instance, a model could very well include a linear dependency on the square of a measured temperature. This underlying assumption of a model class represents one of the main weaknesses of the Experimental Design paradigm. If the postulated model class is incorrect, then the optimal collection of regression data for this (incorrect) model class does not necessarily lead to an optimal model (i.e., with minimum variance on the regression coefficients), and random selection of learning data might produce superior results, since no bias was introduced into the choice of regression data. This explains why Experimental Design usually performs better when the number of regression data samples is restricted, but not necessarily for larger counts. In the absence of any prior knowledge to guide the selection of a model class, the usual approach is to perform a series of *screening experiments*, that is to use Experimental Design iteratively by collecting data in stages and adjusting the model class between iterations if needed, thus allowing for the judicious introduction of higher order terms. Note that in practice, in the absence of prior knowledge, a linear model often performs fairly well since a linear approximation is usually sufficient to capture the main relationships between characteristics and outcome. For

the sake of simplicity, in this work we have assumed some prior knowledge to guide us in the choice of model class.

4.2 Clustering effects

We have already discussed the reason for forming clusters (it allows us to estimate a probability of response). However, there is another reason for this approach. Finding an optimum design in which several thousands of design points are to be selected from an even larger list of candidates implies a significant amount of computer runtime (several days). Clustering brings this runtime down to minutes.

In consequence, the Experimental Design algorithm is tasked with picking the desired number of design clusters from a list of all available clusters. However, at completion, our model is not based on the cluster centers, but on the actual individual records belonging to the selected clusters.

Recall that all clusters were brought to a constant size (50 records) by randomly splitting them into smaller clusters with identical centers and discarding remaining records. This in effect introduces a certain degree of variability into our designs, since for the Experimental Design algorithm all clusters corresponding to a given center (i.e., originating from the same large cluster) are considered identical, while in effect this may not be the case. This variability is obviously small for relatively uniform clusters.

5. MODEL FITTING

5.1 Feature selection

If this were a real campaign rather than a simulation, we would have mailed out solicitations and collected what we believe to be an optimal data set to be used for fitting both our response and donation models. In our context, this simply amounts to setting aside the data to be used to fit our statistical models.

As previously mentioned, we have used logistic regression for the former and a linear regression for the latter. We limited the possible model features (i.e., characteristics) to those previously selected for the Experimental Design. We performed a forward stepwise selection regression for both models. This resulted in two different sets of variables for the response model and the donation model. Note that this feature selection also involves a certain degree of variability.

Using a sample size of 10,000 records with a p-value of 0.2, the following characteristics were selected for the response model: INCOME, PEPSTRFL, NUMPROM, RAMNTALL, NGIFTALL, LASTGIFT, LASTDATE, PGIFT, RFA_2A, and RFA_2F. However, for the donation amount model and with the same p-value, only five characteristics were selected: PEPSTRFL, LASTGIFT, AVGGIFT, RFA_2A, and RFA_2F. The characteristic FISTDATE was not selected for either model. Had we have known this beforehand, we could have produced a more efficient design.

5.2 Sample size

How much data should we collect? This is one of the perennial questions in statistical modeling. Attempting to determine the optimal amount of data to be collected is beyond the scope of this work, but some practical considerations are worth mentioning. One such consideration is the degree of financial risk exposure. The process of data collection implies a cost that can sometimes be significant. An institution may not be willing to take such chances and will set an arbitrary limit to the upfront investment.

On the other hand, setting this cost unreasonably low would result in a poor model that would seriously limit the campaign effectiveness.

We used the following rule of thumb. We require a minimum of 10 records per regression coefficients. Our model has at most 12 characteristics and hence 13 regression coefficients to be estimated. Therefore, we need a minimum of 130 data points. Furthermore, for the donation amount model we can only use the actual responses (not just the solicitations). We expect a response rate of about 5% from past experiences. Therefore, a minimum sample size of about 2,600 prospects would be required. In our simulations, we chose a basis of 2,500 prospects to get round multiples.

6. MODEL EVALUATION

Since we used the KDD Cup 98 data set to evaluate our approach, we begin by providing more details about it. The data contains demographics and donation history related to a fund-raising campaign by a national veteran's organization. In this campaign, solicitations requesting a donation were mailed to prospects who had previously donated money to the organization, but not in the previous twelve months.

For the competition, the data set was split into two parts: a *learning* data set and a *validation* data set. Both sets were made available to the KDD Cup 98 contestants. However, the target fields (response and donation amount) of the *validation* data set were withheld, and revealed only after the contest was over. The *learning* set described 95,412 prospects, of which 4,843 were responders, and the *validation* set described 96,367 prospects, of with 4,873 responders. This amounts to a response rate of about five percent. Note, that this data set was also the subject of a KDD'99 knowledge discovery contest, where participants were invited to apply a range of knowledge discovery techniques to extract unspecified findings of commercial value.

The largest revenue can be obtained by mailing a solicitation request to all prospects in the validation set. This strategy would produce a total of \$76,089.64 for the validation set. However, this is also the most costly strategy: Subtracting \$0.68 per mailing for 96,367 pieces of mail results in a net profit of only \$76,089.64 - \$65,529.56 = \$10,560.08 for the validation set. Note that no model was required. The best theoretical result, given that the learning set has already been collected, can be achieved by soliciting only those prospects in the validation set who will respond. This results in a net profit of \$76,089.64 - \$0.68 * 4,873 = \$72,776.00 for the validation set. The worst theoretical result can be achieved by soliciting only those prospects in the validation set who will not respond, resulting in a total loss of (96,367-4,873) * \$0.68 = \$62,215.92 for the validation set. All published models fell somewhere in between.

In this paper, we have set the problem differently. We did not assume that about half of the data have already been collected (i.e. tagged) for learning purposes. Instead, we wish to decide how much and what data need to be collected in order to fit a model. This decision implies an upfront cost and therefore financial risk exposure. Of course, since the data collection phase also involves soliciting prospects, some revenue will likely be derived as well. As stated, the learning set for the KDD Cup 98 had 95,412 records, implying an up-front cost of 95,412 * \$0.68 = \$64,880.16. This first round of the solicitation campaign had 4,843 responders

who donated a total of \$75,668.70, so that the coincidental profit achieved in the data collection phase was \$75,668.70 - \$64,880.16 = \$10,788.54. This is a significant portion of the total campaign profit, since the best published result achieves a profit of about \$15,000 in the second round of solicitations, giving a total profit of about \$25,800. Other results are itemized in Table 2.

Table 2. Selected results on the KDD Cup 98 data set

Company/ Group	Net profit in 2 nd round of solicitation	Total profit of both rounds of solicitation	Comments/ Reference
Amdocs Ltd	\$15,040.00	\$25,828.54	KDD Cup 99 [3]
SAS Institute	\$14,877.77	\$25,666.31	KDD Cup 99 [4]
Zadrozny & Elkan, UCSD	\$14,741.00	\$25,529.54	Research article [5]
GainSmarts	\$14,712.24	\$25,500.78	KDD Cup 98 Winner [6]
SAS Institute	\$14,662.43	\$25,450.97	KDD Cup 98 2 nd place [7]
Quadstone	\$13,954.47	\$24,743.01	KDD Cup 98 3 rd place [8]
ARIAI/ CARRL	\$13,824.77	\$24,613.31	KDD Cup 98 4 th place
Amdocs Ltd	\$ 13,793.24	\$24,581.78	KDD Cup 98 5 th place
	\$10,560.08	\$21,348.62	Solicit every- body
	-\$53.68	\$10734.86	KDD Cup 98 last place

It is important to put the differences in achieved net profit for the various models into perspective with regards to statistical significance. Rosset and Inger from Amdocs Ltd. write in their KDD Cup 99 article [3] that *a significant difference in model performance should definitely exceed \$500*. Along the same line, Georges and Milley from the SAS institute state in their KDD Cup 99 article [4] that, assuming a two-stage model and a net profit of \$16,000 [for the second round of solicitations], the 95% confidence prediction interval ranges from \$13,600 to \$18,400. They conclude that *the wide interval suggests that there is little statistical difference between [their model] and last year's winning models (and even less between last year's top finishers)*.

To assess the merits of an approach based on Experimental Design, we have compared the total profit generated with that of a Random Design, which is based purely on a random selection of an identical size of learning data. However, both approaches are not entirely deterministic. In the Random Design case, the source of variability originates from the random choice of the learning data. In the Experimental Design case, there are two sources of variability. The first originates from the Fedorov exchange algorithm itself, which is an iterative approximate algorithm involving a random selection order. In practice, those fluctuations were found to be negligible. The second and more significant source of variability in the experimental design arises from our clustering method. As previously described, we used a K-means clustering approach with a constrained minimum cluster size of 50, and we randomly broke larger clusters into smaller clusters of 50 pros-

pects each. Note that these clusters are all identical from the perspective of the Experimental Design algorithm (they all share the same cluster center).

We have addressed the issue of statistical confidence by repeating 50 experiments for each given learning set size (i.e., first round of solicitation). In the Random Design case, one experiment consisted simply in picking the desired number of prospects randomly (from a combined data set of what was originally split into the learning and the validation sets) and fitting and evaluating a model. In the Experimental Design case, an experiment consisted in performing the random splitting of large clusters, computing a near D-optimal design, and fitting and evaluating a model.

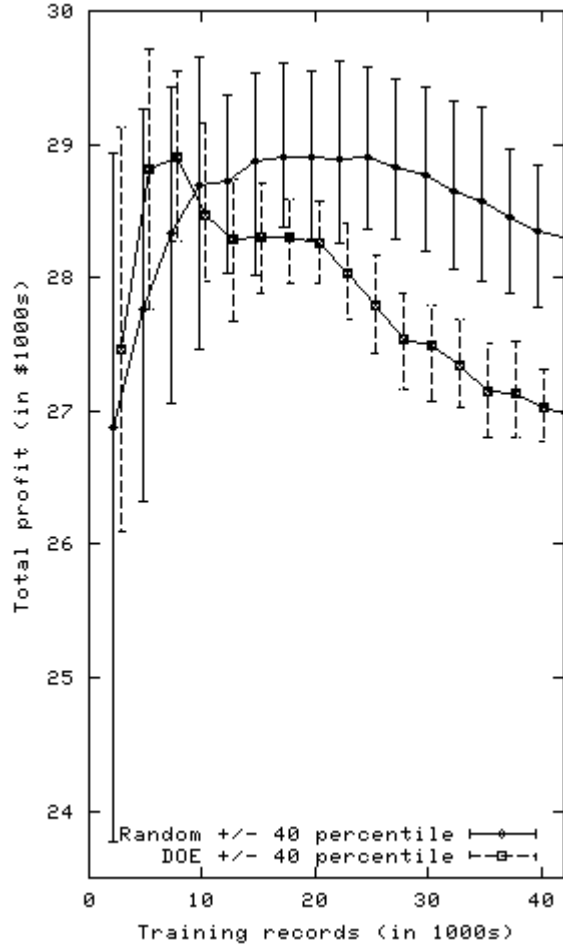


Figure 1. Total net profit of solicitation campaign.

In Figure 1, we report the various total profits achieved over both rounds of solicitation using Experimental Designs (dashed error bars) and Random Designs (solid error bars). We have excluded the top and bottom ten percentiles on account of outliers.

It should be stressed here that the downward trend displayed in Figure 1 should not be interpreted as model error, which is expected to decrease with larger learning data size. Rather, it shows that for both cases, an optimal learning data size exists. This optimum corresponds to the ideal compromise between model accuracy and cost of data collection. Clearly, the Experimental Design approach yields better results and produces smaller variability for

smaller learning set sizes. The crossover point for the KDD Cup 98 data set is at about 10,000 records for the learning set, which of course could not be known prior to the data collection phase. Past this crossover point, the total profit achieved by models based on Experimental Design drops faster than that of the Random Design.

The fact that the approach based on Experimental Design performs worse than a Random design for larger learning data sets can be explained as follows. By choosing a linear model, we have assumed an *exact* linear relationship between characteristics and outcome. The Experimental Design algorithm tends to pick points with the greatest leverage, and thus near the edges of the prospecting universe, to maximize the goodness of fit. In reality, this relationship is rarely *exactly* linear. Therefore, a Random Design becomes preferable because it covers more evenly the prospecting universe.

Finally, as Table 3 shows, the merits of an Experimental Design approach become greater when financial exposure during the first round of a solicitation campaign is of greater concern. Note that reported numbers for entries involving randomness correspond to the average of the various profits achieved over 50 experiments.

Table 3. Total profits for the KDD Cup 98 data set

Method	Training sample size	Upfront cost of first round	Total profit of both rounds of solicitation
Amdocs Ltd KDD Cup 99	95,412	\$64,880.16	\$25,828.54
Experimental Design	10,000	\$6,800	\$28,471.73
Random	10,000	\$6,800	\$28,686.71
Experimental Design	7,500	\$5,100	\$28,911.64
Random	7,500	\$5,100	\$28,329.03
Experimental Design	5,000	\$3,400	\$28,814.58
Random	5,000	\$3,400	\$27,757.05
Experimental Design	2,500	\$1,700	\$27,457.80
Random	2,500	\$1,700	\$26,880.56

7. DISCUSSION

Overall, Experimental Design is clearly beneficial to the modeling of solicitation campaigns. As we have shown, it can allow fundraising institutions to limit upfront financial exposure during the data collection phase, while still allowing for the derivation of well-performing statistical models. But it is important to note that the boost Experimental Design provides occurs mostly when the learning sample size is relatively small. For small learning set samples however, the highly desirable combination of better performance of models than with random data collection, smaller variance (higher confidence in the model performance), and significantly reduced upfront financial exposure can be achieved.

On the other hand, the drawbacks inherent in the method are clear: A significant amount of prior domain knowledge is required to adequately select the relevant characteristics beforehand. Also,

since the crossover in performance between Experimental Design and Random Design cannot be determined in advance, special care has to be taken in selecting the size of the learning data sample. For large data set, a Random Design will likely be superior to an Experimental Design.

A crucial parameter in our argument is the cost of collecting data points. In the case we used for this simulation, this cost, while not negligible, is not overly high. In other applications however, such as a vehicle crash test campaign, this cost would be much higher and the advantages of Experimental Design would be even greater.

On a final note, in this paper we also did not address other typical applications of Experimental Design, such as when one can freely choose certain parameters (e.g. credit card offer parameters), or where one already has collected some data and wishes to selectively collect additional data.

8. ACKNOWLEDGMENTS

The authors wish to thank Fair, Isaac and Company, Inc. for supporting this research.

9. REFERENCES

- [1] Douglas C. Montgomery. Design and Analysis of Experiments, John Wiley & Sons, Inc., 5th edition, 2001.
- [2] V.V. Fedorov. Theory of Optimal Experiments, translated and edited by W.J. Studden and E.M. Klimko, New York: Academic Press, 1972.
- [3] Saharon Rosset and Aaron Inger. KDD-Cup 99: Knowledge Discovery In a Charitable Organization's Donor Database, Amdocs (Israel) Ltd, 8 Hapnina St, Raanana, Israel, 43000 (1999).
- [4] Jim Georges and Anne H. Milley, KDD'99 Competition: Knowledge Discovery Contest, SAS Institute, Jamboree Ctr, Ste 900, 5 Park Plz, Irvine, CA 92614, USA (1999).
- [5] Bianca Zadrozny and Charles Elkan. Obtaining calibrated probability estimates from decision trees and naïve Bayesian classifiers, Department of Computer Science and Engineering, University of California, San Diego, La Jolla, California 92093, USA (2001).
- [6] Jacob Zahavi and Nissan Levin. Urban Science wins the KDD-98 Cup, Tel Aviv University and Urban Science, <http://www.urbanscience.com> [press release, no longer available] (1998).
- [7] The SAS Institute. SAS software takes top medals in KDD data-mining competition, SAS Institute Inc., Cary, NC, USA [press release, no longer available] (1998).
- [8] Quadstone Ltd. KDD Cup 98: Quadstone Take Bronze Miner Award, Quadstone Ltd., <http://www.quadstone.com> [press release, no longer available] (1998).
- [9] J. Kiefer. Optimum Experimental Designs. J. Royal Stat. Soc. B, Vol 21, 272-304 (1959).
- [10] J. Kiefer and J. Wolfowitz. Optimum Designs in Regression Problems. Annals Math. Stat., Vol 30, 271-294 (1959).