Artificial neural networks, classification trees and regression: Which method for which customer base?

Received (in revised form): 4th April, 2004

Roland Linder

earned his doctorate in Human Medicine at the University of Lübeck, Germany. In the early 1990s he started to evaluate artificial neural networks for application in medicine, physics, chemistry etc. In 1996 he joined the Institute of Medical Informatics at the University of Lübeck, and continued improving neural network technology. He has received two awards from the German Association for Pattern Recognition in 1992 and 1998. In 2001 he developed a far-reaching automatic learning software tool.

Jeannine Geier

obtained her Masters degree in human sciences at the University of Vienna, Austria. Subsequently, she worked in market research (politics and social research) at Fessel-GfK (Vienna, Austria). In 2000, Jeannine Geier moved within the GfK-group to the data mining group at dm-plus Direktmarketing AG (St. Gallen, Switzerland). dm-plus is a daughter company of the GfK-group and its core business is address management, campaign management, data warehousing and data mining. dm-plus Direktmarketing AG also providing consulting services at the strategic level in direct marketing.

Mathias Kölliker

did his PhD in Evolutionary and Behavioural Biology at the University of Bern, Switzerland. After one year of teaching and research as a postdoctoral assistant he joined dm-plus Direktmarketing AG as a data mining consultant, where he performed various analyses for direct marketing campaigns in the media as well as in the food and non-food branches. He is now at the Department of Biology, Indiana University, Bloomington, USA as a postdoctoral research fellow supported by the Swiss National Science Foundation.

Abstract The most commonly used modelling methods for targeting customers in direct marketing are artificial neural networks (ANNs), classification trees (CTs) and logistic regression (LR). These methods differ in how rules for the association between purchase behaviour and customer information are derived from the data. The authors investigated the predictive performances of the three methods in a competitive test in a simulated direct marketing scenario. The experimental design consisted of a number of situations comprising varying sample sizes and data complexities. The results show that the performance of all methods increased with the size of the customer base. This relation was less strong for ANNs than for CTs and LR, especially when data complexity was high. As a consequence ANNs outperformed the other methods when sample size was small, but CTs and LR yielded better results when sample size was large — with LR being generally superior to CTs. The combination of the prediction scores of ANNs, CTs and LR into a single model revealed synergistic effects among the three modelling approaches. The combination mostly resulted in better results than any single model. This study shows that ANNs may be especially valuable for small customer bases, but might not be used in isolation for analysing larger customer bases. Irrespective of the size of the customer base and the underlying data complexity, the combination of ANNs, CTs and LR into a single model mostly resulted in the best prediction, suggesting that model combination might be a safe way of maximising predictive performance when the degree of data complexity is unknown (as is the case for most real customer bases).

Roland Linder
Institute of Medical Informatics,
University of Lübeck,
Ratzeburger Allee 160,
D-23538 Lübeck, Germany.

Tel: +49 451 5006633;
Fax: +49 451 556610;
e-mail:
linder@imi.uni-luebeck.de

INTRODUCTION

Customers often face an overwhelming range of available products for satisfying any given need. They are therefore in a position to choose, which puts the producing companies into intense competition for standing out and attracting the customers' attention. The most direct way to achieve this is through direct communication with a customer. In direct marketing households, persons or customers are individually targeted for special offers, announcements of new products, etc.[1] Individual communication is generally expensive, and targeting persons/households in the market — or customers in the customer base — with high interest in an offer is a non-trivial task. To this end, companies now widely collect information about their customers' purchase behaviours because past purchase behaviour may allow one to predict to some extent future purchase behaviour.[2,3] Given the huge amount of customer information that often accumulates, data mining is required to analyse the data and to extract the relevant information from the customer database. Data mining in general, and predictive statistical models specifically, may be useful in any of the scenarios encountered in direct marketing, such as: the segmentation of the customer base into groups of customers requiring distinct direct marketing treatment, for the acquisition of new customers, for the optimisation of the response to mailings, for cross- and/or upselling etc.

Various statistical modelling methods for predicting customer purchase behaviour such as artificial neural networks, (ANNs) classification trees (CTs) and regression models are available today.[4] They all have in common that they allow one to search for significant association rules between a purchase behaviour of interest (eg response to mailings) on the one hand, and purchase history, geographic and sociodemographic information on the other hand. This rule can be used for projecting a purchase behaviour score (eg an estimated response probability) on other customers or on households/persons not in the customer base. Customers/households/persons can subsequently be targeted based on this score, which reflects a prediction for the purchase behaviour of interest.

The modelling methods differ, however, in how the association rules are derived from the data. An important consequence of these differences in approach is that different modelling methods may not necessarily be equally powerful for any given dataset at hand.

The decision as to which statistical method to use is often a rather arbitrary one (eg based on accessibility and/or experience), and the relative strengths and weaknesses of the different statistical approaches are often poorly known and/or considered. Here the authors present a comparison of the relative predictive performances of ANNs, CTs and LR models in a simulated direct marketing scenario. The comparisons are made for different levels of data complexity and sample size in order to assess the generality of potential differences between the three modelling approaches. Finally, an analysis of a combination of the three approaches is presented suggesting that the combination of different modelling approaches may often result in synergistic effects, increasing the predictive performance beyond the level of any single approach.

METHODS

Scenario

Rather than basing the analyses on a real customer base, the authors chose to use simulated data to investigate the differences in predictive performance

among ANNs, CTs and LR models. Simulated data had several advantages over real data for the purpose of this study. The individual relationships between the response variable and the explanatory variables and, thus, the complexity underlying the 'true' association patterns in a given database, could be controlled. This was critical for the main study aim of performing model comparisons in an experimental design including various levels of data complexity and sample size (see below). Also, not only the relative performance of the different statistical modelling approaches could be assessed, but also how closely the predictions came to the underlying true patterns (the simulation rule underlying a database gives the maximum achievable model-fit). On the other hand, however, the authors are conscious of the necessarily 'artificial' nature of simulated data; because simulated data may not achieve the complexities of real world data; the findings relating the predictive performance of statistical models to data complexity may be rather conservative.

For the simulated data, a direct marketing scenario where prospects with high response potential are supposed to be selected for a particular mailing from a large pool of customers/persons was considered. To this end, the response of previously targeted persons/customers was analysed with respect to purchase history as well as sociodemographic and geographic information.[5,6] The response variable is, therefore, of nominal nature; ie a previously targeted person/customer either responded or did not respond to the mailing. The association rules between the nominal response variable and the explanatory variables derived by the statistical models would then be used for targeting the persons/customers in the available pool of persons/customers. Those with the highest model scores (ie response potential) would be targeted in the next mailing.

Simulation

For these databases, the sociodemographic/geographic data was taken from a random sub-sample of an anonymous data collection of a real Swiss population. The purchase history data was simulated and consisted of the following four variables: time of first purchase, monetary turnover, purchase frequency and recency of last purchase. These variables were created by use of random-number generators built into SAS statistical software.[7] In order to render the distributions of the simulated purchase history variables less well-behaved than the output from the random number generator — and thus more challenging for the later analysis — multiple random distributions (normal and/or lognormal) were mixed to produce the final distribution of the variables. The distributions were created so far as possible as to match the authors' experience with real patterns.

Experimental design

The second step in data simulation was the creation of non-random association patterns among the nominal response variable and the explanatory variables. The experimental design consisted of two levels of data complexity and four levels of sample size per level of complexity (Table 1). In the case of low complexity, the relationships between the response variable and the explanatory variables were taken to be purely linear and additive. In the high-complexity databases, about 50 per cent of the variation in the response variable was taken to be due to non-linear (polynomial) and non-additive (interactions) effects. To keep the

Table 1: Experimental design. Varying sample sizes per level of data complexity are shown

1. Low – complexity	2. High – complexity
1.1: n = 1,100	2.1: n = 1,100
1.2: n = 2,200	2.2: n = 2,200
1.3: n = 5,500	2.3: n = 5,500
1.4: n = 11,000	2.4: n = 11,000

simulated databases as natural as possible the authors followed qualitatively their experience with real customer responses to mailings for generating the association rules underlying the database.

The response to the virtual mailing was taken to be approximately 10 per cent overall, and to depend on a large number of customer traits, including their purchase history, sociodemografic traits and geographic information. Table 2 illustrates qualitatively the associations between the response and the explanatory variables underlying the simulated customer databases. The illustration in Table 2 is limited to the 'low- complexity' experimental treatments because the description of interaction effects — especially among traits on a nominal scale — are very tedious and, in the authors' opinion do not add to the understanding of the simulation. As to the 'high-complexity' treatment, linear components are kept qualitatively the same, but non-linear effects — including polynomial and interaction effects — are added. For example, following traditional recency, frequency, monetary (RFM)-thinking,[8] it was assumed that the response increased linearly with the frequency of past purchases in the low-complexity treatments (Table 2). In the high-complexity treatments, however, it involved both a curved association and a dependency on the recency of those purchases. The frequency of recent purchases had a stronger impact on the response than the frequency of ancient purchases.

Each modelling method (ANN, CT, LR) was applied to the eight experimental groups listed in Table 1. Models were built on randomly selected training data (of sample size

Table 2: Traits and simulation patterns underlying response behaviour in our artificial customer database. Please note that we restrict the description of simulation patterns to the main patterns for the sake of clarity. The details may be obtained from the authors on request

Trait	Data-level	Axis	Pattern
A) Purchase behaviour:			
Recency	Interval	Recent – old	
Frequency	Interval	Seldom – often	
Monetary (ie turnover)	Interval	Little – lots	
Time since first contact	Interval	Recent – old	
B) Sociodemography:			
Age–class (10 year)	Nominal/ordinal	Young – old	
Gender	Nominal	Female – male	
Education	Nominal/ordinal	Low – high	
Mobility	Nominal/ordinal	Sessile – mobile	
Housing type	Nominal/ordinal	Single – multiple	
Family status	Nominal	Single – family with children	
C) Geography:			
Size of community	Nominal/ordinal	Small – large	
Industrialisation	Nominal/ordinal	None – strong	

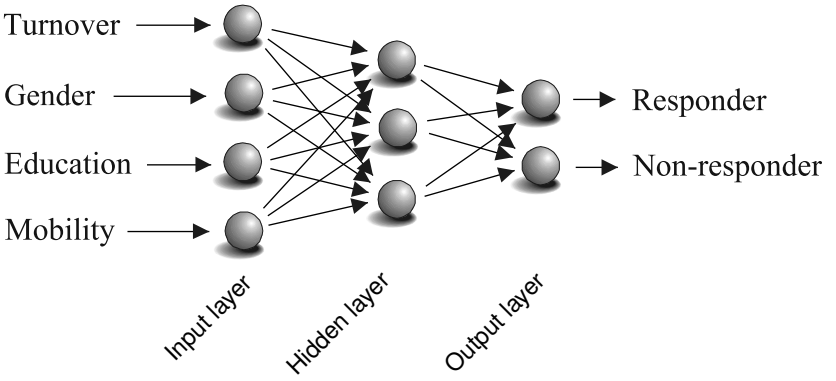

Figure 1: Structure of a standard feed-forward network that could serve for the classification between responders and non-responders (example)

corresponding to the experimental group; see Table 1) and validated on an additional randomly-selected database simulated in the same way and of similar size to the training data.

ANNs

ANNs are computer-based techniques which are frequently used in direct marketing and adjacent fields.[9–16]. ANNs use nonlinear mathematical equations to successively develop meaningful relationships between input and output variables through a learning process. They have a 'training phase' and a 'recall phase'. In the training phase, the relationships between the different input and output variables are established by adaptations of the weight factors assigned to the connections between the layers of artificial neurons. This adaptation is based on rules that are set in the learning algorithm. At the end of the learning process, the weight factors are fixed. In the recall phase, data from cases not previously interpreted by the network are entered and an output is calculated based on the aforementioned, and now fixed, weight factors. Figure 1 gives a diagrammatic representation of a standard feed-forward network. Data are entered at the input neurons and further processed in the hidden layer and output layer.

The simulation features were fed into a feed-forward network implemented as a prototype called 'Approximation and Classification of Medical Data' (ACMD) (Microsoft Visual C++, Office Software International, USA),[17] which is, in principle, based on a multilayer perceptron,[18] using adaptive propagation[19] for automatic training. Group of five networks were arranged as an ensemble, which makes them more robust than single networks and more suitable for training with only a small amount of data.[20] To make the training less susceptible to so-called overfitting,[21] a strategy known as 'early stopping' was used.[22] ACMD comprises a number of further strategies improving both the generalisation performance and accelerating the convergence speed. Several benchmarks give evidence that ACMD provides an excellent tool for training feed-forward networks. For details please see reference 23. ANN's were trained without a feature selection due to the considerable computation time involved.

CT

The Chi-Squared Automatic Interaction Detector (CHAID) — a module of the SPSS analysis software (Answer Tree, SPSS Inc) — was used as a tool for

generating CTs to differentiate and/or filter the optimal target group for campaigns. The principal purpose of the CHAID analysis is to select the customers (or persons) with the highest response probability for a mailing. An additional advantage is that the results can be easily visualised using a tree diagram.

The database comprises the characteristics of the customer; the internal variables (recency, frequency, monetary) and the external variables (eg sociodemographic characteristics). The dependent variable is the focused target group — this group will be explained with the independent variables.

The procedure is to initially find the greatest distinction of the explanatory variables in relation to the dependent variable based on the Chi-squared test. The variable with the greatest difference forms the first level and the particular nodes are the values (eg male/female) which are individually significant. This step will be repeated within each node and at each level until no significant (merge level = 0.05) differences among sub-groups can be found or the sub-groups are too small for further splitting.

To compare or to evaluate each model SPSS generates a score, which corresponds in this scenario to the proportion of test-mailing responders within each subgroup. This score will be computed for every record (for each individual customer) and indicates the affinity to the defined target group.

LR

Logistic models can be considered a special class of linear regression models. The binary response variable violates normality assumptions of general regression models, however. A link function linearising parameter estimates and a binomial error definition is required to properly incorporate the binary nature of the response variable. The parameter estimates for each explanatory variable are found by iterative maximum likelihood algorithms using reweighted least-squares (see, eg reference 24). Logistic regression analysis was carried out using SAS Statistical Software, PROC LOGISTIC program. The model was set up with a logit link and binomial errors.

A large number of socioeconomic and purchase variables and an arbitrary choice of interaction effects were initially entered as explanatory variables. The automatic step-wise model selection option available in PROC LOGISTIC was used to reduce model complexity towards a final model containing statistically significant ($p < 0.05$) variables only. The parameter estimates from the final model were then used to project the predicted response probabilities on the validation data. Using parameter estimates from logistic models for prediction yields the probability scores on the linear scale. To obtain the predictions as a probability score bound between 0 and 1, the linearised predictions have to be transformed back by use of the inverse logit function, which is

$$\hat{y}_p = \frac{1}{(1 + e^{-\hat{y}_l})} \quad (1)$$

(see, eg SAS manual, 1999). $\hat{y}$ is the prediction from the model, the subscript l indicates the linear scale and the subscript p the probability scale.

Combining predictions from the different modelling methods

By combining the predictions of the three modelling methods studied here, the authors tested for synergistic effects where the combination of the three approaches might yield a better prediction than any single method as a stand-alone. Their approach to

combining the predictions was to use the predictions from the ANN, CT and LR as independent variables in a subsequent regression analysis.

Assessing predictive performance: Hit probabilities and receiver operating characteristics analysis

One crude way to judge the predictive ability of the models for this specific goal is simply to calculate the number of real responders (ie hits) among the 50 per cent of the customer base with highest model scores. Because this figure provides a crude estimate of model fit only, and is only valid for the specifics of the mentioned scenario, the authors also performed a receiver operating characteristic (ROC) analysis providing more general and detailed information. A ROC curve is obtained by systematically varying the threshold between both classes (responders, non-responders) and plotting the true-positive fraction (TPF, Equation 2) against the false-positive fraction (FPF, Equation 3), as obtained from the model scores (see Figure 2).[24] Afterwards, the area under the curve (AUC) value can be derived through numeric integration and used as a reliable characteristic for the exactness of fit. An AUC value near 0.5 suggests no discrimination, ie, one might as well flip a coin. An AUC value close to 1.0 is considered outstanding discrimination.[26] The ROC curve displays information of model fit that is comparable to the information contained in lift/gain charts. (See, for example references 27 and 28 for examples of lift charts.)

$$\text{True-positive fraction (TPF)} = \frac{\text{No. true predicted responders}}{\text{No. actual responders}} \quad (2)$$

$$\text{False positive fraction (FPF)} = \frac{\text{No. false predicted responders}}{\text{No. actual non-responders}} \quad (3)$$

RESULTS

Predictive performance

Under the low-complexity scenarios, the analysis of hit probabilities did not reveal clear cut tendencies for differences among modelling methods (Figure 3a). For all methods, a rough general positive relation between hit probability and sample size was observed. More subtle differentiation is provided by the analysis of AUC values (Table 3). Using AUC-values as a criterion for the assessment of predictive performance, ANN tended to outperform both CT and LR when sample sizes were low and LR produced the best result when sample sizes were large.

As to the high-complexity scenarios, the analysis of hit probabilities showed that ANN outperformed both CT and LR in predictive performance when customer bases were small (1,100 and 2,200). This effect seemed to be mostly due to the low performance of CT and LR models when sample sizes were low and an increase in predictive performance of CT and LR models with sample size. Hit probabilities for both CT and LR increased gradually from the $n = 1{,}100$ to the $n = 11{,}000$ scenario, while hit probabilities of ANN's were pretty much independent of sample size. This basic pattern of differences among modelling techniques was confirmed in the AUC analysis (Table 3).

Combining the predictions of the three approaches into a single regression analysis revealed synergistic effects among ANN, CT and LR. A synergistic effect is present if the prediction of the model combination outperforms each single model. In terms of hit probabilities, a synergistic effect could be detected for the high-complexity scenario with 11,000 customers. Synergistic effects may be too subtle to detect with the rough assessment method of hit probabilities, however.

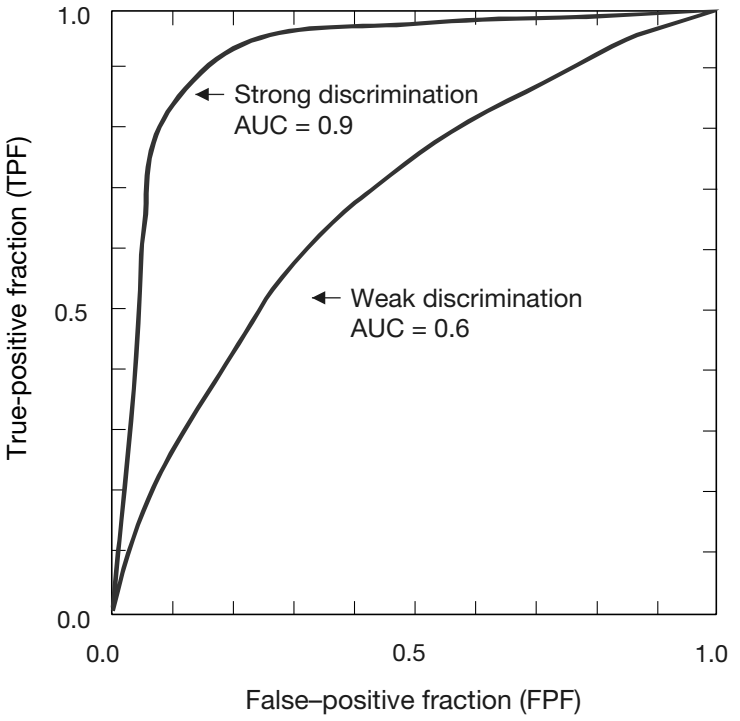

Figure 2: Receiver operating characteristics (ROCs) for the differentiation between responders and non-responders. True-positive fraction (TPF) indicates the percentage of correctly classified responders, false-positive fraction (FPF) is the percentage of misclassified non-responders

Table 3: Receiver operating characteristic analyses of the validation data: area under the curve values between 0.5 and 1.0 (the higher the values, the better the prediction). 'Simulation' means a prediction based on the underlying rules of generating the scenario, ie the maximum achievable model-fit

Scenario			Approach			
Complexity	Samples	Simulation	ANN	CT	LR	Combination
Low	1,100	0.742	0.704	0.647	0.695	0.697
Low	2,200	0.742	0.708	0.657	0.703	0.709
Low	5,500	0.748	0.731	0.705	0.731	0.734
Low	11,000	0.763	0.738	0.744	0.754	0.757
High	1,100	0.818	0.770	0.677	0.750	0.777
High	2,200	0.824	0.789	0.746	0.782	0.798
High	5,500	0.834	0.788	0.744	0.812	0.802
High	11,000	0.827	0.787	0.814	0.818	0.828

Indeed, in terms of AUC values the model combination outperformed all single model approaches in all but one case (see Table 3: exception is 'high-complexity, 5,500 samples').

Overfitting

As described in the Method section, all models were built on a training dataset and validated on an equally-sized validation dataset. The analysis of predictive performance (see above) was based on the performance of the models on the validation data only. Situations may occur where not enough data may be available to split a customer base into a training and a validation dataset. In such a situation, it is important to know which modelling technique is least susceptible to

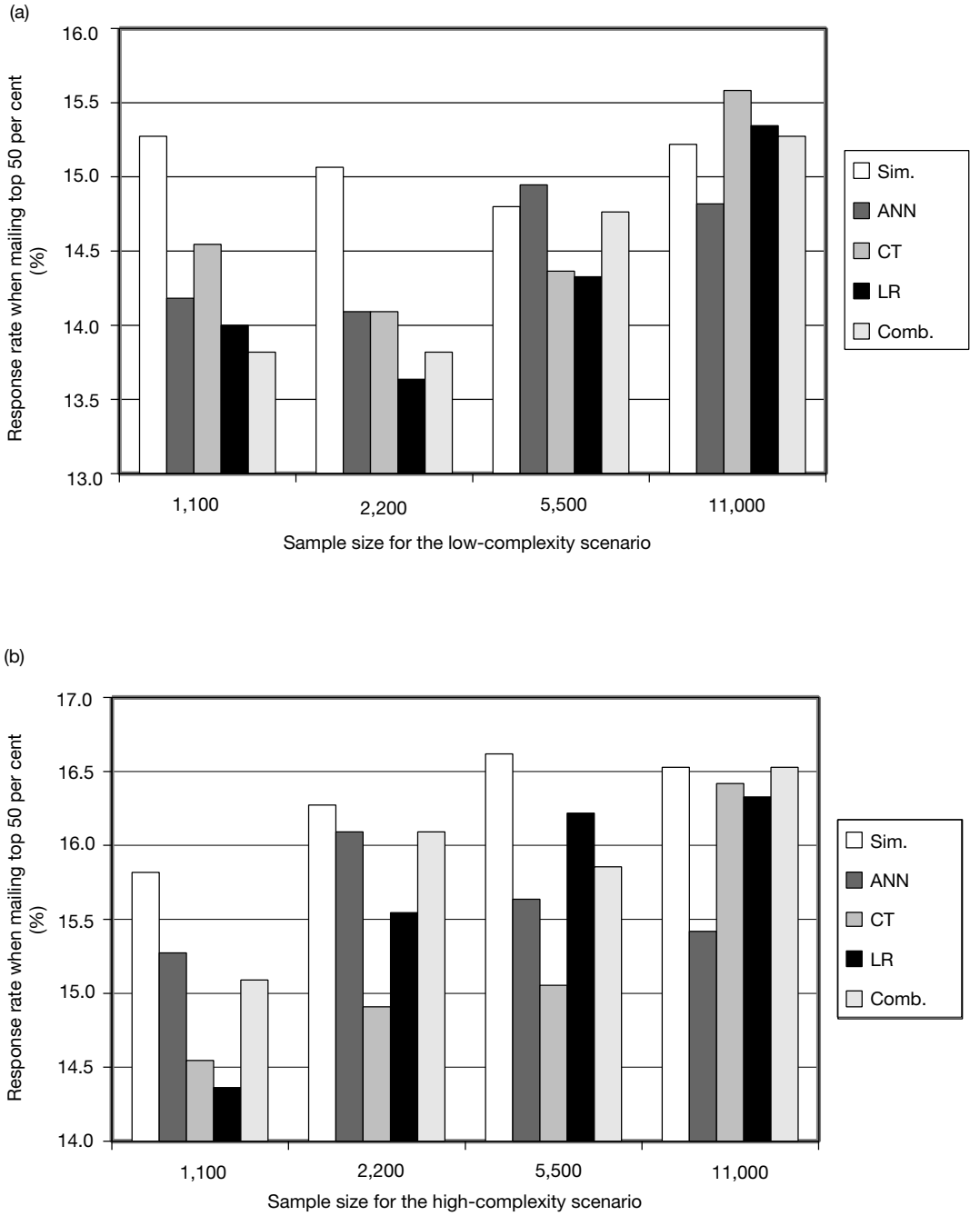

Figure 3: Hit probability charts for (a) the low-complexity scenarios and (b) the high-complexity scenarios when mailing the top 50 per cent most promising customers.
'Sim.' means the simulation based on the underlying rules of generating the scenario, ie the maximum achievable model-fit; 'Comb.' is the prediction when combining the results of ANN, CT and LR. Note that the overall response rate in the simulation was 10 per cent.

Table 4: Determination of overfitting by the difference $ROC_{validation\ data} - ROC_{training\ data}$ (low values mean little overfitting). 'Simulation' means a prediction based on the underlying rules of generating the scenario, ie the maximum achievable model-fit

Scenario			Approach			
Complexity	Samples	Simulation	ANN	CT	LR	Combination
Low	1,100	0.013	0.062	0.102	0.062	0.079
Low	2,200	0.028	0.057	0.100	0.073	0.081
Low	5,500	0.017	0.029	0.075	0.042	0.050
Low	11,000	0.006	0.029	0.022	0.024	0.097
High	1,100	0.000	0.022	0.083	0.075	0.053
High	2,200	0.000	0.005	0.034	0.037	0.027
High	5,500	–0.018	–0.004	0.063	0.015	0.035
High	11,000	–0.008	–0.010	–0.019	0.009	–0.005

overfitting resulting in biased models and predictions. Confirming a general expectation, the degree of overfitting mostly, and for all modelling approaches, decreased with increasing size of the customer base (Table 4). CT and LR showed more stability problems than ANN when the size of the customer base was small. When sample sizes were large, the stabilities of ANN, CT and LR were comparable. The stability of the model combination was generally intermediate to the single model approaches (Table 4), with the notable exception of one case where the combination resulted in more overfitting than in any of the single approaches (see Table 4: 'low-complexity, 11,000 samples').

DISCUSSION

Although a number of studies drawing a comparison between different modelling methodologies already exist,[29–33] as far as the authors are aware, this is the first time that the interdependencies between a certain method on one hand and the level of sample size and data complexity on the other hand have been investigated in a systematic way. This subject could only be investigated sufficiently by simulating the response behaviour of the customers in a (virtual) test mailing. So the authors were allowed not only to control the complexities of the underlying 'true' patterns, but also to assess how close the different model predictions come to the maximum achievable model-fit.

Overall, all three methods investigated — artificial neural networks (ANN), classification trees (CTs) and logistic regression (LR) — achieved surprisingly good results compared to the achievable model-fit. It stays open whether the methods were so good or the simulated data too easy. But, despite the general comparability between the results, some useful conclusions can be drawn with respect to differences in performance among the three modelling methods.

ANN

Contrary to expectation, the relative usefulness of ANN was demonstrated to be independent from training by data of linear or nonlinear complexity. In both cases the strength of the ANN was a prediction based on a fairly small sample size for training, 1,100 or 2,200 customers, respectively, in the present study. For the ROC analysis as well as for predicting the hit probability, the ANN outperformed the more traditional approaches. Moreover, particularly for the small sized situations, ANN results for the training data were demonstrably closer to the validation results than could be considered for the other methods.

Thus, at low sample sizes, where the splitting of a customer base into training and validation data becomes difficult, ANNs might be the approach of choice: ANN outperformed both CT and LR in predictive accuracy and simultaneously suffered least from overfitting. Contrary to popular belief, it did not take much effort to build them when using a modern, far-reaching, automatic learning software tool like ACMD. One might speculate if the advantage of ANNs might be even more obvious when using a feature selection (as done by CT and LR). By setting weights of redundant or irrelevant inputs to zero, however, ANNs are surprisingly capable of ignoring irrelevant explanatory variables even without feature selection.

CT/LR

CT/LR achieved the best results when specifying these predictors with 5,500 or 11,000 customers. According to the ROC analysis, the ROC curves for CHAID comprised fewer coordinates than the other methods due to the lower granularity of the CHAID model. This circumstance might possibly have influenced the AUC calculation, leading to slightly lower AUC values. In contrast to ANN, CHAID and LR represent explicit knowledge on relevant explanatory variables after model specification, so they are suitable for large databases and are useful for transparent interpretations of the results. The observation that both CT and LR showed a stronger dependency of predictive performance on the size of the customer base used for training than ANN might also be of interest.

Model combination

Combining the prediction results using a subsequent regression analysis, revealed synergistic effects among modelling approaches, as expected if the different statistical approaches used in ANNs, CTs and LRs to derive association rules among the response and the explanatory variables extract somewhat different patterns from the data. The predictions of the combination model achieved results comparable to, or mostly better than, the best single results. The detected synergistic effects may be of a rather subtle nature, however, because they turned up in a systematic way in the more accurate AUC analysis only, and not as clearly in the hit probability analysis. The stability of the combined model was generally in the range of the single models, but decreased in one case (see Table 4). This suggests that model combination may sometimes reduce the stability of the prediction.

The as-accurate-as-possible prediction of customer behaviour is critical for the efficient targetting in direct marketing. For an objective and reliable comparison of ANNs, CTs and LRs in their predictive performance, the models were developed 'blindly', ie excluding the modelist's experience and knowledge to influence the outcome. Of course, the modelist's knowledge in a particular direct marketing situation is, in reality, of utmost importance in the choices of explanatory variables when building a model. Acquisition of such knowledge implies that inferences can be drawn from the carefully built models. ANNs, CTs and LRs differ in the ease with which inferences about the relationships between the response and the explanatory variables may be drawn. CTs are probably most appealing for an intuitive and visual understanding of the relations involved. Inferences from LRs can be gained by interpreting the regression parameter estimates, which may require some statistical experience. Finally, ANNs are often thought of as a

'black-box' for inference, because information about specific relations between the response and the explanatory variables are not easily obtained and displayed. However, some promising neural rule extraction algorithms exist to explain why a specific class label was assigned to a certain case.[34–35] Another criticism uttered against ANN is the extensive computational burden. In the present study, using a commercial notebook (2.8 GHz), learning with 1,100 samples took 19 minutes. Applying the trained ANN takes just a few seconds.

CONCLUSIONS

These analyses show that ANNs may be powerful tools for modelling customer behaviours when available sample sizes are small. They outperformed CTs and LR in both accuracy and stability. Because of an overall increase in the predictive performance of CTs and LRs with sample size, which was not observed equally for ANNs, these two methods should be used when sample sizes become larger. It would be interesting to incorporate additional modelling approaches, like genetic algorithms or case-based reasoning, in similar future comparisons. Moreover, the simulated scenario should be extended to predicting continuous variables instead of categories. All three methods employed here are capable of modelling continuous response variables. The authors also demonstrate the potential benefits of combining different modelling approaches into a single model. Model combination mostly increased predictive performance beyond any single approach, and model stability generally remained comparable. Model combination might, therefore, often be a safe approach for maximising predictive performance when the complexity underlying the data at hand is unknown. To the authors, given the computing power generally available today, the simultaneous use of more than one modelling method seems practicable — at least for relatively small databases.

This paper is mostly concerned with the predictive performance of ANNs, CTs and LRs. An additional criterion for choosing a particular method might also be the ease with which it can be visualised, intuitively understood and thus explained to statistically non-trained marketers in the field. Predictive performance for targetting customers versus inferences/interpretations drawn from a model may be considered different objectives often with different optimal statistical solutions. Again, using more than one method simultaneously may help to get both maximal predictive performance (eg ANNs or model combination for small sample sizes) and inferences that can be visualised (CTs or LRs).

Acknowledgments

The authors thank dm-plus Direktmarketing AG for providing an anonymous, random sample of sociodemographic information from the Swiss population.

References

1 Hughes, A. M. (2000) 'Strategic Database Marketing: The Masterplan for Starting and Managing a Profitable Customer-Based Marketing Programme', 2nd edn, McGraw-Hill, New York, NY.
2 *Ibid*.
3 David Shepard Associates (1999) 'The New Direct Marketing, How to Implement a Profit-Driven Database Marketing Strategy', 3rd edn, McGraw-Hill, New York, NY.
4 *Ibid*.
5 Hughes (2000) *op. cit*.
6 David Shepard Associates (1999) *op cit*.
7 SAS Institute Inc (1999) SAS for Windows, version 8.02 *SAS Institute*, Cary, NC.
8 Hughes (2000) *op cit*.
9 Baesens, B., Setiono, R., Mues, C. and Vanthienen, J. (2003) 'Using neural network rule extraction and decision tables for credit-risk evaluation', *Management Science*, Vol. 49, No. 3, pp. 312–329.

10 Baesens, B., Verstraeten, G., Van den Poel, D. *et al.* (2003) 'Bayesian network classifiers for segmented customer relationship management strategies', *European Journal of Operational Research*, Vol. 114, pp. 346–353.
11 Baesens, B., Viaene, S., Van den Poel, D. *et al.* (2002) 'Using Bayesian neural networks for repeat purchase modelling in direct marketing', *European Journal of Operational Research*, Vol. 138, No. 1, pp. 191–211.
12 Fish, K. E., Barnes, J. H. and Aiken, M. (1995) 'Artificial neural networks: A new methodology for industrial market segmentation', *Industrial Marketing Management*, Vol. 24, pp. 431–438.
13 Pan, Z., Liu, X. and Mejabi, O. (1997) 'A neural-fuzzy system for forecasting', *Journal of Computational Intelligence in Finance*, Vol. 5, No. 1, pp. 7–15.
14 Viaene, S., Baesens, B., Van den Poel, D. *et al.* (2001) 'Wrapped input selection using multilayer perceptrons for repeat- purchase modeling in direct marketing', *International Journal of Intelligent Systems in Accounting, Finance and Management*, Vol. 10, No. 2, pp. 115–126.
15 Wilson, R. and Sharda, R. (1994) 'Bankruptcy prediction using neural networks', *Decision Support Systems*, Vol. 11, No. 3, pp. 545–557.
16 Zahavi, J. and Levin, N. (1997) 'Applying neural computing to target marketing', *Journal of Direct Marketing*, Vol. 11, No. 1, pp. 5–22.
17 Linder, R. and Pöppl, S. J. (2001) 'ACMD: A practical tool for automatic neural net based learning', *Lecture Notes in Computer Science*, Vol. 2199, pp. 168–173.
18 Rumelhart, D. E., Hinton, G. E. and Williams, R. J. (1986) 'Learning representations by back-propagating errors', *Nature*, Vol. 323, pp. 533–536.
19 Linder, R., Wirtz, S. and Pöppl, S. J. (2000) 'Speeding up backpropagation learning by the APROP algorithm', in Bothe, H. and Rojas, R., *Proceedings of the Second International ICSC Symposium on Neural Computation (NC 2000)*, ICSC Academic Press, Millet Alberta, Canada, pp. 122– 128.
20 Bishop, C. M. (1995) 'Neural Networks for Pattern Recognition', Clarendon Press, Oxford, UK.
21 Amari, S., Murata, N., Müller, K. *et al.* (1997) 'Asymptotic statistical theory of overtraining and cross-validation', *IEEE Transactions On Neural Networks*, Vol. 8, No. 5, pp. 985–986.
22 Prechelt, L. (1998) 'Automatic early stopping using cross validation: Quantifying the criteria', *Neural Networks*, Vol. 11, pp. 761–767.
23 Linter and Pöppl (2001) *op cit.*
24 SAS Institute Inc (1999) *op cit.*
25 Metz, C. E. (1978) 'Basic principles of ROC analysis', *Seminars Nucl. Med.*, Vol. 8, No. 4, pp. 283–298.
26 Hosmer, D.W. and Lemeshow, S. (2001) 'Applied Logistic Regression'. 2nd edn, John Wiley & Sons, New York, NY.
27 Potharst, R., Kaymak, U. and Pijls, W. H. L. M. (2001) *Neural Networks for Target Selection in Direct Marketing*, Technical Report ERS-2001-14-LIS, Erasmus University, Rotterdam, the Netherlands.
28 Kölliker, M. (2001) 'Zielgenau dank Data Mining', *Marketing & Kommunikation*, Vol. 5, pp. 38–40.
29 Balakrishnan, P., Cooper, M., Jacob, V. and Lewis, P. (1996) 'Comparative performance of the FSCL neural net and K-means algorithm for market segmentation', *European Journal of Operation Research*, Vol. 93, No. 10, pp. 346–357.
30 Hruschka, H. (1993) 'Determining market response functions by neural network modeling. A comparison to econometric techniques', *European Journal of Operational Research*, Vol. 66, No. 1, pp. 27– 35.
31 Hruschka, H. and Natter, M. (1999) 'Comparing performance of feedforward neural nets and K-means for cluster-based market segmentation', *European Journal of Operational Research*, Vol. 114, pp. 346– 353.
32 Saad, E. Prokhorov, D. and Wunsch, D. (1998) 'Comparative study of stock trend prediction using time delay, recurrent and probabilistic neural networks', *IEEE Transactions on Neural Networks*, Vol. 9, No. 6, pp. 1456–1470.
33 Wiedmann, K.-P. and Buckler, F. (2001) *Neuronale Netze im Marketing-Management*, Gabler, Wiesbaden, Germany.
34 Duch, W., Adamczak, R. and Grabczewski, K. (1998) 'Extraction of logical rules from neural networks', *Neural Processing Letters*, Vol. 7, pp. 211–219.
35 Vaughn, M. L. (1999) 'Derivation of the multilayer perceptron weight constraints for direct network interpretation and knowledge discovery', *Neural Networks*, Vol. 12, pp. 1259–1271.
36 Setiono, R. (2000) 'Generating concise and accurate classification rules for breast cancer diagnosis', *Artificial Intelligence in Medicine*, Vol. 18, No. 3, pp. 205–219.