DATA MINING IN DIRECT MARKETING DATABASES

PETER VAN DER PUTTEN
Sentient Machine Research, Amsterdam
and Dept. of Computer Science, Leiden University
E-mail: putten@cs.leidenuniv.nl

In direct marketing large amounts of customer data are collected that might have some complex, non linear relation to customer behavior. Data mining techniques can offer insight in these relations. In this paper we give a basic introduction in the application of data mining to direct marketing. Best practices for data selection, algorithm selection and evaluation of results are described and illustrated with a number of real world cases. We suggest two lines of research which we consider important to put data mining in the hands of the marketeer: automating data mining techniques and integration of data mining in an open knowledge management framework.

1 Introduction

In marketing, there are two opposed approaches to communication: mass marketing and direct marketing. In mass marketing, a single communication message is broadcast to all potential customers through media such as print, radio or television. Such an approach always implies a high waste: only a small proportion of the customers will actually buy the product. Now that competition increases and markets get more fragmented the problem of waste worsens. Moreover, in spite of huge investments in market research and media planning, it is still hard to quantify the benefits of mass marketing.

These developments have led to an increased popularity of direct marketing, especially in the sectors of finance, insurance and telecommunication. The ultimate goal of direct marketing is cost-effective, two-way, one-to-one communication with individual customers. For this it is essential to learn present and predict future customer preferences. In today's chaotic business environment, customer preferences change dynamically and are too complex to derive straightforwardly.

Data mining, the continuous analysis of customer behavior patterns, may offer a flexible solution to this problem.[1,2,3] In this paper we will give a practical introduction to data mining for direct marketing purposes.

We will not discuss any theoretical algorithmic issues, nor will we describe experiments in detail. We only aim to offer managers a tutorial style introduction to current data mining best practices: how is data mining commonly applied and evaluated, and which data and algorithms are most appropriate, given common direct marketing tasks.

In the first part we will describe the data mining process in a direct marketing context. A case from insurance is added to give an impression of the practical issues related to data mining projects, including the evaluation of data mining results. In the second part we will focus on lessons learned with respect to the selection of data and algorithms, based on eight data mining projects carried out in co-operation with the Dutch association for direct marketing, sales promotion and distance selling. We conclude with suggesting directions for research.

2 Data mining

Data mining can be defined as the extraction of valuable patterns that are hidden in large amounts of customer data.[4] The whole process of steps involved in data mining is often referred to as the knowledge discovery cycle (see figure 1). This includes definition of the objectives, selection and preparation of the data and evaluation of the results with technical and business criteria.

Within the loop of a single project, it is not uncommon to go through the knowledge discovery cycle a number of times. For instance, by doing data mining analysis one might discover that some important data was not selected or was not prepared in the appropriate format. By performing different data mining projects repeatedly, an organization starts to learn more and more about customers. This can be considered to be a second loop of learning.

The current success of data mining in businesses is facilitated by a number of technical factors. Growing amounts of customer data are collected and made accessible in corporate data warehouses. Powerful new data analysis algorithms are discovered by researchers from statistical pattern recognition and artificial intelligence fields such as machine learning, neural networks and evolutionary computation. Today, ordinary office computers are powerful enough to run these advanced data mining algorithms.

In a direct marketing context, two prototypical data mining objectives can be distinguished: *prediction* and *description*. Prediction involves predicting unknown or future customer behavior from known customer attributes. Description aims at discovering human interpretable patterns in the data. Best practice application of prediction and description in direct marketing are described below. For a detailed example we refer to the insurance case described in the next section.

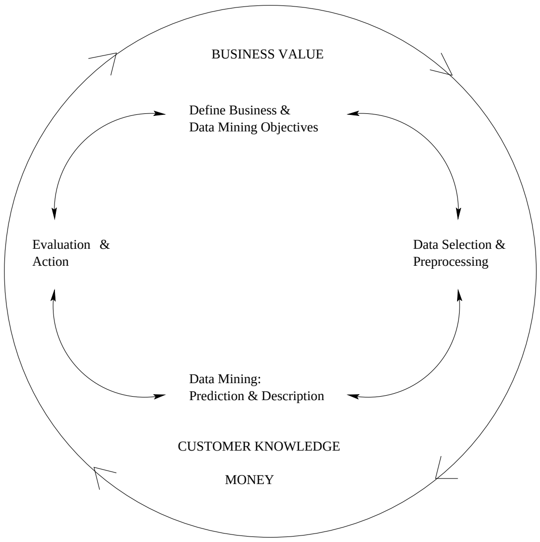

Figure 1. The Knowledge Discovery Cycle.

2.1 Prediction

The classical case in direct marketing for prediction is response modeling. Usually, the relative number of customers that responds to direct mail is very low (5% or less). Predictive models can be built to identify the prospects most likely to respond. Historical data about previous mailings or product ownership are used to construct the model. If such information is unavailable, for instance when selecting prospects for a new product, a test mailing is performed to collect information. The resulting model can be applied to filter prospects from the existing customer base or from address lists acquired from list brokers. Although response analysis is by far the most common type of predictive modeling, other applications, such as predicting customer retention or estimating turnover potential, are promising for direct marketing as well.

2.2 Description

A shortcoming of prediction is that it produces models that are black boxes. A response prediction model is perfect for selecting interesting prospects. But it does not give the marketeer more insight into the reasons why these customers respond.

There are a several data mining techniques for descriptive modeling. The main subtasks we will discuss here are profiling and segmentation.

Profiling, also known as deviation detection, is used to discover for which attribute-value pairs a selection differs significantly compared to the database as a whole or other reference groups. For example, we considered profiling useful for mining the Dutch national media survey (SUMMO). Every questionnaire contained hundreds of questions on media interests, product consumption and socio-demography, so it was infeasible to construct the profile of deviating attribute values manually. By using profiling for an analysis of vodka drinkers we found that vodka drinkers are more often students, drink more Bacardi rum and are more frequent visitors of cinemas.

In segmentation, the goal is to discover subgroups in data. Customers within a cluster should resemble each other as much as possible, where as the clusters themselves should differ as much as possible. For example, in the vodka case we found out that the average vodka drinker does not exist. Instead, subgroups were found that could be described as "cocktail drinking teenagers", "young student couples" and "travelling salesmen".

3 Insurance case

The business objective in the insurance case was to expand the market for an existing consumer product, a caravan insurance, with only moderate cost investment.[a] We identified two data mining objectives: selecting individual prospects and describing existing customers.

3.1 Data selection and preprocessing

Each customer was characterized by a selection of 95 attributes. The attributes could be divided in two groups. The product usage attributes defined the product portfolio of an individual customer, so these attributes can be considered to be internal (company owned), behavioral attributes. We also purchased socio-demographic survey data that had been collected on zip code level. All customers belonging to the same zip code area have the same value for these attributes. This included information on education, religion, marital status, profession, social class, house ownership and income. The selection of attributes to be used was made based on expert domain knowledge and exploratory data analysis (correlation with attributes to be predicted).

[a] Some details in the insurance case have been changed for reasons of confidentiality, without loss of generality.

A number of preprocessing steps were taken, some of which are provided directly by the data mining environment we used.[b] Most numerical attributes were transformed to categorical values. For each attribute, normalization factors were computed so that all attributes had the same standard deviation. Missing values were identified so that the algorithms that were going to be used could handle these values correctly.

3.2 Construction and evaluation of the response prediction model

To select prospects we constructed a model to predict the attribute "owns a caravan policy" given all other attributes. A random sample, the training set, was drawn from the customer base. The training set was used to construct a nearest neighbor prediction model. In a nearest neighbor model, the prediction for a given customer is computed by averaging over the response of the most similar customers from the training set.[5] Similarity between customers is computed by combining the distance per attribute, taking several influencing factors into account, such as attribute types (categorical or numerical), normalization factors and missing values. Missing values for the attributes of the customer for which prediction was to be made were simply ignored (don't care values); missing values for customers from the training set resulted in some constant distance for this attribute (don't know values). The relative importance of different attributes was computed by running a profile analysis of the policy owners compared to all customers in the training set. A number of attributes were assigned very low importance, so the actual number of attributes taken into account to compute the resemblance was reduced.

The prediction model was applied to the testing set, a random sample of customers, disjoint with the training set. For each test customer, the response score was computed to indicate the potential to buy a caravan insurance.

Given a low response rate of 6.5%, the naive prediction model which scores all records as non-respondents, already achieves 93.5% classification performance. So this standard data mining measure, which counts the relative numbers of cases for which the classes were predicted correctly, did not suffice. This suggests that other evaluation criteria were needed to evaluate the accuracy of the prediction model.

For this kind of analysis, often a cumulative response chart is used (see figure 2). All records are ordered from left to right on the x-axis with respect to their predicted response rate. If only the top 25% is mailed, the cumulative response $RespCum_m$ (relative number of respondents in mail selection) is

[b] The DataDetective^{TM} data mining environment.

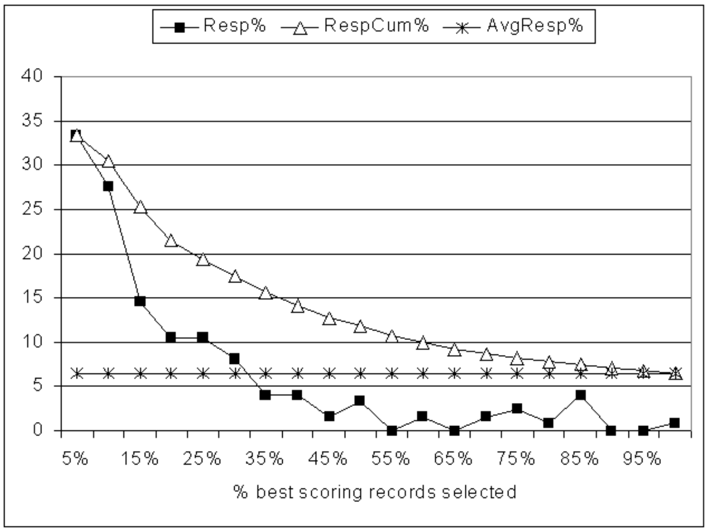

Figure 2. Cumulative Response Chart Insurance Data.

21.5%, which is more than 3 times higher than the response level $RespCum_r$ achieved when records are selected randomly. Another way to evaluate the model is shown in figure 3. Here the relative part of all respondents that is found is plotted:

$$TotResp_m = \frac{RespCum_m * s}{RespCum_r}$$

with s selection size). If the top 25% customers are selected, nearly 75% of all respondents are found (!). The Gain measure:

$$Gain_m = \frac{(RespCum_m - RespCum_r) * s}{RespCum_r}$$

indicates when the gains of using a prediction model start to decrease.

The optimal mail selection size depends on the gain, cost and profit per mail piece, marketing strategy, mining costs etc. Assume that in the insurance case the strategy was just to mail the top 25% of the customers with the highest response score and we wanted to maintain the same absolute number of respondents. So the customer base we had to select from had to contain 100/75 % customers. We only had to mail $25 * 100/75$ % = 33.33% of the amount of customers we would have had to mail without a prediction model.

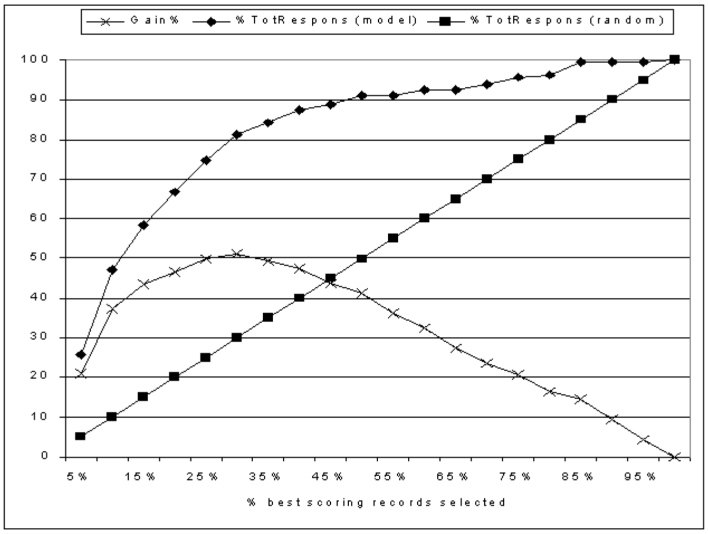

Figure 3. Gain Chart Insurance Data.

Considering that mass mailings to millions of customers at 1 US$ per mail piece are common,[6] reductions in cost can be substantial.

3.3 Descriptive segmentation of respondents

Contrary to prediction, description does not aim to offer results that can be measured and interpreted in a single, objective manner. There exist few standard algorithm-independent error measures for segmentation techniques, and even less for profile or rule discovery. The quality of the description rather depends on the way the marketeer interprets the descriptions, the conclusions that are drawn and the actions that are taken, which is, of course, subjective and hard to measure.

In the insurance example, we wanted to find segments of caravan insurance owners which should be marketed with a different tone of voice, different offers etc. For this we used an iterative projection algorithm, which projected all owners from high (95) dimensional space on to 2 dimensions in a number of steps (see figure 4). In each step, every customer performs a nearest neighbor search to find resembling customers and moves a little bit in that direction. When the projection process had converged, we built customer profiles of the

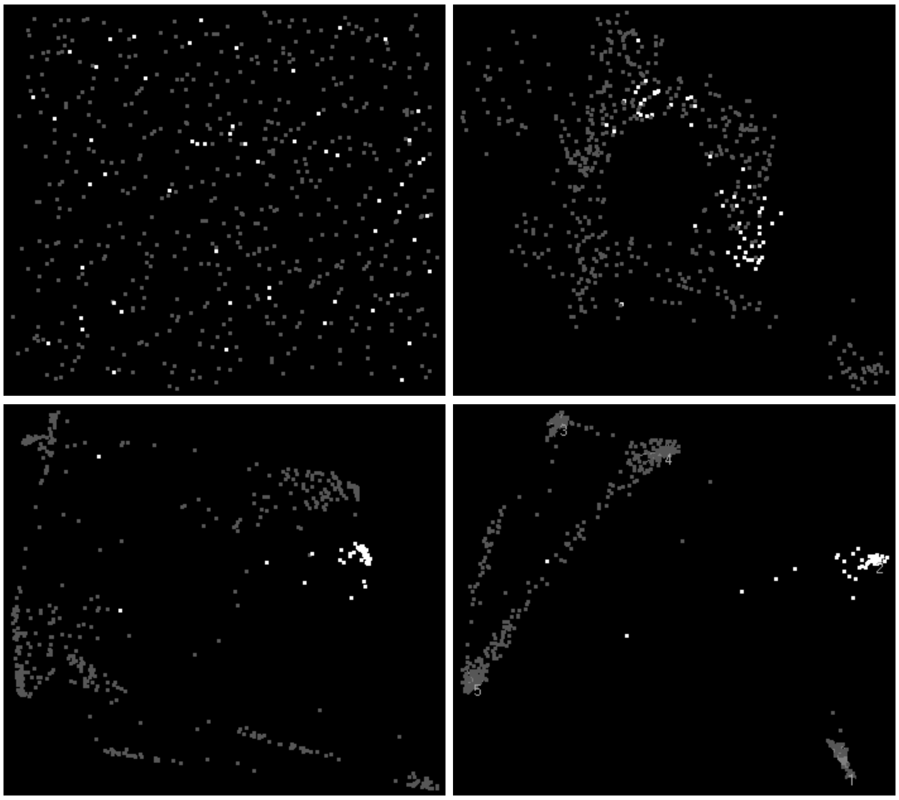

Figure 4. Iterative nearest neighbor projection reveals clusters in data. The figure at the right hand bottom side shows the end result. The user assigns the points to clusters depending on the final state of the projection and the resulting customer profiles. The white points belong to cluster 2.

resulting clusters. For instance, if we compared cluster 2 to all customers and the other clusters, we found out that the cluster contained relatively loyal customers (two car insurance policies, more often in a high turnover category), who where living in residential areas with a high rate of privately owned houses and who were more often belonging to a religious, senior citizen family structure.

Organization	Sector	Business Objective
Fund 1	charity	segmentation fund members
Fund 2	charity	upgrading, estimation donation potential
NS	transport	identifying suspects for prolongation of railways discount card
ABN AMRO	banking	selecting prospects for a new service (balance check by telephone)
VSB	banking	cross selling saving programs to current account holders
Readers Digest	publishing	selecting prospects from a list broker database
Centraal Beheer	insurance	selecting potential converters from a list broker database
NV Databank	list broker	selecting prospects for marketing of a new product

Figure 5. Cases in the DMSA project.

4 Data Mining in Marketing Practice: the DMSA cases

In this section we will propose some best practices for data selection and data mining algorithm selection, based on data mining experiences from a variety of direct marketing projects.

Developing a single data mining algorithm that offers the best peak performance (e.g. predictive accuracy) on all possible data mining problems might seem a good approach for research. However, a recent theorem, known as the "No Free Lunch Theorem", seems to suggest that such an approach will probably not be succesful. This theorem states that measured over all possible data sets and cost functions, there can be no single algorithm that performs best .[7] So it makes more sense to identify the requirements for which a certain class of algorithms performs better.

Furthermore, researchers sometimes assume that the data set is given. However, the choice of data is probably even more important than the algorithm used. In the specific context of marketing, it might be possible to develop best practices for collecting appropriate data.

These issues have been the focus of a research project Sentient Machine Research has performed in co-operation with the Dutch Organization for Direct Marketing, Sales Promotion and Distance Selling (DMSA). To be more exact, the objective of the research was to identify under which circumstances and for which data the relatively "new" data mining algorithms such as neural networks, rule induction and evolutionary algorithms performed "better". We compared these algorithms to classical statistical techniques such as linear regression and discriminant analysis. Experiments were performed in eight real world data mining projects for several organizations, including banks, insurance companies, publishers, a railway company and charities[6] (see figure 5). In the next subsection, we would like to share some of the lessons learned from this project.

4.1 Collecting the right data

A general result was that the data used was the most important factor for the success of the data mining projects. Data collection was always a critical constraint in project planning. Data preparation amounted to up to 80% of total work invested. The requirement for what kind of data to use depends primarily on the nature of the data mining task at hand.

For prediction tasks, the data should possess as much predictive power as possible. Firstly, the number of attributes plays an important role. The more attributes are used, the higher the probability becomes that intricate non-linearities between attributes occur which are important for predicting customer behavior. On the other hand, the so called "Curse of Dimensionality"[8] limits the amount of attributes that can be used. If the number of attributes increases, the density of the data set in pattern space drops exponentially and complexity of models can grow linearly or worse. Complex models (i.e. a large number of parameters) have a higher chance of overfitting to the training data and will not perform well on new data (low generalization).

Secondly, the type of attributes to be used is of importance. The best data to use is internal behavioral customer data which relates directly to the products to be marketed or the customer behavior to be predicted. Examples are product usage, information requests, response, account balances etc. Traditional marketing attributes such as social class, education, age and gender do not suffice to predict modern customer behavior. Empirical support for these claims is shown in figure 6. Highest gains were achieved in projects based on data as described above.

For description tasks, however, raw summaries of product usage etc. do not suffice to inspire marketeers. Descriptive attributes, such as socio-demographic attributes that are collected at zip code level, should be added. These attributes typically possess much less predictive power, but offer more insight to marketeers.

4.2 Choosing the right algorithm

In the DMSA project, we roughly distinguished between adaptive pattern recognition techniques such as neural networks, rule induction and genetic

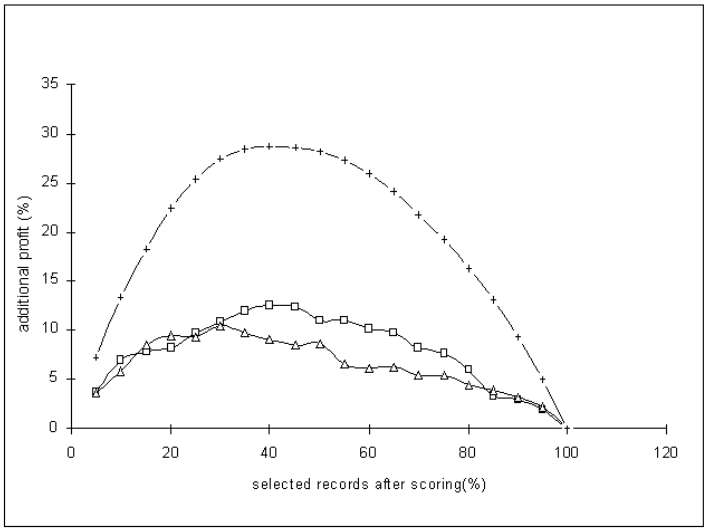

Figure 6. Chart with gains of several projects. Including internal data results in higher gains for response prediction tasks.

algorithms and classical linear statistical techniques such as regression and discriminant analysis. The main advantage of adaptive techniques in general is that these techniques are able to model highly non linear relationships. Furthermore, it is often claimed that these algorithms are less parametric, i.e. make less assumptions about the relation between customer attributes and predicted behavior. For linear regression for instance, this relation is assumed to be linear, which is a pretty tough assumption.

However, what is often not mentioned is that adaptive techniques require implicit assumptions about the data as well. Also, it is sometimes claimed that a high noise level might linearize the non-linear relations that are present in real world marketing data. Practical aspects such as speed, amount of data preparation needed and ease of use are also important. For example, linear regression is still one of the fastest algorithms around when it comes to scoring large data sets.

Overall we found that when the number of attributes was large enough for non-linearities to occur, neural networks performed best in "creaming the crop": selecting the top customers in a prediction task (figure 7). We achieved cumulative response rates that were up to twice as high as the runner up al-

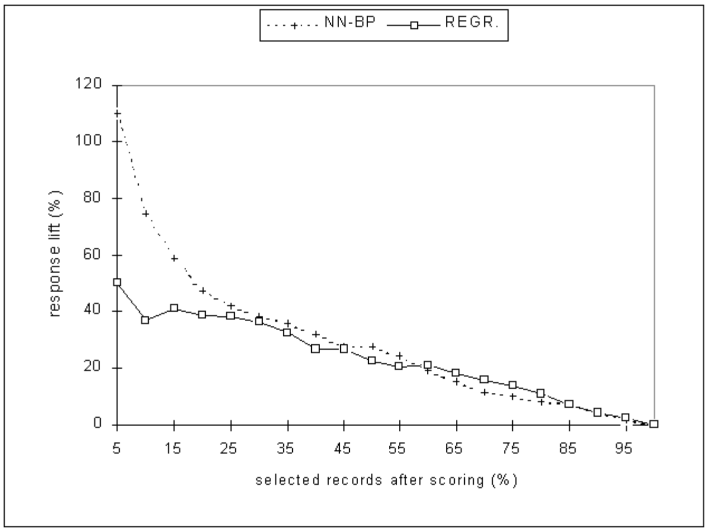

Figure 7. Creaming the crop with backpropagation neural networks. This graph displays the lift $RespCum_m/RespCum_r$ for linear regression versus backpropagation neural networks for one of the DMSA projects.

gorithms, which can correspond to considerable savings. For larger selections, the advantage of neural networks diminishes. This is reasonable to expect, because the relation between customer attributes and response for customers that have an average score is very weak. So if the marketeer is mainly interested in two-way, high quality relations with top customers, which are likely to be interested in the offer, application of adaptive techniques for data mining makes sense. If a marketeer is more interested in performing large bulk mailings with a high response in absolute numbers, wants to mail a relatively large proportion of the customer base, and is willing to accept a high rate of non response, classical techniques such as linear regression will suffice and will be faster to apply.

5 From Data Mining to Knowledge Discovery

The goal of data mining should be the transition to a learning, customer oriented organization. Customer analysis should be performed on a regular basis, or customers should even be monitored on line. Prediction should not be limited to response analysis, but should be a business activity aimed at modeling higher level customer attributes, needs and attitudes, such as "willing to take risk", "high probability of churn" or "early adopter".

Reality is still different. A general problem of most data mining projects is that data collection is a major bottleneck, data mining algorithms still require a lot of parameterization and data preparation and reusability of data mining results is poor. To shorten the knowledge discovery cycle we suggest two important directions for research: automating data mining techniques and integration of data mining in an open knowledge management framework.

5.1 Automating data mining algorithms

Although current data mining algorithms generally require less parameters to be set and less data preparation to be performed than classical statistical algorithms such as linear regression, users still need to have a low level understanding of how a specific algorithm works.

We identify several possible directions to solve this problem. First, in a practical approach, one could identify heuristics for making reasonable choices for specific applications, data mining techniques and steps in the knowledge discovery cycle. These best practices could be properly documented in some kind of data mining methodological framework, or ideally, these best practices are incorporated in intelligent assistants which guide the user through data preparation and prediction processes.

There are also less heuristic and more general algorithmic approaches, which are sometimes referred to as meta learning methods. These algorithms learn to make choices which were normally made by the data mining analyst. For example in ensemble learning, a set of models with different parameter settings is constructed, and the predictions are combined linearly. Recently, various theoretically founded methods were developed to set the coefficients of this linear combination.[9] Another algorithmic approach is Bayesian learning. In Bayesian learning, there is no single parameter setting, but rather a probability distribution over the possible parameter values, which is updated if new data arrives.[8]

5.2 Combining knowledge management and data mining

Whereas it might be fruitful to aim at automating the data mining algorithms, research into the direction of a more open and integrated framework for storing and reusing data mining processes and results would be useful as well. In cognitive psychology there is an important lesson known as the Learning Paradox: "He who knows nothing can learn nothing". There have only been a few attempts to introduce knowledge management to data mining.[10,11] In our view research must not focus on a a single analysis or algorithm, but on a number of analyses in a learning organization. For this it is required that analyst and domain knowledge is represented, a library of data mining results is maintained and intelligent assistance is offered based on this knowledge base.

6 Conclusion

Data mining can be a helpful tool for managers and organizations to cope with a dynamically changing and complex business environment. We identified best practices for application of data mining for direct marketing, selection of data and algorithms and evaluation of results. The key to successful application of data mining will be integration in information infrastructure and business processes.

References

1. R.P. ter Heide. Neural Networks applied to Direct Marketing. In *Proceedings SNN Symposium Amsterdam*, 1995.
2. Charles X. Ling and Chenghui Li. Data mining for direct marketing: Problems and solutions. In *Proceedings 4th International Conference on Knowledge Discovery in Databases (KDD-98)*, New York, 1998.
3. Michael J.A. Berry and Gordon Linoff. *Data mining techniques for marketing, sales, and customer support.* John Wiley & Sons, New York, NY, 1997.
4. Usama M. Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy Uthurusamy. *Advances in Knowlege Discovery and Data Mining.* MIT Press, Cambridge, Mass., 1996.
5. Belur V. Dasarathy. *Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques.* IEEE Computer Society Press, Los Alamitos, CA, 1991.
6. E. Wagenaar. *Data Mining in Marketing Databases.* DMSA, Amsterdam, NL, 1997. In Dutch.
7. D.H. Wolpert and W.G. MacReady. No free lunch theorems for search. Technical Report SFI-TR-95-02-010, The Santa Fe Institute, 1996. ftp://ftp.santafe.edu/pub/wgm/nfl.ps.
8. Christopher M. Bishop. *Neural Networks for Pattern Recognition.* Oxford University Press, Oxford, UK, 1995.
9. Michiel C. van Wezel, Walter A. Kosters, and Joost N. Kok. Maximum

likelihood weights for a linear ensemble of regression neural networks. In *Proceedings ICONIP'98*, Kitakyushu, Japan, 1998.

10. R. Wirth, C. Shearer, U. Grimmer, Th. Reinartz, J. Schlösser, Chr. Breitner, R. Engels, and G. Lindner. Process-oriented tool support for kdd. In *Proceedings of the 1st European Symposium on Principles of Data Mining and Knowledge Discovery*, Trondheim, Norway, 1997.
11. Robert Engels, Guido Lindner, and Rudi Studer. A guided tour through the data mining jungle. In *Proceedings of the 3rd International Conference on Knowledge Discovery in Databases (KDD-97)*, Newport Beach, CA, 1997.