Available online at www.sciencedirect.com
SCIENCE DIRECT®
European Journal of Operational Research 173 (2006) 781–800
EUROPEAN JOURNAL OF OPERATIONAL RESEARCH
www.elsevier.com/locate/ejor

The impact of preprocessing on data mining: An evaluation of classifier sensitivity in direct marketing

Sven F. Crone [a], Stefan Lessmann [b,*], Robert Stahlbock [b]

[a] Department of Management Science, Lancaster University, Lancaster LA1 4YX, United Kingdom
[b] Institute of Information Systems, University of Hamburg, Von-Melle-Park 5, 20146 Hamburg, Germany

Received 15 November 2004; accepted 18 July 2005
Available online 15 November 2005

Abstract

Corporate data mining faces the challenge of systematic knowledge discovery in large data streams to support managerial decision making. While research in operations research, direct marketing and machine learning focuses on the analysis and design of data mining algorithms, the interaction of data mining with the preceding phase of data preprocessing has not been investigated in detail. This paper investigates the influence of different preprocessing techniques of attribute scaling, sampling, coding of categorical as well as coding of continuous attributes on the classifier performance of decision trees, neural networks and support vector machines. The impact of different preprocessing choices is assessed on a real world dataset from direct marketing using a multifactorial analysis of variance on various performance metrics and method parameterisations. Our case-based analysis provides empirical evidence that data preprocessing has a significant impact on predictive accuracy, with certain schemes proving inferior to competitive approaches. In addition, it is found that (1) selected methods prove almost as sensitive to different data representations as to method parameterisations, indicating the potential for increased performance through effective preprocessing; (2) the impact of preprocessing schemes varies by method, indicating different 'best practice' setups to facilitate superior results of a particular method; (3) algorithmic sensitivity towards preprocessing is consequently an important criterion in method evaluation and selection which needs to be considered together with traditional metrics of predictive power and computational efficiency in predictive data mining.

Keywords: Data mining; Neural networks; Data preprocessing; Classification; Marketing

1. Introduction

In competitive consumer markets, data mining faces the growing challenge of systematic knowledge discovery in large datasets to achieve

* Corresponding author. Tel.: +49 40 42838 5500; fax: +49 40 42838 5535.
E-mail addresses: s.crone@lancaster.ac.uk (S.F. Crone), lessmann@econ.uni-hamburg.de (S. Lessmann), stahlboc@econ.uni-hamburg.de (R. Stahlbock).

0377-2217/$ - see front matter
doi:10.1016/j.ejor.2005.07.023

operational, tactical and strategic competitive advantages. As a consequence, the support of corporate decision making through data mining has received increasing interest and importance in operational research and industry. As an example, direct marketing campaigns aiming to sell products by means of catalogues or mail offers [1] are restricted to contacting a certain number of customers due to budget constraints. The objective of data mining is to select the customer subset most likely to respond in a mailing campaign, predicting the occurrence or probability of purchase incident, purchase amount or interpurchase time for each customer [2,3] based upon observable customer attributes of varying scale. Traditionally, response modelling has utilised transactional data consisting of continues variables to predict purchase incident focusing on the recency of the last purchase, the frequency of purchases and the overall monetary purchase amount, referred to as recency, frequency and monetary value (RFM)-analysis [2]. The continuous scale of these attributes together with their limited number has facilitated the use of conventional statistical methods, such as logistic regression.

Recently, progress in computational and storage capacity has enabled the accumulation of ordinal, nominal, binary and unary demographic and psychographic customer centric data, inducing large, rich datasets of heterogeneous scales. On the one hand, this has advanced the application of data driven methods like decision trees (DT) [4], artificial neural networks (NN) [2,5,6], and support vector machines (SVM) [7], capable of mining large datasets. On the other hand, the enhanced data has created particular challenges in transforming attributes of different scales into a mathematically feasible and computationally suitable format. Essentially, each customer attribute may require special treatment for each algorithm, such as discretisation of numerical features, rescaling of ordinal features and encoding of categorical ones. Applying a variety of different methods, the phase of data preprocessing (DPP) represents a complex prerequisite for data mining in the process of knowledge discovery in databases [8].

Aiming to maximise the predictive accuracy of data mining, research in management science and machine learning is largely devoted to enhancing competing classifiers and the effective tuning of algorithm parameters. Classification algorithms are routinely tested in extensive benchmark experiments, evaluating the impact on predictive accuracy and computational efficiency, using preprocessed datasets; e.g. [9–11]. In contrast to this, research in DPP focuses on the development of algorithms for particular DPP tasks. While feature selection [12–14], resampling [15,16] and the discretisation of continuous attributes [17,18] are analysed in some detail, few publications investigate the impact of data projection for categorical attributes and scaling [19,20]. More importantly, interactions on predictive accuracy in data mining are not been analysed in detail, especially not within the domain of corporate direct marketing.

To narrow this gap in research and practice, we seek to investigate the potential of DPP in a real world scenario of response modelling, predicting purchase incident to identify those customers most likely to respond to a mailing campaign in the publishing industry. We analyse the impact of different DPP schemes across a selection of established data mining methods. Due to the questionable usefulness of traditional statistical techniques in large scale data mining settings [21,22] and mixed scaling levels of customer attributes, we confine our analysis to data driven methods of C4.5 DT, NN and SVM.

The remainder of the paper is organised as follows: We begin with a short overview of the classification methods of DT, NN and SVM used. Next, the task of DPP for competing methods for scaling, sampling and coding is discussed in Section 3. Conducting a structured literature review, we exemplify that the influence of DPP is widely overlooked to motivate our further analysis. This is followed by the case study setup of purchase incident modelling for direct marketing in Section 4 and the experimental results providing empirical evidence for the significant impact of DPP on classification performance in Section 5. Conclusions are given in Section 6.

2. Classification algorithms for data mining

2.1. Multilayer perceptrons

NN represent a class of statistical methods capable of universal function approximation, learning non-linear relationships between independent and dependent variables directly from the data without previous assumptions about the statistical distributions [23]. Multilayer perceptrons (MLP) represent a prominent class of NN [24–26], implementing a paradigm of supervised learning methods which is routinely used in academic and empirical classification and data mining tasks [27–29].

The architecture of a MLP, as shown in Fig. 1, consists of several layers of nodes u_j fully interconnected through weighted acyclic arcs w_{ij} from each preceding layer to the following, without lateral connections or feedback [27]. The information is processed from left to right, using nodes in the input layer to forward input vector information to the hidden layer. Each hidden node j calculates a weighted linear combination $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{o}$ of its input vector $\boldsymbol{o}$, weighting each input activation o_i of node i in the preceding layer with the transposed matrix $\boldsymbol{w}^{\mathrm{T}}$ of the trainable weights w_{ij} including a trainable constant θ_j. The linear combination is transformed by means of a bounded, non-decreasing, non-linear activation functions in each node [21] to model different network behaviour. The processed results are forwarded to the nodes in the output layer, which compute an output vector of the classification results for each presented input pattern.

MLP learn to separate classes directly from presented data, approximating a function $g(\boldsymbol{x})$: $X \rightarrow Y$ by iteratively adapting $\boldsymbol{w}$ after presentation of an input pattern to minimise a given objective function $e(\boldsymbol{x})$ using a learning algorithm. Each node forms a linear hyperplane that partitions feature space into two half-spaces, whereby the non-linear activation function models a graded response of indicated class membership depending on the distance of $\boldsymbol{x}$ to each node hyperplane [27]. Nodes in successive hidden layers form convex regions as intersections of these hyperplanes. Output units form unisons of the convex regions into arbitrarily shaped, convex, non-convex or disjoint regions. The successive combination creates a complex decision boundary that separates feature space into polyhedral sets or regions, each one being assigned to a different class of Y. The desired output of class membership may be coded using a single output node $y_i = \{0; 1\}$ or using n nodes for multiple classifications $y_i = \{(0,1);(1,0)\}$, respectively. Moreover, the choice of the output function allows the prediction of binary class memberships as well as the more suitable conditional probability of class membership to rank each customer instance (see Section 4.3).

Being universal approximators, NN should theoretically be capable of processing any continuous input data or categorical attributes of ordinal, nominal, binary or unary scale [19] to learn any

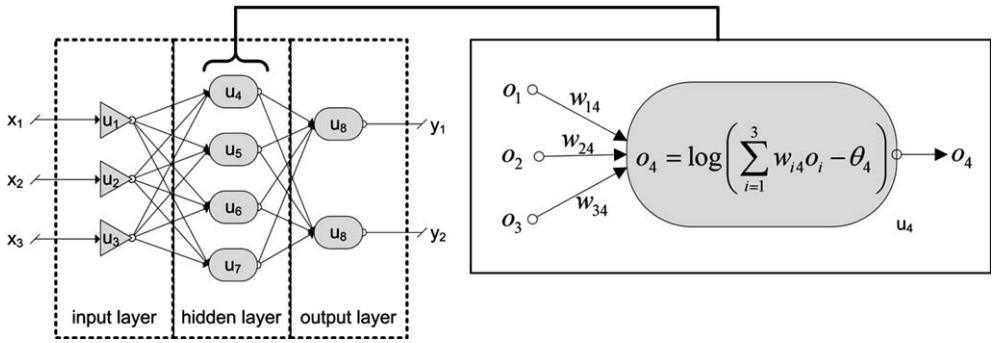

Fig. 1. Three layered MLP showing the information processing within a node, using a weighted sum as input function, the logistic function as sigmoid activation function and an identity output function.

non-linear decision boundary to a desired degree of accuracy. However, best practices suggest scaling of continuous and categorical input to $[-1; 1]$, output data to match the range of the activation functions, i.e. $[0; 1]$ or $[-1; 1]$, and avoidance of ordinal coding [19] to facilitate learning speed and robustness. Despite their significant attention and application, only limited research on the impact of DPP decisions of scaling, coding and sampling on data mining performance exists.

2.2. Decision trees

DT are intuitive methods for classifying a pattern through a sequence of rules or questions, in which the next question depends on the answer on a current question. They are particularly useful for categorical data, as rules do not require any notion of metric. A variety of different DT paradigms exists, such as ID3, C4.5, CART or CHAID. A popular approach to DT modelling induces decision trees based on the information theoretical concept of entropy [30]. Depending upon the proportion of examples of class -1 and $+1$ in the sample, a tree is split into nodes on the attribute which maximises the expected reduction of entropy. The tree is constructed with recursive partitioning of successive splits. A rule set can be formulated by derivation of a rule for each path from the tree's root to a leaf node. Due to the recursive growing strategy, DT tends to overfit the training data, constructing a complex structure of many internal nodes. Consequently, overfitting is controlled through retrospective pruning procedures for deleting redundant parts of rules [30,31]. Extending the case of binary classification, DT permit the prediction of a conditional probability of class membership using the concentration of class $+1$ records within a node as a ranking criterion. DT are robust to continuous or categorical attributes in the sense that appropriate split criteria for each scaling type exist [31].

2.3. Support vector machines

The original SVM can be characterised as a supervised learning algorithm capable of solving linear and non-linear binary classification problems. Given a training set with m patterns $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{m}$, where $\boldsymbol{x}_i \in X \subseteq \Re^n$ is an input vector and $y_i \in \{-1, +1\}$ its corresponding binary class label, the idea of support vector classification is to separate examples by means of a maximal margin hyperplane [32]. That is, the algorithm strives to maximise the distance between examples that are closest to the decision surface. It has been shown that maximising the margin of separation improves the generalisation ability of the resulting classifier [33]. To construct such a classifier one has to minimise the norm of the weight vector $\boldsymbol{w}$ under the constraint that the training patterns of each class reside on opposite sides of the separating surface (see Fig. 2). Since $y_i \in \{-1, +1\}$ we can formulate this constraint as

$$y_i((\boldsymbol{w} \cdot \boldsymbol{x}_i) + b) \geqslant 1, \quad i = 1, \ldots, m. \tag{1}$$

Examples which satisfy (1) with equality are called support vectors since they define the orientation of the resulting hyperplane.

To account for misclassifications, that is examples where constraint (1) is not met, the so called soft margin formulation of SVM introduces slack variables ξ_i [32]. Hence, to construct a maximal margin classifier one has to solve the convex quadratic programming problem (2).

$$\begin{aligned} &\min_{w,b,\xi} \frac{1}{2}\|\boldsymbol{w}\| + C\sum_{i=1}^{m}\xi_i \\ &\text{s.t.: } y_i((\boldsymbol{w} \cdot \boldsymbol{x}_i) + b) \geqslant 1 - \xi_i, \quad i = 1, \ldots, m. \end{aligned} \tag{2}$$

C is a tuning parameter which allows the user to control the trade off between maximising the mar-

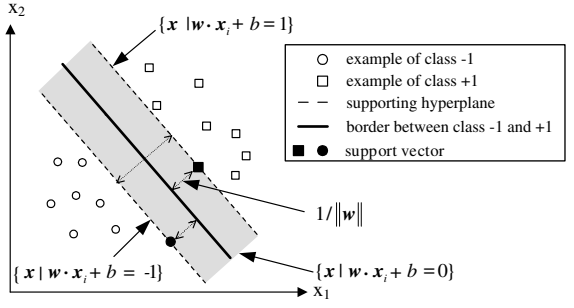

Fig. 2. Linear separation of two classes -1 and $+1$ in two-dimensional space with SVM classifier [34].

gin (first term in the objective) and classifying the training set without error. The primal decision variables $\boldsymbol{w}$ and b define the separating hyperplane, so that the resulting classifier takes the form

$$y(\boldsymbol{x}) = sgn((\boldsymbol{w}^* \cdot \boldsymbol{x}) + b^*), \quad (3)$$

where $\boldsymbol{w}^*$ and b^* are determined by (2).

To construct more general non-linear decision surfaces SVM implement the idea to map the input vectors into a high-dimensional feature space via an a priori chosen non-linear mapping function Φ. Constructing a separating hyperplane in this feature space leads to a non-linear decision boundary in the input space. Expensive calculation of dot products $\Phi(\boldsymbol{x}_i)\cdot\Phi(\boldsymbol{x}_j)$ in a high-dimensional space can be avoided by introducing a kernel function $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \Phi(\boldsymbol{x}_i)\cdot\Phi(\boldsymbol{x}_j)$ [32].

SVM requires specific postprocessing to model conditional class membership probabilities; see e.g. [35]. However, a ranking of customer instances, as is usually required in direct marketing, can be produced by removing the sign function in (3). This gives the distance of an example to the separating hyperplane which is directly related to the confidence of correct classification [35]. Therefore, customer instances that are further apart from the separating surfaces receive a higher ranking.

Research of SVM in conjunction with DPP focuses mainly on data reduction and feature selection in particular, e.g. [36–38]. While some work on the influence of scaling and discretisation of continuous attributes [39–41] exists, the effect of coding of categorical attributes has to our best knowledge not been investigated.

3. Data preprocessing for predictive classification

3.1. Current research in data preprocessing

The application of each data mining algorithm requires the presence of data in a mathematically feasible format, achieved through DPP. Consequently, DPP represents a prerequisite phase for data mining in the process of knowledge discovery in databases. DPP tasks are distinguished in data reduction, aiming at decreasing the size of the dataset by means of instance selection and/or feature selection, and data projection, altering the representation of the data, e.g. mapping continuous variables to categories or encoding nominal attributes [8]. While some of these are imperative for the valid application of a method, such as scaling for NN, others appear to be more general to facilitate method performance in general.

To evaluate the impact of DPP methods on classification accuracy and to derive best practices within the domain, we conduct a structured literature review of publications in corporate data mining applications of classification within the related domains of target selection in direct marketing, including case-based analyses as well as comparative papers evaluating various algorithms on multiple datasets [9]. We analyse each publication regarding the methods applied, whether parameter tuning was conducted, and which DPP methods of data reduction and projection could be observed. The results of our analysis are presented in Table 1.

Our review documents the emphasis on evaluating and tuning competing classification algorithms in a particular data mining task or dataset. In addition, it shows only limited documentation and almost no competitive evaluation of DPP issues within data mining applications. Only 47% of all studies use and document data reduction approaches while only 64% consider data projection in general. Only a single publication provides information on the treatment of categorical attributes, although categorical variables are used and documented in 71% of all studies and commonly encountered in the application and the data mining domain in general. In contrast, information on the respective procedures for parameter tuning is provided in 16 out of 19 publications. Most strikingly, across all surveys only a single DPP technique is applied, ignoring possible alternatives without evaluation or justification. In data projection, only [10,6] evaluate models incorporating discretised as well as standardised alternatives of continuous attributes in their study. Standardisation of continuous attributes are routinely included in experimental setups [10], particularly of NN, their use appears scarce. While the necessity of DPP for data reduction is motivated by the size of the individual dataset, all three authors

Table 1
Data preprocessing activities within publications on corporate data mining

	Input type[a,b]	Methods[c]	Parameter tuning	Data reduction[d]		Data projection		
				FS	RS	Continuous attributes		Categories
						Standardisation	Discretisation	Coding
[2]	2	BMLP, LR, LDA, QDA	X			X		
[42]	1	MLP, LR, CHAID	X			X		
[43]	2	MLP, RBF, LR, GP, CHAID	X		X			
[44]	3	MLP, LR, LDA	X	X				
[4]	2	CHAID, CART			X			
[6]	2	MLP, LR	X	X	X	X	X	
[9]	2	LVQ, RBF, 22 DT, 9 SC	X					X
[45]	2	LDA, LR, KNN, KDE, CART, MLP, RBF, MOE, FAR, LVQ	X				X	
[3]	1	MLP		X		X		
[7]	2	LSSVM	X	X		X		
[11]	2	LR, LS-SVM, KNN, NB, DT	X			X	X	
[10]	1	LDA, QDA, LR, BMLP, DT, SVM, LSSVM, TAN, LP, KNN	X				X	
[46]	2	LR, MLP, BMLP	X	X				
[47]	2	LSSVM, SVM, DT, RL, LDA, QDA, LR, NB, IBL	X			X		
[48]	1	DT, MLP, LR, FC	X					
[49]	1	FC	X			X		

[a] Type 1: only continuous; 2: continuous and categorical; 3: only categorical.

[b] Some publications provide no detailed information about the type or scaling level of their variables. Considering the fact that demographic customer data consist mostly of categorical variables, we assume that any experiment that includes demographic customer information together with transaction oriented data has to deal with continuous as well as categorical variables. Binary variables are considered as categorical ones.

[c] BMLP: Bayesian learning MLP, CART: classification and regression tree, CHAID: Chi-square automatic interaction detection, FAR: fuzzy adaptive resonance, FC: fuzzy classification, GP: genetic programming, IBL: instance based learning, KDE: kernel density estimation, KNN: K-nearest neighbor, LDA: linear discriminant analysis, LP: linear programming, LR: logistic regression, LVQ: learning vector quantisation, MLP: multilayer perceptron, MOE: mixture of experts, NB: Naïve Bayes, QDA: quadratic discriminant analysis, RBF: radial basis function NN, RL: rule learner, SC: statistical classifiers (e.g. LDA, LR, etc.), LSSVM: least squares SVM, TAN: tree augmented Naïve Bayes.

[d] FS: feature selection; RS: resampling.

that make use of instance selection techniques evaluate only one single procedure.

As the choices of DPP depend on the individual dataset used, the lack of DPP may be contributed to the use of ready preprocessed, 'toy' datasets. However, we may conclude that the potential impact of DPP decisions on the predictive performance of classification methods has neither been analysed nor systematically exploited. Particular recommendations exist for selected algorithm classes, which must not hold for other methods. However, only a single DPP scheme is utilised to compare classifier performance, possibly biasing the evaluation results. Consequently, the suitability of different DPP approaches for different methods within a specific task, as well as the sensitivity of data mining algorithms towards DPP in general, requires further investigation. We present an overview of the relevant methods in data reduction and data projection for DPP, which will later be evaluated in a comprehensive experimental setup.

3.2. Data reduction

Data reduction is performed by means of feature selection and/or instance selection. Feature selection aims at identifying the most relevant, explanatory input variables within a dataset [14]. In addition to improving the performance of the predictors, feature selection facilitates a better understanding of the underlying process that generated the data. Also, reducing the feature-vector condenses the size of the dataset, accelerating the task of training a classifier and thereby increasing computational efficiency [13]. Feature selection methods are categorised as wrappers and filters [50]. While filters make use of designated methods for feature evaluation and construction, e.g. principal component analysis [51] and factor analysis [52], wrappers utilise the particular learning algorithm to assess selected feature subsets heuristically by means of the resulting prediction accuracy. In general, wrapper-based approaches have proven more popular for direct marketing applications; see e.g. [3,7,12]. Feature selection appears to be well researched and established in data mining practice as for enhancing individual methods [13,14]. Therefore we limit our experiments on the effects of less analysed DPP choices, disregarding the impact of feature selection from further analysis.

The selection of data instances through resampling techniques often represents a prerequisite for data mining, establishing computational feasibility on large datasets or ensuring unbiased classification on imbalanced datasets. Particularly in empirical domains of corporate response modelling, such as direct marketing, fraud detection, etc., the number of instances in the interesting minority class is significantly smaller than of the majority class. For example, the number of customers who respond to a mail offer is usually very small compared to the overall size of a solicitation [4,5,46] so that the target class distributions are highly skewed. These imbalances obstruct classification methods by biasing the classifier towards the majority class [53] requiring specific DPP treatment to diminish negative effects. Popular approaches to account for imbalances without modifying the classifier are random oversampling of the minority class or random undersampling of the majority class, respectively [54,55]. Additionally, sophisticated techniques have recently been proposed, e.g. the removal of noisy, borderline and redundant training instances of the majority class [16] or the creation of new members of the minority class as a mixture of two adjacent class members [15].

3.3. Data projection

Data projection aims at transforming raw data into a feasible, beneficial representation for a particular classification algorithm. It comprises techniques of value transformation, e.g. mapping of categorical variables and discretisation or scaling of continuous ones. Working with large attribute sets of mixed scale, data mining routinely encounters mixtures of categorical and continuous attributes. Consequently, the combination of different data projection approaches offers vast degrees of freedom in the DPP stage.

Continuous attributes may be preprocessed using various forms of discretisation or standardisation, of which we present the most common variants. Discretisation or binning represents a transformation of continuous attributes into a limited set of values (bins), thereby suppressing noise and removing outlier values. Each raw value x_i is uniquely mapped to a particular symbol s_i, e.g. $s_i = 1$ for $x_{\min} < x_i \leqslant x_{c1}$, $s_i = 2$ for $x_{c1} < x_i \leqslant x_{c2}$, $s_i = 3$ for $x_{c2} < x_i \leqslant x_{\max}$, thus deriving a set of artificially created ordinal attributes from metric variables. With a higher quantity of used symbols, more details of the original attributes are captured in the transformed dataset. Obviously, the resulting dataset depends on the definition of the critical boundaries x_c between two adjacent symbols. As an unfavourable choice of values may lead to a loss of meaningful information [40,41], the DPP choice of discretisation is not without risk. Popular variants of discretisation are analysed [18], confirming their relevance for classifier performance. Alternatively, standardisation of continuous attributes (4) ensures that all scaled attributes values $\hat{x}_i$ reside in a similar numerical range [21]:

$$\hat{x}_i = \frac{x_i - \bar{x}_i}{\sigma_{x_i}} \tag{4}$$

Table 2
Schemes for encoding categorical attributes

Ordinal raw value	N encoding			$N-1$ encoding		Thermometer encoding			Ordinal encoding
	x_1	x_2	x_3	x_1	x_2	x_1	x_2	x_3	x_1
High	1	0	0	0	0	1	0	0	1
Medium	0	1	0	1	0	1	1	0	2
Low	0	0	1	1	1	1	1	1	3

with mean $\bar{x}_i$ and standard deviation σ_{x_i} of all realisations of attribute x_i, this approach is sensitive to outlier values but avoids the creation of additional features that increase the dimensionality of the dataset.

While variants for data projection of continuous attributes receive selected attention, variants for numerical mapping of categorical attributes or data conversion are largely neglected. Several encoding schemes are feasible, which are exemplified in Table 2 for three ordinal values on a N encoding, $N-1$ encoding, thermometer code and ordinal encoding scheme using one to three binary (dummy) variables [8,19,56].

After mapping original data by means of reasonable transformation rules and encoding schemes, scaling procedures transform values of each variable into an interval being appropriate to a particular classification algorithm. Typical intervals are $[-1;1]$ and $[0;1]$, either with binary values only or with real values, depending on the encoding scheme.

4. Case study of data preprocessing in direct marketing

4.1. Experimental setup

We analyse the impact of individual DPP choices on classification performance in a structured experiment, based upon the characteristics of an empirical dataset from a previous direct mailing campaign conducted in the publishing industry. The objective is to evaluate customers for cross-selling, identifying those most likely to buy an additional magazine subscription from all customers already subscribed to at least one periodical. The original campaign contacted 300,000 customers, of which 4019 ordered a new subscription. The response rate of 1.4% is considered representative for the application domain. The dataset characterises each customer instance by 28 attributes of nominal scale, e.g. flags identifying email, previous merchandising treatment, etc., categorical scale, such as age group, order channel, etc., and continuous scaling level, including the total number of subscriptions, number of cancellations, overall revenue, etc. The binary target variable identifies a customer as one of the 4019 responders (1) or as a non-responder (-1). The significantly skewed target class distribution and the mixed scaling level of potentially valuable customer attributes poses particular challenges to be addressed using DPP. Therefore, projection of categorical attributes, discretisation or scaling of continuous ones as well as resampling are of primary importance. Regarding the moderate number of attributes, the wealth of previous research and the scope of our analysis, we omit feature selection from our study.

An explorative analysis reveals the presence of outlier values in some of the continuous attributes, e.g. customer instances with 253 inactive subscriptions in contrast to and average of 0.8. As binning may diminish the effect of outliers while scaling remains sensitive to extreme values, we create two sets of experiments implementing discretisation as in [18] versus standardisation. For categorical attributes we consider the four encoding schemes of Table 2. To evaluate possible effects of scaling into different intervals, we run two sets of experiment setups, scaling all attributes to $[0;1]$ and $[-1;1]$, respectively. Finally, we evaluate the impact of over- and undersampling [54] to counter class imbalance between responders and

Table 3
Identification of experimental setups—sampling, encoding and scaling of attributes

	Oversampling								Undersampling							
	N		$N-1$		Temperat.		Ordinal		N		$N-1$		Temperat.		Ordinal	
	0	−1	0	−1	0	−1	0	−1	0	−1	0	−1	0	−1	0	−1
Experiment #ID																
Discretisation	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	#11	#12	#13	#14	#15	#16
Standardisation	#17	#18	#19	#20	#21	#22	#23	#24	#25	#26	#27	#28	#29	#30	#31	#32
No. of attributes[a]																
Discretisation	117	117	90	90	117	117	29	29	117	117	90	90	117	117	29	29
Standardisation	88	88	70	70	88	88	29	29	88	88	72	72	88	88	29	29

[a] Varying attribute numbers result from applying different encoding schemes (see Table 2).

non-responders, aiming to increase classifier sensitivity for the economically relevant minority class 1.

The resulting 32 experiments (Table 3) are evaluated applying a hold-out method, requiring three disjoint datasets for training, validation and testing. While training data is used to parameterise each classifier, the second set is used for model selection and to prevent overfitting through early stopping for NN. The trained and selected classifiers are tested out-of-sample on an unknown hold-out set to evaluate their classification performance as an indication of their ability to generalise on unknown data. To ensure comparability all datasets contain the same records over all experiments, differing only in data representation according to the respective DPP treatment. To separate balanced datasets, we randomly select 65,000 records for the test set, leading to a statistically representative asymmetric class distribution of 1.4% responders (912 class 1) to 98.6% non-responders (64,088 class −1). In order to facilitate full usage of the remaining 3107 responders, 66.6% (2072) are randomly assigned to the training set with 33.3% (1035) assigned to the validation set. Using strategies of oversampling versus undersampling, different sizes of the training and validation datasets are created through resampling of responders and non-responders until equally distributed class sizes are achieved. In undersampling, 2072 records of non-responders are randomly chosen for the training set until their number equals that of responding customers, with 1035 records for the validation set, respectively. For oversampling, 20,000 and 10,000 records of inactive customers are randomly chosen for the training and validation set, while responders are randomly duplicated to equal the number of non-responders in each set. The size of the individual data subsets is chosen to balance the objective of learning to accurately predict responders from the training set while keeping datasets computationally feasible. The resulting datasets are summarised in Table 4.

Table 4
Dataset size and structure for the empirical simulation—over-/undersampling approaches

Data subset	Data partition (number of records)			
	Oversampling		Undersampling	
	Class 1	Class −1	Class 1	Class −1
Training set	20,000	20,000	2072	2072
Validation set	10,000	10,000	1035	1035
Test (hold-out) set	912	64,088	912	64,088

4.2. Method parameterisation

Each experimental setup is evaluated using different parameterisations for each classifier to account for possible interactions between method tuning and the effects of the multifactorial design of sampling, coding and scaling on predictive performance.

With regard to the large degrees of freedom and the considerable computational time of over 3 hours for MLP training, we conduct a pre-experimental sensitivity analysis to heuristically identify a suitable subset of parameters from hidden nodes,

activation functions, learning algorithms, etc. We limit the experiments to architectures using $n_i = 25$ hidden nodes and two sets of activation function in the hidden layer $\text{act}_j = \{\tanh, \log\}$, using a softmax output-function on the two nodes in the output layer to model the conditional probability of class membership for each pattern in order to rank each customer instance according to its probability of belonging to class 1. Each NN is initialised four times and trained up to a maximum of 10,000,000 iterations, evaluating the performance on the validation set after every epoch for early stopping. We apply the Delta–Bar–Delta learning rule, using autoadaptive learning parameters for each weight w_{ij} to further limit the degrees of freedom. For SVM modelling, we consider alternative regularisation parameters C in the range $\log(C) = \{-3, -2, -1, 0\}$ and kernel parameters $\log(\sigma^2) = \{-3, -2\}$, derived from a previous grid search for a Gaussian kernel function. The selection of the Gaussian kernel is motivated by previous results [57] and a pre-experimental analysis, indicating computational infeasibility of polynomial kernels with training times of over 72 hours on the oversampled datasets. Degrees of freedom in C4.5 parameterisation are mainly concerned with pruning, to guide the process of cutting back a grown tree for better generalisation. We consider the standard pruning procedure together with reduced-error pruning and vary the confidence threshold in the range of $\{0.1, 0.2, 0.25, 0.3\}$ [58].

We compute a total of 768 classifiers for each data subset, relating to 256 results per NN, SVM and DT each, and corresponding to 32 groups of 8 observations per dataset and method, i.e. 384 results for each scaling effect, 384 experiments per sampling effect, 192 experiments per coding effect of categorical attributes and 384 experiments of coding continuous variables. This leads to a total of 2304 classification results evaluated across three performance measures in order to test the effect of factors and factor combinations independent of method parameterisation. All experiments are carried out on 3.6 GHz Pentium IV workstation with 4GB main memory. The WEKA software library [58] is used to model tree classifiers, taking an average of 4 minutes to build a DT. In contrast, parameterising SVM takes on average 20 minutes per experiment for undersampling and 2 hours for oversampling using the LIBSVM package [59]. MLP are trained using Neural Works Professional II+, taking 25 minutes for undersampling and on average 3 hours, depending on the early stopping of each initialisation. In total, experimental runtime consists of 34 days excluding pre-experiments, setup and evaluation.

4.3. Performance metrics for method evaluation

A variety of performance metrics exists in data mining, direct marketing and machine learning, permitting an evaluation of DPP effects by alternative performance metrics. As certain metrics provide biased results for imbalanced classification [60], we limit potential biases by evaluating the impact of DPP on three alternative performance metrics established in business classification problems [57]. Classifier performance is routinely assessed using a confusion matrix of the predicted and actual class memberships (see Table 5).

Performance metrics calculate means of the correctly classified records within each class to obtain a single measure of performance such as arithmetic (AM) or geometric mean (GM) classification rates

$$\text{AM} = \frac{1}{2}\left(\frac{h_{00}}{h_{0.}} + \frac{h_{11}}{h_{1.}}\right); \quad \text{GM} = \sqrt{\left(\frac{h_{00}}{h_{0.}} \cdot \frac{h_{11}}{h_{1.}}\right)}. \quad (5)$$

While these performance metrics assess only the capability of a binary classifier to separate the classes without error, they do not take a classifier's ability to rank instances by their probability of class membership into consideration. As direct marketing applications need to identify customers ranked by the highest propensity to buy, given a

Table 5
Confusion matrix for binary classification problem with output domain $\{-1, +1\}$

		Predicted class		$\sum$
		-1	$+1$	
Actual class	-1	h_{00}	h_{01}	$h_{0.}$
	$+1$	h_{10}	h_{11}	$h_{1.}$
	$\sum$	$h_{.0}$	$h_{.1}$	L

varying constraint of the size of a possible mailing campaign, a lift analysis reflects a more appropriate approach to evaluate response models [53,61,62]. Using a classifier to score customers according to their responsiveness from most likely to least likely buyers, the lift reflects the redistribution of responders after the ranking, with superior classifiers showing a high concentration of actual buyers in the upper quantiles of the ranked list. Hence, the lift evaluates a classifier's capability to identify potential responders and measures the improvement over selecting customers for a campaign at random. Given a ranked list of customers S with known class membership a lift index is calculated as

$$\text{Lift} = (1.0 \cdot S_1 + 0.9 \cdot S_2 + \cdots + 0.1 \cdot S_{10}) \Big/ \sum_{i=1}^{10} S_i \tag{6}$$

with S_i denoting the number of responders in the ith decile of the ranked listed. An optimal lift provides a value of 1 with $S_1 = \sum_i S_i < 10\%$, while a random selection of customers would result in a lift of 50% [53].

We evaluate the impact of DPP on classifier performance using the performance metrics of AM, GM and lift index. As individual classifiers use particular error metric to guide their parameterisation processes, such as early stopping of NN on AM, or the selection of a best parameterisation on the validation set, this may induce an additional bias if evaluated on a inconsistent metric. To confirm the robustness of our experiments and the appropriateness of analysing the results using a single performance metric, we analyse Spearman's rho non-parametric correlations between the individual metrics across all experiments and all datasets. The analysis reveals consistent, positive correlations significant at a 0.01 level, indicating a mean correlation of 0.775 between GM, AM and lift index across all datasets of training, validation and test for each method. Consequently, the use of an arbitrary performance metric seems feasible, utilising the AM for parameterisation where the lift metric is inapplicable as an objective function. The lift is used for out of sample evaluation across all methods to reflect the business objective. In order to adhere to space restrictions and to present results in a coherent manner for both the direct marketing and the machine learning domains, unless otherwise stated we provide results using the out-of-sample lift index. However, all presented results on the impact of DPP upon the classification performance also hold for alternative performance metrics.

5. Experimental results

5.1. Impact of data preprocessing across classification methods

We calculate the lift index of SVM, NN and DT across 32 experimental designs of different DPP variants and across three datasets of training, validation and test data, visualised in Fig. 3.

To quantify the impact and significance of each DPP candidate on the classification performance of different methods, we conduct a multifactorial analysis of variance with extended multi comparison tests of estimated marginal means across all methods and for each of the three methods separately. The experimental setup assures a balanced factorial design, modelling each DPP variant as different factor treatment of equal cell sizes. Sampling, scaling, coding of continuous attributes, coding of categorical attributes and the method are modelled as fixed main effects to test whether the factor levels show different linear effects on the dependent variables, the classification lift index on the training, validation and test datasets. In addition, we investigate ten 2-fold, ten 3-fold, five 4-fold and one 5-fold non-linear interaction effects between factors. We consider factor effects as relevant if they prove consistently significant at a 0.01 level of significance using Pillai's trace statistic across all datasets. In addition, a factor needs to prove significant for the individual test set to indicate an consistent out-of-sample impact independent of the data sample. We disregard a significant Box's test of equality and a significant Levene statistic of indifferent group variances due to the large dataset, equal cell sizes across all factor-level-combinations and ex postanalysis of the residuals revealing no violations of the

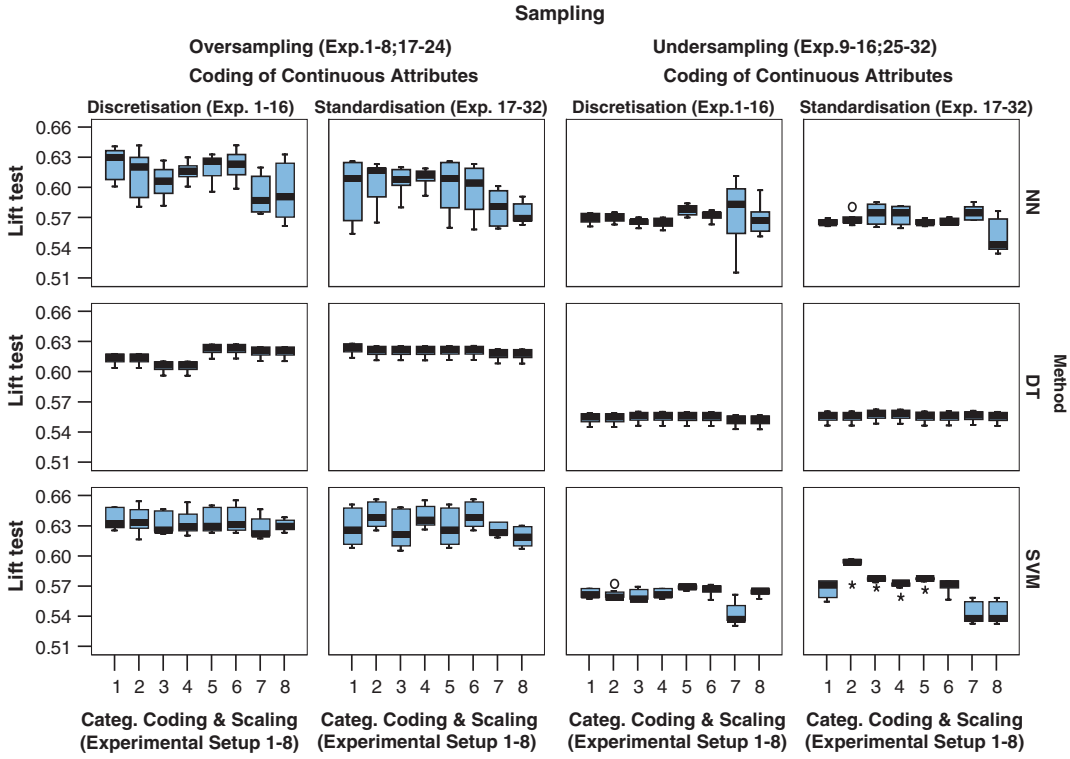

Fig. 3. Boxplots of lift performance on the test sets for NN, DT and SVM across 32 experimental setups of sampling, scaling, coding of categorical and coding of continuous attributes. Boxplots provide median and distributional information, additional symbols of stars and circles indicate outliers and extreme values. Higher lift values indicate increased accuracy.

underlying assumptions. The individual contribution of each main factor and their interactions to explaining a proportion of the total variation is measured by a partial eta squared statistic (η), with larger values relating to higher relative importance. To contrast the impact of each factor levels within each factor we conduct a set of posthoc multi comparison tests using Tamhane's T2 statistics, accounting for unequal variances in the factor cells. This evaluates the positive or negative impact of each factor level on the classification accuracy of lift across the data subsets by estimated marginal means, $\text{mm}_i = \{\text{training; validation, test}\}$, with positive impacts indicating increased accu-

Table 6
Significance of DPP main effects by individual datasets and individual methods using Pillai's trace

Factors	Significance by dataset				Significance by method		
	All	Train	Valid	Test	NN	SVM	DT
Method	0.000**	0.000**	0.000**	0.000**	–	–	–
Scaling	0.077	0.011*	0.092	0.343	No	No	No
Sampling	0.000**	.000**	0.000**	0.000**	Yes	Yes	Yes
Continuous coding	.000**	0.000**	0.000**	0.153	Yes	No	Yes
Categorical coding	0.000**	0.000**	0.000**	0.000**	Yes	Yes	Yes

* Significant at the 0.05 level (2-tailed).
** Highly significant at the 0.01 level (2-tailed).

racy and vice versa. Table 6 presents a summary of the findings by dataset across all methods and for each method individually.

The main factors of sampling ($\eta = 0.958$), method choice ($\eta = 0.392$) and coding of categorical attributes ($\eta = 0.108$) prove significant at a 0.01 level in the order of their relative impact, while the effect of scaling and the coding of continuous attributes prove just insignificant. In addition, all two-way interactions of the significant main effects led by sampling * method ($\eta = 0.404$) and one three-way interaction of method * sampling * categorical prove significant. This confirms a significant impact of DPP through different levels of sampling, coding of categorical attributes and coding of continuous attributes on out-of sample model performance for the case study dataset. In addition, the significant impact proves consistent across alternative methods. However, no significant impact of different scaling ranges for continuous and categorical variables can be validated.

In order to determine the size and positive or negative direction of each DPP choice upon classification performance, we analyse the treatments of the significant factors in more detail. In addition, the analysis indicates interaction effects between the used classification methods and selected DPP factor levels of varying significance and impact. As this indicates method specific reactions to individual DPP factor levels, we need to analyse the impact of the factor effects in separate multifactorial ANOVA analyses for each method.

5.2. Impact of sampling on method performance

To further investigate the significant impact of over- versus undersampling we analyse the estimated marginal means of the classification performance for NN, SVM and DT separately. Regarding undersampling, the results across NN, SVM and DT are consistent and confirm an increased performance across training and validation datasets and a severely decreased performance on the test set. The impact of undersampling versus oversampling for NN is estimated at $mm_{NN} = \{0.088; 0.081; -0.035\}$, indicating a -3.5% drop in lift accuracy, for SVM at $mm_{SVM} = \{0.071; 0.078; -0.068\}$ and for DT at $mm_{DT} = \{0.035; 0.033; -0.063\}$. As already a 1% increase in out-of-sample accuracy is regarded as economically relevant due to the highly asymmetric costs in the problem domain, the use of undersampling would induce a significant monetary loss. In addition, the marginal means in Fig. 4 indicate a stronger impact of undersampling on SVM and DT than on NN.

Our analysis clearly identifies undersampling as suboptimal to oversampling across all methods, leading to significantly increased yet irrelevant in-sample performance at the cost of decreased out-of-sample performance regardless

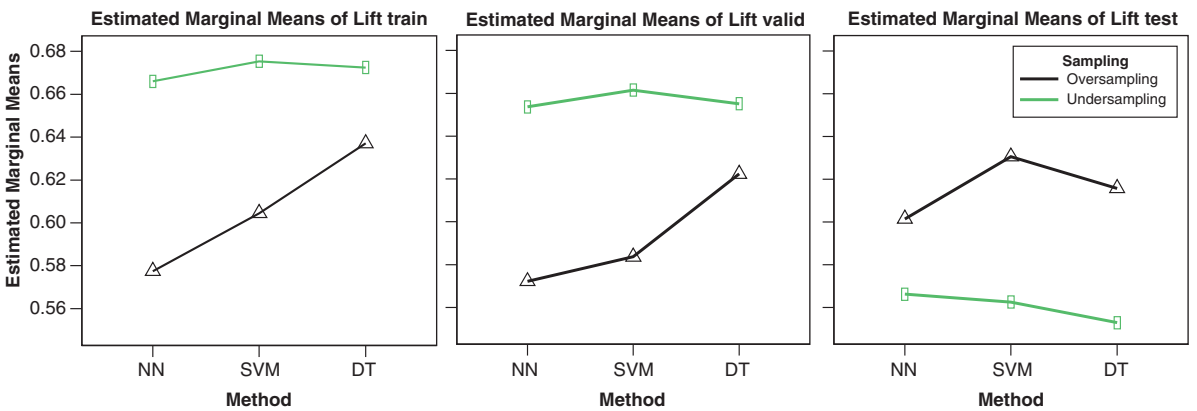

Fig. 4. Estimated marginal means plots of the test set performance of two sampling factor treatments of oversampling (△) and undersampling (□) across different classification methods of NN, SVM and DT.

Table 7
Spearman's rho non-parametric correlation coefficients between datasets for sampling variants

Spearman's rho		NN correlations			SVM correlations			DT correlations		
		Train	Valid	Test	Train	Valid	Test	Train	Valid	Test
Oversampling	Train	1.000	0.912**	0.858**	1.000	0.594**	0.762**	1.000	0.778**	0.775**
	Valid	0.912**	1.000	0.786**	0.594**	1.000	0.803**	0.778**	1.000	0.671**
	Test	0.858**	0.786**	1.000	0.762**	0.803**	1.000	0.775**	0.671**	1.000
Undersampling	Train	1.000	0.985**	−0.307**	1.000	0.878**	−0.540**	1.000	0.970**	−0.626**
	Valid	0.985**	1.000	−0.329**	0.878**	1.000	−0.631**	0.970**	1.000	−0.639**
	Test	−0.307**	−0.329**	1.000	−0.540**	−0.631**	1.000	−0.626	−0.639	1.000

* Correlation is significant at the 0.05 level (2-tailed).
** Correlation is highly significant at the 0.01 level (2-tailed).

of the classification method. The selective increase on in-sample performance indicates overfitting instead of learning to generalising for unseen instances from the training data. Regardless of any computational advantages of undersampling due to the reduced sample size, undersampling seems inapplicable in contrast to the time demanding oversampling for the case study dataset. In addition to the inferior accuracy, undersampling induces inconsistencies in selecting 'best' candidate parameterisations for each method. A correlation analysis confirms high correlations between training, validation and test performance for oversampling in contrast to a negative correlation on the out of sample test set for undersampling, see Table 7.

Consequently, classifiers with a high performance on out-of-sample data cannot reliably be selected based upon superior in-sample performance, indicating undersampling as unsuitable for the given imbalanced classifications problem. In contrast, oversampling promises a valid and reliable selection of favourable SVM, NN or DT parameterisations on the validation set to facilitate a high out of sample performance. Considering the lack of generalisation and suboptimal results, we exclude undersampling from further analysis.

5.3. Impact of coding on method performance

After eliminating the dominating factor level of undersampling from the analysis design, we evaluate the effects of coding of categorical and continuous variables across the three methods. Only the coding of categorical variables remains significant for SVM ($\eta = 0.066$). A multiple comparison test confirms a negative impact of ordinal encoding on SVM lift performance of $mm_{SVM} = \{-0.014; -0.002; -0.009\}$ in contrast to a homogeneous subset of all other categorical coding schemes of N, $N-1$ and temperature showing no significant impact. This seems particularly surprising, considering the induced multicollinearity through N encoding. Considering the insignificant differences on classification performance by discretisation or standardisation of continuous attributes, we derive that SVM perform indifferent of binning of metric variables, scaling in different intervals, and N, $N-1$ or temperature encoding of categorical attributes on the given dataset.

In contrast to SVM, both the coding of continuous attributes ($\eta = 0.173$) and the coding of categorical attributes ($\eta = 0.131$) have a significant impact on NN out-of-sample accuracy at a 0.01 level, while no interaction of both coding schemes is observed. An analysis of the marginal means reveals a negative impact of standardisation of continuous variables $mm_{NN} = \{-0.011; -0.009; -0.014\}$ in contrast to discretisation. As with SVM, a multiple comparison test of individual factor levels of categorical coding reveals two homogeneous subsets and a significant, negative impact of ordinal encoding on lift accuracy of $mm_{NN} = \{-0.013; -0.006; -0.024\}$. The negative impact of ordinal coding is considerably larger than for SVM, confirming NN sensitivity to ordinal coding [19]. The impacts of all other factor levels of N, $N-1$ and temperature coding prove

insignificant. Scaling of variables remains insignificant for NN performance. These results seem interesting, considering the frequent assumption that NN learning may benefit from metric variables, and that the limited research conducted by [19] indicates the benefits of scaling to $[-1;1]$ intervals. More specifically, it indicates a dataset specific need for analysis of DPP choices in using NN.

For DT only categorical coding of attributes ($\eta = 0.350$) and its interaction with different continuous codings ($\eta = 0.280$) prove significant, while the main effects of continuous coding or scaling are not significant. In contrast to SVM and NN, an analysis of the marginal means provides inconsistent results, indicating a small but significant decrease in performance of $N-1$ coding of $\mathrm{mm}_{\mathrm{DT}} = \{-0.004; -0.001; -0.004\}$ in contrast to N-coding, a significant increase in performance of temperature encoding of $\mathrm{mm}_{\mathrm{DT}} = \{0.003; 0.004; 0.004\}$ in contrast to N-coding and no significant impact of ordinal encoding. This is attributed to an observed interaction effect of categorical with continuous encoding, as apparent in Fig. 5 at method DT. While no impact is apparent for standardised continuous attributes, a strong negative effect of N and $N-1$ encoding becomes visible for discretised continuous attributes, contrasted by a strong positive effect on the accuracy using temperature or ordinal coding.

In contrast, the plots of marginal means show no interaction between coding categorical and continuous attributes for NN and SVM, with consistently inferior classification results of standardisation for NN but not for SVM. While the impact of scaling remains statistically insignificant for all methods, our analysis indicates that scaling to the interval $[0;1]$ consistently improves out of sample accuracy across NN and SVM, while leaving DT unaffected. However, these results are just insignificant at a 0.05 level. In addition, interactions of scaling, continuous coding and categorical coding emerge for NN. For all standardised and discretised attributes of interval scale, all categorical coding schemes improve test lift when scaled to $[0,1]$. However, N encoding of discretised attributes displays pre-eminent performance when scaled to $[-1;1]$, while scaling to $[0,1]$ decreases out of sample accuracy by 1.5%. In contrast, SVM and DT are generally unaffected by these interaction effects.

5.4. Implications of data preprocessing impact on method performance

As a conclusion from the analysis across various alternative architectures and parameterisations, we determine undersampling to be inferior DPP alternative for NN, SVM and DT. Ordinal coding of categorical variables appears to be a

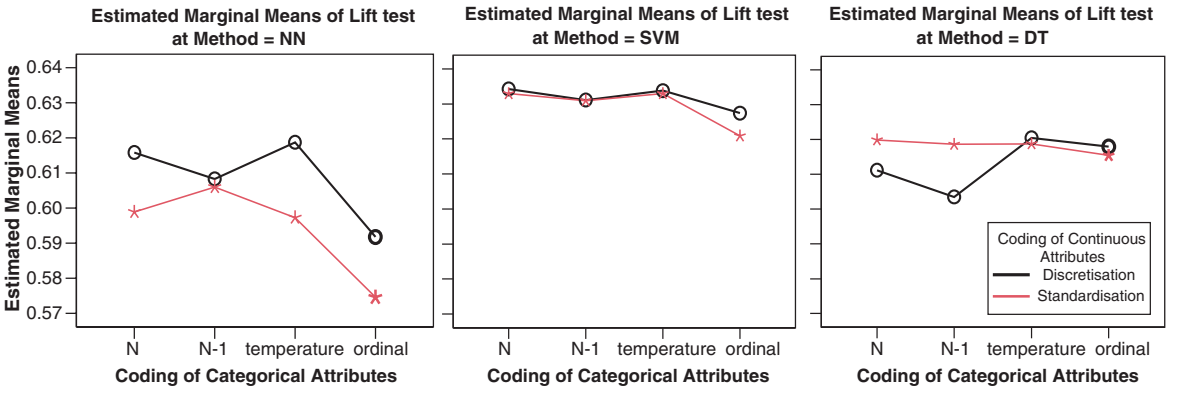

Fig. 5. Plots of the estimated marginal means of lift performance on the test set resulting from continuous coding schemes of discretisation (○) and standardisation (∗) across different categorical coding schemes of N, $N-1$, temperature and ordinal encoding, for each method of NN, SVM and DT.

suboptimal DPP choice for SVM and NN but has no effect on DT classification. Standardisation of continuous attributes is inferior to discretisation for NN, given the case study dataset induced by outliers in the data. As neither temperature scaling, N nor $N-1$ coding of categorical attributes show a significant impact on classification performance across datasets and methods, we propose the use of $N-1$ encoding. $N-1$ encoding reduces the size of the input vector, resulting in a lower dimensional classification domain and increased computationally efficiency through reduced training time. Accordingly, we propose standardisation of continuous attributes to reduce input vector length in the lack of negative effect on SVM or DT performance, but not for NN. On the contrary, discretisation of attributes paired with $N-1$ encoding should be avoided for DT. While scaling to $[0,1]$ generally suggests slightly increased performance across all methods and other DPP choices, this in combination with the computationally motivated preference of $N-1$ encoding would simultaneously avoid significantly decreased NN-performance resulting from the interaction effect with scaling for discretised attributes. To summarise, NN provide best results on the given dataset when continuous data is discretised to categorical scale, N-encoded and scaled to $[-1;1]$ using oversampling. In contrast, SVM benefit from standardised continuous attributes, $N-1$ encoding of categorical attributes and scaling to $[0,1]$ while DT are indifferent and may use the same scheme as SVM.

We conclude that in avoiding undersampling and ordinal coding, SVM as NN offer a robust out-of-sample performance equal or better to DT, which is not significantly influenced by preprocessing through different coding or scaling of variables. However, these findings suggest method specific best practices in using DPP to facilitate out of sample performance for different classification methods. Moreover, it implies that different learning classifiers may produce suboptimal results if they are all evaluated on a single, identical dataset with a single, implicit decision for DPP. Therefore, we eliminate the impact of different method parameterisations and evaluate DPP impact on a selected 'best' architecture for NN, SVM and DT.

5.5. Impact of data preprocessing on best classifier architectures

After analysing the effect of DPP across different parameterisations of each method, we omit the impact of modelling decisions from our analysis by selecting a single 'best' architecture for NN, SVM and DT. We select the method setup from the experiments 1–6 and 17–22, avoiding biased results from suboptimal DPP methods of undersampling and single number encoding found in our preceding analysis. In addition, we identify a single architecture setup for each method based upon the highest mean lift performance on the validation data subset. For NN, we select a topology of 25 hidden nodes in a single hidden layer using a hyperbolic tangent activation function. We apply a DPP scheme from experiment setup #2, discretising continuous variables and scaling all $N-1$ encoded attributes to $[-1,1]$, leading to a lift performance of 0.640 on the test set. For SVM, we select DPP scheme #19, standardising continuous variables, encoding all categorical as $N-1$ and scaling them to $[0,1]$. For DT we apply the same DPP scheme #19, resulting in an out-of-sample lift of 0.619. SVM demonstrate best performance, achieving a lift of 0.645 on the test set.

However, these results are based upon our preceding analysis of different DPP variants across all methods and the individual matching of DPP to method. To relate our findings to the effects of DPP on the validity and reliability of results provided in incomplete case studies from our literature analysis, we need to simulate the effect of choosing a single, arbitrary DPP combination of scaling and coding. Consequently, we analyse the lift performance of the 12 dominant DPP setups for SVM, NN and DT across all three data subsets. A successive multivariate ANOVA reveals limited differences of the classification performance between SVM, NN and DT at a 0.05 level. Although an average SVM lift of 0.634 outperforms the mean NN lift of 0.627 by 0.7% and a DT mean lift of 0.616 by 1.8% on the out-of-sample test set, these results prove not significant. An analysis of estimated marginal mean reveals two homogeneous subgroups. DT perform significantly inferior on out-of-sample than NN or SVM, with $\mathrm{mm}_{\mathrm{DT}} =$

$\{0.049; 0.043; -0.011\}$ and $mm_{DT} = \{0.021; 0.042; -0.018\}$, respectively. While the mean performances of SVM and NN are significantly different across training and validation datasets, no significant difference can be confirmed in out-of-sample accuracy (see Fig. 6).

We conclude that SVM and NN significantly outperform DT on the case study dataset, representing a valuable monetary benefit considering the costs attributed to the imbalanced classes in the case study domain. However, neither SVM nor NN significantly outperform each other across different choices of coding of continuous attributes, coding of categorical attributes or scaling. The lack of significant differences between SVM and NN accuracy seems unsurprising in the light of recent publications inconsistently identifying one method as superior over the other, presenting a different winner from one empirical case study to the next. Our experiments indicate one potential influence: the variance induced by different DPP choices towards the out-of-sample performance of NN and SVM. An analysis of the variance of the out-of-sample performances of each method induced by DPP reveals a significant difference, confirmed by Levene's test of equality at a 5% level. While NN provide a reduce mean performance, they also show a reduced variance of the classification performance across competing DPP, indicating more robust results in comparison with increased DPP sensitivity of SVM. SVM provide not only a larger variance of the results, but also promise a higher maximum performance against the risk of a lower minimum performance than NN. Two thirds of the 95% interval of NN lift ranges, from 0.622 to 0.633, overlap with the SVM results from 0.629 to 0.640. Therefore, SVM incorporate all potential NN performances and most mean performances within their range of results, depending on an individual DPP choice. In contrast, the DT interval of 0.611–0.622 clearly proves inferior considering not only mean performance but also robustness of performance across DPP choices. The results prove consistent across different performance metrics of lift, arithmetic mean classification accuracy and geometric mean classification accuracy, provided in Fig. 6. This implies that comparing in-sample and out-of-sample performance between SVM and NN based upon a particular, arbitrarily motivated DPP choice of coding and scaling on a given dataset may lead to arbitrary results of superior performance of a method, favouring either SVM

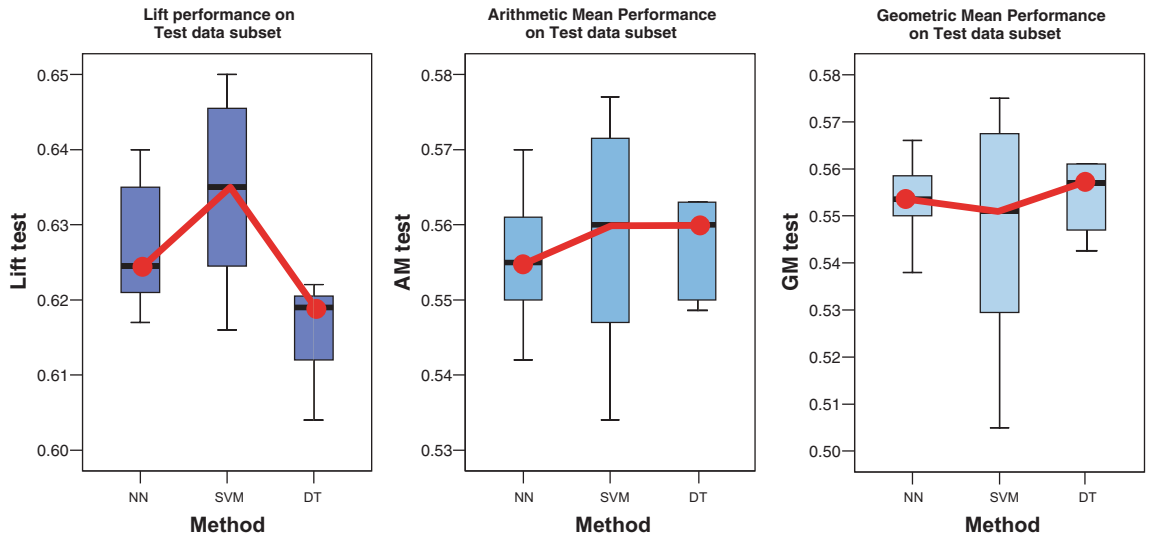

Fig. 6. Boxplots of performances on test data subset for different methods of NN, SVM and DT, displaying mean, across performance measures of lift, AM and GM (from left to right). The estimated marginal means are connected across boxes to highlight mixed patterns of method superiority across performance metrics.

or NN. Although these results are not valid across all possible datasets, they support the importance of DPP decisions with regard to model evaluation. As a consequence, the individual performance of SVM or NN may be increased by evaluating alternative coding, scaling and novel sampling schemes.

Moreover, the variation induced by DPP choices for each classification method is larger than the differences between the methods mean performance. In particular, the impact of DPP on NN and SVM accounts for 50–70% of the variation in accuracy induced by selecting optimal NN architectures, with an average increase of 0.016 through selecting the correct activation function, or SVM parameters, with the impact of selecting significant σ- and C-parameters between 0.004 and 0.021. Considering the variability of performances for SVM and NN depending on adequate DPP, an analysis of alternative preprocessing methods may prove more beneficial in increasing classifier performance than the evaluation of alternative classification methods also sensitive to preprocessing decisions. It is generally accepted within data mining as in operational research, that to derive sound classification results on empirical datasets, alternative candidate methods need to be evaluated, as no single method may be considered generally superior. In addition, our experimental results suggest that avoiding the evaluation of different DPP variants in the experimental designs may limit the validity and reliability of results regarding method performances, possibly leading to an arbitrary method preference.

6. Conclusions

We investigate the impact of different DPP techniques of attribute scaling, sampling, coding of categorical and continuous attributes on classifier performance of NN, SVM and DT in a case-based evaluation of a direct marketing mailing campaign. Supported by a multifactorial analysis of variance, we provide empirical evidence that DPP has a significant impact on predictive accuracy. While certain DPP schemes of undersampling prove consistently inferior across classification methods and performance metrics, others have a varying impact on the predictive accuracy of different algorithms.

Selected methods of NN and SVM prove almost as sensitive to different DPP schemes as to the evaluated method parameterisations. In addition, the differences in mean out-of-sample performance between both methods prove small and insignificant in comparison to the variance induced by evaluating different DPP schemes within each method. This indicates the potential for increased algorithmic performance through effective, method specific preprocessing. Furthermore, an analysis of DPP approaches may not only increase classifier performance of SVM and NN, it may even indicate a higher marginal return in analysing the individual classifiers regarding different DPP alternatives than the conventional approach of evaluation competing classification methods on a single, preprocessed candidate dataset of DPP. Consequently, the choice of a 'superior' algorithm may be supported or even replaced by the evaluation of a 'best' preprocessing approach. Additionally, the performance of NN and SVM across DPP schemes falls within a similar range of predictive accuracy. This suggests that if a dataset is preprocessed in a particular way to facilitate performance of a specific classifier, the results of other classifiers may be negatively biased or produce arbitrary results of method performance. If arbitrary DPP schemes are selected, method evaluation may exemplify the superiority of an arbitrary algorithm, lacking validity and reliability and leading to inconsistent research findings. If however different DPP schemes are evaluated to facilitate the performance of a favoured classifier, the results may even be biased towards prove of his dominance.

The single case-based analysis of DPP prohibits generalised conclusions of enhanced method performance. Considering the almost prohibitive runtime of our experiments on a single dataset, the evaluation on a variety of dissimilar datasets may be infeasible. Additional research may extend the analysis towards a larger set of DPP schemes for selected methods and across different artificial and empirical datasets. However, the significant impact on this representative case raises questions for the validity and reliability of current method

selection practices. The presented results justify the structured analysis of competing sampling, coding and scaling methods—currently neglected from systematic analysis—in order to derive valid and reliable results of the performance of classification methods.

References

[1] E.L. Nash, The Direct Marketing Handbook, second ed., McGraw-Hill, New York, 1992.

[2] B. Baesens, S. Viaene, D. Van den Poel, J. Vanthienen, G. Dedene, Bayesian neural network learning for repeat purchase modelling in direct marketing, European Journal of Operational Research 138 (1) (2002) 191–211.

[3] S. Viaene, B. Baesens, D. Van den Poel, G. Dedene, J. Vanthienen, Wrapped input selection using multilayer perceptrons for repeat-purchase modeling in direct marketing, International Journal of Intelligent Systems in Accounting, Finance and Management 10 (2) (2001) 115–126.

[4] D. Haughton, S. Oulabi, Direct marketing modeling with CART and CHAID, Journal of Direct Marketing 11 (4) (1999) 42–52.

[5] J. Zahavi, N. Levin, Issues and problems in applying neural computing to target marketing, Journal of Direct Marketing 11 (4) (1999) 63–75.

[6] J. Zahavi, N. Levin, Applying neural computing to target marketing, Journal of Direct Marketing 11 (4) (1999) 76–93.

[7] S. Viaene, B. Baesens, T. Van Gestel, J.A.K. Suykens, D. Van den Poel, J. Vanthienen, B. De Moor, G. Dedene, Knowledge discovery in a direct marketing case using least squares support vector machines, International Journal of Intelligent Systems 16 (9) (2001) 1023–1036.

[8] D. Pyle, Data Preparation for Data Mining, Morgan Kaufmann, San Francisco, 1999.

[9] T.-S. Lim, W.-Y. Loh, Y.-S. Shih, A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms, Machine Learning 40 (3) (2000) 203–228.

[10] B. Baesens, T. Van Gestel, S. Viaene, M. Stepanova, J. Suykens, J. Vanthienen, Benchmarking state-of-the-art classification algorithms for credit scoring, Journal of the Operational Research Society 54 (6) (2003) 627–635.

[11] S. Viaene, R.A. Derrig, B. Baesens, G. Dedene, A comparison of state-of-the-art classification techniques for expert automobile insurance claim fraud detection, Journal of Risk and Insurance 69 (3) (2002) 373–421.

[12] Y.S. Kim, W.N. Street, G.J. Russell, F. Menczer, Customer targeting: A neural network approach guided by genetic algorithms, Management Science 51 (2) (2005) 264–276.

[13] S. Piramuthu, Evaluating feature selection methods for learning in data mining applications, European Journal of Operational Research 156 (2) (2004) 483–494.

[14] J. Yang, S. Olafsson, Optimization-based feature selection with adaptive instance sampling, Computers and Operations Research, in press.

[15] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: Synthetic minority over-sampling technique, Journal of Artificial Intelligence Research 16 (2002) 321–357.

[16] M. Kubat, S. Matwin, Addressing the curse of imbalanced training sets: One-sided selection, in: Proceedings of the 14th International Conference on Machine Learning, 1997.

[17] P. Berka, I. Bruha, Empirical comparison of various discretization procedures, International Journal of Pattern Recognition and Artificial Intelligence 12 (7) (1998) 1017–1032.

[18] U.M. Fayyad, K.B. Irani, On the handling of continuous-valued attributes in decision tree generation, Machine Learning 8 (1) (1992) 87–102.

[19] W.S. Sarle, Neural Network FAQ, 2004, Downloadable from website ftp://ftp.sas.com/pub/neural/FAQ.html.

[20] S. Zhang, C. Zhang, Q. Yang, Data preparation for data mining, Applied Artificial Intelligence 17 (5/6) (2003) 375–381.

[21] C.M. Bishop, Neural Networks for Pattern Recognition, Oxford University Press, Oxford, 1995.

[22] J.A.K. Suykens, J. Vandewalle, Nonlinear Modeling: Advanced Black-box Techniques, Kluwer, Dordrecht, 1998.

[23] K.A. Smith, J.N.D. Gupta, Neural networks in business: Techniques and applications for the operations researcher, Computers and Operations Research 27 (11–12) (2000) 1023–1044.

[24] K.A. Krycha, U. Wagner, Applications of artificial neural networks in management science: A survey, Journal of Retailing and Consumer Services 6 (1999) 185–203.

[25] B.K. Wong, V.S. Lai, J. Lam, A bibliography of neural network business applications research: 1994–1998, Computers and Operations Research 27 (11–12) (2000) 1045–1076.

[26] B.K. Wong, T.A. Bodnovich, Y. Selvi, Neural network applications in business: A review and analysis of the literature (1988–1995), Decision Support Systems 19 (4) (1997) 301–320.

[27] R.D. Reed, R.J. Marks, Neural Smithing: Supervised Learning in Feedforward Artificial Neural Networks, MIT Press, Cambridge, 1999.

[28] J.P. Bigus, Data Mining with Neural Networks: Solving Business Problems from Application Development to Decision Support, McGraw-Hill, New York, 1996.

[29] M.W. Craven, J.W. Shavlik, Using neural networks for data mining, Future Generation Computer Systems 13 (2–3) (1997) 211–229.

[30] J.R. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann, San Mateo, 1993.

[31] R.O. Duda, P.E. Hart, D.G. Stork, Pattern Classification, second ed., Wiley, New York, 2001.

[32] N. Cristianini, J. Shawe-Taylor, An Introduction to Support Vector Machines and Other Kernel-based Learn-

ing Methods, Cambridge University Press, Cambridge, 2000.
[33] V.N. Vapnik, The Nature of Statistical Learning Theory, Springer, New York, 1995.
[34] C.J.C. Burges, A tutorial on support vector machines for pattern recognition, Data Mining and Knowledge Discovery 2 (2) (1998) 121–167.
[35] J.C. Platt, Probabilities for support vector machines, in: A. Smola, P. Bartlett, B. Schölkopf, D. Schuurmans (Eds.), Advances in Large Margin Classifiers, MIT Press, 1999, pp. 61–74.
[36] J. Weston, S. Mukherjee, O. Chapelle, M. Pontil, T. Poggio, V. Vapnik, Feature selection for SVMs, in: Proceedings of the Annual Conference on Neural Information Processing Systems, 2000.
[37] G. Fung, O.L. Mangasarian, Data selection for support vector machine classifiers. in: Proceedings of the 6th International Conference on Knowledge Discovery and Data Mining, 2000.
[38] H. Fröhlich, A. Zell, Feature subset selection for support vector machines by incremental regularized risk minimization, in: Proceedings of the International Joint Conference on Neural Networks, 2004.
[39] C. Edwards, B. Raskutti, The effect of attribute scaling on the performance of support vector machines, in: 17th Australian Joint Conference on Artificial Intelligence, 2004.
[40] R. Kumar, A. Kulkarni, V.K. Jayaraman, B.D. Kulkarni, Symbolization assisted SVM classifier for noisy data, Pattern Recognition Letters 25 (4) (2004) 495–504.
[41] R. Kumar, V.K. Jayaraman, B.D. Kulkarni, An SVM classifier incorporating simultaneous noise reduction and feature selection: Illustrative case examples, Pattern Recognition 38 (1) (2005) 41–49.
[42] R. Potharst, U. Kaymak, W. Pijls, Neural networks for target selection in direct marketing, Technical Report ERS-2001-14-LIS, Erasmus Research Institute of Management (ERIM), Erasmus University Rotterdam, Rotterdam, 2001, Downloadable from website http://ideas.repec.org/p/dgr/eureri/200177.html.
[43] A.E. Eiben, T.J. Euverman, W. Kowalczyk, E. Peelen, F. Slisser, J.A.M. Wesseling, Comparing adaptive and traditional techniques for direct marketing, in: 4th European Congress on Intelligent Techniques and Soft Computing, 1996.
[44] P.M. West, P.L. Brockett, L.L. Golden, A comparative analysis of neural networks and statistical methods for predicting consumer choice, Marketing Science 16 (4) (1997) 370–391.
[45] D. West, Neural network credit scoring models, Computers and Operations Research 27 (11–12) (2000) 1131–1152.
[46] G. Cui, M.L. Wong, Implementing neural networks for decision support in direct marketing, International Journal of Market Research 46 (2) (2004) 235–254.
[47] T. van Gestel, J.A.K. Suykens, B. Baesens, S. Viaene, J. Vanthienen, G. Dedene, B. de Moor, J. Vandewalle, Benchmarking least squares support vector machine classifiers, Machine Learning 54 (1) (2004) 5–32.
[48] S. Madeira, J.M. Sousa, Comparison of target selection methods in direct marketing, in: European Symposium on Intelligent Technologies, Hybrid Systems and their implementation on Smart Adaptive Systems, 2002.
[49] J.M. Sousa, U. Kaymak, S. Madeira, A comparative study of fuzzy target selection methods in direct marketing, in: International Conference on Fuzzy Systems, 2002.
[50] R. Kohavi, G.H. John, Wrappers for feature subset selection, Artificial Intelligence 97 (1–2) (1997) 273–324.
[51] I.T. Jolliffe, Principal Component Analysis, second ed., Springer, Berlin, 2002.
[52] R.L. Gorsuch, Factor Analysis, second ed., L. Erlbaum Associates, Hillsdale, 1983.
[53] C.X. Ling, C. Li, Data mining for direct marketing: Problems and solutions, in: Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining, 1998.
[54] N. Japkowicz, S. Stephen, The class imbalance problem: A systematic study, Intelligent Data Analysis 6 (5) (2002) 429–450.
[55] G.M. Weiss, Mining with rarity: A unifying framework, ACM SIGKDD Explorations Newsletter 6 (1) (2004) 7–19.
[56] M. Smith, Neural Networks for Statistical Modeling, International Thomson Computer Press, London, 1996.
[57] S. Lessmann, Solving imbalanced classification problems with support vector machines, in: Proceedings of the International Conference on Artificial Intelligence, 2004.
[58] I.H. Witten, F. Eibe, Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann, San Francisco, 1999.
[59] C.-C. Chang, C.-J. Lin, LIBSVM—A Library for Support Vector Machines, 2001, Downloadable from website http://www.csie.ntu.edu.tw/~cjlin/libsvm.
[60] F. Provost, T. Fawcett, R. Kohavi, The case against accuracy estimation for comparing induction algorithms, in: Proceedings of the 5th International Conference on Machine Learning, 1998.
[61] J. Banslaben, Predictive modelling, in: E.L. Nash (Ed.), The Direct Marketing Handbook, second ed., McGraw-Hill, New York, 1992.
[62] M.J.A. Berry, G. Linoff, Data Mining Techniques: For Marketing, Sales and Customer Relationship Management, second ed., Wiley, New York, 2004.