Customer-Centric Decision Support

A Benchmarking Study of Novel Versus Established Classification Models

Classification analysis contributes to the support of several corporate decision making tasks. In particular, the domain of customer relationship management comprises a variety of respective applications, which involve estimating some aspects of customer behavior. Whereas classical statistical techniques as well as decision tree models are routinely employed to approach such tasks, the use of modern techniques is still in its infancy. A major obstacle prohibiting a wider adoption of novel methods may be seen in the fact that their potential to improve decision quality in customer-centric settings has not yet been investigated. Therefore, this paper contributes to the literature by conducting an empirical study that compares the performance of modern to that of established classifiers as to their predictive accuracy and economic consequences. The observed results provide strong evidence for the value of modern techniques and identify one approach which appears to be particularly well suited for solving customer-centric classification problems.

DOI 10.1007/s12599-010-0094-8

The Authors

Dr. Stefan Lessmann (✉)
Prof. Dr. Stefan Voß
Institut für Wirtschaftsinformatik
Universität Hamburg
Von-Melle-Park 5
20146 Hamburg
Deutschland
lessmann@econ.uni-hamburg.de,
url: http://iwi.econ.uni-hamburg.de,
stefan.voss@uni-hamburg.de

Received: 2009-01-20
Accepted: 2009-10-26
Accepted after two revisions
by Prof. Dr. Buhl.
Published online: 2010-03-09

This article is also available in German in print and via http://www.wirtschaftsinformatik.de: Lessmann S, Voß S (2010) Unterstützung kundenbezogener Entscheidungsprobleme. Eine Analyse zum Potenzial moderner Klassifikationsverfahren. WIRTSCHAFTSINFORMATIK. doi: 10.1007/s11576-010-0216-4.

Electronic Supplementary Material
The online version of this article (doi: 10.1007/s12599-010-0094-8) contains supplementary material, which is available to authorized users.

1 Introduction

The field of data mining embraces techniques and tools to analyze large, heterogeneous datasets and uncover hidden patterns that may prove valuable to support decision making. In corporate contexts, data mining can be employed to, e.g., confirm the efficacy of business processes, gain a better understanding of customer behavior, needs and preferences, and, more generally, identify opportunities for gaining competitive advantage.

Classification analysis belongs to the branch of directed data mining (Berry and Linoff 2004, p. 7) and aims at estimating the probability of events on the basis of past observations. For example, many corporate applications of classification aim at solving operational planning problems in (analytical) customer relationship management like assessing the credit worthiness of loan applicants, identifying an appropriate target group for direct-mailing campaigns, detecting fraud, e.g., in the financial or insurance industry, or identifying customers at the risk of churning prior to defection (Lessmann and Voß 2008, p. 237 ff). These planning tasks are referred to as *customer-centric classification problems* throughout this paper.

The development of novel methods to solve classification problems enjoys ongoing popularity in data mining and related disciplines, so that a large number of alternative methods are available. Not surprisingly, algorithmic advancements are usually not adopted immediately in corporate practice, where classical techniques like logistic regression or decision tree approaches prevail (Cui and Curry 2005, p. 595; Friedman 2006, p. 180). However, a wider consideration of novel classification methods could be justified since several data mining software systems already support such techniques and some evidence for their superiority has been provided in the literature. Consequently, it is desirable to examine whether encouraging results from other scientific domains also hold true for the field of customer-centric classification. More specifically, an empirical

proof is needed to scrutinize whether novel techniques offer economic advantage over their traditional counterparts.

Clearly, a redesign of corporate planning processes for the use of more advanced classifiers requires some initial investments to be made. In particular, an upgrade of an existing data mining system or even the purchase of a new software package may be necessary. Moreover, expenditures for attaining the required know-how to master new methods have to be considered. An accurate estimation of respective costs is of pivotal importance. However, costs can be expected to depend strongly upon the particular business, i.e., differ substantially from company to company, and should be relatively easy to anticipate. Therefore, investment costs are not considered in this paper. Instead, possible revenue increases are emphasized that may be achievable by employing novel classification methods. For example, higher predictive accuracy, and thus higher decision quality, may help to avoid some bad risks in consumer lending and thereby increase a company's profits. Consequently, the paper strives to assess the economic value derived from the use of novel classification methods within customer-centric applications. To that end, an empirical benchmark experiment is undertaken, which contrasts several established and novel classifiers regarding a monetary accuracy measure. Therefore, the study facilitates appraising the merit of novel methods within the considered domain as well as an identification of particularly suitable techniques.

The paper is organized as follows: The next section elaborates the experiment's motivation in detail and reviews the related literature. Section 3 explains the experimental design, before empirical results are provided in Sect. 4. The paper concludes with a summary and discussion of the main findings (Sect. 5) as well as limitations and opportunities for future research (Sect. 6). The Appendix contains further details concerning experimental design.

2 Related Literature and Motivation

Techniques for solving classification problems enjoy ongoing popularity in data mining as well as adjacent disciplines like statistics and machine learning. Specifically, a common undertaking is to develop novel algorithms, e.g., to account for special requirements of a particular – possibly novel – application. The development of a new or the modification of an existing procedure is usually accompanied by an empirical evaluation to verify the efficacy of the proposed approach, whereby 'efficacy' is routinely measured in terms of the accuracy of a model's predictions.[1]

Benchmarking experiments are a popular way to complement mainly algorithmic-centric research by contrasting several alternative classification models in different applications. Early studies include Curram and Mingers (1994), Weiss and Kapouleas (1989) as well as the well-known Statlog project (King et al. 1995). One of the largest experiments has been conducted by Lim et al. (2000); more recent results are presented by Caruana and Niculescu-Mizil (2006). An advantage of benchmarking experiments stems from the fact that they facilitate an independent assessment of autonomously developed classification models and, thereby, a verification and confirmation of previous results. Such replications are an imperative part of empirical research (Fenton and Neil 1999, p. 680; Ohlsson and Runeson 2002, p. 217). In contrast to an independent evaluation, empirical assessments carried out by the developers of a new technique, i.e., within the paper that initially proposes the method, bear the risk of being overly optimistic. In these cases encouraging results may – to some extent – be due to the particular expertise of the developers but not be reproducible by others.

In addition to *general* benchmarks that comprise multiple techniques and data from various domains, several comparative studies target clearly defined methodological sub-problems. Respective research includes classification with ensemble methods (Bauer and Kohavi 1999; Dietterich 2000; Hamza and Larocque 2005; Hothorn and Lausen 2005; Sohn and Shin 2007; Wang et al. 2009) or a particular method in general (Meyer et al. 2003; van Gestel et al. 2004), the effect of skewed class distributions (Batista et al. 2004; Burez and van den Poel 2009; Hulse et al. 2007) or asymmetric misclassification costs (Ting 2002; Weiss and Provost 2003) as well as the effect of dataset size (Perlich et al. 2003) and alternative accuracy indicators (Caruana and Niculescu-Mizil 2004; Ferri et al. 2009). Furthermore, benchmarks are carried out in the context of special application domains to identify particularly appropriate techniques (Cooper et al. 1997; Khoshgoftaar and Seliya 2004; Lessmann et al. 2008; Liu et al. 2003; Zickus et al. 2002). This paper belongs to the latter category.

Taking an Information Systems perspective, applications of classification models associated with corporate planning tasks are most relevant. In particular, the field of analytical customer relationship management embodies multiple decision problems that can effectively be addressed by means of classification analysis. A literature survey of respective tasks and solutions can be found in Lessmann and Voß (2008, p. 237 ff) as well as Ngai et al. (2009), and for particular sub-domains in Bose and Xi (2009) as well as Crook et al. (2007). In general, the papers discussed there reflect the abovementioned situation: a novel solution for a particular planning problem (e.g., credit scoring) is proposed and empirically compared with selected – mainly traditional – benchmark methods. To that end, one or more datasets are considered which represent either artificial or real-world classification problems. Consequently, comparisons of several state-of-the-art techniques are scarce. Moreover, the relatively small scope of many experiments (e.g., the number and size of datasets as well as the number and type of benchmark methods) may prohibit a generalization of observed results. Considering the importance of (analytical) customer relationship management (Hippner 2006, p. 362) and customer-centric classification problems, respectively, it is desirable to obtain a more holistic picture of alternative classifiers' competitive performance within this domain. Benchmarking studies like those carried out in other fields contribute towards achieving this goal. However, only very few comparative experiments are dedicated to customer-centric classification, with Baesens et al. (2003); Burez and van den Poel (2009) and Viaene et al.

[1] This is also shown by a scientometric analysis of papers published at the *International Conference on Machine Learning* between 1999–2003 (Demšar 2006, p. 3 ff).

(2002) being noteworthy exceptions. Burez and van den Poel (2009) consider the problem of churn prediction and examine six real-world applications. The size of the datasets employed is a remarkable distinction of this experiment. However, the authors emphasize the negative effect of imbalanced class distributions as commonly encountered in customer attrition analysis. Therefore, they restrict their study to only two classification models. Baesens et al. (2003) conduct a large-scale experiment in the field of credit-scoring which comprises 17 classifiers and eight datasets. Analogously, seven techniques are compared in a case of automobile insurance claim fraud detection in Viaene et al. (2002). Both studies leave out ensemble classifiers, which had not received much attention at the time these studies were conducted. However, they are nowadays considered as most powerful off-the-shelf classifiers (Hamza and Larocque 2005, p. 632). A more severe problem may be seen in the fact that all three studies employ proprietary data.[2] Consequently, a replication of results as well as a comparison of future methods with those considered in these papers on identical data is impossible.

In summary, it may be concluded that the question whether specific classification methods are particularly well suited for decision problems in the realms of customer relationship management and whether novel techniques perceptibly improve decision quality has not yet received sufficient attention. Its importance follows directly from the relevance of respective planning tasks in corporate practice, e.g., for risk-management in consumer lending, for targeting in direct-marketing and the mail-order industry or for a proactive identification of customers at risk of churning. Therefore, this paper contributes to the literature by conducting a large-scale benchmark of established versus novel classification methods in customer-centric applications. In particular, the following characteristics enable a clear distinction from previous endeavors: (1) All datasets employed in the study represent customer-centric planning tasks and are publicly available. The former is to facilitate the generalization of findings to a certain extent, whereas the latter permits a replication of results by other researchers. (2) A large number of alternative classifiers are considered, so that the benchmark embraces methods which are currently used in corporate practice as well as cutting-edge techniques. (3) Prior to assessment, all classifiers are tuned to a particular problem instance in a fully automatic manner. Consequently, the comparison is fair as well as representative. Specifically, it must not be assumed that in-depth expertise with every classifier is available in all corporate applications. Therefore, an autonomous adaptation of methods to tasks does not only embody a key Information Systems objective, but also facilitates a realistic appraisal of the methods' predictive potential. (4) The experimental design incorporates several repetitions as well as statistical tests particularly suited for comparing classifiers. Both factors are meant to secure the validity and reliability of the study's results. (5) The economic consequences of employing a classifier within a particular decision context serve as major indicator of predictive accuracy. (6) A large number of experiments to scrutinize and secure the generalizability of empirical results are conducted. For example, the degree to which results depend upon task-specific characteristics as well as the particular selection of tasks itself is appraised.

Due to the features mentioned above, the study enables a pre-selection of techniques which appear especially well suited for practical implementation. Moreover, the experimental design can be re-used in practical applications to identify the best method for a particular task. That is, the study's setup may be considered a reference model or best-practice for assessing alternative classifiers. Furthermore, future experiments, e.g., to benchmark novel methods yet to be developed, may reference the results of this study and re-use the datasets employed.

3 Experimental Design

3.1 Classification Analysis

The term classification describes the process and the result of a grouping of objects into a priori known classes. The objects are characterized by measurements or attributes that are assumed to affect class membership. However, the concrete relationship between attribute values and class is unknown and needs to be estimated from a sample of example cases, i.e., objects with known class (Izenman 2008, p. 237 ff). The resulting decision function is termed a classifier or, synonymously, a classification model and enables predicting the class membership of novel objects, i.e., cases where only the attribute values are known. Therefore, the primary objective of classification is to forecast the class membership of objects to the highest degree possible.

Several customer-centric classification problems require a distinction between one economically relevant class, e.g., bad credit risks or customers at the risk of churning, and an alternative group (Lessmann and Voß 2008, p. 233). Consequently, attention is restricted to two-class problems in this paper.

Classification methods are developed in several scientific disciplines including statistics and machine learning. The former are commonly based upon probabilistic considerations and strive to estimate class membership probabilities. This principle is exemplified by, e.g., the well-known logistic regression. Machine learning methods commonly operate in a completely data-driven fashion without distributional assumptions and try to categorize objects, e.g., by means of rule-induction, into groups. Decision tree methods are established representatives of this approach. Regarding more recent classification techniques, two main branches can be distinguished (Friedman 2006, p. 175). These include support vector machines, which are motivated by statistical learning theory (Vapnik 1995) and are particularly well suited for coping with a large number of explanatory factors (i.e., attributes), as well as ensemble methods that ground on the principle of combining a large number of individual classification models to improve predictive accuracy. Representatives of this category differ mainly in terms of their approach to construct complementary classifiers. Therefore, the individual classifiers, commonly termed base models, must show some diversity to improve the accuracy of the full model. A comprehensive description of classification as well as traditional and contemporary models can be found in standard textbooks like, e.g., Hastie et al. (2009) and Izenman (2008).

With regard to the large variety of alternative classifiers a selection has to be made for the present study, whereby the setup should comprise methods which

[2] Two of the eight datasets used by Baesens et al. (2003) are publicly available.

Table 1 Classification methods of the benchmarking study

Classifier	Description
Multivariate statistical classifiers	
Naive Bayes (NBayes)	Approximates class-specific probabilities under the assumption of all attributes being statistically independent.
Linear discriminant analysis (LDA)	Approximates class-specific probabilities by means of multivariate normal distributions assuming identical covariance matrices. This assumption yields a linear classification model, whose parameters are estimated by means of maximum likelihood procedures.
Quadratic discriminant analysis (QDA)	In comparison to LDA, assumptions concerning covariance matrices are less restrictive, resulting in a quadratic model.
Logistic regression (LogReg)	Approximates class-membership probabilities by means of a logistic function, whose parameters are determined through maximum likelihood estimation.
K-nearest neighbor classifier (K-NN)	Classifies an object into the class prevailing among its K nearest neighbors.
Decision trees	
C4.5	Induces a decision tree by partitioning the training data so as to maximize the reduction of entropy within tree nodes.
CART	Similar to C4.5, but organizes data partitioning according to the *Gini* coefficient rather than information entropy.
Support vector machine type methods	
Linear support vector machine (LSVM)	Separates objects by means of a linear hyperplane whose normal and intercept follow from solving a mathematical program that maximizes the distance between objects of adjacent classes.
Support vector machine with radial basis function kernel (RBFSVM)	Extends LSVM by projecting data into a feature space of higher dimension prior to separation. As a consequence, a nonlinear decision surface is constructed in the original input space.
Relevance vector machine (RVM)	Modification of RBFSVM according to Tipping (2000) that circumvents problems associated with selecting parameters of the classifier.
Ensemble-methods	
Bagging (Bag-base classifier)	Several base classifiers are derived from bootstrap samples of the training data. The base models' class predictions are averaged to form the final forecast. It is implemented with the following base classifiers in this study: NBayes, LogReg, C4.5 and CART.
Random-Forest (RF)	Extends bagging by constructing several CART classifiers from bootstrap samples whereby only a randomly selected subset of attributes is considered to split a node in individual CART trees. This modification is supposed to increase diversity among base classifiers.
Boosting (SGB)	Constructs an ensemble of classifiers in an iterative manner. The base classifiers to be appended to the collection are built so as to avoid the errors of the current ensemble. Specifically, Friedman's (2002) stochastic gradient boosting is employed in this study.

are currently popular in corporate practice as well as state-of-the-art techniques. Concerning the latter, an additional constraint is imposed. In particular, only methods that have been implemented in some data mining software package are considered. This is to ensure that all considered classifiers can in principle be utilized in corporate practice with acceptable efforts. In other words, the constraint serves the objective to conduct a pre-selection of candidate methods for practical data mining applications.

The group of "novel classifiers" consists of ensemble methods and support vector machine type techniques, which have been developed in the mid-nineties and made available in software systems, respectively. The chosen methods are shown and described in **Table 1**.

Some classifiers may not be used off-the-shelf but require the user to determine some parameters. The approach to identify suitable settings is described in Appendix I.

3.2 Decision Problems

The benchmarking study comprises nine publicly available datasets that represent real-world decision problems in customer-centric data mining. Four datasets stem from the UCI Machine Learning Repository (Asuncion and Newman 2007), whereas the remaining tasks are selected from the annual Data Mining Cup competition[3] organized by Prudsys AG.

The datasets *Australian* and *German Credit* represent classification problems from the field of credit scoring. The binary target variable indicates whether a customer has defaulted on a loan. Direct-marketing tasks are represented by five datasets: *Adult, Coil, DMC 2000, 2001, 2004*. The Adult datasets is concerned with the prediction of US households' annual income (below/above

[3] http://www.data-mining-cup.de.

Table 2 Characteristics of the datasets employed in the benchmarking study

	No. of cases	No. of attributes	A priori probability of the economically relevant class[a]	A priori probability of the alternative class
AC	690	14	44.49	55.51
GC	1000	24	30.00	70.00
Adult	48,842	14	23.93	76.07
Coil	9822	86	5.97	94.03
DMC 2000	38,890	96	5.87	94.13
DMC 2001	28,128	106	50.00	50.00
DMC 2002	20,000	101	10.00	90.00
DMC 2004	40,292	107	20.43	79.57
DMC 2005	50,000	119	5.80	94.20

[a]Each dataset contains a single class that is of primary interest from an economical perspective. For example, identifying defaulting customers can be considered most relevant in credit scoring. In other words, an error in detecting a bad risk is more costly than the reverse error of denying credit to a solvent customer. The important class is encoded as class 1 in this study; the alternative group as class 0

$ 50.000), whereby demographic and socio-demographic census data is provided to construct a classification model. A respective analysis could be part of prospect management (Haas 2006), e.g., to identify an appropriate target group for a marketing campaign. The *Coil* dataset has been employed within a previous classification competition (Putten and Someren 2000). The objective is to predict whether a customer is interested in purchasing an insurance policy. An analogous question is considered in *DMC 2000*. The data stems from the mail-order industry and characterizes the response behavior observed within a customer acquisition campaign. Another decision problem from the catalog industry is considered in *DMC 2001*. Specifically, the aim is to distinguish between customers who receive a detailed product catalog and those who are only informed about product innovation. This is to reduce costs of serving less prosperous customers and thereby optimize mailing efficiency. Finally, the tendency to return items previously ordered from a catalog is examined in *DMC 2004*. In view of high costs associated with managing returns, a timely identification of "high-return" customers is desirable to deal with them in a more efficient manner.

DMC 2002 represents a typical churn prediction problem. In particular, after privatization of the German energy market, strategies for sustaining customers have become imperative in this sector. The data is provided by a German utility company that strives to proactively identify customers at risk of abandoning their relationship in order to sustain them by, e.g., offering a discounted price.

The problem of fraud detection in online businesses is explored in *DMC 2005*. Specifically, the task's objective is to identify high-risk customers whose payment options in online transactions should be restricted.

The main characteristics of the datasets are summarized in **Table 2**. More detailed information (e.g., concerning customer attributes) are available at the Data Mining Cup website and the aforementioned sources, respectively.

3.3 Assessing Predictive Accuracy

Evaluating a classification model's predictive performance requires selecting an appropriate indicator of forecasting accuracy and a procedure to simulate a real-world application of the method. The monetary consequences resulting from using a classifier are considered as primary accuracy indicator in this study. To that end, correct/wrong class predictions are weighted with profits and misclassification costs, respectively and aggregated over all cases of a dataset to obtain an overall utility measure. This procedure as well as the employed costs/profits are described in detail in Appendix II, whereas the influence of alternative accuracy indicators on the competitive performance of classifiers is explored in Sect. 4.2.1.

A method's real-world application is usually simulated by randomly partitioning the data into two disjoint sets. Then, a classification model is built from the first set (training data) and applied to the cases of the second dataset (test data). Since these examples have not been employed during training, they enable an unbiased assessment of the classifier. This split-sample strategy is adopted in the present study, using a ratio of 60:40 to partition all datasets into training and testing set. In order to decrease variance, the partitioning is repeated ten times and performance estimates are averaged over the resulting random test sets.

4 Empirical Results

4.1 Comparisons in Terms of Profits/Misclassification Costs

The benchmarking study strives to shed light on the question whether alternative classification methods exert a substantial effect on decision making quality. To that end, the first experiment contrasts the monetary consequences arising from employing the classifiers within the considered applications.[4] Respective results are shown in **Table 3**, where performance estimates are averaged over ten randomly drawn test sets and the corresponding standard deviations are given in square brackets. The second row of **Table 3** indicates whether a task's objective is cost minimization (C) or profit maximization (P). The best result per dataset is highlighted in bold.

The empirical results demonstrate that predictive performance (i.e., profits and misclassification costs) varies considerably across alternative classification models. For example, the standard deviation

[4]The procedure to compute this measure is documented in Appendix II, together with the costs/profits of false/correct classifications per dataset.

Table 3 Monetary assessment of alternative classification models

Classifier (see **Table 1**)	*AC*		*GC*		*Adult*		*Coil*		*DMC 2000*		*DMC 2001*		*DMC 2002*		*DMC 2004*		*DMC 2005*	
	C		C		C		P		P		P		P		P		P	
NBayes	31.6	(2.0)	323.3	(23.2)	1,623.7	(33.2)	190.1	(11.8)	−44,084	(2,025)	4,473,860	(45,217)	519,427	(1,585)	8,435	(68)	256,545	(763)
LDA	19.1	(2.1)	**271.7**	(16.5)	1,968.0	(28.6)	190.1	(11.8)	−80,972	(2,267)	4,523,052	(59,375)	519,209	(1,540)	8,897	(59)	255,080	(669)
QDA	20.7	(2.7)	318.5	(28.9)	1,920.7	(35.3)	190.1	(11.8)	−78,463	(2,090)	4,610,222	(42,933)	519,907	(2,150)	8,429	(65)	255,031	(795)
LogReg	19.8	(1.9)	276.5	(15.3)	1,513.3	(28.7)	190.1	(11.8)	−78,267	(1,958)	4,550,985	(47,987)	522,467	(1,113)	10,156	(340)	**257,914**	(805)
K-NN	22.9	(4.4)	368.9	(30.9)	1,621.6	(33.4)	190.1	(11.8)	−35,911	(3,331)	4,313,006	(50,971)	519,427	(1,532)	8,912	(276)	254,324	(935)
CART	26.9	(3.6)	366.2	(44.4)	1,495.1	(34.1)	190.1	(11.8)	−64,628	(16,416)	4,583,786	(58,340)	518,060	(2,102)	9,665	(303)	253,171	(1,369)
C4.5	23.6	(2.3)	340.7	(40.2)	1,428.2	(43.0)	190.1	(11.8)	−22,309	(4,115)	4,475,442	(48,231)	520,729	(1,255)	9,005	(70)	254,669	(1,165)
LSVM	19.6	(2.0)	272.9	(20.5)	1,518.3	(29.9)	190.1	(11.8)	−41,948	(1,331)	4,582,977	(48,026)	522,209	(1,226)	10,254	(76)	257,688	(837)
RBFSVM	19.2	(2.5)	297.9	(28.3)	1,490.3	(18.0)	190.1	(11.8)	−1,504	(2,026)	4,625,056	(46,791)	520,988	(1,571)	**10,310**	(83)	254,629	(1,311)
RVM	19.2	(2.2)	273.5	(17.1)	1,514.6	(28.4)	190.1	(11.8)	−42,948	(1,745)	4,585,178	(48,785)	522,249	(1,188)	10,261	(73)	257,798	(854)
Bag-NBayes	32.7	(2.3)	330.3	(12.9)	1,624.7	(25.1)	185.8	(10.3)	−43,919	(1,631)	4,510,367	(48,651)	520,018	(1,663)	8,391	(77)	256,233	(1,378)
Bag-LogReg	19.1	(1.8)	296.7	(20.8)	1,512.0	(21.3)	180.5	(10.3)	−42,899	(3,106)	4,622,923	(50,466)	**522,715**	(1,524)	10,306	(80)	257,535	(1,183)
Bag-CART	19.6	(3.6)	318.3	(24.6)	1,432.5	(18.8)	186.6	(11.9)	−171	(1,874)	4,627,051	(50,497)	521,310	(1,494)	10,089	(49)	256,656	(1,311)
Bag-C4.5	19.3	(3.0)	300.3	(26.5)	**1,366.3**	(16.5)	177.5	(7.1)	−975	(2,346)	4,636,133	(43,545)	522,067	(1,260)	10,151	(69)	257,233	(1,240)
RF	19.6	(3.4)	293.3	(21.0)	1,414.6	(19.9)	**199.2**	(9.1)	**6,649**	(3,072)	**4,657,196**	(40,472)	520,454	(1,276)	10,219	(69)	256,184	(810)
SGB	**19.0**	(3.0)	288.5	(20.6)	1,368.2	(29.0)	192.2	(9.0)	−1,278	(2,294)	4,619,028	(50,472)	521,264	(1,233)	10,161	(66)	257,560	(895)
Mean[a]	*22.0*	*(4.5)*	*308.6*	*(31.3)*	*1,550.8*	*(173.9)*	*189.0*	*(4.8)*	*−35,852*	*(30,114)*	*4,562,266*	*(87,586)*	*520,781*	*(1,366)*	*9,603*	*(772)*	*256,140*	*(1,479)*
		21%		*10%*		*11%*		*3%*		*84%*		*2%*		*0%*		*8%*		*1%*
Best vs. C4.5	*4.6*	*19%*	*69.0*	*20%*	*61.9*	*4%*	*9.0*	*5%*	*28,958*	*130%*	*181,753*	*4%*	*1,986*	*0%*	*1,306*	*14%*	*3,245*	*1%*
Best vs. LogReg	*0.8*	*4%*	*4.8*	*2%*	*147.0*	*10%*	*9.0*	*5%*	*84,916*	*108%*	*106,211*	*2%*	*249*	*0%*	*154*	*2%*	*0*	*0%*

[a]The mean is computed per dataset across all classifiers. The respective standard deviation is calculated analogously and shown in square brackets. In addition, the (percentage) coefficient of variation (i.e., standard deviation over mean) is given. Being a normalized measure, this figure facilitates comparing classifier-induced performance variation across datasets

of misclassification costs for *AC* is 4,5. This translates into a 21% deviation from the mean costs across all methods. Respective statistics are given for all datasets in the row *Mean* in **Table 3**.

More formally, significant performance variations may be detected by means of the *Friedman*-Test (Demšar 2006, p. 9ff.). Specifically, the test's null-hypothesis that no significant differences exists between classifiers can be rejected with high probability (>0.9999) for the results of **Table 3**. Therefore, it may be concluded that the selection of a particular classification model has a significant impact upon predictive performance and, thus, decision quality.

It has been reported that considering only a single model, predominantly logistic regression, is still a common practice in many (marketing) applications (Cui and Curry 2005, p. 595; Lemmens and Croux 2006, p. 276). Considering previous results, such practices should be revised. In particular, the last row of **Table 3** gives the relative difference between the best method per dataset and logistic regression and C4.5, respectively. Apparently, improvements of some percent over these well established techniques are well achievable and would translate into noteworthy profit increases/cost reductions in practical applications. For example, a two-percent increase in predictive performance leverages profits of € 106,211 in the case of *DMC 2001*. Consequently, to solve a given decision problem, a comparison of multiple alternative classifiers should axiomatically be undertaken, i.e., to identify the most suitable candidate model.

Having established the need for classifier benchmarks in general, the following experiment strives to clarify whether more attention should be devoted to novel classifiers within such comparisons. To that end, the results of **Table 3** are normalized so as to ensure that all figures range from zero to one, with one denoting highest performance for a given dataset. Afterwards, the mean performances for the two groups of established versus novel classifiers are calculated, whereby the former are represented by statistical and nearest neighbor methods as well as decision tree classifiers and the latter by support vector machines and ensembles. (see **Table 1**). The results of the comparison are shown in **Fig. 1**.

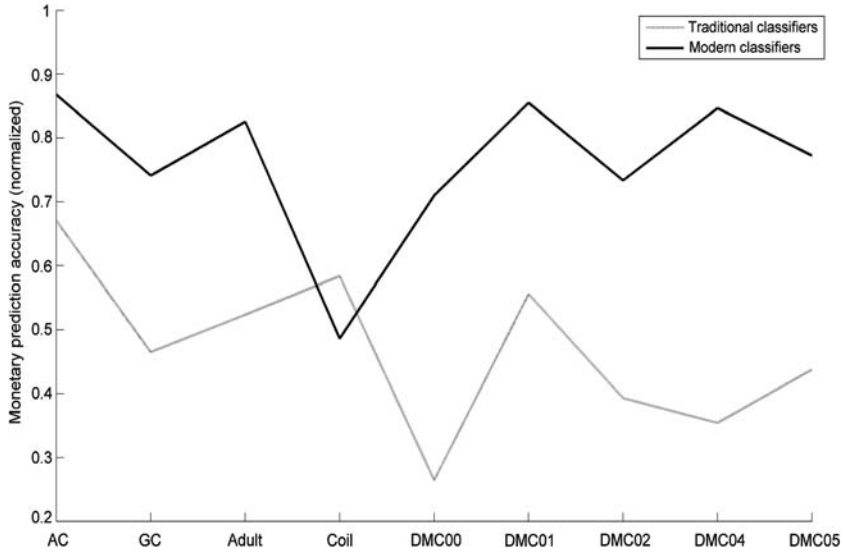

Fig. 1 Comparison of established versus novel classifiers

Fig. 1 shows a clear trend towards modern methods. The latter deliver superior predictive performance in eight out of nine cases. A one-tailed T-test for paired samples confirms that the mean performance of novel classifiers is significantly higher than those of classical techniques.[5] Therefore, a stronger consideration of modern classification models in practical application appears well justified. This means, they should be considered alongside traditional methods when selecting a classifier for a given decision problem.

However, it has to be scrutinized whether the previous results facilitate more concrete recommendations, i.e., with respect to which particular techniques should be considered, or in other words appear most suitable for customer-centric classification. To shed some light upon this question, the subsequent experiment performs a statistical test, the *Nemenyi*-Test, for all possible pairwise combinations of classifiers. The test checks whether performance differences between two techniques are statistically significant (Demšar 2006, p. 9 ff). The *Nemenyi*-Test is based upon differences in classifier rankings. A ranking is obtained by ordering all classifiers according to their performance from best (rank one) to worst (rank sixteen), and averaging the resulting ranks across all datasets. The test results are presented in **Fig. 2**, which depicts all classifiers' mean ranks in ascending order. Hence, a low (mean) rank indicates superior forecasting accuracy. The horizontal lines represent significance thresholds. That is, the rightmost end of each line indicates from which mean rank onwards the corresponding classifier significantly outperforms an alternative method at the 5%-level.

The classifier ranking re-emphasizes the superiority of novel methods in the sense that they achieve better (lower) ranks than their traditional counterparts with very few exceptions. In view of the widespread use of decision trees in practical applications, the performance of the two representatives C4.5 and CART is particularly disappointing and casts doubt upon their appropriateness for the domain considered here. In contrast, the logistic regression classifier is well competitive to any new method.

Predictions of overall best performing classifier, SGB, are significantly better than some alternative methods. However, a significant difference between SGB and, e.g., logistic regression cannot be detected. In all cases where classifier performances do not differ significantly, it must be concluded that the empirical results do not provide sufficient evidence for judging whether the observed perfor-

[5] The empirical results allow rejecting the null-hypothesis that the mean forecasting accuracy between both groups does not differ with high probability (>0.9999).

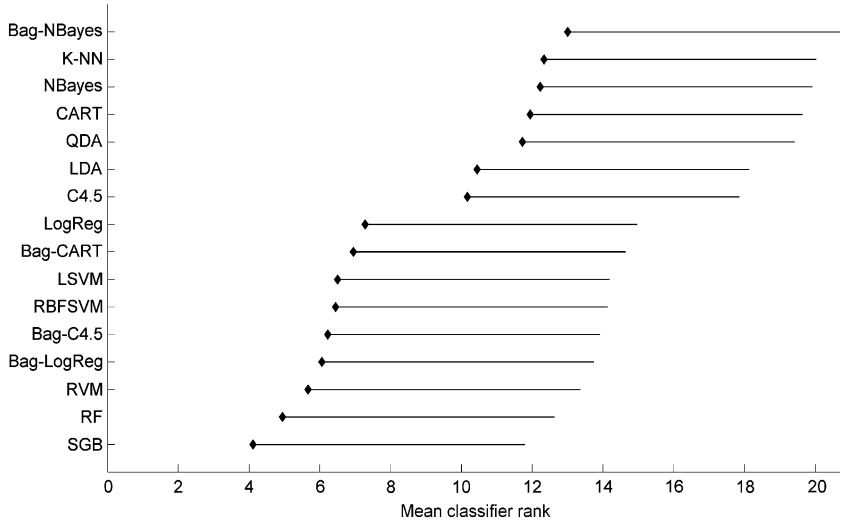

Fig. 2 Results of the Nemenyi-Test with significance level $\alpha = 0.05$

mance (i.e., rank) differences are systematic or random.[6]

Since several pairwise comparisons remain insignificant, the following experiments aim at appraising the stability and generalizability of observed results and clarifying whether, e.g., SGB or the runner-up RF are really well suited for customer-centric classification.

4.2 External Validity of Benchmarking Results

4.2.1 The Impact of Alternative Accuracy Indicators

Previous experiments have employed a monetary accuracy indicator to measure a classification model's predictive power. Although this figure can be considered highly relevant in corporate planning contexts, it depends upon the specific decision problem and especially upon the (estimated) costs/profits of wrong/correct classifications (see **Table 7** in Appendix II). Consequently, the study's results, i.e., the performance ranking of alternative classification models, may differ if other accuracy indicators are employed.

Several statistics have been proposed in the literature to assess a classifier's (predictive) performance. In particular, three main branches can be distinguished: Some indicators ground on a discrete (crisp) classification of objects into classes and measure – in different ways – the accuracy of such a categorization. Among others, this group includes the well-known classification accuracy (percentage of correct classification) and its inverse, classification error, as well as different averages of class-individual accuracy/error-rates. The monetary indicator used above also belongs to this category. Other accuracy indicators take the distance between a classifier's prediction and an object's true class into account. Representatives of this type include, e.g., the mean squared error measure which is commonly employed to assess the forecasting accuracy of regression models. Finally, some accuracy indicators assess a model's capability to rank objects according to their probability of being member of a particular class. The area under a receiver-operating-characteristics curve (AUC) is probably the most widely used ranking measure (Fawcett 2006, p. 873).

In order to explore the influence of alternative accuracy indicators on the study's results, the previous analysis (see Sect. 4.1) is repeated with nine alternative indicators. The selection of different candidate measures is guided by Caruana and Niculescu-Mizil (2006) and includes discrete, distance-based and ranking indicators. Specifically, classification accuracy (CA), the arithmetic (A-Mean) and geometric mean (G-Mean) of class-specific classification accuracy, the F-measure,[7] the Lift-Index, which is particularly popular in marketing contexts, with thresholds of 10 and 30%, the mean cross-entropy (MXS), the root mean squared error (RMSE) and the AUC are considered. A more detailed description of these measures is given by Caruana and Niculescu-Mizil (2004, p. 77–78) as well as Crone et al. (2006, p. 790).

To take account of different measurement ranges among different indicators, all results are normalized to the interval [0, 1], whereby a value of one represents an optimal prediction. Consequently, nine performance estimates (one per indicator) are obtained per classifier and dataset. Due to normalization, these can be averaged to give an aggregated performance measure per classifier and dataset. This aggregated accuracy indicator is referred to as mean normalized performance (MNP) in the remainder of the paper.

Ranking classifiers according to MNP and averaging over datasets, the statistical comparison of performance differences (**Fig. 2**) can be repeated, whereby mean ranks are now computed in terms of MNP. Respective results are shown in **Fig. 3**.

A comparison of **Figs. 2** and **3** reveals minor differences within the two classifier rankings. However, the strong trend towards novel method persists and SGB once more achieves the overall best result, significantly outperforming QDA and all other competitors with higher rank. Moreover, the similarity of the two rankings (**Figs. 2** and **3**) may be confirmed by means of a correlation analysis (**Table 4**). In particular, the ranking of classifiers across all datasets is determined for each accuracy indicator individually. Subsequently, correlations between all possible pairs of rankings are computed to appraise the degree of correspondence between classifier performances in terms of different accuracy indicators.[8] A strong positive correlation (>0.6) can be observed in most cases and most correlations are statistically signif-

[6] Note that the reverse conclusion, i.e., that two classifiers perform alike if they do not differ significantly, is statistically incorrect. Whereas rejecting the null-hypothesis facilitates the conclusion that the alternative hypothesis is correct with high probability (i.e., 1-significance level), a failure to do so does not allow to draw any conclusions regarding the correctness of the alternative hypothesis.

[7] The F-measure is widely used in the field of Information Retrieval. Basically, it is calculated as the weighted harmonic mean of precision and sensitivity (Caruana and Niculescu-Mizil 2004, p. 78), whereby a weight of 0.5 is used in this study.

[8] *Kendall*'s Tau is used to assess ranking correlation.

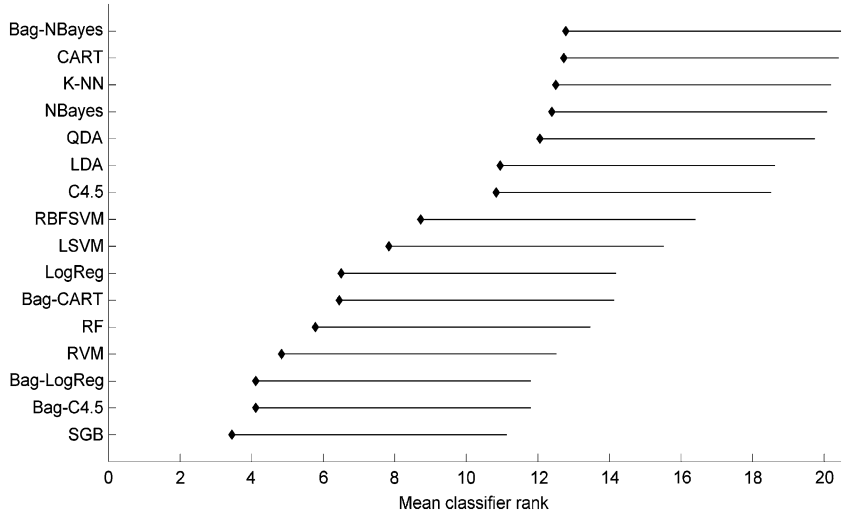

Fig. 3 Results of the Nemenyi-Test with significance level $\alpha = 0.05$ on the basis of a MNP-based ranking of classifiers across all datasets

icant. Especially the monetary accuracy indicator is highly correlated with many alternative indicators, which follows from the first column of **Table 4**.

In view of the results of **Fig. 3** and **Table 4**, it may be concluded that the previous findings concerning the general suitability of novel classification models and the particular appropriateness of SGB do not depend upon the employed accuracy indicators. In other words, the comparative performance of alternative classifiers is similar over a wide range of candidate indicators.

4.2.2 The Impact of Problem Characteristics

In addition to alternative accuracy indicators, the particular characteristics of the decision problem could affect a classification model's performance. For example, it is possible that some classifiers excel in specific circumstances, but fail in others. If such patterns exist, they could offer valuable information concerning the requirements of a classification model and thus complement the study's results. Furthermore, classifiers' sensitivity (or robustness) towards problem-specific characteristics is important for assessing the external validity of the observed results, i.e., to what extent they may be generalized to other tasks.

Within classification analysis, each decision task is represented by a collection of classes, objects and measurements (i.e., attributes). All classifiers operate on this abstract level, i.e., a dataset of examples. Consequently, the characteristics of a respective dataset are an important determinant of classification performance. In particular, **Table 2** identifies dataset size, the number of (customer) attributes as well as the (im-)balance of class distributions as possible drivers of classification performance. In addition, the complexity of the prediction task itself can be considered a fourth determinant characterizing a particular decision problem. The latter may be approximated by the means of the forecasting accuracy that can be achieved with a simple classifier.

In order to scrutinize the importance of problem specific characteristics for classifier performance, different levels for each factor need to be defined to, e.g., decide when to consider the number of attributes as "large". This step entails some uncertainty. **Table 5** shows a grouping that strives to achieve a balance between an accurate separation on the one hand and a manageable number of factor levels on the other. It should be noted that the predictive accuracy of Naïve Bayes has been employed to classify decision problems according to their complexity.[9]

Table 5 represents the basis for the following experiment: For each factor level, we compute the rank a classifier would achieve if only datasets (problems) of the respective level were incorporated in the study. For example, consider the factor dataset size and the factor level small. In this case, a classifier's performance is calculated as the average of its ranks on *AC, GC* and *Coil*, whereby the MNP is employed as performance measure to account for the possible effect of alternative accuracy measures.

Results of the experiment (**Fig. 4**) indicate that no factor has a substantial impact on the classifier ranking. On the contrary, a high degree of similarity can be observed between the rankings resulting from different factor levels. That is, a classifier which performs well in, e.g., high dimensional settings (no. of attributes = high) is likely to also achieve good results when only a small number of attributes is available. The LDA classifier may be seen as an exception since its performance shows some variation between the two cases. However, considering all factors, such variations between one method's ranks across different factor levels are uncommon and occur most likely within the group of traditional methods and LDA in particular. Concerning the group of novel techniques, the RF classifier shows some irregularity, whereas the performances of the other techniques display little if any variation.

Furthermore, the correlation among classifier rankings across different factor levels can once more be appraised in terms of Kendall's Tau. In particular, correlations are strongly positive (>0.61) and statistically significant on the 1%-level without exception. Therefore, it may be concluded that the problem specific characteristics of **Table 5** (and their categorization) have little effect on the ranking of competing classifiers.

4.2.3 The Impact of Dataset Selection

The benchmarking study embraces a comparably large number of datasets to secure the representativeness of empirical results. Nonetheless, it is important to examine the stability of results with respect to dataset selection, i.e., assess the likelihood of observing a similar ranking of classifiers when working with other data. In order to shed some light upon this question, a bootstrapping experiment is conducted. A random sample of nine datasets is drawn from the given decision problems with replacement. Consequently, some datasets will appear mul-

[9] The categorization is based upon the AUC since this criterion is independent of dataset characteristics like, e.g., the skewness of class distributions or the asymmetry of error costs (Fawcett 2006, p. 873). Thus, it appears well suited for assessing the general complexity of a classification task.

Table 4 Correlation between classifier rankings across different accuracy indicators

	Costs/profits	CA	A-Mean	G-Mean	F-measure	$Lift_{10\%}$	$Lift_{30\%}$	RMSE	MXE	AUC
Costs/profits	1.00									
CA	0.78	1.00								
A-Mean	0.78	0.92	1.00							
G-Mean	0.78	0.94	0.96	1.00						
F-measure	0.78	0.92	0.99	0.95	1.00					
$Lift_{10\%}$	0.63	0.81	0.80	0.83	0.79	1.00				
$Lift_{30\%}$	0.66	0.62	0.62	0.62	0.62	0.64	1.00			
RMSE	0.48	0.51	*0.45*	*0.47*	*0.46*	**0.35**	**0.29**	1.00		
MXE	0.66	0.77	0.73	0.72	0.74	0.66	0.73	0.49	1.00	
AUC	0.62	0.66	0.62	0.64	0.61	0.72	0.77	**0.37**	0.71	1.00

All correlations not explicitly highlighted are significant at the 1%-level. Italic face indicates that a correlation is significant at the 5%-level, whereas bold face highlights insignificant correlations

Table 5 Factors and factor level of dataset specific characteristics

	Dataset size		No. of attributes		Class imbalance		Task complexity	
	Value	Group	Value	Group	Value[a]	Group	Value	Group
AC	690	Small	14	Small	44.49	Small	0.82	Small
GC	1000	Small	24	Small	30.00	Small	0.76	Small
Adult	48,842	Large	14	Small	23.93	Medium	0.89	Small
Coil	9822	Small	86	Large	5.97	Large	0.69	Large
DMC 2000	38,890	Large	96	Large	5.87	Large	0.78	Small
DMC 2001	28,128	Medium	106	Large	50.00	Small	0.62	Large
DMC 2002	20,000	Medium	101	Large	10.00	Large	0.60	Large
DMC 2004	40,292	Large	107	Large	20.43	Medium	0.75	Small
DMC 2005	50,000	Large	119	Large	5.80	Large	0.65	Large

[a] A priori probability of class 1

tiple times in the sample, whereas others are neglected. Subsequently, a classifier ranking is produced for the respective sample using the MNP as indicator of predictive accuracy. Since the resulting ranking is based upon a random sample of datasets, deviation from the previous results (**Fig. 3**) may occur. Specifically, large deviations would indicate that classifier performance varies considerably with the particular selection of benchmarking datasets.

The procedure is repeated 1,000 times, each time with a different random (bootstrap) sample of datasets. Thus, 1,000 ranks are obtained per classifier and the average of these ranks is depicted in **Fig. 5**. The horizontal lines represent an interval of one standard deviation around the mean.

On average, the bootstrapping experiment gives the same ranking as observed on the original nine datasets (see **Fig. 3**). Although the standard deviations indicate that the selection of datasets affects the classifier ranking moderately, it does not influence the trend towards novel methods. This follows mainly from the large gap between the RBFSVM classifier and C4.5, which represents a border between the novel and most traditional classifiers. In view of the standard deviations observed within the experiment, it appears unlikely that this gap will be surmounted if other data is employed. In this respect, the analysis provides little evidence for dependence between classifier ranking and dataset selection. In other words, it is likely that a similar precedence of alternative classification models can be observed on other data and in other applications, respectively. Consequently, the previous recommendation to intensify the use of novel classifiers in corporate applications can be maintained.

5 Summary and Discussion

The paper is concerned with the design and the results of an empirical benchmarking experiment of established versus novel classification models in customer-centric decision problems. In particular, we have explored whether recently proposed techniques offer notable improvements over more traditional counterparts within the considered application context. The observed results allow the conclusion that this is the case for the datasets and methods employed in the study. Specifically, the predictions of modern methods proved to be much more accurate on average. Moreover, additional experiments have confirmed the robustness of this finding with respect to accuracy indicators, problem characteristics and dataset selection.

It is arguable which recommendations should be derived from the observed re-

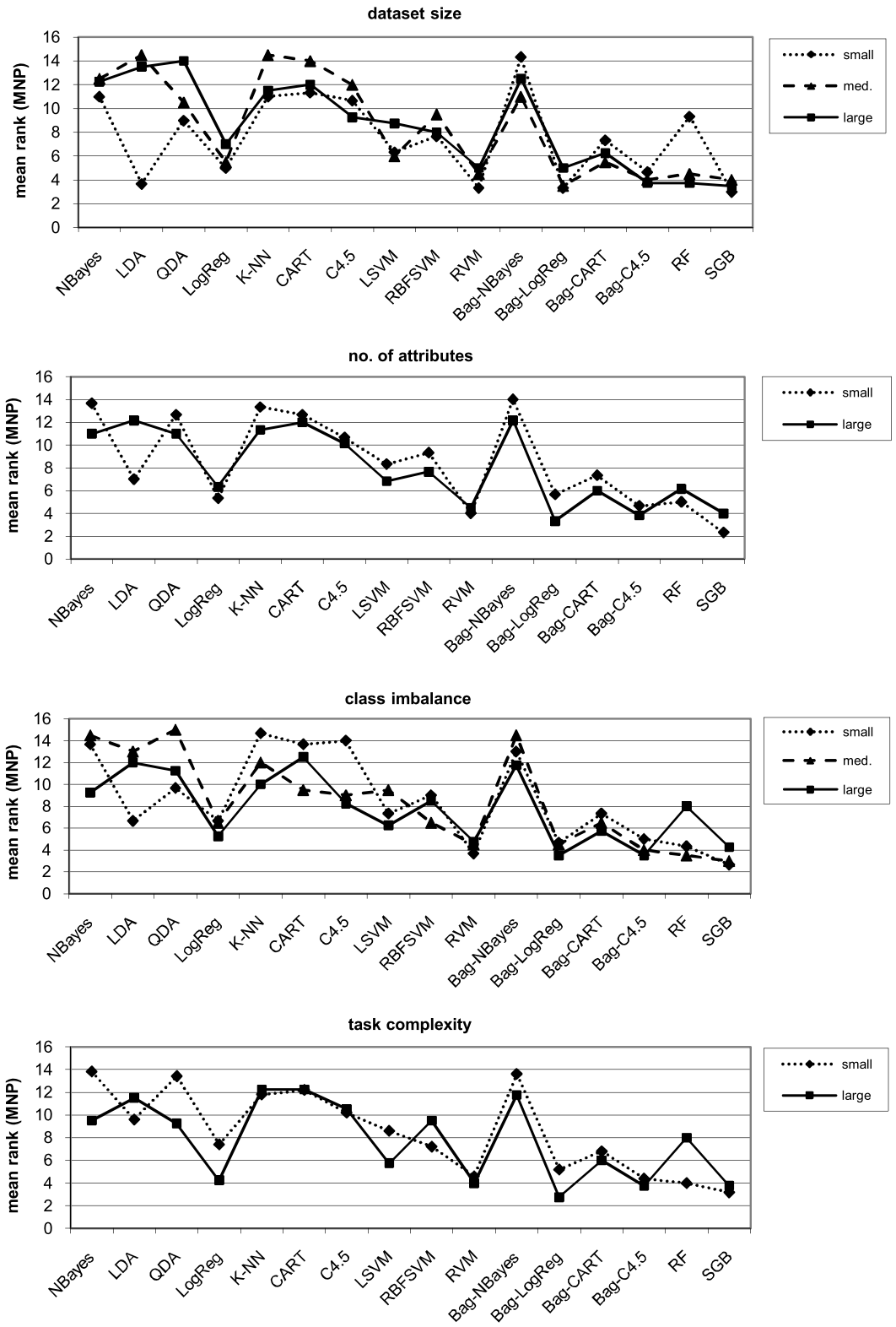

Fig. 4 Impact of problem specific characteristics on classifier ranking in terms of MNP

sults. It is widely established that even marginal improvements in predictive accuracy can have a tremendous impact on profits/costs in customer-centric data mining (Baesens et al. 2002, p. 193; Neslin et al. 2006, p. 205; van den Poel and Lariviere 2004, p. 197–198). For example, Reichheld and Sasser (1990, p. 107) estimate that a 5% reduction of customer churn rates facilitates an 85% increase in revenue within the financial service sector. Lemmens and Croux (2006, p. 281)

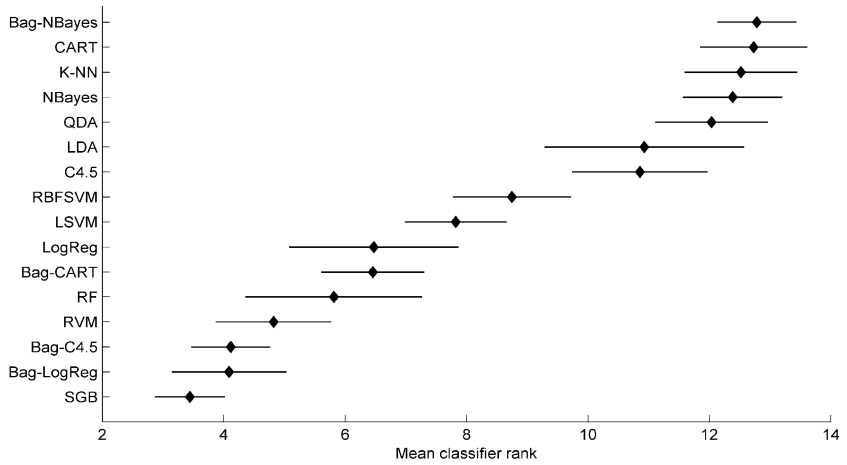

Fig. 5 Mean rank and standard deviation in terms of MNP per classifier across 1000 bootstrap samples of nine datasets

establish a similarly strong effect within the mobile telecommunications industry. In view of this and the comparative differences between alternative classifiers, a careful evaluation of multiple candidates should generally precede the application of a particular model. That is, decision processes that routinely employ a particular method without considering alternatives should be revised. Concerning novel techniques, it is natural that these are not employed immediately in practice. In this sense, the study evidences that possible concerns regarding the maturity of novel methods are unjustified and that techniques like bagging, boosting or random forests are well suited for corporate applications. Consequently, they should be considered within a classifier selection stage. This is especially true if an available data mining system supports such methods and they have up to now remained unused because of, e.g., lack of experience.[10]

If respective software is still unavailable, the profitability of an investment into a system upgrade or a novel package may be appraised by means of classical capital budgeting techniques. The study supports such an endeavor in two ways: First, a detailed description how to assess the potential of novel classifiers (i.e., the profit increases/cost reductions derived from their utilization) has been provided. Secondly, it may be feasible to re-use the above results to quantify monetary advantages in a more direct manner. Specifically, if a company faces a decision problem similar to, e.g., *DMC 2001* (i.e., similar in terms of objective and data characteristics), currently approaches the task by means of a decision tree classifier, and considers replacing this technique with, e.g., RF, then it may be acceptable to approximate the advantage of the latter with 4%, i.e., the result observed here (see **Table 3**). In other words, the anticipated utility of the new method could be estimated without any additional experimentation on the basis of the results presented in this study. In view of the fact that all considered decision problems stem from the field of customer-centric classification and span a variety of applications within this domain, conclusions by analogy may be feasible to some degree. Then, it would suffice to determine the costs associated with licensing the novel software and training users to make a qualified and informed investment decision.

Whereas performance comparisons between the two groups of novel and traditional classifiers have provided strong evidence for the superiority of the former, it is debatable whether recommendations for particular techniques are warranted. The empirical results are less clear in this respect (e.g., Fig. 2) and it is obvious that there is no dominant approach which excels in all possible (customer-centric) classification problems. However, it seems justified to highlight the performance of the SGB classifier. This method has delivered consistently good results within all experiments. In particular, SGB achieves highest predictive accuracy within the monetary comparison (**Fig. 2**), the benchmark in terms of MNP (**Fig. 3**) and the bootstrapping experiment (**Fig. 5**). Moreover, **Fig. 4** evidences its robustness towards dataset characteristics. On the basis of these results, it seems likely that SGB will perform well in other customer-centric settings. Hence, data mining practitioners may want to pay particular attention to this approach.

6 Limitations and Future Research

A critical discussion of observed results (e.g., with respect to their generalizability) and an experiment's limitations is a pivotal part of empirical research. The results of Sect. 4.2 indicate that several risks that may impede a benchmarking study's external validity in general can be rejected in the present case. However, a vast number of customer-centric decision problems exist and it is likely that their quantity will continue to grow. Several drivers like company size, customer homogeneity/heterogeneity etc. may affect the structure of corresponding decision problems, so that it is questionable whether the trend in classification models' comparative performance as observed in this study will persist in all possible environments. Therefore, decision makers are well advised to carefully examine the similarity between their applications and the tasks considered here, before courses of actions are derived from the study's results.

The study assesses classifiers solely in terms of their predictive accuracy and especially according to profits/costs derived from their deployment. Although these measures are of key importance in decision making, this scope leaves out other factors that influence a method's suitability alongside forecasting performance and may thus be too narrow. For example, computational complexity may be relevant and traditional methods generally possess an advantage in this respect. However, considering the focal application domain, the speed of model building and application seems less important. Marketing campaigns are routinely planned with some lead time, leaving sufficient time to train computer intensive

[10] For example, support vector machines as well as some ensemble classifiers and SGB in particular are supported in current releases of SAS Enterprise Miner and SPSS PASW Modeler (formally Clementine). Furthermore, there are several specialized packages like Salford Systems, KXEN or DTREG, which also support respective techniques.

classifiers. Moreover, the computationally more expensive techniques like support vector machines or ensembles offer multiple opportunities for organizing model building in a parallel fashion.

Contrary to resource intensity, a classifier's usability and comprehensibility must be seen as key requirements in corporate data mining. In fact, the objective to detect *understandable* patterns in data is stressed in virtually any definition of data mining. Clearly, novel and more complex methods suffer some limitations in that aspect, which may be seen as the main reason for their conservative use in corporate practice. With regard to usability, it has been shown that a fully-automated model building and calibration procedure delivers promising results. Hence, respective reservations appear unfounded. The interpretability of a model's predictions, on the other hand, is often not given if employing one of the modern classifiers. However, post-processing procedures are available to overcome this obstacle and clarify a model's internal mechanisms and predictive behavior, respectively (Barakat and Bradley 2007, p. 733 ff; Breiman 2001, p. 23 ff; Friedman 2001, p. 1216 ff; Martens et al. 2009, p. 180 ff; 2007, p. 1468 ff). In particular, these techniques enable drivers for customer behavior to be discerned, i.e., explain why a model classifies a customer as churner. Such insight may suffice to satisfy constraints associated with model comprehensibility in many applications. However, the area would benefit from future research to, e.g., develop a formal taxonomy for assessing classifiers' interpretability. This would nicely complement the accuracy-based evaluation presented in this paper, whereby appraising the (monetary) value of higher comprehensibility, or, similarly, lower computational complexity, will represent a major challenge.

In addition to classification, a data mining process embraces several preceding and succeeding tasks. Especially data pre-processing activities may substantially affect the predictive performance of a classification model (Crone et al. 2006, p. 792 ff). This aspect has not been examined in the present study. Consequently, it would be interesting to explore the influence of alternative pre-processing techniques on classifier's accuracy. For example, pre-processing could refer to, e.g., missing value imputation, attribute transformation and/or feature selection. The results of respective experiments, which could also concentrate on steps succeeding classification within a data mining process, would amend the findings of this study and may facilitate an assessment and comparison of the relative importance of different data analysis tasks and, thereby, help to design resource efficient data mining processes. This would indeed be a significant contribution to the field.

In general, one may argue that economical constraints have not yet received sufficient attention within the data mining community. Thus, increasing the awareness of business requirements of real-world decision contexts is a worthwhile undertaking. For example, constraints associated with resource availability could be taken into account when building a classification model to further enhance its value. Considerations along this line have been put forward within the young field of *utility-based data mining* (Weiss et al. 2008) and the great potential of research at the interface between data mining and corporate decision making has been exemplified in recent work of Boylu et al. (2009) as well as Saar-Tsechansky and Provost (2007). The underlying idea of an economically driven data analysis has also been adopted in this paper, i.e., by assessing classification models in terms of their monetary consequences in real-world decision contexts. Whereas clarifying the economic potential of novel classifiers has been the study's major objective, it may also help to increase awareness of the potential and challenges of utility-based data mining within Information Systems and, thereby, motivate future research within this discipline.

References

Asuncion A, Newman DJ (2007) UCI machine learning repository. http://archive.ics.uci.edu/ml/. Accessed 2009-09-02

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

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

Barakat NH, Bradley AP (2007) Rule extraction from support vector machines: a sequential covering approach. IEEE Transactions on Knowledge and Data Engineering 19(6):729–741

Abstract

Stefan Lessmann, Stefan Voß

Customer-Centric Decision Support

A Benchmarking Study of Novel Versus Established Classification Models

Classification analysis is an important tool to support decision making in customer-centric applications like, e.g., proactively identifying churners or selecting responsive customers for direct-marketing campaigns. Whereas the development of novel classification algorithms is a popular avenue for research, corresponding advancements are rarely adopted in corporate practice. This lack of diffusion may be explained by a high degree of uncertainty regarding the superiority of novel classifiers over well established counterparts in customer-centric settings. To overcome this obstacle, an empirical study is undertaken to assess the ability of several novel as well as traditional classifiers to form accurate predictions and effectively support decision making. The results provide strong evidence for the appropriateness of novel methods and indicate that they offer economic benefits under a variety of conditions. Therefore, an increase in use of respective procedures can be recommended.

Keywords: Data mining, Customer relationship management, Decision support, Classification models

Batista G, Prati RC, Monard MC (2004) A study of the behavior of several methods for balancing machine learning training data. ACM SIGKDD Explorations Newsletter 6(1):20–29

Bauer E, Kohavi R (1999) An empirical comparison of voting classification algorithms: bagging, boosting, and variants. Machine Learning 36(1–2):105–139

Berry MJA, Linoff G (2004) Data mining techniques: for marketing, sales and customer relationship management, 2nd edn. Wiley, New York

Bose I, Xi C (2009) Quantitative models for direct marketing: a review from systems perspective. European Journal of Operational Research 195(1):1–16

Boylu F, Aytug H, Koehler GJ (2009) Principal-agent learning. Decision Support Systems 47(2):75–81

Breiman L (2001) Random forests. Machine Learning 45(1):5–32

Burez J, van den Poel D (2009) Handling class imbalance in customer churn prediction. Expert Systems with Applications 36(3):4626–4636

Caruana R, Niculescu-Mizil A (2004) Data mining in metric space: an empirical analysis of supervised learning performance criteria. In: Kim W, Kohavi, R, Gehrke, J, DuMouchel W (eds) Proc. 10th ACM SIGKDD intern. conf. on knowledge discovery and data mining, Seattle

Caruana R, Niculescu-Mizil A (2006) An empirical comparison of supervised learning algorithms. In: Cohen, WW, Moore A (eds) Proc. 23rd intern. conf. on machine learning, Pittsburgh

Cooper GF, Aliferis CF, Ambrosino R, Aronis J, Buchanan BG, Caruana R, Fine MJ, Glymour C, Gordon G, Hanusa BH, Janosky JE, Meek C, Mitchell T, Richardson T, Spirtes P (1997) An evaluation of machine-learning methods for predicting pneumonia mortality. Artificial Intelligence in Medicine 9(2):107–138

Crone SF, Lessmann S, Stahlbock R (2006) The impact of preprocessing on data mining: an evaluation of classifier sensitivity in direct marketing. European Journal of Operational Research 173(3):781–800

Crook JN, Edelman DB, Thomas LC (2007) Recent developments in consumer credit risk assessment. European Journal of Operational Research 183(3):1447–1465

Cui D, Curry D (2005) Predictions in marketing using the support vector machine. Marketing Science 24(4):595–615

Curram SP, Mingers J (1994) Neural networks, decision tree induction and discriminant analysis: an empirical comparison. Journal of the Operational Research Society 45(4):440–450

Demšar J (2006) Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning Research 7:1–30

Dietterich TG (2000) An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization. Machine Learning 40(2):139–157

Fawcett T (2006) An introduction to ROC analysis. Pattern Recognition Letters 27(8):861–874

Fenton N, Neil M (1999) A critique of software defect prediction models. IEEE Transactions on Software Engineering 25(5):675–689

Ferri C, Hernández-Orallo J, Modroiu R (2009) An experimental comparison of performance measures for classification. Pattern Recognition Letters 30(1):27–38

Friedman JH (2001) Greedy function approximation: a gradient boosting machine. The Annals of Statistics 29(5):1189–1232

Friedman JH (2002) Stochastic gradient boosting. Computational Statistics & Data. Analysis 38(4):367–378

Friedman JH (2006) Recent advances in predictive (machine) learning. Journal of Classification 23(2):175–197

Gestel T, van, Suykens JAK, Baesens B, Viaene S, Vanthienen J, Dedene G, Moor B, De, Vandewalle J (2004) Benchmarking least squares support vector machine classifiers. Machine Learning 54(1):5–32

Haas A (2006) Interessentenmanagement. In: Hippner H, Wilde KD (eds) Grundlagen des CRM. Gabler, Wiesbaden, pp 443–472

Hamza M, Larocque D (2005) An empirical comparison of ensemble methods based on classification trees. Journal of Statistical Computation and Simulation 75(8):629–643

Hastie T, Tibshirani R, Friedman JH (2009) The elements of statistical learning, 2nd edn. Springer, New York

Hippner H (2006) Komponenten und Potenziale eines analytischen Customer Relationship Management. In: Chamoni P, Gluchowski P (eds) Analytische Informationssysteme, 3rd edn. Springer, Heidelberg, pp 361–384

Hothorn T, Lausen B (2005) Bundling classifiers by bagging trees. Computational Statistics & Data. Analysis 49(4):1068–1078

Hsu C-W, Chang C-C, Lin C-J (2003) A practical guide to support vector classification. Arbeitspapier, Department of Computer Science and Information Engineering, National Taiwan University

Hulse JV, Khoshgoftaar TM, Napolitano A (2007) Experimental perspectives on learning from imbalanced data. In: Ghahramani Z (ed) Proc. 24th intern. conf. on machine learning, Corvallis

Izenman AJ (2008) Modern multivariate statistical techniques. Springer, Heidelberg

Khoshgoftaar TM, Seliya N (2004) Comparative assessment of software quality classification techniques: an empirical case study. Empirical Software Engineering 9(3):229–257

King RD, Feng C, Sutherland A (1995) StatLog: comparison of classification algorithms on large real-world problems. Applied Artificial Intelligence 9(3):259–287

Lemmens A, Croux C (2006) Bagging and boosting classification trees to predict churn. Journal of Marketing Research 43(2):276–286

Lessmann S, Voß S (2008) Supervised classification for decision support in customer relationship management. In: Bortfeldt A, Homberger J, Kopfer H, Pankratz G, Strangmeier R (eds) Intelligent Decision Support. Gabler, Wiesbaden, pp 231–253

Lessmann S, Baesens B, Mues C, Pietsch S (2008) Benchmarking classification models for software defect prediction: a proposed framework and novel findings. IEEE Transactions on Software Engineering 34(4):485–496

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

Liu C-L, Nakashima K, Sako H, Fujisawa H (2003) Handwritten digit recognition: benchmarking of state-of-the-art techniques. Pattern Recognition 36(10):2271–2285

Martens D, Baesens B, van Gestel T, Vanthienen J (2007) Comprehensible credit scoring models using rule extraction from support vector machines. European Journal of Operational Research 183(3):1466–1476

Martens D, Baesens B, van Gestel T (2009) Decompositional rule extraction from support vector machines by active learning. IEEE Transactions on Knowledge and Data Engineering 21(2):178–191

Meyer D, Leisch F, Hornik K (2003) The support vector machine under test. Neurocomputing 55(1–2):169–186

Neslin SA, Gupta S, Kamakura W, Lu J, Mason CH (2006) Defection detection: measuring and understanding the predictive accuracy of customer churn models. Journal of Marketing Research 43(2):204–211

Ngai EWT, Xiu L, Chau DCK (2009) Application of data mining techniques in customer relationship management: a literature review and classification. Expert Systems with Applications 36(2):2592–2602

Ohlsson MC, Runeson P (2002) Experience from replicating empirical studies on prediction models. In: Proc. 8th intern. software metrics symposium, Ottawa

Perlich C, Provost F, Simonoff JS, Cohen WW (2003) Tree induction vs. logistic regression: a learning-curve analysis. Journal of Machine Learning Research 4(2):211–255

Poel D, van den, Lariviere B (2004) Customer attrition analysis for financial services using proportional hazard models. European Journal of Operational Research 157(1):196–217

Putten P, Someren M (2000) CoIL Challenge 2000: the insurance company case. Working paper, Sentient Machine Research, Amsterdam

Reichheld FF, Sasser WE (1990) Zero defections: quality comes to service. Harvard Business Review 68(5):105–111

Saar-Tsechansky M, Provost F (2007) Decision-centric active learning of binary-outcome models. Information Systems Research 18(1):4–22

Sohn SY, Shin HW (2007) Experimental study for the comparison of classifier combination methods. Pattern Recognition 40(1):33–40

Ting KM (2002) An instance-weighting method to induce cost-sensitive trees. IEEE Transactions on Knowledge and Data Engineering 14(3):659–665

Tipping ME (2000) The relevance vector machine. In: Solla SA, Leen TK, Müller K-R (eds) Advances in neural information processing systems, vol 12. MIT Press, Cambridge, pp 652–658

Vapnik VN (1995) The nature of statistical learning theory. Springer, New York

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

Wang S-j, Mathew A, Chen Y, Xi L-f, Ma L, Lee J (2009) Empirical analysis of support vector machine ensemble classifiers. Expert Systems with Applications 36(2):6466–6476

Weiss SM, Kapouleas I (1989) An empirical comparison of pattern recognition, neural nets and machine learning classification methods. In: Sridharan NS (eds) Proc. 11th intern. joint conf. on artificial intelligence, Detroit

Weiss GM, Provost F (2003) Learning when training data are costly: the effect of class distribution on tree induction. Journal of Artificial Intelligence Research 19:315–354

Weiss GM, Zadrozny B, Saar-Tsechansky M (2008) Guest editorial: special issue on utility-based data mining. Data Mining and Knowledge Discovery 17(2):129–135

Zickus M, Greig AJ, Niranjan M (2002) Comparison of four machine learning methods for predicting PM10 concentrations in Helsinki. Finland Water, Air, & Soil Pollution Focus 2(5):717–729

[Anhang] [Online only]

Appendix I: Model Selection

Some classification models exhibit additional parameters, which allow adapting the model to a particular task and need to be tuned by users prior to applying the classifier. For example, the number of neighboring objects considered within distance calculations has to be determined for the k-NN classifier. Such parameters are tuned in a fully-automatic fashion. In particular, grounding on recommendations from the literature, a set of candidate values has been defined for each parameter. Subsequently, all possible combinations of parameter settings have been evaluated empirically by means of five-fold cross-validation (Izenman 2008, p. 121) on training data. Specifically, this evaluation has been conducted per classifier and dataset. The parameter combination achieving maximal forecasting accuracy has been retained and a respective classifier has been built on the full training dataset to predict the test data. Considering a classifier with, e.g., two parameters and three alternative settings each, this procedure requires 3*3 parameter combinations * 5-fold cross-validation*10 randomly sampled train/test sets + 10 final models = 460 classification models to be constructed and assessed per dataset. This approach is advantageous because every classifier is carefully adapted to each individual decision task. Consequently, a competitive and representative estimate of its performance is obtained. The list of candidate parameter settings is shown in Tab. 6.

[Tab06]

Tab. 6 Parameter settings considered during model selection

Classifier	No. of parameters	Parameter	Candidate settings
NBayes	0		
LDA	In general, these classifiers do not require any model selection. However, due to numerical difficulties, optimization problems underlying these techniques cannot be solved for high dimensional settings with many correlated attributes. To alleviate such difficulties, a backward feature elimination procedure is employed.		
QDA			
LogReg			
K-NN	1	No. of nearest neighbors	[1;3;5;7;9]
C4.5	1	Confidence of pruning strategy	[0,1; 0,2; 0,25; 0,3]
CART	0		
LSVM	1	Regularization constant	$2^{[-6, -5, \ldots, 16]*}$
RBFSVM	>=3	Regularization constant	A radial basis function kernel with one additional parameter is employed. Therefore, an overall number of two parameters is considered during model selection. To simultaneously tune these two, a heuristic pattern search method is implemented. On average, the algorithm requires 21 steps to detect a local optimum within the parameter space.
		Kernel function for nonlinear data transformation	
		Individual parameters of the Kernel function.	
RVM	0		
Bagging	1	No. of base classifiers within the ensemble	[5; 25; 50]
RF	2	No. of CART decision trees	[50; 100; 250; 500]
		No. of attributes drawn at random each time a node is split in an individual tree	[0,5; 1; 2]*M, with M being the square root of the number of attributes
SGB	1	No. of boosting iterations	[5; 10; 25]

* It is common practice to consider an exponential scale to search over a large range of candidate values (Hsu et al. 2003, p. 5).

Appendix II: Monetary Assessment of Classification Models

In order to compute the monetary consequences of employing a given classification model in a given decision context, information concerning the profits/costs associated with correct/false class predictions is required. Such information is provided within the dataset descriptions for most of the tasks considered here, with AC, Adult and Coil being an exception. Lacking any information on costs/profits for these datasets, it is assumed that all correct classifications are associated with zero costs, whereas incorrect predictions in one class are defined as being equivalent with the a priori probability of the alternative class. This assumption is common within the literature and ensures that errors within the minority class are punished more severely. The resulting profits/costs are depicted in Tab. 7.

[Tab07]

Tab. 7 Profits and costs of correct and incorrect class predictions

		Predicted class	
	True class	-1	+1
AC	-1	0	0,4449
	+1	0,5551	0
GC	-1	0	1
	+1	5	0
Adult	-1	0	0,2393
	+1	0,7607	0
Coil	-1	0	0,06
	+1	0,94	0
DMC 2000	-1	0	-6
	+1	-95	95
DMC 2001	-1	1.110	662
	+1	-265	-25
DMC 2002	-1	72	66,3
	+1	0	43,8
DMC 2004	-1	1	-1
	+1	-1	1
DMC 2005	-1	15	13
	+1	-25	2

In addition, a crisp classification of objects into classes is required to calculate an overall (monetary) indicator of predictive accuracy. However, most classifiers produce a continuous estimate of an object's likelihood of belonging to a particular class (Fawcett 2006, p. 863). Consequently, a post-processing is needed to obtain discrete class predictions. Specifically, objects are assigned to class 1 if their estimated probability of belonging to this class exceeds a certain threshold, whereas all other objects are assigned to the alternative class. Therefore, a profit/cost-based classifier assessment requires an approach to determine dataset-dependent thresholds. In fact, this task is particularly simple in practical applications because the number of objects to be classified as members of class 1 (e.g., the number of customers to be solicited in a mailing campaign) is usually given exogenously. For example, decision makers may have a pre-defined budget for a marketing campaign that allows contacting N customers. Then, the objective of classification analysis is to estimate (for all customers) whether they are likely to show an intended response if solicited. Subsequently, the N customers with highest response probability form the campaigns' target group. In other words, the sought threshold value is the prediction of the customer/object with N-highest estimate.

Since information on budget constraints is unavailable for the datasets employed in this study, the procedure outlined above is simulated. In particular, it is assumed that the number of objects to be classified as class 1 is equivalent to the a priori probability of class 1 in the training data. Using this probability, it is straightforward to calculate the number of test set instances that are to be assigned to the positive class. Once all objects have been classified, correct/false classifications are weighted with the values of Tab. 7 to produce a model's final performance indicator.