Technical Report

14/2009

D-optimal plans for variable selection in data bases

Julia Schiffner, Claus Weihs

Technische
Universität
Dortmund

SFB 475
Komplexitätsreduktion in
multivariaten Datenstrukturen

Sonderforschungsbereich 475
Universität Dortmund
44221 Dortmund
Germany

D-optimal Plans for Variable Selection in Data Bases

Julia Schiffner and Claus Weihs

Lehrstuhl Computergestützte Statistik, Fakultät Statistik, Technische Universität Dortmund,
D-44221 Dortmund, Germany

Abstract. This paper is based on an article of Pumplün et al. (2005a) that investigates the use of Design of Experiments in data bases in order to select variables that are relevant for classification in situations where a sufficient number of measurements of the explanatory variables is available, but measuring the class label is hard, e. g. expensive or time-consuming.

Pumplün et al. searched for D-optimal designs in existing data sets by means of a genetic algorithm and assessed variable importance based on the found plans. If the design matrix is standardized these D-optimal plans are almost orthogonal and the explanatory variables are nearly uncorrelated. Thus Pumplün et al. expected that their importance for discrimination can be judged independently of each other. In a simulation study Pumplün et al. applied this approach in combination with five classification methods to eight data sets and the obtained error rates were compared with those resulting from variable selection on the basis of the complete data sets. Based on the D-optimal plans in some cases considerably lower error rates were achieved. Although Pumplün et al. (2005a) obtained some promising results, it was not clear for different reasons if D-optimality actually is beneficial for variable selection. For example, D-efficiency and orthogonality of the resulting plans were not investigated and a comparison with variable selection based on random samples of observations of the same size as the D-optimal plans was missing. In this paper we extend the simulation study of Pumplün et al. (2005a) in order to verify their results and as basis for further research in this field.

Moreover, in Pumplün et al. D-optimal plans are only used for data preprocessing, that is variable selection. The classification models are estimated on the whole data set in order to assess the effects of D-optimality on variable selection separately. Since the number of measurements of the class label in fact is limited one would normally employ the same observations that were used for variable selection for learning, too. For this reason in our simulation study the appropriateness of D-optimal plans for training classification methods is additionally investigated.

It turned out that in general in terms of the error rate there is no difference between variable selection on the basis of D-optimal plans and variable selection on random samples. However, for training of linear classification methods D-optimal plans seem to be beneficial.

1 Introduction

In classification the aim is to predict the class membership of an object based on measurements on other more readily available variables. In order to be able to learn a classification rule the class label has to be known for a set of training observations. In some cases measuring the class of an object may be so hard, e.g. expensive, destructive, or time-consuming, that it poses a problem to obtain enough measurements of the class label to learn a classification rule.

Since the number of possible measurements k of the response is limited and much smaller than n, the number of observations of the explanatory variables, one has to decide for which observations of the explanatory objects the corresponding class label should be measured. It is clear that the selected observations should be as informative as possible for classification. But two questions automatically arise, namely *What does 'informative' mean?* and *How can the k most informative observations be found?*.

The meaning of the term 'informative' depends on the purpose. In classification normally one would select observations that carry much information about discrimination and thus result in a low error rate. Here like in the article of Pumplün et al. (2005a), we also want to identify observations that are informative for variable selection. The problem in variable selection is that the predictor variables normally are not independent of each other. It would be desirable if importance could be assessed for each variable independently of the others and it would be helpful if the predictors were at least uncorrelated.

In order to to find informative observations for classification and especially variable selection Design of Experiments is applied. Design of Experiments allows for obtaining as much information as possible about a dependent variable by means of comparatively few measurements. Here, we focus on *D-optimal* plans because of their relationship to orthogonal designs and uncorrelated variables. In Section 2 we review some facts about Design of Experiments, especially D-optimal and orthogonal designs and their relationship.

Pumplün et al. (2005a) searched for D-optimal plans in data sets by means of a genetic algorithm and used them as basis for variable selection. Here, we also employ D-optimal plans for training of classification methods. In Section 3 the approach is described in detail.

In Section 4 we describe a simulation study that we carried out in order to investigate the appropriateness of D-optimal plans as basis for variable selection and training of classification methods. The results are given in Section 5. Finally, in Section 6 a summary is given.

2 Design of Experiments: Orthogonality and D-Optimality

In this Section in order to introduce the notation and to explain the approach of Pumplün et al. (2005a) we review some well-known facts about orthogonal and D-optimal designs and their relationship. This Section is based on the textbooks of Weihs and Jessenberger (1999) and Hinkelmann and Kempthorne (2005, 2008).

Design of Experiments (DoE) aims at obtaining as much information as possible about the relationship between influential factors and a dependent variable by means of as few experiments as possible. In general an experiment can be a production process in industry, a medical examination, or a computer simulation. In case of k experiments and p influential factors the factor values where the measurements $\boldsymbol{y} = (y_1, \ldots, y_k)'$ of the dependent variable are taken are coded in the *(main effect) design matrix* $\boldsymbol{X} \in \mathbb{R}^{k \times p}$.

In order to assess which influential factors have an effect on the dependent variable usually a *screening plan* is used. Screening plans permit to investigate the impact of many factors at the same time in order to assess their importance by means of relatively few experiments. Let $\boldsymbol{y} = \boldsymbol{Z}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ a linear screening model. $\boldsymbol{Z} := (\boldsymbol{1}_k, \boldsymbol{X})$ denotes the *(extended) design matrix* where $\boldsymbol{1}_k := (1, \ldots, 1)'$, $\boldsymbol{\beta} := (\beta_0, \ldots, \beta_p)'$ the parameter vector, and $\boldsymbol{\epsilon} := (\epsilon_1, \ldots, \epsilon_k)'$ random noise. We assume that $\mathrm{E}(\boldsymbol{\epsilon}) = 0$ and $\mathrm{Cov}(\boldsymbol{\epsilon}) = \sigma^2 \boldsymbol{I}_k$ where $\sigma^2 > 0$. Since it is desirable to assess the importance of a single factor independently of the others usually *orthogonal designs* are used.

Definition 1 (Orthogonal Design). A design with design matrix $\boldsymbol{Z}$ is called *orthogonal* iff $\boldsymbol{Z}'\boldsymbol{Z}$ is a diagonal matrix, that is $\boldsymbol{Z}'_{\cdot i}\boldsymbol{Z}_{\cdot j} = 0$ for all $i \neq j$ with $i, j \in \{1, \ldots, p+1\}$.

The least-squares estimate of $\boldsymbol{\beta}$ in the screening model is given as $\hat{\boldsymbol{\beta}} = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{y}$ and its covariance matrix is $\mathrm{Cov}(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{Z}'\boldsymbol{Z})^{-1}$. In case of an orthogonal design the estimated factor effect $\hat{\beta}_j = {}^1\!/\!_{\|\boldsymbol{X}_{\cdot j}\|_2^2} \cdot \boldsymbol{X}'_{\cdot j}\boldsymbol{y}$, $j = 1, \ldots, p$, depends only on the corresponding influential factor $\boldsymbol{X}_{\cdot j}$. Additionally, the estimates $\hat{\beta}_j$ are pairwise uncorrelated. Thus the importance of factors can be assessed independently of each other.

The orthogonal influential factors $\boldsymbol{X}_{\cdot j}$ themselves are not necessarily (empirically) uncorrelated because their means need not to be zero. For centered data, $\boldsymbol{X}^{\star}_{\cdot j} := \boldsymbol{X}_{\cdot j} - \bar{\boldsymbol{X}}_{\cdot j}$, $j = 1, \ldots, p$, however, it holds that

$$r_{\boldsymbol{X}_{\cdot i}, \boldsymbol{X}_{\cdot j}} = r_{\boldsymbol{X}^{\star}_{\cdot i}, \boldsymbol{X}^{\star}_{\cdot j}} = \frac{\boldsymbol{X}^{\star\prime}_{\cdot i}\boldsymbol{X}^{\star}_{\cdot j}}{\|\boldsymbol{X}^{\star}_{\cdot i}\|_2 \|\boldsymbol{X}^{\star}_{\cdot j}\|_2} = \cos(\alpha_{\boldsymbol{X}^{\star}_{\cdot i}, \boldsymbol{X}^{\star}_{\cdot j}}).$$

Thus in case of a centered orthogonal design where $\cos(\alpha_{\boldsymbol{X}^\star_{\cdot i},\boldsymbol{X}^\star_{\cdot j}}) = {}^{\pi}\!/_{2}$ the predictors $\boldsymbol{X}^\star_{\cdot i}$ and $\boldsymbol{X}^\star_{\cdot j}$ as well as $\boldsymbol{X}_{\cdot i}$ and $\boldsymbol{X}_{\cdot j}$ are pairwise uncorrelated.

Orthogonal plans do not always exist. They can easily be constructed if for each factor two levels are specified and coded as -1 and $+1$. For example (fractional) factorial designs exist if k is a power of 2 or Plackett-Burman designs are available if k is a multiple of 4. If orthogonal designs do not exist or if the influential factors should not be coded often D-optimal plans are used.

For a linear screening model the *information matrix* is given as $\boldsymbol{Z}'\boldsymbol{Z}$.

Definition 2 (D-optimal Design). A design with design matrix $\boldsymbol{Z}$ is called *D-optimal* iff $\boldsymbol{Z}'\boldsymbol{Z}$ is nonsingular and the determinant of the information matrix $\det(\boldsymbol{Z}'\boldsymbol{Z})$ is maximal on the set of all $k \times (p+1)$ design matrices. The value $D(\boldsymbol{Z}) := \det(\boldsymbol{Z}'\boldsymbol{Z})$ is called the *D-value.*

D-optimal designs, in general, need not to be orthogonal. The relationship between orthogonality and D-optimality is given by the following result (see e. g. Box, 1952): Let $\boldsymbol{M}$ be a matrix with full column rank. If the diagonal entries of $\boldsymbol{M}'\boldsymbol{M}$ are fixed $\det(\boldsymbol{M}'\boldsymbol{M})$ is maximal if all off-diagonal entries are zero, that is if the column vectors of $\boldsymbol{M}$ are orthogonal.

If the standardized design matrix $\boldsymbol{Z}^* := (\boldsymbol{1}_k, \boldsymbol{X}^*)$ with

$$\boldsymbol{X}^* := \left(\frac{\boldsymbol{X}_{\cdot 1} - \bar{\boldsymbol{X}}_{\cdot 1}}{\sqrt{{}^1\!/_k \sum_{i=1}^k (\boldsymbol{X}_{i1} - \bar{\boldsymbol{X}}_{\cdot 1})^2}}, \dots, \frac{\boldsymbol{X}_{\cdot p} - \bar{\boldsymbol{X}}_{\cdot p}}{\sqrt{{}^1\!/_k \sum_{i=1}^k (\boldsymbol{X}_{ip} - \bar{\boldsymbol{X}}_{\cdot p})^2}} \right)$$

is considered the diagonal entries of the information matrix $\boldsymbol{Z}^{*\prime}\boldsymbol{Z}^*$ are fixed since $\boldsymbol{Z}^{*\prime}_{\cdot j} \cdot \boldsymbol{Z}^*_{\cdot j} = \|\boldsymbol{Z}^*_{\cdot j}\|_2^2 = k$ for $j = 1, \dots, p+1$ and D-optimality of the standardized design thus leads to orthogonal and uncorrelated column vectors $\boldsymbol{Z}^*_{\cdot j}$. Note that this in general is not the case for the unstandardized design matrix $\boldsymbol{Z}$.

The quality of a found design with design matrix $\boldsymbol{Z}$ can be assessed by means of the *D-efficiency*

$$D_{\text{eff}}(\boldsymbol{Z}) = \frac{D(\boldsymbol{Z})^{1/p}}{D(\boldsymbol{Z}_{\text{opt}})^{1/p}} \tag{1}$$

where $\boldsymbol{Z}_{\text{opt}}$ is the design matrix of a design with maximal possible D-value. If a (standardized) orthogonal design exists it is D-optimal with D-value $D(\boldsymbol{Z}^*_{\text{opt}}) = k^p$. Thus the distance to exact orthogonality can be measured as

$$D_{\text{eff}}(\boldsymbol{Z}^*) = \frac{D(\boldsymbol{Z}^*)^{1/p}}{k}. \tag{2}$$

The primary purpose of D-optimal designs does not consist in the relationship to orthogonality, but is to minimize the uncertainty about the unknown model coefficients $\boldsymbol{\beta}$. More precisely, D-optimal designs minimize the volume of the confidence ellipsoid for $\boldsymbol{\beta}$ since $D(\boldsymbol{Z}^*)^{-0.5}$ is proportional to the volume. In case of a standardized design matrix where D-optimality results in orthogonality the confidence ellipsoid is a sphere. Actually, one is more interested in plans that reduce the uncertainty about the prediction of $\boldsymbol{y}$, but these plans generally are more difficult to construct than D-optimal designs. However, it is possible to prove that D-optimality approximately guarantees a minimal prediction interval for predictions of the dependent variable $\boldsymbol{y}$ (see Kiefer and Wolfowitz, 1960). It is difficult to find the absolute D-optimal plan in the set of all $k \times (p+1)$ design matrices. In order to restrict the number of possible plans usually candidate points are specified and the best plan that can be built from these candidate points is sought.

3 Design of Experiments in Data Bases for Variable Selection and Classification

In this Section the approach of Pumplün et al. (2005a) is described. In order to find observations that are informative for variable selection in classification a D-optimal plan is searched in a data set. The selected observations are used as basis for variable selection. The main effect design matrix $\boldsymbol{X}$ is considered.

In contrast to ordinary experimental design not arbitrary $k \times p$-matrices can be chosen as plans, but the $\binom{n}{k}$ possible designs of size k, where normally $k \ll n$, that can be obtained from the observed data. That is in our case the whole data set can be regarded as a set of candidate points. The D-optimal plan is the one with the highest D-value of the $\binom{n}{k}$ possible plans. For a standardized design matrix $\boldsymbol{X}^*$ with fixed diagonal entries k of the information matrix the maximal D-value is $D(\boldsymbol{X}^*_{\text{opt}}) = k^p$.

3.1 Search for a D-optimal Design in Data Bases

Pumplün et al. (2005a) propose to carry out a heuristic search for D-optimal plans by means of the following genetic algorithm:

1. outer loop: repeat 100 times

```
    (a) randomly choose 10 plans,
    (b) inner loop: repeat 10 times
        i. compute the D-value of each plan,
        ii. locally optimize the best plans by mutation or cross-over,
2. return the best plan / the set of best plans in each iteration.
```

A plan consists of k observations. First, a set of 10 plans is chosen randomly and for each plan the D-value is calculated. In order to optimize the set of designs the plans with lowest D-values are replaced with newly-constructed plans that are similar to the plans with high D-values. Two methods are used for construction of plans, namely mutation and cross-over. Details can be found in Pumplün et al. (2005a).

In a variation of this algorithm not only the overall best plan is returned, but also a set of plans with maximal D-value resulting from every new start of the algorithm, that is the 100 iterations in the outer loop.

The standardized design matrix $\boldsymbol{X}^*$ is used to compute the D-value in order to make sure that D-optimality leads to orthogonality and to uncorrelated explanatory variables (see Section 2). The orthogonality of the designs resulting from this algorithm is investigated in the simulation study described in Subsection 5.2. The corresponding matrix $\boldsymbol{X}$ is used as basis for variable selection and/or training of classification methods.

3.2 Variable Selection

The feature selection problem consists in finding the subset of v most relevant variables for discrimination for a fixed $v < p$. Two different simple variable selection methods were used in Pumplün et al. (2005a).

Correlation. The empirical correlation coefficients $r_{\boldsymbol{X}_{\cdot j},\boldsymbol{y}}$ of the explanatory variables $\boldsymbol{X}_{\cdot j}$, $j = 1, \ldots, p$, and the class variable $\boldsymbol{y}$ are calculated. The v variables with largest absolute values are selected. If the correlations of two or more predictors is equal a random selection is performed. A disadvantage of this approach may be that correlation coefficients measure only linear dependencies and thus important variables may be missed.

Tree. Tree-based feature selection using the gini index is employed. First the classification tree is learned. Each time a variable occurs in the tree it is assigned the weight 2^{-d} where d denotes the depth of the corresponding node. The value 2^{-d} is chosen because a binary tree can have at most 2^d nodes on the d-th level. Variables occurring early in the tree are considered more important than those close to the leaves. The measure of feature importance is the sum of the weights of each variable. If less than v variables are included in the tree, the remaining features are selected randomly. Decision trees are known to be unstable, that is minor changes in the data may lead to completely different decision trees and thus to distinct values of variable importance.

The quality of the variable selection methods is investigated in the simulation study and the results are given in Section 5.1.

Pumplün et al. (2005a) propose two variable selection schemes. On the one hand variable selection is done based on the plan with the highest D-value (called *vs doptimal*). On the other hand variable selection on the basis of a set of 100 plans with highest D-values resulting from the 100 iterations in the outer loop of the genetic algorithm is considered. For each variable the relative frequency of selection in the individual plans is calculated and taken as a measure for variable importance. This procedure is called *vs doptimal it* and expected to give more stable results.

3.3 Training

In Pumplün et al. (2005a) D-optimal plans are used only for variable selection, but the classification methods are trained on the complete data set. This is useful in order to assess separately if D-optimal plans are beneficial for variable selection. But in our situation where the number of measurements of the class label is limited one would normally use the D-optimal plans also as basis for training, too.

In the literature there are several approaches that use D-optimal designs in data bases. Rüping and Weihs (2009) describe a kernelization of experimental design in data bases and use kernelized D-optimal plans for training of a support vector machine for regression problems. Their approach significantly outperforms competing algorithms.

Choueiki and Mount-Campbell (1999) employ the D-optimality criterion to select data for training a neural network for function approximation in situations when measuring the outcome is expensive, hazardous, or time-consuming. They show that as long as the training data is chosen according to the D-optimality criterion the network is able to generalize well. The performance criterion used is the mean squared error.

Manolov (1990) proposes a sequential design of training observations for classification in order to find points near the true decision boundary. A D-optimal plan is taken as initial experimental design. Using these points a first approximation of the decision boundary is obtained. Based on this approximation new points are generated that, first, have to lie on the approximation of the decision boundary and, second, together with the old points form a D-optimal plan. Then points that are farthest from the decision boundary are rejected and a new approximation of the decision boundary is calculated.

In the linear model framework the performance and estimation criterion is the mean squared error. In classification things are more complicated. The performance is usually assessed by the error rate. For estimation of classification rules there are many different criteria, for example ML-estimation.

In order to verify if using D-optimal plans for training of classification methods is beneficial five different methods are applied in this paper, namely LDA, QDA, CART, linear SVM, and SVM with radial basis kernel. We expect training based on D-optimal plans mainly to work for LDA and linear SVM for the following reasons.

Optimal plans are model-dependent. The screening model the D-optimal plans are based on (see Section 2) reappears for both methods, linear SVM and LDA in case of two classes with equal prior probabilities, since the classification rule can be written as

$$\hat{y} = \operatorname{sign}(\beta_0 + \boldsymbol{\beta}_1' \boldsymbol{x}).$$

In case of linear SVMs the estimation criterion is the size of the margin between the two classes. In case of LDA for two classes with equal prior probabilities the classification rule originally is given as

$$\hat{y} = \operatorname{sign}\big(- {}^1\!/_2 (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) + (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}^{-1} \boldsymbol{x} \big)$$

and can be rewritten as above with

$$\beta_0 \coloneqq -{}^1\!/_2 (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$$

and

$$\boldsymbol{\beta}_1 \coloneqq \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2).$$

The class means and the covariance matrix are estimated by ML-estimation. But if we suppose that the two classes are coded e. g. as -1 and $+1$ then the estimate of the coefficient vector $\boldsymbol{\beta}_1$ from least squares is proportional to $\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ (see Hastie et al., 2001).

Altogether, the classification models for LDA and linear SVM are very similar in form to the screening model in Section 2. Moreover, LDA has strong relations to least squares estimation and thus fits best into the linear model framework the approach of Pumplün et al. originates from.

4 Simulation Study

The simulation study described in this Section is based on the simulations in Pumplün et al. (2005a). The same eight data sets as in Pumplün et al. are used. Six data sets (balance, breast, diabetes, iris, liver, and wine) are taken from the UCI machine learning repository (Murphy and Aha, 1994). The balance data set is artificial, the other five as well as the remaining two non-public data sets (business and medicine) are real-world data. All data sets constitute two-class problems. If the number of classes originally was larger either observations of some classes were omitted or very similar classes (like in the iris data set the species 'virginica' and 'versicolor') were combined. Table 1 shows the numbers of observations n and the dimensionalities p of the data sets.

data set	n	p
balance	576	4
breast	683	9
diabetes	768	8
iris	150	4
liver	345	6
wine	178	13
business	157	13
medicine	6610	18

Table 1: Numbers of observations and variables.

All data sets are standardized. Misclassification rates are estimated by means of tenfold cross-validation. In each of the ten training data sets D-optimal plans of different sizes are searched for. For the purpose of comparison random samples of the same size as well as the whole training data sets are considered as basis for variable selection and/or estimation of classification models.

In Pumplün et al. plans of size $k = p + 1$ were used for variable selection. However, these contain far too few observations for training of some classification methods, e. g. QDA, therefore we try five different numbers of observations, namely $k \in \{p+1, 2(p+1), 0.1n, 0.25n, 0.5n\}$ with k rounded to whole numbers. Usage of all training observations for variable selection and/or estimation corresponds to $k \approx 0.9n$. Table 2 shows the number of observations in the plans searched for in the eight data sets.

data set	$p+1$	$2(p+1)$	$0.1n$	$0.25n$	$0.5n$	$0.9n$
balance	5	10	58	144	288	518
breast	10	20	68	171	342	615
diabetes	9	18	77	192	384	691
iris	5	10	15	38	75	135
liver	7	14	34	86	172	311
wine	14	28	18	44	89	160
business	14	28	16	39	78	141
medicine	19	38	661	1652	3305	5949

Table 2: Number k of observations per plan (rounded to integers).

A genetic algorithm as described in Section 3.2 is used to search for the D-optimal plans. Similar parameters as in Pumplün et al. are chosen, solely the population size is increased from 10 to 100:

- number of iterations in the outer loop: 100,
- number of iterations in the inner loop: 10,
- population size: 100,
- proportion of plans that are replaced by crossover of the best plans: 0.4,
- probability of a mutation: 0.01.

As already described in Section 3.2 two different criteria, *correlation* and *tree*, are used for variable selection. Three different numbers v of variables are tried, namely $v \in \{0.25p, 0.5p, 0.75p\}$ with v rounded to whole numbers. For the purpose of comparison we also calculate the error rates based on all variables. Table 3 shows the numbers of selected variables for the eight data sets.

Depending on the plan(s) selected from the training data we obtain several variable selection schemes called:

data set	$0.25p$	$0.5p$	$0.75p$	p
balance	1	2	3	4
breast	2	4	7	9
diabetes	2	4	6	8
iris	1	2	3	4
liver	2	3	4	6
wine	3	6	10	13
business	3	6	10	13
medicine	4	9	14	18

Table 3: Number v of selected variables (rounded to integers).

- *vs doptimal*: variable selection using the *correlation* or the *tree* criterion based on an (almost) D-optimal plan of size k,
- *vs doptimal it*: variable selection based on a set of 100 (almost) D-optimal plans of size k. (The number of plans in the set depends on the number of iterations in the outer loop of the genetic algorithm.)

Additionally, for comparison purposes, in Pumplün et al. (2005a) two other variable selection schemes are considered called:

- *no vs*: no variable selection, that is $v = p$,
- *vs standard*: variable selection on the whole training data set.

Here, we add

- *vs random var*: variables are chosen randomly (in order to assess the quality of the variable selection criteria),
- *vs random obs*: variable selection based on a random sample of k training observations (counterpart to *vs doptimal*),
- *vs random obs it*: variable selection based on a set of 100 random samples of size k (counterpart to *vs doptimal it*).

All schemes (except *no vs* and *vs random var*) are used in conjunction with both the *correlation* and the *tree* criterion.

Five different classification methods are applied, namely Lda, Qda, Cart, linear Svm (Svmdot), and Svm with radial basis kernel (Svmrbf).

In Pumplün et al. (2005a) variable selection is done based on the selected plans, but the classification models are estimated on the basis of all training observations. Note that this does not make sense in our situation where it is not possible to measure the class labels of all observations, but that this is helpful in order to assess separately if D-optimal plans are beneficial for variable selection. In this paper, also the appropriateness of D-optimal plans as basis for training of classification methods is investigated and thus the classification models are additionally estimated on the found plans instead of all training observations. According to the variable selection schemes described above the following variants are considered:

- *est all*: classification models are estimated on the whole training data set,
- *est doptimal*: classification models are estimated on the D-optimal plan,
- *est random obs*: classification models are estimated on a random sample,
- *est doptimal it*: classification models are estimated on a set of 100 D-optimal plans,
- *est random obs it*: classification models are estimated on a set of 100 random samples.

We employ these estimation schemes without variable selection and in conjunction with the corresponding variable selection schemes, e. g. *vs doptimal* combined with *est doptimal*, in short *vs est doptimal*.

The procedure described above can be summarized as follows:
For i in 1 to 10:

1. Select a plan of size k / a set of 100 plans of size k (with $k \in \{p+1, 2(p+1), 0.1n, 0.25n, 0.5n\}$) in the i-th training data set (either D-optimal plans or random samples) or keep all training observations ($k \approx 0.9n$).
2. Use the *correlation* and the *tree* criterion to select v variables (with $v \in \{0.25p, 0.5p, 0.75p\}$) based on the found plans or keep all variables ($v = p$).
3. Estimate classification models (LDA, QDA, CART, SVMDOT, SVMRBF) on the basis of the found plan(s) or the whole training data set.
4. Predict the class labels on the i-th test data set.

The R software (R Development Core Team, 2008) is used for the simulation study. The following packages are employed:

- MASS (Venables and Ripley, 2002) for LDA and QDA,
- rpart (Therneau and Atkinson, 2009) for CART, and
- e1071 (Dimitriadou et al., 2009) for SVMDOT as well as SVMRBF.

5 Results

This Section consists of five parts. Firstly, in Subsection 5.1 variable selection based on the whole set of training observations is considered. Then, the found D-optimal and random plans are investigated with respect to D-efficiency and correlations between the explanatory variables in Section 5.2. In Section 5.3 the appropriateness of D-optimal plans as basis for variable selection is assessed and subsequently in Section 5.4 D-optimal plans as basis for training of classification methods are studied. Finally, in Section 5.5 we investigate if usage of D-optimal plans is beneficial for both, variable selection and training in combination.

5.1 Variable Selection and Estimation Based on All Training Observations.

In a first step, we investigate if variable selection is beneficial for some data sets, that is if the error rates decrease or remain constant if the number of explanatory variables is reduced. For this purpose, the error rates resulting from *vs standard* in conjunction with the *correlation* and *tree* criteria and the error rates based on all variables (*no vs*) are compared. Note that the dimensionality of many data sets is rather small. Therefore, we cannot expect that variable selection will be beneficial for all data sets.

In a second step the quality of the two variable selection criteria is considered. As described in Section 3.2 the *tree* criterion is known to be instable while the *correlation* criterion captures only linear dependencies and thus maybe important variables will be missed. In order to check this error rates resulting from *vs standard* based on the two criteria and *vs random var* where variables are selected randomly are compared.

Figures 1 and 2 show the cross-validated error rates for the eight data sets and different numbers of (selected) variables $v \in \{0.25p, 0.5p, 0.75p, p\}$, the five classification methods and two variable selection criteria. Additionally, the error rates resulting from random variable selection are plotted in gray.

Wine and mainly iris constitute very simple classification problems, since error rates close to zero are achieved. The results obtained on the iris data sets will not be used in the following because the classification problem is too simple to detect small differences between classification or variable selection methods. Often CART and QDA show slightly worse performance than the other classification methods, particularly QDA on the

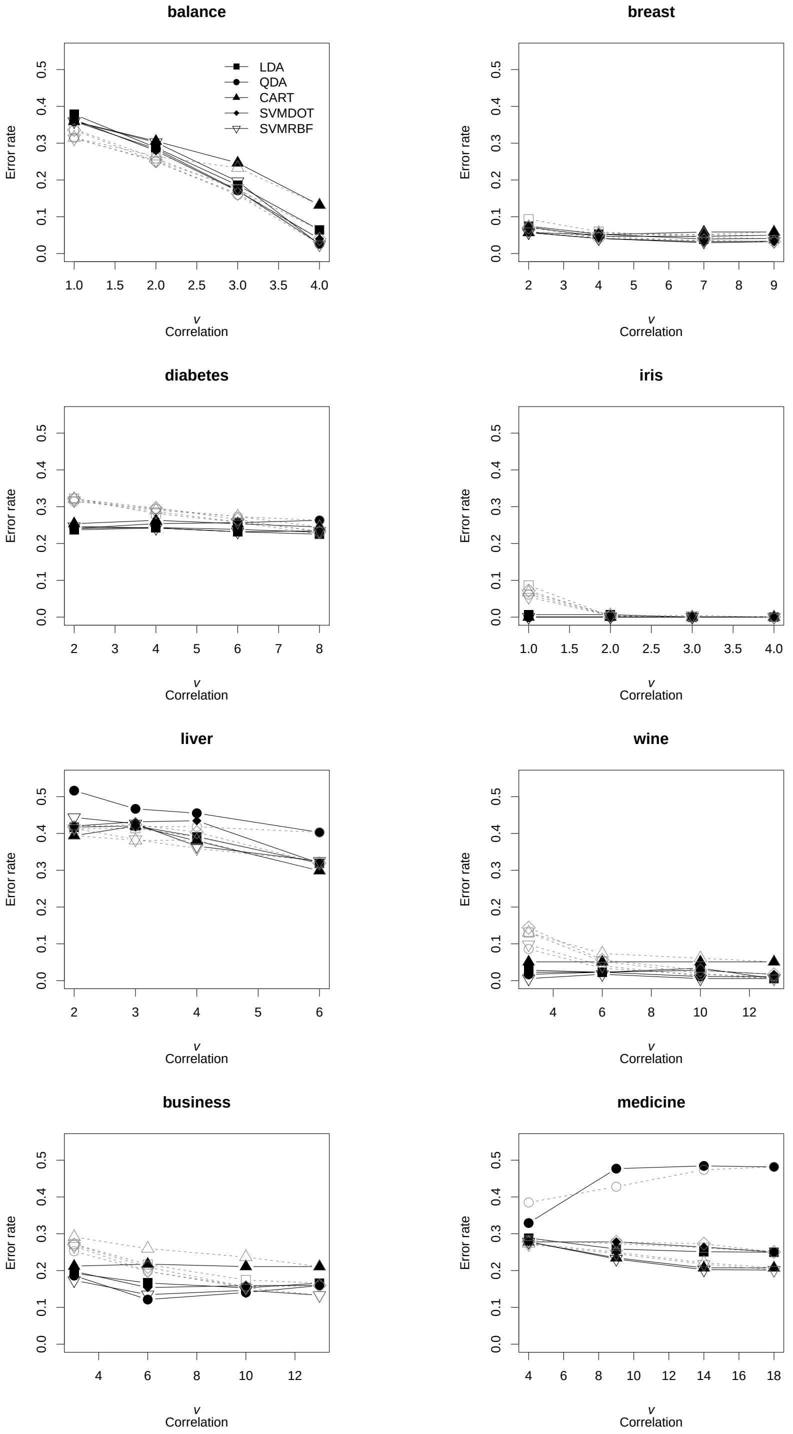

Fig. 1: Cross-validated error rates resulting from *vs standard* with the *correlation* criterion (black), *no vs* (black), and *vs random var* (gray) for different numbers of variables $v \in \{0.25p, 0.5p, 0.75p, p\}$.

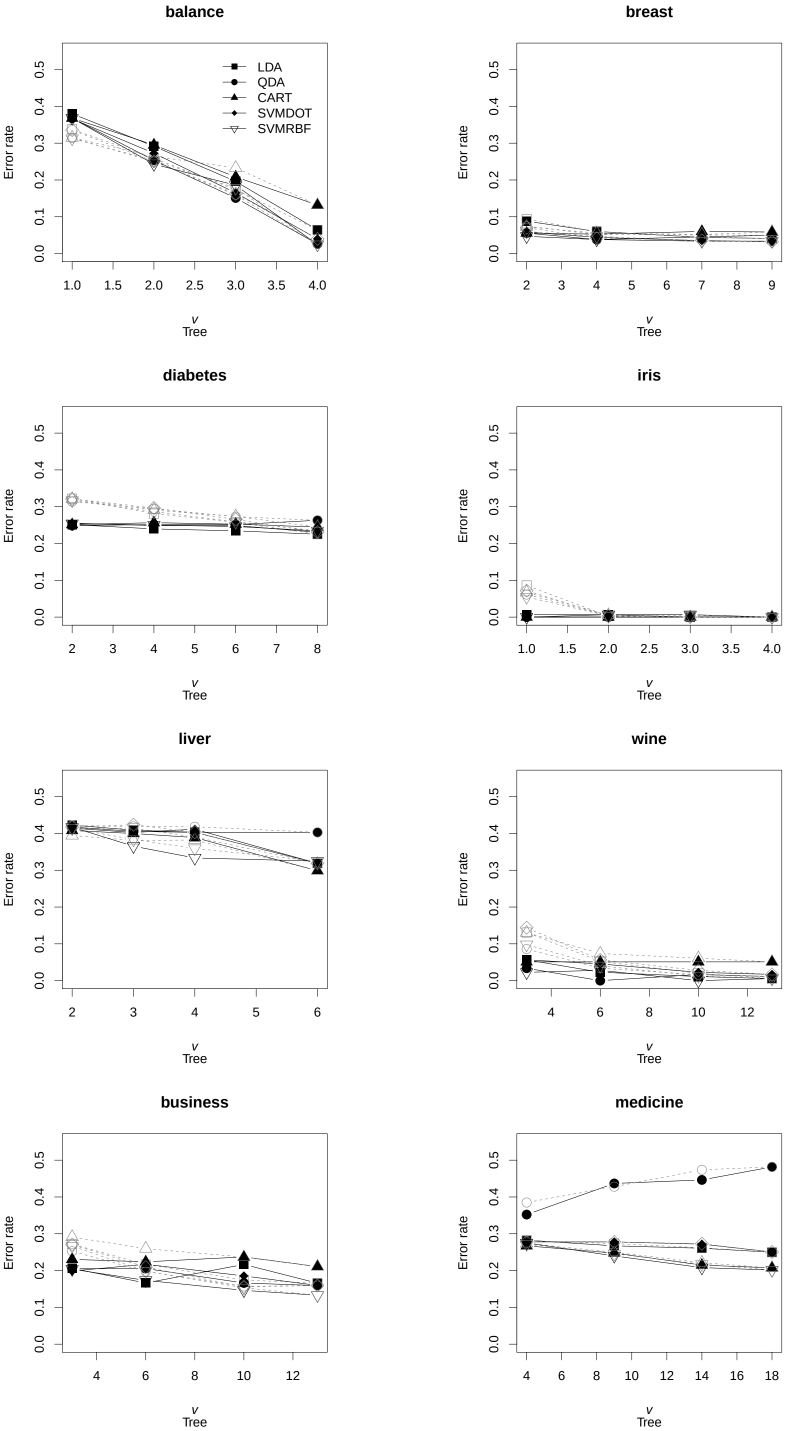

Fig. 2: Cross-validated error rates resulting from *vs standard* with the *tree* criterion (black), *no vs* (black) and *vs random var* (gray) for different numbers of variables $v \in \{0.25p, 0.5p, 0.75p, p\}$.

medicine data set. Variable selection by means of the *correlation* and *tree* criterion is beneficial for most data sets (breast, diabetes, iris, wine, business, and medicine (particularly for QDA)). Only for the liver data set and most notably for the balance data set the error rates grow considerably if the number of variables is reduced. The *correlation* and *tree* criteria yield similar results.

Usage of the *correlation* and *tree* criteria leads to lower error rates than random variable selection for the diabetes, iris, wine, and business data sets. Primarily for low numbers of selected variables the criteria prove to be effective (see the plots for diabetes, iris an business data sets). For medicine and breast similar results are obtained by both methods. *vs standard* performs rather worse than random variable selection for balance and liver data. Note that for these two data sets variable selection in general is not beneficial.

In order to obtain a consolidated result for all data sets we calculate the mean relative deviance

$$\mathrm{MRD}(\text{method 1}, \text{method 2}) = \frac{1}{m} \sum_{i=1}^{m} \frac{e_1^i - e_2^i}{e_2^i}$$

where e_1^i and e_2^i denote the error rates obtained by means of method 1 and 2 on the i-th data set. Since the difference $e_1^i - e_2^i$ is divided by e_2^i an increase of a low error rate is considered worse than an increase of a high error rate and it is taken into account that a decrease of a low misclassification rate is harder to obtain than of a high one. Additionally, we calculate the relative frequencies that method 2 is better than method 1 as

$$\frac{1}{m} \sum_{i=1}^{m} I(e_1^i > e_2^i)$$

where I denotes the indicator function.

Table 4 shows the mean relative deviance MRD(*vs standard*, *no vs*) for the *correlation* and *tree* criteria and the five classification methods. The results of the iris data sets are not used.

Variable selection leads to an increase of the error rates on average. The mean relative deviance MRD(*vs standard*, *no vs*) is largest for a low number v of selected variables. Table 4 is helpful to detect differences between the two variable selection criteria. If the number of variables is small the *correlation* criterion seems to work better than the *tree* criterion. The error rates resulting from CART change only slightly if the number of variables is reduced. The main reason is that usually not all variables are used in a CART-tree and thus CART implicitly selects variables anyway.

correlation	v	Lda	Qda	Cart	Svmdot	Svmrbf	total
	$0.25p$	1.50	1.98	0.34	1.42	2.22	1.49
	$0.5p$	1.05	1.55	0.26	1.01	2.04	1.18
	$0.75p$	1.02	0.79	0.17	0.61	1.03	0.72
tree	v	Lda	Qda	Cart	Svmdot	Svmrbf	total
	$0.25p$	2.29	2.18	0.36	1.73	2.69	1.85
	$0.5p$	1.08	1.10	0.25	1.24	1.98	1.13
	$0.75p$	0.54	0.74	0.16	0.59	0.83	0.57

Table 4: Mean relative deviance MRD(*vs standard*, *no vs*) of the error rates resulting from *vs standard* and *no vs*.

Table 5 shows the relative frequencies that *vs standard* yields the same or lower error rates than *no vs* for different numbers of selected variables and the five classification methods (exclusive of the iris data set).

correlation	v	Lda	Qda	Cart	Svmdot	Svmrbf	total
	$0.25p$	0.00	0.29	0.29	0.00	0.14	0.14
	$0.5p$	0.00	0.57	0.29	0.14	0.00	0.20
	$0.75p$	0.29	0.57	0.43	0.29	0.43	0.40
tree	v	Lda	Qda	Cart	Svmdot	Svmrbf	total
	$0.25p$	0.00	0.29	0.29	0.00	0.00	0.11
	$0.5p$	0.00	0.57	0.29	0.00	0.00	0.17
	$0.75p$	0.00	0.57	0.14	0.00	0.14	0.17

Table 5: Relative frequency that *vs standard* reaches the same or lower error rates than *no vs*.

Using the *correlation* criterion on the average the error rates decrease or remain constant slightly more often than with the *tree* criterion. Mainly Qda benefits from variable selection.

Table 6 shows the mean relative deviance MRD(*vs random var*, *vs standard*) for standard and random variable selection (exclusive of the iris data set). Random variable selection on the average increases the error rates obtained by *vs standard*, particularly if the number of selected variables is small.

correlation	v	LDA	QDA	CART	SVMDOT	SVMRBF	total
	$0.25p$	0.64	0.66	0.33	0.88	2.47	1.00
	$0.5p$	0.26	0.15	0.10	0.28	0.28	0.21
	$0.75p$	-0.01	0.08	0.04	0.03	0.21	0.07
tree	v	LDA	QDA	CART	SVMDOT	SVMRBF	total
	$0.25p$	0.26	0.34	0.32	0.33	0.57	0.36
	$0.5p$	0.24	0.07	0.09	0.04	0.15	0.12
	$0.75p$	0.08	0.04	0.04	0.02	0.03	0.04

Table 6: Mean relative deviance MRD(*vs random var*, *vs standard*) of the error rates resulting from *vs standard* and *vs random var*.

In Table 7 the relative frequencies that *vs standard* yields smaller error rates than *vs random var* are presented. As you can see this is the case for the majority of the data sets for both criteria and for almost all combinations of v and classification methods.

correlation	v	LDA	QDA	CART	SVMDOT	SVMRBF	total
	$0.25p$	0.57	0.71	0.71	0.57	0.57	0.63
	$0.5p$	0.71	0.57	0.71	0.57	0.71	0.66
	$0.75p$	0.57	0.57	0.71	0.43	0.71	0.60
tree	v	LDA	QDA	CART	SVMDOT	SVMRBF	total
	$0.25p$	0.57	0.86	0.71	0.71	0.57	0.69
	$0.5p$	0.71	0.57	0.71	0.43	1.00	0.69
	$0.75p$	0.43	0.71	0.57	0.57	0.86	0.63

Table 7: Relative frequency that *vs standard* yields smaller error rates than *no vs*.

Finally in this Subsection the selection frequencies of variables resulting from usage of the *correlation* and *tree* criteria as well as random variable selection are investigated. As already described in Section 3.2 the *correlation* criterion provides an ordering of the whole variable set, whereas when using the *tree* criterion no information about the importance of variables that are not included in the tree is available. For this reason, in order to reach the desired number v of variables the missing variables are selected randomly if not enough variables are in the tree.

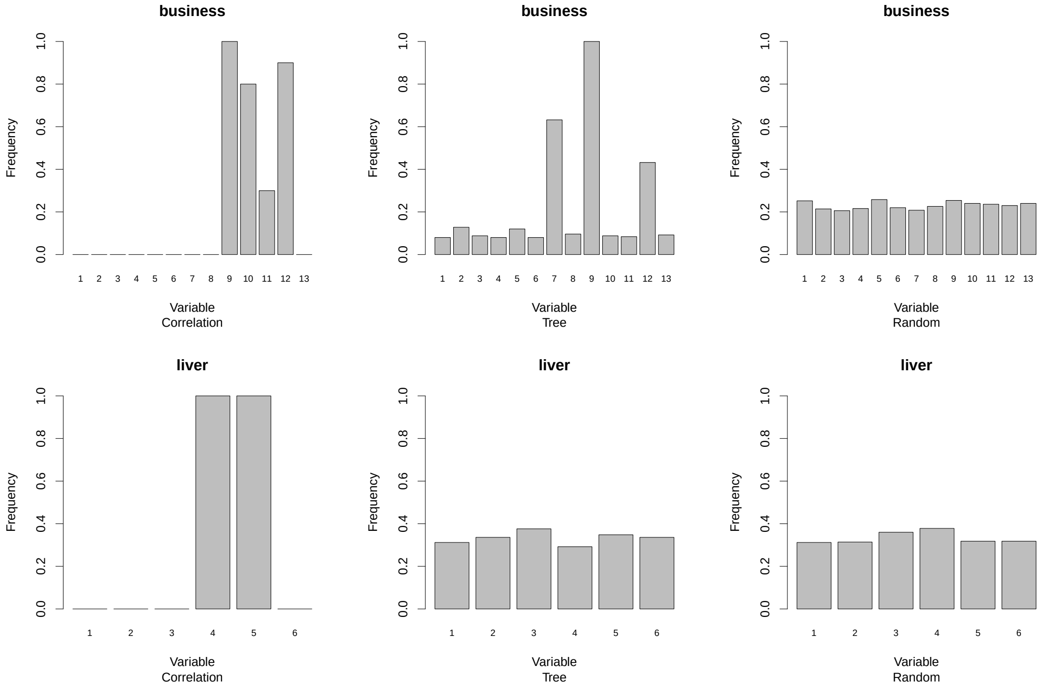

Fig. 3: Selection frequencies of variables using the *correlation* and the *tree* criterion as well as random variable selection in the business and liver data sets for $v = 0.25p$.

Figure 3 shows the relative selection frequencies of the variables in the business and the liver data sets. The desired numbers of variables are $v = 0.25p$, respectively, which results in 2 in case of the liver data set and 3 in case of the business data set. By means of the two criteria different variable subsets are selected. As expected variable selection by means of the *tree* criterion is more unstable than with the *correlation* criterion, particularly in case of the liver data set. Using the example of the business data set you can see that the CART-trees did not always include 3 variables and therefore some variables were randomly selected.

In general we can state that variable selection results only in a small increase of the error rates (except for the balance and liver data sets). The *correlation* criterion seems to work better than the *tree* criterion, particularly for a low number of variables.

5.2 Plans

In order to assess the importance of variables for classification independently of each other the aim was to find (almost) D-optimal plans that are nearly orthogonal and thus lead to uncorrelated predictors. For this reason we firstly investigate the D-efficiencies and correlations between the explanatory variables in the data sets in order to assess how much improvement is possible by means of a D-optimal plan. Second, D-efficiencies and correlations of the resulting D-optimal and randomly chosen plans are investigated.

D-efficiency and Correlations in the Data Sets. The eight data sets can be regarded as designs of size n, respectively. As described in Section 4 the data sets were standardized and therefore the maximal D-value if all pairs of explanatory variables were uncorrelated is $D(\boldsymbol{X}^*_{\text{opt}}) = n^p$.

Table 8 shows the D-efficiencies $D_{\text{eff}}(\boldsymbol{X}^*) = \frac{D(\boldsymbol{X}^*)^{1/p}}{n}$ of the individual data sets. The explanatory variables in balance, which is an artificial data set, already are almost exactly orthogonal to each other and thus uncorrelated, whereas the variables in the iris data set are not.

data set $\boldsymbol{X}^*$	$D_{\text{eff}}(\boldsymbol{X}^*)$	$\bar{r}$	s_r
balance	0.999	-0.021	0.024
breast	0.458	0.602	0.132
diabetes	0.856	0.147	0.160
iris	0.300	0.290	0.661
liver	0.785	0.265	0.194
wine	0.555	0.085	0.351
business	0.438	0.066	0.363
medicine	0.507	0.096	0.255

Table 8: D-efficiencies of the eight data sets under investigation.

Table 8 additionally shows the means $\bar{r}$ and standard deviations s_r of the empirical correlation coefficients $r_{\boldsymbol{X}_{\cdot i},\boldsymbol{X}_{\cdot j}}$ between all pairs of explanatory variables $\boldsymbol{X}_{\cdot i}$ and $\boldsymbol{X}_{\cdot j}$, $i, j = 1, \ldots, p$, $i \neq j$. Beside balance also business, wine, and medicine have a mean correlation close to zero. But due to the high standard deviations the D-efficiencies for these data sets are rather low.

Figure 4 shows parallel boxplots of the correlations between all pairs of explanatory variables in the eight data sets, sorted by D-efficiency in descending order.

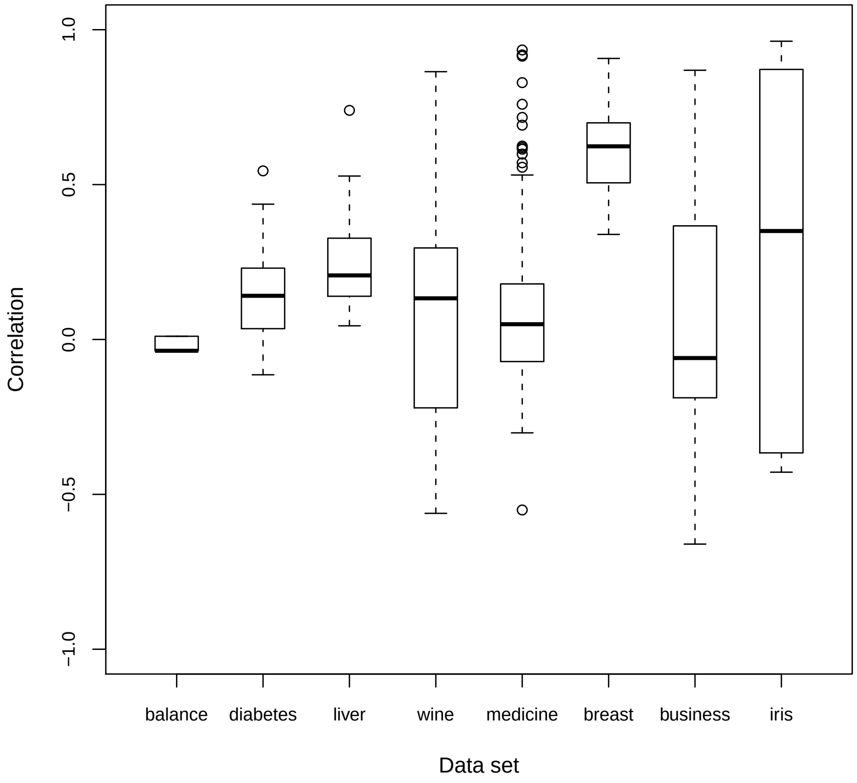

Fig. 4: Parallel boxplots of the correlations between all pairs of explanatory variables, ordered by D-efficiency.

D-efficiency and Correlation of the Resulting Plans. In this paragraph firstly the D-efficiencies of the resulting D-optimal and randomly chosen plans as well as the correlations between the explanatory variables in the plans are investigated.

Figure 5 shows the mean D-efficiencies and standard deviations of the D-optimal plans and random samples for each number k of observations per plan and the eight data sets. In general D-optimal plans possess higher mean D-efficiencies with smaller variances than random samples. The differences in D-efficiency between D-optimal and random plans are largest when the number of observations k in the plans is small. For most data sets the D-efficiencies first increase with rising k. Often there exists an optimal number of observations with highest D-efficiency and if k gets larger we approach the true D-efficiencies of the data sets (see Table 8) that are, with the exception of the balance data set, smaller.

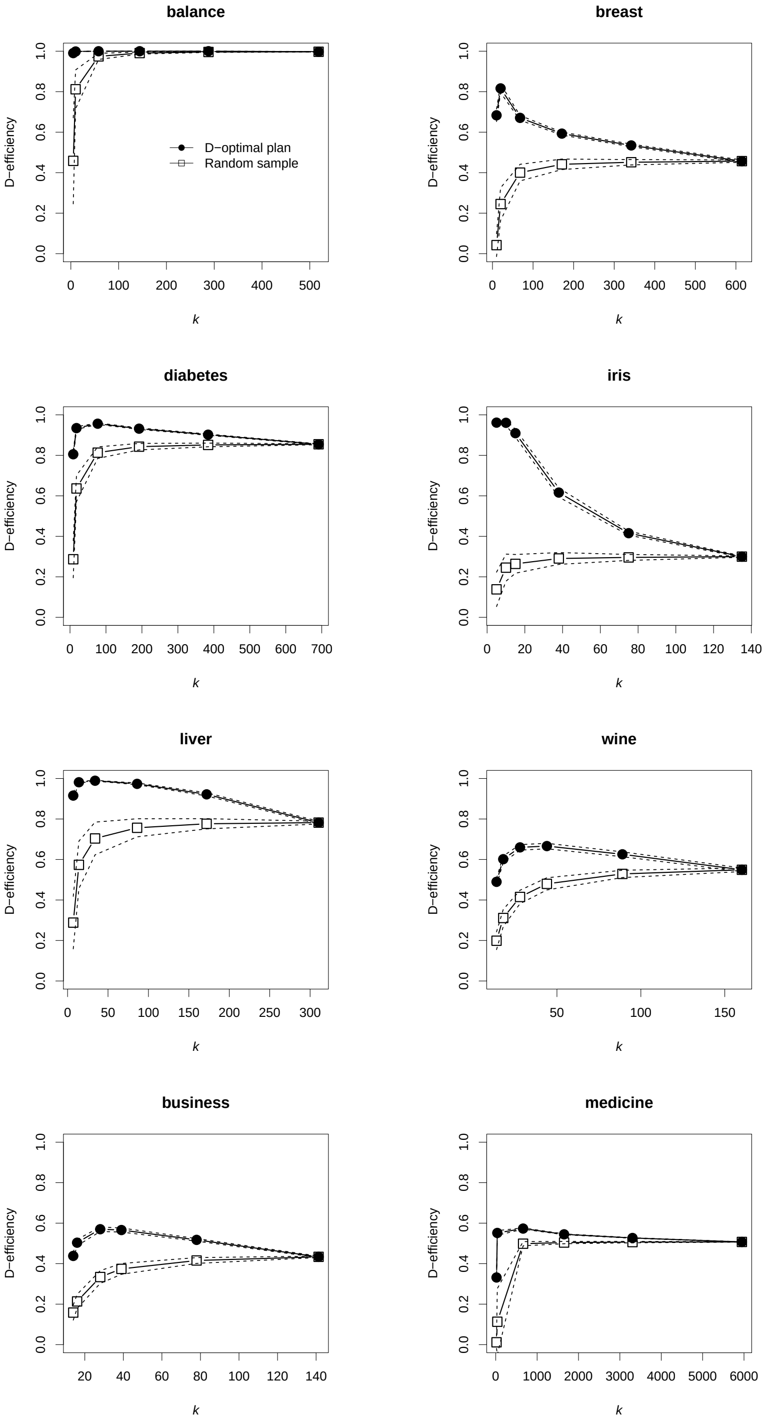

Fig. 5: Mean D-efficiencies and standard deviations of D-optimal and randomly chosen plans for different numbers $k \in \{p+1, 2(p+1), 0.1n, 0.25n, 0.5n, 0.9n\}$ of observations per plan.

The D-efficiencies of the D-optimal plans are very different for distinct data sets. For the balance, diabetes, iris, and liver data sets D-efficiencies of 1 are reached, whereas e. g. for the medicine data set the D-efficiencies are lower than 0.6 for all k. The parameters of the genetic algorithm were selected in an ad-hoc fashion and maybe by means of optimized parameters higher D-efficiencies could be achieved for some data sets. However, depending on the candidate points (see Section 2) that are determined by the particular data set it may not always be possible to reach orthogonality.

Parallel boxplots of the correlations between the explanatory variables in D-optimal plans and random samples are given in Figures 6 and 7. In the D-optimal plans the correlations between the predictors are considerably closer to zero than in the random samples. Moreover, the variances of the correlations are much smaller.

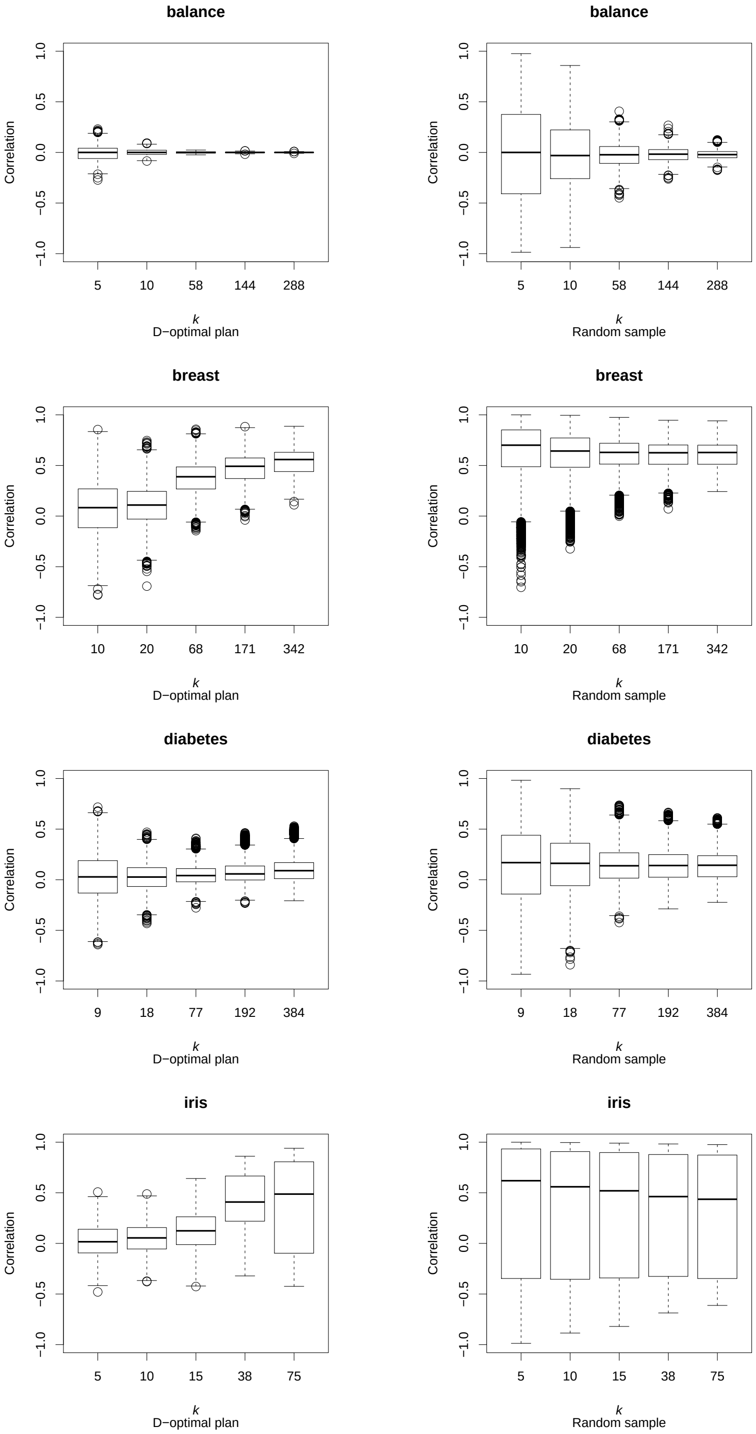

Fig. 6: Correlations between all pairs of explanatory variables for D-optimal and random plans for different numbers $k \in \{p+1, 2(p+1), 0.1n, 0.25n, 0.5n\}$ of observations per plan.

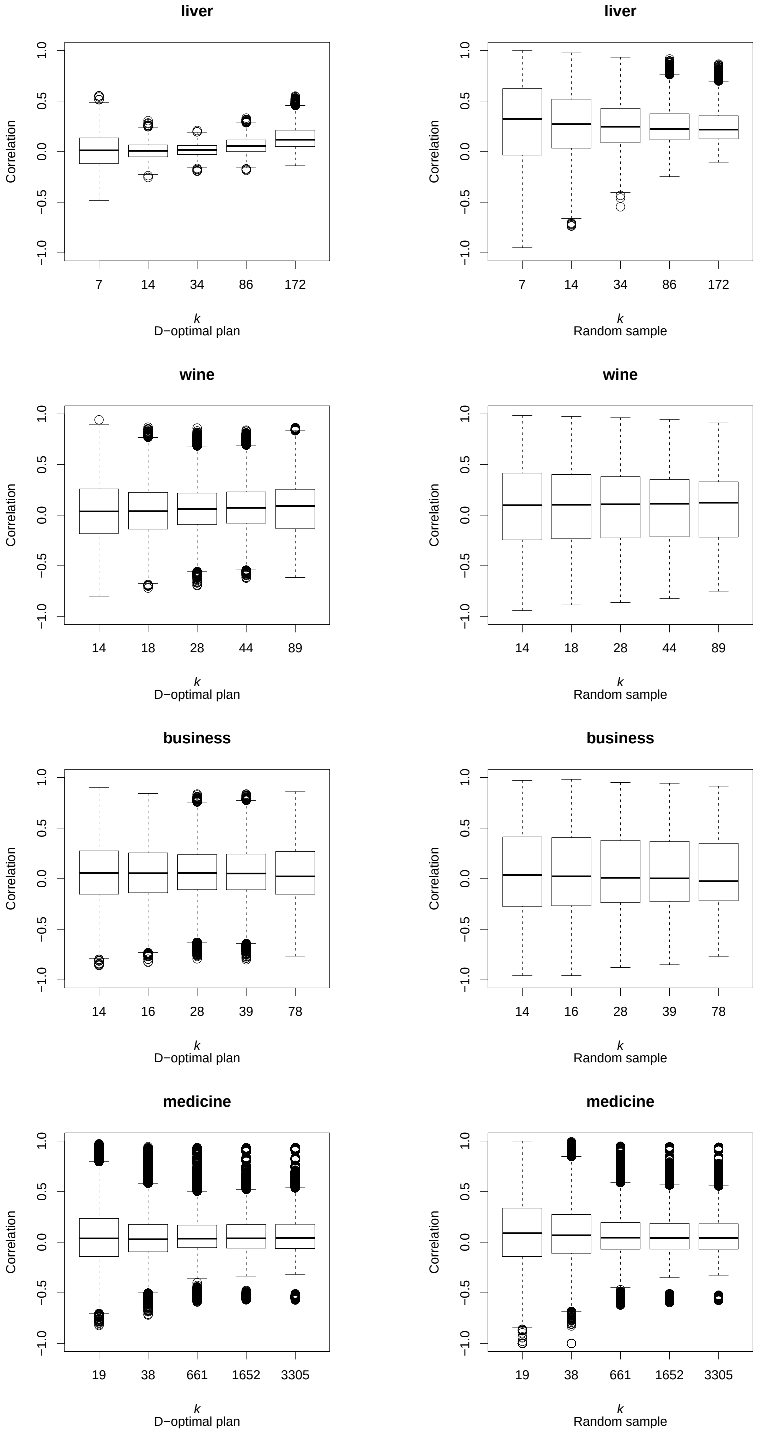

Fig. 7: Correlations between all pairs of explanatory variables for D-optimal and random plans for different numbers $k \in \{p+1, 2(p+1), 0.1n, 0.25n, 0.5n\}$ of observations per plan.

As described in Section 3 Pumplün et al. proposed to use not only the plan with highest D-value for variable selection and/or training, but also a set of 100 plans with highest D-values. In this paragraph we investigate if this approach makes sense. Remember that we are in a situation where the class labels of only few training observations can be measured. One might assume that probably the 100 D-optimal plans in one set highly overlap such that the proportion of observations for which the class label has to be determined gets not too large.

In order to check this we calculate the proportions of observations involved in variable selection and/or training of the classification methods. Since the error rates are estimated by means of tenfold cross-validation the average size of the training data sets is $0.9n$. Hence, for a single D-optimal plan or random sample the mean proportion of observations involved in variable selection and/or training is ${}^{k}\!/_{0.9n}$. We also calculate the mean proportions of observations in the sets of D-optimal plans and the sets of randomly chosen plans. Mean values over all eight data sets are given in Table 9. Additionally, values for single data sets are presented in Figure 8.

			k		
	$p+1$	$2(p+1)$	$0.1n$	$0.25n$	$0.5n$
doptimal/random obs	0.04	0.07	0.11	0.28	0.56
doptimal it	0.68	0.77	0.92	0.99	1.00
random obs it	0.79	0.90	1.00	1.00	1.00

Table 9: Mean proportions of observations involved in variable selection and/or training.

If a set of D-optimal plans is used a lower proportion of observations than in case of randomly chosen plans is required. However, the proportions are much too large. Alone for $k = p+1$ we have to measure the class labels of more than 50% of the observations and for $k = 0.1n$ already on average almost all training observations appear in a set of plans. Therefore, the proposition of Pumplün et al. does not make sense and we do not pursue this approach in the following.

In general we can state that the search for D-optimal plans in the data set was successful. By means of the genetic algorithm we found D-optimal plans with considerably larger D-efficiency than the random samples and the complete data sets. Moreover, for some data sets a D-efficiency of 1, that is exact orthogonality, was achieved.

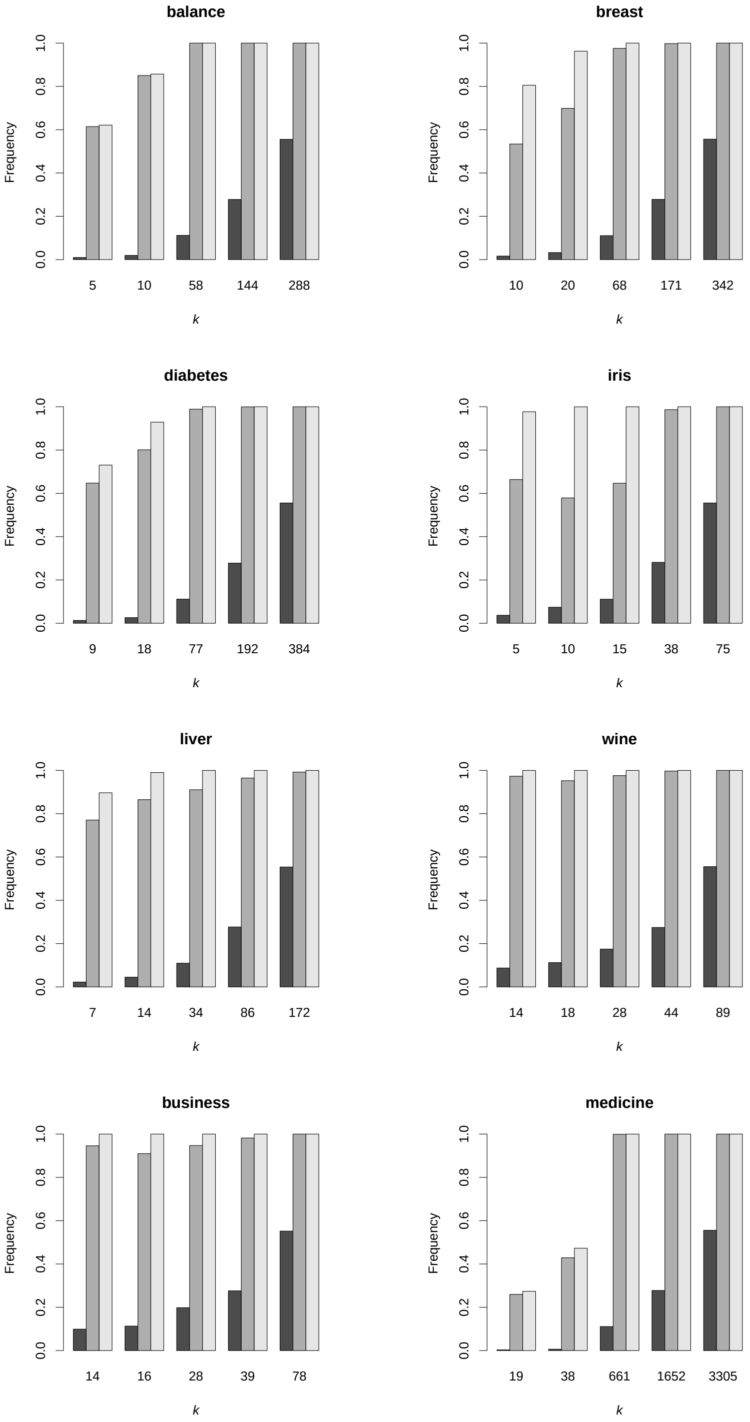

Fig. 8: Mean proportions of observations involved in variable selection and/or training (*doptimal*/*random obs*: black, *doptimal it*: dark gray, *random obs it*: light gray).

5.3 D-optimal Plans as Basis for Variable Selection

In this Section we investigate if D-optimal plans are beneficial for variable selection. In order to assess the effect of D-optimality on the error rate separately only variable selection is done based on D-optimal plans, but the whole training data set is used for estimation of the classification models.

Since the correct variable subsets are unknown correctness can only be measured by means of the error rate. In order to assess if variable selection on D-optimal plans yields lower error rates than on random samples, as in Subsection 5.1, the error rates resulting from *vs doptimal* and *vs random obs* are compared in terms of mean relative deviance MRD(*vs random obs*, *vs doptimal*) and relative frequencies. Subsequently, as the error rates depend of many other different factors like the data set, the classification method, the criterion for variable selection and so on, a linear model is fitted and an analysis of variance is carried out in order to assess which factors are important and if variable selection on D-optimal plans yields significantly lower error rates.

Table 10 shows the mean relative deviance MRD(*vs random obs*, *vs doptimal*) for different numbers of observations per plan, numbers of variables and classification methods (exclusive of the iris data set). A value larger than zero indicates that on average D-optimal plans result in lower error rates than random samples. As you can see the values vary around zero and the proportions of positive and negative values are approximately equal. Hence *vs doptimal* seems not to be beneficial for variable selection.

In Table 11 relative frequencies that variable selection on D-optimal plans results in lower error rates than on random samples are given. The values vary around 0.5, therefore also this table does not provide an indication that D-optimality is beneficial for variable selection.

correlation	v	Lda	Qda	Cart	Svmdot	Svmrbf	total
$k = p + 1$	$0.25p$	0.20	-0.17	0.05	-0.10	0.06	0.01
	$0.5p$	0.10	-0.01	0.02	-0.07	-0.03	0.00
	$0.75p$	0.09	0.02	0.02	-0.09	-0.01	0.00
$k = 0.1n$	$0.25p$	0.13	-0.09	-0.08	0.09	0.07	0.02
	$0.5p$	0.10	-0.04	-0.05	-0.06	0.09	0.01
	$0.75p$	0.09	0.02	0.02	-0.09	-0.01	0.17
$k = 0.5n$	$0.25p$	-0.02	0.08	0.02	0.04	0.06	0.03
	$0.5p$	-0.01	0.02	0.05	-0.01	-0.07	-0.01
	$0.75p$	-0.03	0.01	0.01	0.04	0.05	0.02
tree	v	Lda	Qda	Cart	Svmdot	Svmrbf	total
$k = p + 1$	$0.25p$	-0.02	0.02	0.10	-0.09	0.06	0.01
	$0.5p$	0.12	-0.05	-0.10	0.01	-0.04	-0.01
	$0.75p$	0.10	-0.03	-0.05	0.11	0.20	0.07
$k = 0.1n$	$0.25p$	-0.09	0.02	0.07	-0.01	-0.06	-0.02
	$0.5p$	0.36	0.13	-0.09	0.02	-0.15	0.05
	$0.75p$	-0.12	0.10	0.06	-0.01	0.17	0.04
$k = 0.5n$	$0.25p$	-0.04	0.10	-0.05	0.05	-0.09	-0.01
	$0.5p$	0.03	0.18	0.00	-0.05	-0.06	0.02
	$0.75p$	0.14	-0.10	-0.05	-0.07	0.01	-0.02

Table 10: Mean relative deviance MRD(*vs random obs*, *vs doptimal*) of the error rates resulting from *vs doptimal* and *vs random obs*.

correlation	v	LDA	QDA	CART	SVMDOT	SVMRBF	total
$k = p + 1$	$0.25p$	0.71	0.29	0.43	0.29	0.43	0.43
	$0.5p$	0.57	0.57	0.43	0.43	0.14	0.43
	$0.75p$	0.14	0.14	0.57	0.43	0.29	0.31
$k = 0.1n$	$0.25p$	0.43	0.29	0.43	0.43	0.86	0.49
	$0.5p$	0.71	0.43	0.29	0.29	0.57	0.46
	$0.75p$	0.43	0.57	0.43	0.71	0.14	0.46
$k = 0.5n$	$0.25p$	0.29	0.86	0.57	0.43	0.57	0.54
	$0.5p$	0.29	0.71	0.43	0.29	0.29	0.40
	$0.75p$	0.29	0.43	0.29	0.43	0.57	0.40
tree	v	LDA	QDA	CART	SVMDOT	SVMRBF	total
$k = p + 1$	$0.25p$	0.43	0.71	0.71	0.43	0.71	0.60
	$0.5p$	0.43	0.14	0.29	0.43	0.43	0.34
	$0.75p$	0.29	0.43	0.29	0.71	0.86	0.51
$k = 0.1n$	$0.25p$	0.43	0.57	0.71	0.14	0.29	0.43
	$0.5p$	0.86	0.57	0.29	0.29	0.29	0.46
	$0.75p$	0.14	0.71	0.86	0.29	0.71	0.54
$k = 0.5n$	$0.25p$	0.43	0.71	0.43	0.57	0.29	0.49
	$0.5p$	0.71	0.86	0.29	0.43	0.14	0.49
	$0.75p$	0.43	0.14	0.14	0.29	0.57	0.31

Table 11: Relative frequency that the error rates resulting from *vs doptimal* are lower than the error rates obtained with *vs random obs*.

Using the examples of the business and the liver data sets we assess if different variables are chosen on the basis of D-optimal and random plans. For both data sets the D-efficiency of D-optimal plans is much larger than of random samples (see Subsection 5.2). The mean D-efficiency of D-optimal plans in liver actually is 1, hence the predictors are uncorrelated.

Figures 9 and 10 show the selection frequencies of variables. Variable selection on D-optimal plans and random samples is less stable than based on all training observations due to the lower number of observations that are used. The same variables are considered most important on both types of plans as well as on the whole training data, for example variable 9 in the business data set or variable number 5 in the liver data set. The selection frequencies based on all training observations and on the basis of random samples are very similar. On the basis of D-optimal plans slightly different variables are selected, for example in the business data set variable number 2. Hence D-optimality introduces a bias into variable selection, but this bias does not seem to be beneficial for the error rate.

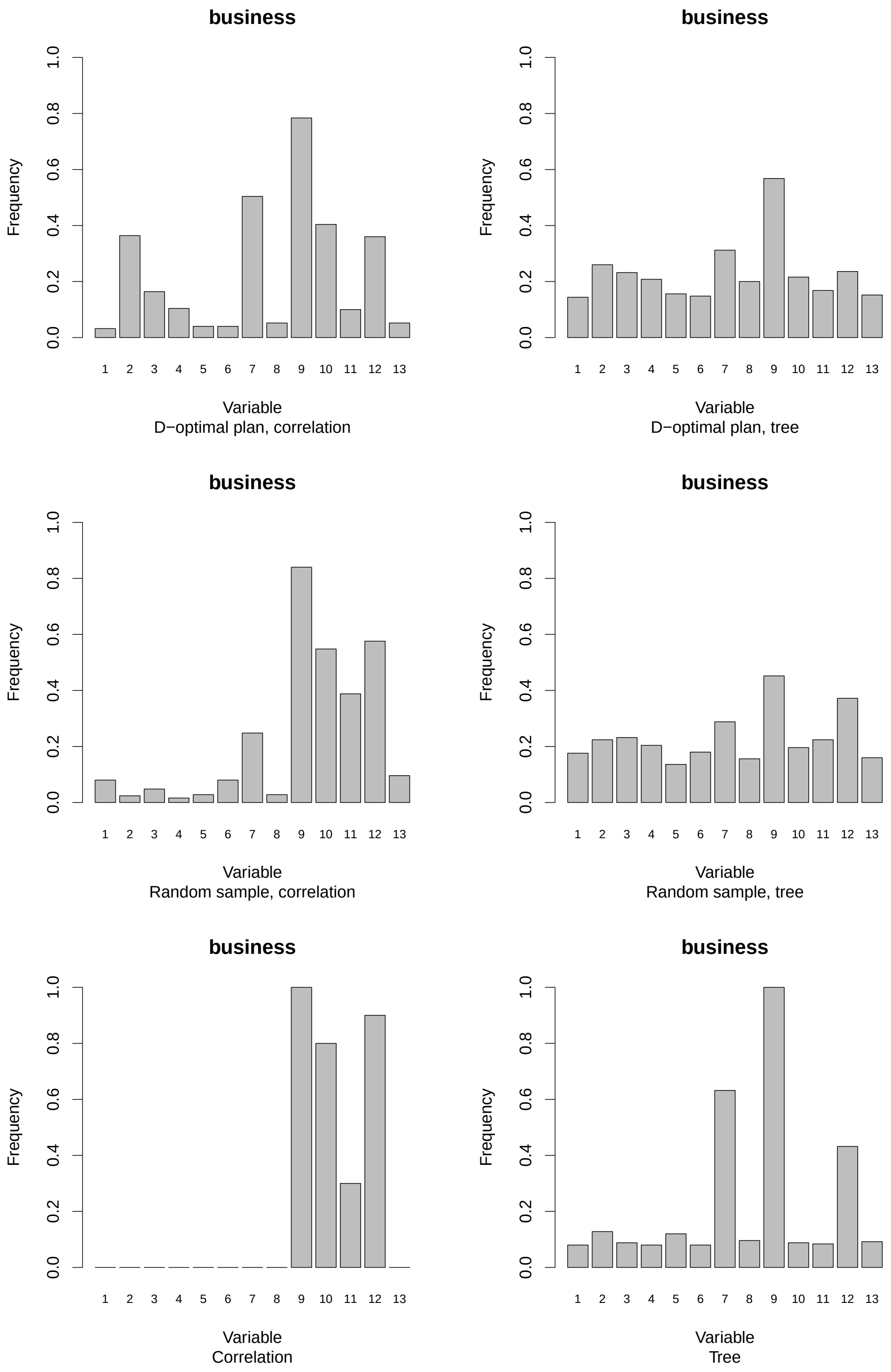

Fig. 9: Selection frequencies of variables using the *correlation* and the *tree* criterion as well as random variable selection on the business data set for $v = 0.25p = 3$.

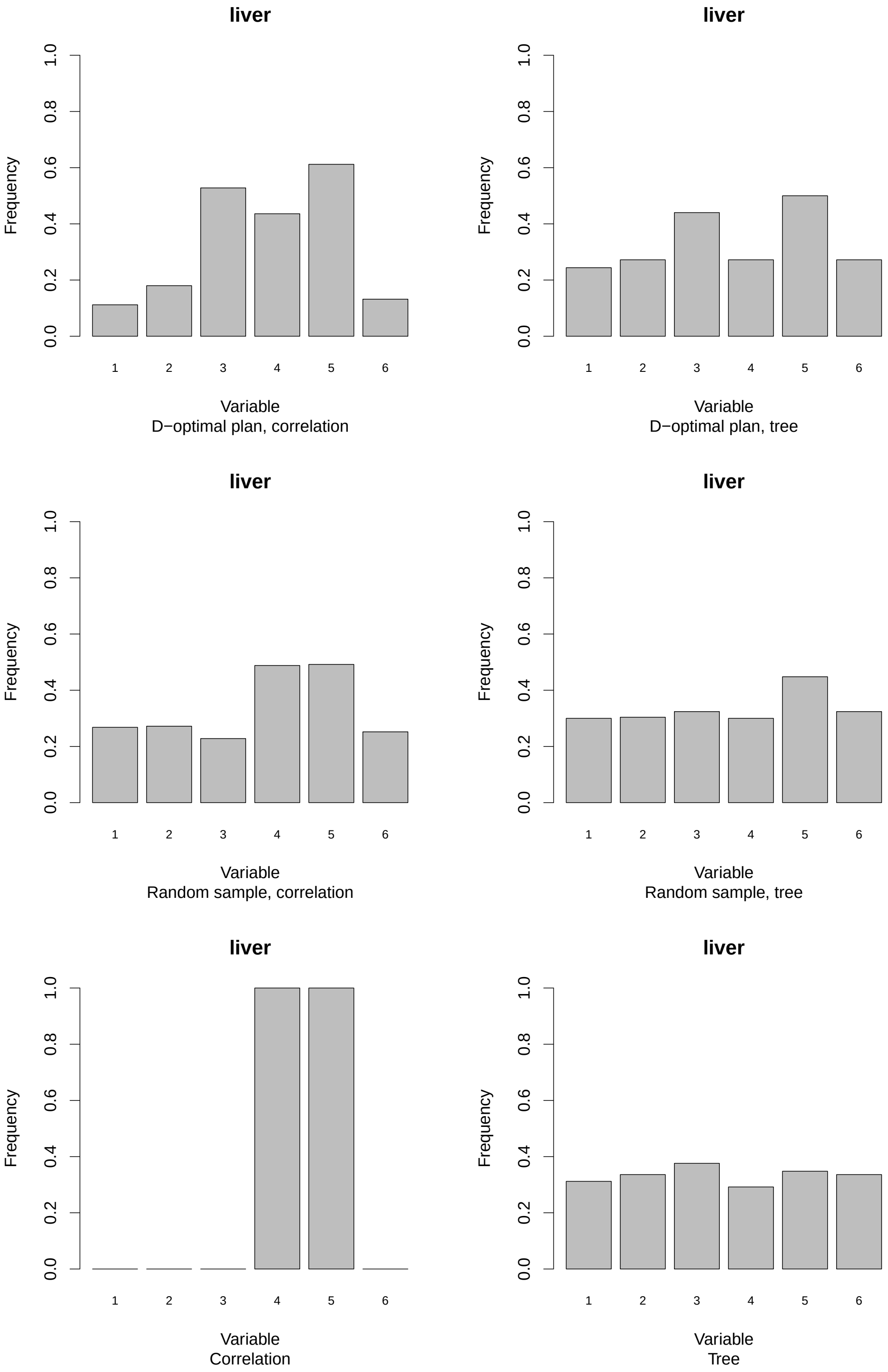

Fig. 10: Selection frequencies of variables using the *correlation* and the *tree* criterion as well as random variable selection on the liver data set for $v = 0.25p = 2$.

The error rates depend on the type of *plan* (D-optimal or random), the *number of observations* per plan and the *number of selected variables*, the selection *criterion*, the *classification method*, and the *data set*. In order to assess which influential factors have an effect on the error rate we fit a linear model including an intercept term, main effects, and two-factor interactions. As response we take the logit of the error rates. A factor effect of 1 then means an increase of the proportion of error and correctness rate by the factor 2.7. Deviation from means coding (see Hosmer and Lemeshow, 2000) of the influential factors is used. Random samples are coded as -1 while D-optimal plans are coded as $+1$. Thus if the estimated effect of the factor type of plan is negative this means that D-optimal plans are beneficial since the error rate is decreased. Deviation from means coding leads to a block-diagonal information matrix, hence the main effects and the two-factor interactions are not confounded. The results on the iris data set are not used.

As can be seen in the analysis of variance table of the main effects (Table 12) the type of plan used for variable selection does not have an effect on the error rate. All other influential factors, particularly the data sets and the number of selected variables, have a considerable impact.

influential factor	df	F-statistic	p-value
plan	1	0.01	*0.91*
number of observations	4	43.87	0.00
number of selected variables	2	1049.64	0.00
criterion	1	111.73	0.00
classification method	4	161.38	0.00
data set	6	11030.22	0.00
residuals	1956		

Table 12: Analysis of variance table of the main effects. Adjusted R^2 is 0.993.

In addition a separate model is fitted for linear classification methods (LDA and SVM-DOT). But for linear classification methods the type of plan does not have a significant effect on the error rate as well (see Table 13).

Figure 11 shows boxplots of the error rates in the D-optimal and random plans for all classification methods and especially for linear classification methods. These confirm the results of the analyses of variance.

influential factor	df	F-statistic	p-value
plan	1	1.19	*0.28*
number of observations	4	28.04	0.00
number of selected variables	2	685.01	0.00
criterion	1	96.70	0.00
classification method	4	4.6804	0.03
data set	6	14783.70	0.00
residuals	741		

Table 13: Analysis of variance table of the main effects for linear classification methods LDA and SVMDOT. Adjusted R^2 is 0.997.

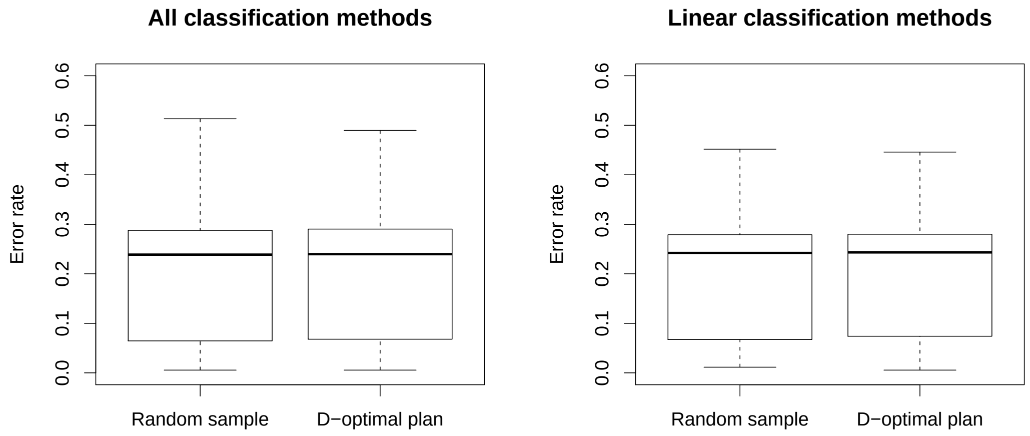

Fig. 11: Error rates resulting from variable selection on random samples and on D-optimal plans.

Finally, linear models are fitted and analyses of variance are carried out for single data sets (see Table 14). The type of plan has a significant effect for liver and business. But the signs of the estimated factor effects are different. That is in case of the liver data set D-optimal plans are beneficial, whereas for the business data set better results are obtained by means of a random plan.

data set	effect of plan type	F-statistic	p-value	R^2
balance	0.012	1.06	0.30	0.996
breast	0.022	0.80	0.37	0.999
diabetes	0.043	2.17	0.14	0.996
liver	-0.128	13.71	*0.00*	0.958
wine	-0.027	1.09	0.30	0.990
business	0.058	3.62	*0.06*	0.990
medicine	0.012	0.85	0.36	0.994

Table 14: Estimated factor effects and F-tests for single data sets (F-distribution with 1 and 234 degrees of freedom).

5.4 D-optimal Plans as Basis for Training

In this Section the appropriateness of D-optimal plans for training of classification methods is investigated. No additional variable selection is done.

The plans of size $p+1$ and $2(p+1)$ often contain too few observations for training of the classification methods. Particularly for QDA due to singular within-class covariance matrices error rates for $k = p+1$ are not available. For $k = 2(p+1)$ we obtained results only for three data sets.

k	LDA	QDA	CART	SVMDOT	SVMRBF	total
$p+1$	0.00	–	-0.04	-0.07	-0.28	-0.10
$2(p+1)$	0.28	0.08	-0.21	0.14	-0.29	-0.01
$0.1n$	0.17	0.17	0.03	0.06	-0.14	0.05
$0.25n$	0.11	0.01	-0.02	0.09	-0.06	0.03
$0.5n$	0.03	0.00	0.00	0.15	0.05	0.05

Table 15: Mean relative deviance MRD(*est random obs*, *est doptimal*) (exclusive of the iris data set) of the error rates resulting from *est doptimal* and *est random obs*.

First, the error rates resulting from *est doptimal* and *est random obs* are compared. Table 15 shows the mean relative deviance MRD(*est random obs*, *est doptimal*) for different numbers of observations per plan and the five classification methods under consideration. Except for SVMRBF most values in Table 15 are positive, that is *est doptimal* results in lower error rates than *est random obs*. Particularly for k of medium size, $0.1n$ and $0.25n$, D-optimal plans for training seem to be beneficial.

k	LDA	QDA	CART	SVMDOT	SVMRBF	total
$p+1$	0.43	–	0.71	0.29	0.14	0.39
$2(p+1)$	1.00	0.67	0.29	0.43	0.14	0.48
$0.1n$	1.00	0.60	0.71	0.71	0.43	0.70
$0.25n$	1.00	0.43	0.43	0.86	0.57	0.66
$0.5n$	0.86	0.57	0.43	0.71	0.57	0.63

Table 16: Relative frequency that training the classification methods based on a D-optimal plan results in lower error rates than training on the basis of a random sample.

Table 16 which shows the relative frequencies that D-optimal plans result in lower error rates than random samples confirms these observations. Lda benefits the most, whereas for Svmrbf the error rates rather worsen if D-optimal plans are used for training.

As described in Subsection 5.3 a linear model is fitted and an analysis of variance is carried out. The influential factors are the type of *plan*, the *number of observations* per plan, the *classification method*, and the *data set*. The model again contains an intercept term, main effects and two-factor interactions. Due to the missing error rates of Qda orthogonality is lost and therefore factor effects are confounded if all available results are used. In order to avoid this problem we also omit the remaining results for Qda and thus again obtain an orthogonal design.

influential factor	df	F-statistic	p-value
plan	1	3.93	*0.05*
number of observations	4	339.69	0.00
classification method	3	99.91	0.00
data set	6	177.43	0.00
plan ∗ number of observations	4	3.68	0.01
plan ∗ classification method	3	7.48	0.00
plan ∗ data set	6	3.79	0.00
number of observations ∗ classification method	12	2.93	0.00
number of observations ∗ data set	24	10.23	0.00
classification method ∗ data set	18	12.83	0.00
residuals	198		

Table 17: Analysis of variance table. Adjusted R^2 is 0.984.

Table 17 shows the results of the analysis of variance. Although the other influential factors have a considerably larger effect on the error rate the type of plan also seems to have an impact. The estimated effect of the plan type is 0.040. This means that training on the basis of a D-optimal plan slightly increases the error rates.

If we conduct an analysis of variance especially for linear classification methods (LDA und Svmdot) we obtain a contrary result (see Table 18). The type of plan seems to have an impact on the error rate as well, but the estimated factor effect is -0.054, that is D-optimality rather helps to reduce the error rate in case of linear classification methods. This is what we expected in Section 3.3.

influential factor	df	F-statistic	p-value
plan	1	5.59	*0.02*
number of observations	4	352.54	0.00
classification method	1	29.34	0.00
data set	6	213.16	0.00
plan $*$ number of observations	4	3.10	0.02
plan $*$ classification method	1	3.00	0.09
plan $*$ data set	6	3.62	0.00
number of observations $*$ classification method	4	8.27	0.00
number of observations $*$ data set	24	10.30	0.00
classification method $*$ data set	6	1.18	0.32
residuals	82		

Table 18: Analysis of variance table for linear classification methods. Adjusted R^2 is 0.994.

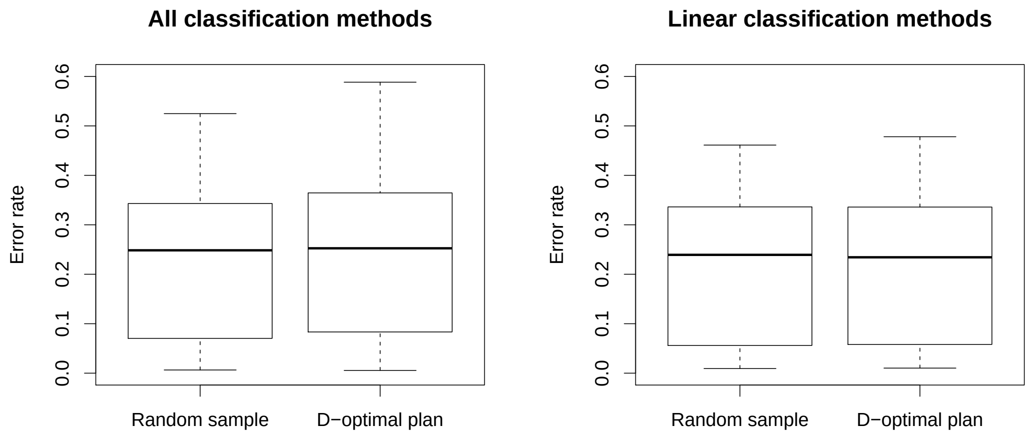

Fig. 12: Error rates resulting from training on random samples and on D-optimal plans.

This can also be seen in the boxplots in Figure 12.

Table 19 shows the results of analyses of variance for single data sets. For all data sets except of balance, wine, and medicine the type of plan seems to have an effect on the error rate. For the liver data set D-optimal plans are beneficial, whereas for breast, diabetes, and business random samples result in lower error rates.

data set	effect of plan type	F-statistic	p-value	R^2
balance	-0.023	1.57	0.23	0.996
breast	0.198	5.59	*0.04*	0.957
diabetes	0.075	26.40	*0.00*	0.999
liver	-0.110	9.39	*0.01*	0.989
wine	0.025	0.34	0.57	0.988
business	0.128	9.53	*0.01*	0.985
medicine	-0.016	1.71	0.22	1.000

Table 19: Estimated factor effects and F-tests for single data sets (F-distribution with 1 and 12 degrees of freedom).

5.5 D-optimal Plans as Basis for Variable Selection and Training

In this Section firstly the error rates resulting from variable selection and training on the basis of D-optimal plans *vs est doptimal* and random samples *vs est random obs* are compared.

Table 20 shows the mean relative deviance MRD(*vs est random obs*, *vs est doptimal*) for different numbers of observations and variables, classification methods and variable selection criteria.

correlation	v	Lda	Qda	Cart	Svmdot	Svmrbf	total
$k = p + 1$	$0.25p$	0.14	-0.19	-0.06	-0.18	-0.22	-0.10
	$0.5p$	-0.01	–	-0.07	-0.01	-0.26	-0.09
	$0.75p$	0.28	–	-0.07	-0.09	-0.23	-0.03
$k = 0.1n$	$0.25p$	0.09	-0.10	0.04	-0.04	-0.06	-0.01
	$0.5p$	0.12	0.17	0.03	0.08	-0.16	0.04
	$0.75p$	0.07	0.16	0.05	-0.03	-0.11	0.02
$k = 0.5n$	$0.25p$	0.07	0.02	0.03	0.02	0.04	0.04
	$0.5p$	0.02	-0.05	0.04	0.08	0.06	0.03
	$0.75p$	0.12	0.00	0.00	0.06	-0.02	0.03
tree	v	Lda	Qda	Cart	Svmdot	Svmrbf	total
$k = p + 1$	$0.25p$	-0.12	-0.12	-0.04	-0.09	-0.17	-0.11
	$0.5p$	-0.03	-0.53	0.03	-0.11	-0.22	-0.11
	$0.75p$	0.34	–	0.03	0.11	-0.24	0.06
$k = 0.1n$	$0.25p$	-0.10	-0.12	0.04	-0.05	-0.09	-0.06
	$0.5p$	0.12	0.33	0.02	0.01	-0.12	0.07
	$0.75p$	0.11	0.08	0.01	-0.03	-0.09	0.01
$k = 0.5n$	$0.25p$	-0.03	0.02	0.01	-0.03	-0.06	-0.02
	$0.5p$	0.13	0.10	-0.05	-0.09	-0.08	0.00
	$0.75p$	0.10	-0.15	-0.07	0.02	0.05	-0.01

Table 20: Mean relative deviance MRD(*vs est random obs*, *vs est doptimal*) (exclusive of the iris data set) of the error rates resulting from *vs est doptimal* and *vs est random obs*.

In Table 21 the relative frequencies that the error rates resulting from *vs est doptimal* are lower than the error rates obtained by means of *vs est random obs* are given.

correlation	v	Lda	Qda	Cart	Svmdot	Svmrbf	total
$k = p + 1$	$0.25p$	0.57	0.14	0.14	0.29	0.14	0.26
	$0.5p$	0.57	–	0.29	0.43	0.29	0.39
	$0.75p$	0.71	–	0.29	0.57	0.43	0.50
$k = 0.1n$	$0.25p$	0.57	0.43	0.71	0.57	0.57	0.57
	$0.5p$	0.86	0.80	0.86	0.86	0.14	0.70
	$0.75p$	0.71	0.75	0.57	0.57	0.57	0.62
$k = 0.5n$	$0.25p$	0.57	0.71	0.43	0.29	0.71	0.54
	$0.5p$	0.43	0.57	0.57	0.71	0.57	0.57
	$0.75p$	0.57	0.71	0.43	0.43	0.29	0.49
tree	v	Lda	Qda	Cart	Svmdot	Svmrbf	total
$k = p + 1$	$0.25p$	0.29	0.50	0.29	0.57	0.43	0.41
	$0.5p$	0.57	0.00	0.43	0.29	0.29	0.37
	$0.75p$	1.00	–	0.57	0.71	0.00	0.57
$k = 0.1n$	$0.25p$	0.57	0.14	0.71	0.43	0.29	0.43
	$0.5p$	0.71	0.43	0.43	0.57	0.43	0.51
	$0.75p$	0.71	0.80	0.57	0.43	0.43	0.58
$k = 0.5n$	$0.25p$	0.29	0.43	0.57	0.43	0.29	0.40
	$0.5p$	0.71	0.71	0.57	0.29	0.43	0.54
	$0.75p$	0.57	0.29	0.14	0.43	0.57	0.40

Table 21: Relative frequency that the error rates resulting from *vs est doptimal* are lower than the error rates obtained by means of *vs est random obs*.

Again for Qda due to the low number of training observations for $k = p+1$ not all error rates are available. In Table 20 the values vary around zero. For $k = p + 1$ most values are negative and for Svmrbf also MRD(*vs est random obs*, *vs est doptimal*) is mainly smaller than zero. That is D-optimal plans rather increase the error rate in these cases. In Table 21 we can see that mainly for k of medium size and a large number of variables v variable selection and training on D-optimal plans may be beneficial. Often for Lda the largest relative frequencies are observed.

As in the previous Sections a linear model is fitted to the results. As in Subsection 5.3 the influential factors are the type of *plan* (D-optimal or random), the *number of observations* per plan and the *number of selected variables*, the selection *criterion*, the *classification method*, and the *data set*. Again the results obtained for QDA are omitted. Table 22 shows the results of the analysis of variance.

The type of plan influences the error rates. Since the estimated effect of the plan type is 0.036 D-optimal plans rather increase the error rate.

influential factor	df	F-statistic	p-value
plan	1	8.44	*0.00*
number of observations	4	639.94	0.00
number of variables	2	126.02	0.00
criterion	1	58.27	0.00
classification method	3	179.32	0.00
data set	6	1093.77	0.00
residuals	1551		

Table 22: Analysis of variance table of the main effects. Adjusted R^2 is 0.971.

The results of the analyses of variance for linear classification methods are given in Table 23. As in the previous Subsection 5.4 the estimated factor effect -0.031 is negative for linear classification methods. Hence, for linear classification methods variable selection and training on D-optimal plans seems to be beneficial.

influential factor	df	F-statistic	p-value
plan	1	2.99	*0.08*
number of observations	4	330.33	0.00
number of variables	2	114.71	0.00
criterion	1	72.43	0.00
classification method	1	0.86	0.35
data set	6	394.70	0.00
residuals	741		

Table 23: Analysis of variance table of the main effects for linear classification methods. Adjusted R^2 is 0.980.

data set	effect of plan type	F-statistic	p-value	R^2
balance	-0.032	4.42	*0.04*	0.988
breast	0.237	83.32	*0.00*	0.974
diabetes	0.084	12.84	*0.00*	0.987
liver	-0.275	35.59	*0.00*	0.874
wine	0.090	15.40	*0.00*	0.982
business	0.138	27.82	*0.00*	0.965
medicine	0.009	0.28	0.60	0.997

Table 24: Estimated factor effects and F-tests for single data sets (F-distribution with 1 and 183 degrees of freedom).

Table 24 shows the estimated factor effects of the plan type and the results of the F-test for single data sets. For all data sets except medicine the plan type seems to have an impact on the error rate. For the majority of data sets D-optimal plans lead to slightly worse error rates than random samples. Only for the balance and business data sets variable selection and training on D-optimal plans is rather beneficial.

6 Summary

In this paper the appropriateness of D-optimal plans for variable selection and training of classification methods is investigated. In our simulation study it turned out that for most data sets the application of a variable selection method is beneficial and that the used criteria *correlation* and *tree* are suitable to detect important variables for classification, where *correlation* is slightly better.

D-optimal plans did not turn out to be useful for variable selection since the type of plan used as basis for selection does not have a significant effect on the error rate. Although on the example of two data sets we could see that based on D-optimal plans slightly different variables are found than on the basis of random samples, these differences do not seem to have an influence on the error rate.

In contrast, D-optimal plans as basis for training rather have a small effect on the error rate. Unfortunately, by means of D-optimal plans the error rate is increased. However, for linear classification methods D-optimal plans as basis for training of the classification methods seem to be beneficial. But note that the other influential factors, e. g. the data set, the classification method, or the number of training observations have a much larger impact.

The same applies for D-optimal plans as basis for variable selection and training. The error rates are slightly increased by D-optimal plans if all classification methods are taken into account. But for linear classification methods the error rates are rather improved.

Acknowledgments

The financial support of Deutsche Forschungsgemeinschaft (SFB 475, “Reduction of complexity in multivariate data structures”) is gratefully acknowledged.

Bibliography

G. E. P. Box. Multi-factor designs of first order. *Biometrika*, 39:49–57, 1952.

M. H. Choueiki and C. A. Mount-Campbell. Training data development with the D-optimality criterion. *IEEE Transactions on Neural Networks*, 10(1):56–63, 1999.

E. Dimitriadou, K. Hornik, F. Leisch, D. Meyer, and A. Weingessel. *e1071: Misc Functions of the Department of Statistics (e1071), TU Wien*, 2009. R package version 1.5-19.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer Series in Statistics. Springer, New York, 2001.

K. Hinkelmann and O. Kempthorne. *Design and Analysis of Experiments*, volume 2: Advanced Experimental Design. Wiley, Hoboken, New Jersey, 2005.

K. Hinkelmann and O. Kempthorne. *Design and Analysis of Experiments*, volume 1: Introduction to Experimental Design. Wiley, Hoboken, New Jersey, second edition, 2008.

D. W. Hosmer and S. Lemeshow. *Applied Logistic Regression.* Wiley, New York, second edition, 2000.

J. Kiefer and J. Wolfowitz. The equivalence of two extremum problems. *Canadian Journal of Mathematics*, 12:363–366, 1960.

N. E. Manolov. Design of experiment supplying training sample for pattern recognition. In V. V. Fedorov, W. G. Müller, and I. K. Vuchkov, editors, *Model Oriented Data Analysis*, pages 113–120, Heidelberg, 1990. Physica-Verlag.

P. M. Murphy and D. W. Aha. UCI repository of machine learning databases, 1994.

C. Pumplün, S. Rüping, and C. Weihs. D-optimal plans in observational studies. Technical Report 44/2005, SFB 475, Complexity reduction in multivariate data structures, Technische Universität Dortmund, 44221 Dortmund, Germany, 2005a. URL `http://www.statistik.tu-dortmund.de/sfb-tr2005.html`.

C. Pumplün, C. Weihs, and A. Preusser. Experimental design for variable selection in data bases. In C. Weihs and W. Gaul, editors, *Classification – The Ubiquitous Challenge*, pages 192–199, Berlin, 2005b. Springer.

R Development Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2008. URL `http://www.R-project.org`.

S. Rüping and C. Weihs. Kernelized design of experiments. Technical Report 02/2009, SFB 475, Complexity reduction in multivariate data structures, Technische Universität Dortmund, 44221 Dortmund, Germany, 2009. URL `http://www.statistik.tu-dortmund.de/sfb-tr.html`.

T. M. Therneau and B. Atkinson. R port by B. Ripley. *rpart: Recursive Partitioning*, 2009. URL `http://CRAN.R-project.org/package=rpart`. R package version 3.1-44.

W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer, New York, fourth edition, 2002. URL `http://www.stats.ox.ac.uk/pub/MASS4`.

C. Weihs and J. Jessenberger. *Statistische Methoden zur Qualitätssicherung und -optimierung in der Industrie*. Wiley-VCH, Weinheim, 1999.