D-optimal plans in observational studies *

Constanze Pumplün † Stefan Rüping ‡
Katharina Morik ‡ Claus Weihs †

October 11, 2005

Abstract

This paper investigates the use of Design of Experiments in observational studies in order to select informative observations and features for classification. D-optimal plans are searched for in existing data and based on these plans the variables most relevant for classification are determined. The adapted models are then compared with respect to their predictive accuracy on an independent test sample. Eight different data sets are investigated by this method.

Keywords:
D-optimality, Genetic Algorithm, Prototypes, Feature Selection

1 Introduction

Due to technological advances, large data bases exist in all branches of industry. Hence it is of particular interest to use these data to determine the variables most important with respect to a certain effect. The aim is to distinguish a subset of the data containing all the necessary important information for classification, so called prototypes. Here, the approach is made

*This work has been supported by the Deutsche Forschungsgemeinschaft, Sonderforschungsbereich 475.

†Fachbereich Statistik, Universität Dortmund, 44221 Dortmund, Germany
‡LS Informatik VIII, Universität Dortmund, 44221 Dortmund, Germany

by Statistical Experimental Design, in particular by the D-optimality criterion. As the search for the D-optimal plan in the data is too time consuming, plans with high D-value are constructed from the data by a genetic algorithm. Based on the plan with the highest D-value the relevance of the factors is determined and a subset of the most relevant factors is selected. With these variables a model is adapted using either linear discriminant analysis or recursive partitioning and regression trees or the support vector machine with a linear kernel or the support vector machine with a radial basis kernel. The resulting error rates on an independent test set are then compared with those using the above methods on the complete data set.

2 D-optimal plans

Let $y = F_X\beta + \varepsilon$ be a linear screening model, $y = (y_1, ...y_n)^t$ denoting the result, $F_X := (\mathbf{1}, X)$ the extended design matrix, $\mathbf{1} := (1, ..., 1)^t$, where X is the $(n \times m)$ design matrix, $\beta = (\beta_1, ...\beta_{m+1})^t$ the vector of unknown coefficients and $\varepsilon := (\varepsilon_1, ..., \varepsilon_n)^t$ the error vector with independent $\varepsilon_i \sim \mathcal{N}(0, \sigma^2), 1 \leq i \leq n$ (cp. [1]). The information matrix of this experiment is defined by $F_X^t F_X$ (cp. [1]). It is well known that the least squares estimate of β, denoted by $\hat{\beta}$, results in $\hat{\beta} = (F_X^t F_X)^{-1} F_X^t y$, with covariance matrix $\sigma^2 (F_X^t F_X)^{-1}$.

In general, a design X is called D-optimal, if it maximizes $|F_Y^t F_Y|$, on the set of all possible $(n \times m)$ design matrices Y. The value $|F_Y^t F_Y|$ is called the D-value of the design matrix Y. For a given design matrix $Y \in \mathrm{Mat}_{\mathbb{R}}(n \times m)$ the $100(1-\alpha)$ per cent confidence region, $0 \leq \alpha \leq 1$, for all components of β forms an ellipsoid, whose volume is proportional to $|F_Y^t F_Y|^{-\frac{1}{2}}$. $|F_Y^t F_Y|^{-\frac{1}{2}}$ is also called the generalized variance of the parameter estimates (cp. [1, 9]). Hence, D-optimal plans minimize the volume of this ellipsoid, i.e. they minimize the generalized variance of the parameter estimates with respect to all comparable designs. Thus, for a D-optimal plan, the estimate $\hat{\beta}$ is the best possible. In the following we look for D-optimal plans in already existing data sets.

3 Heuristic Search for D-optimal Plans

The complete search for a D-optimal plan in the data is often impossible as the data sets are too big and complex. Currently, fast algorithms are only known for the special case of binary matrices having entries {-1,1} with pairwise orthogonal columns [5]. In this case, e.g. Plackett-Burman plans are D-optimal. Although no analysis of the complexity of the general D-optimal plan problem has been done yet, it has been proved that a similar problem, the computation of the MCD estimator, is NP-complete [2]. This is not a proof for our problem, but it indicates that the D-optimal plan problem may not be solved efficiently. Hence, we follow a heuristic approach.

The idea is to use a genetic algorithm to construct an almost D-optimal plan from the data, i.e. a plan with minimum generalized variance. Genetic algorithms, as general purpose optimization algorithms, have been shown to provide good solutions for a variety of practical optimization problems [4].

The algorithm runs as follows: a plan consists of d+1 observations. First, a finite set of plans is chosen at random. For each of these plans, the D-value is computed. The plans with the lowest D-values are replaced by plans similar to those plans with highest D-value, in order to locally optimize the set of plans. Two methods for construction are used. The first one is called *mutation* and consists of randomly selecting an old plan with probability proportional to its D-value (this is called roulette selection in genetic algorithms) and then replacing one of the observations in this plan by an observation drawn at random from all observations except the ones in this plan. The other method for generating a new plan is called *two-point crossover*. Given two plans $(x_1, \ldots, x_{d+1})$ and $(x'_1, \ldots, x'_{d+1})$, two indices i and j with $0 \leq i < j \leq d+1$ are randomly selected. The new plan then consists of $(x_1, \ldots, x_i, x'_{i+1}, \ldots, x'_j, x_{j+1}, \ldots, x_{d+1})$. For $i = 0$, in this vector $x_1, \ldots, x_i$ are omitted and for $j = d+1$, in this vector $x_{j+1}, \ldots, x_{d+1}$ are omitted. These two methods may be seen as a special way of local optimization of the existing set of plans and are repeated for a fixed number of times. Then the whole procedure is repeated from the start to account for bad starting values due to the random initialisation. In the end, the algorithm returns the plan with the highest D-value.

In a variation of this algorithm, not only the overall best plan is returned, but several best plans are returned, namely the plans with maximum D-value from every new start of the algorithm.

The complete algorithm used in this paper can be described as follows:

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

The actual parameters in steps 1, (a) and (b) may be varied for each data set. In general, a large number of iterations in 1 and small numbers in (a) and (b) prove to be both effective in terms of the obtained D-value and computationally efficient. In the experiments, 100 iterations in 1, 10 plans in (a) and 10 iterations in (b) were used. In a prior investigation, a comparison with a complete search on small data sets showed that the optimal solution is approximated efficiently by this genetic algorithm.

4 Feature Selection

4.1 General Feature Selection Methods

The feature selection problem in this case consists of finding the subset of d most important factors for the prediction, where d is fixed. In general, feature selection methods can be classified as wrapper methods or filter methods [6].

Filter methods use a fixed measure of feature importance to select the most important features on a data set independently from the applied learning algorithm. The most popular measure of feature importance is the absolute correlation between a factor and the result y. Several other methods, e.g. based on cross-entropy, have also been proposed.

Wrapper methods repeatedly construct a classification model on subsets of the factors, in order to assess the predictive performance of the variables. Hence, these methods often perform better than filter methods as they take information about the classifier into account. On the other hand, they are obviously much more computer intensive and the evaluation of the predictive performance is problematic for small data sets, as e.g. the observations in D-optimal plans.

Two methods of feature selection are used in our experiments, one based on classification trees and one based on correlation. Both methods give a ranking of factor importance, from which the top d factors are selected.

4.2 Tree-based Feature Selection

Tree-based feature selection is a more complex filter method based on the gini index as it is employed in the construction of classification trees [3]. First, a classification tree is learned on the available data. Each time a variable occurs in this tree, it is assigned a weight of 2^{-p}, where p is the depth of the corresponding node. The weight 2^{-p} is chosen, because the pth level can have at most 2^p nodes. As variables occuring early in the tree are more important than those close to the leaves, they are assigned more weight than the latter. The measure of feature importance is the sum of the weights of each variable.

5 Feature Selection on D-optimal Plans

Here, the search for D-optimal plans is confined to feature selection. The idea underlying this approach is that the selected observations in the plan are taken as optimal if they minimize the correlation between different factors and hence allow to assess the importance of one factor independently of the others. While this is strictly true for the least squares estimator of a linear model (see Section 2), we hope that other model classes and feature selection procedures benefit also from this selection. This is in part motivated by the empirical observation that feature selection by linear classifier weights also has a positive impact on other classification models [7].

A second advantage of this approach is an increase in computational efficiency in the feature selection, because only a small subset of instances is used. However, the drawback is that an additional step in the plan search has to be executed, so an improvement in computer time may only be expected for complex feature selection schemes.

Two feature selection schemes are compared, on the one hand feature selection based only on the plan with the maximal D-value found (called "fs doptimal" in the tables below), and on the other hand a feature selection for a set of plans with high D-values, where the basic feature selection step was executed for each plan and the acutal set of features was selected based on how often each feature was selected in these steps (fs doptimal it). While

the former approach is computationally more efficient, the latter is expected to give more robust results. We also compared the results to the feature selection on all available examples (fs standard). Note that in all versions the final classifier was learned on all observations in contrast to the feature selection which is performed on certain subsets.

The complete algorithm runs as follows:

1. Input: a set of examples $(x_i, y_i)_{i=1...n}$, the desired number of factors $k \leq m$ fixed.
2. Search for the best plan / set of best plans in (x_i, y_i) using the genetic algorithm.
3. Use the feature selection method to either
   (a) select the k most relevant factors, based on the best plan or
   (b) compute the factor weights for each available plan and select the k factors with the highest sum of weights.
4. Estimate a classification model on all observations with the selected factors from either
   (a) the best plan,
   (b) all plans or
   (c) all observations.
5. Return the constructed model.

6 Experimental Results

To validate the performance of the algorithms, 6 data sets, only with continuous attributes from the UCI library [8] (balance, breast-cancer, diabetes, iris, liver and wine), are chosen plus 2 additional non-public real-world data sets (business and medicine). In all these data sets, there are 2 classes to determine. The following table summarizes these data sets.

Name	Size	Dimension
balance	576	4
breast-cancer	683	9
diabetes	768	8
iris	150	4
liver	345	6
wine	178	13
business	157	13
medicine	6610	18

As each data set was evaluated by the tree-based feature selection and the correlation-based feature selection, only 3 out of 16 tables are shown in the following. For the complete evaluation see the appendix. The first column indicates the feature selection method used to select the d most important variables the adapted model is based on. The rows "all features" denote the relative error on the test sample of the estimated classification model based on all features and one of the 4 methods lda: linear discriminant analysis, rpart: recursive partitioning and regression trees, svmdot: the support vector machine with a linear kernel, svmrbf: the support vector machine with a radial basis kernel. The last column holds the number of observations used to determine the d most relevant factors. The percentage results are the differences with respect to the relative error of "all features" in the corresponding cell, if the relative error of "all features" differs from zero. If the relative error of "all features" equals zero, the percentage results denote the absolute differences. Figures 1 - 3 indicate that using only a very small set of observations for determining the importance of the respective factors for classification, the error rate using linear discriminant analysis as well as recursive partitioning and regression trees or the support vector machine with a linear kernel can be improved. There is one case where the improvement is 20.6% of an error rate of 0.05 in figure 2. This motivates the use of the term prototype, as i.e. in figure 2 only 10 observations out of 683 are needed to improve the error rate 0.03 by nearly 10%.

Figure 4 gives a survey, how often each feature selection method gives the best results.

It is surprising that both support vector machines yield best results, using only the plan with the highest D-value for variable selection in at least as many cases as if using the standard feature selection. The support vector machine with a linear kernel results in the best classification model in 44%

	d=4	d=7	d=10	nr. obs.
		lda		
all features	0.16	0.16	0.16	157
fs standard	12.11%	3.68%	19.47%	157
fs doptimal	31.84%	11.58%	-4.21%	14
fs doptimal it	12.11%	7.89%	24.21%	141
		rpart		
all features	0.20	0.20	0.20	157
fs standard	-19.02%	-9.82%	-9.82%	157
fs doptimal	27.81%	3.07%	-9.82%	14
fs doptimal it	-15.75%	-9.82%	-9.82%	141
		svmdot		
all features	0.14	0.14	0.14	157
fs standard	36.53%	37.13%	8.98%	157
fs doptimal	87.72%	36.53%	-4.49%	14
fs doptimal it	32.04%	36.83%	13.77%	141
		svmrbf		
all features	0.13	0.13	0.13	157
fs standard	32.81%	23.44%	18.75%	157
fs doptimal	56.25%	23.13%	33.13%	14
fs doptimal it	27.81%	32.81%	19.06%	141

Figure 1: Correlation-based Feature Selection on the data set business

of all cases studied. Also considering the linear discriminant analysis and the recursive partitioning and regression trees, using all plans with a high D-value increases the percentage of being the best choice only by approximately 3%.

7 Conclusion

Comparing the error rates resulting from different feature selection methods, the method based on the plan with the highest D-value yields particular good results, especially for support vector machines. Depending on the feature selection method (tree, correlation), the D-optimal selection may result in even smaller error rates than using all observations (see also Appendix).

	d=2	d=5	d=8	nr. obs.
		lda		
all features	0.04	0.04	0.04	683
fs standard	89.10%	18.45%	0.00%	683
fs doptimal	74.00%	14.72%	0.00%	10
fs doptimal it	48.17%	18.45%	0.00%	467
		rpart		
all features	0.05	0.05	0.05	683
fs standard	11.91%	-5.91%	-20.60%	683
fs doptimal	64.75%	-8.78%	-20.60%	10
fs doptimal it	11.86%	-5.87%	-5.91%	467
		svmdot		
all features	0.03	0.03	0.03	683
fs standard	71.43%	23.79%	0.07%	683
fs doptimal	61.79%	14.36%	-9.50%	10
fs doptimal it	47.50%	14.22%	0.07%	467
		svmrbf		
all features	0.03	0.03	0.03	683
fs standard	116.57%	38.80%	5.50%	683
fs doptimal	116.49%	49.80%	10.99%	10
fs doptimal it	83.35%	16.57%	5.50%	467

Figure 2: Correlation-based Feature Selection on the data set breast-cancer

This is only a first step in this direction. Future work might be to investigate the D-efficiency of the constructed plans and its connection to the results. Furthermore, one might look for designs by a different optimality criterion. Also feature selection methods which are more specific, depending on the classification method, could be used, in order to further improve the results.

	d=2	d=4	d=7	nr. obs.
		lda		
all features	0.22	0.22	0.22	768
fs standard	4.09%	7.05%	1.75%	768
fs doptimal	44.54%	19.36%	9.38%	9
fs doptimal it	10.57%	9.37%	0.59%	495
		rpart		
all features	0.26	0.26	0.26	768
fs standard	-0.48%	1.55%	-2.52%	768
fs doptimal	2.58%	8.07%	-1.01%	9
fs doptimal it	1.53%	3.03%	3.04%	495
		svmdot		
all features	0.23	0.23	0.23	768
fs standard	4.56%	5.15%	-0.01	768
fs doptimal	12.59%	12.03%	2.85%	9
fs doptimal it	9.17%	8.62%	1.14%	495
		svmrbf		
all features	0.25	0.25	0.25	768
fs standard	-0.01%	-4.78%	-4.25%	768
fs doptimal	12.21%	9.52%	4.22%	9
fs doptimal it	5.83%	-4.24%	-1.62%	495

Figure 3: Correlation-based Feature Selection on the data set diabetes

	lda	rpart	svmdot	svmrbf
fs standard	46.43%	38.71%	32.00%	34.62%
fs doptimal	25.00%	29.03%	44.00%	34.62%
fs doptimal it	28.57%	32.26%	24.00%	30.77%

Figure 4: Percentages of best results for the observation selection methods over all data sets and all values of d (Feature Selection: Correlation)

8 Acknowledgments

The financial support of the Deutsche Forschungsgemeinschaft (SFB 475,'Reduction of Complexity in Multivariate Data Structures') is gratefully acknowledged.

References

[1] A.C. Atkinson and A.N. Donev. *Optimum Experimental Design*. Clarendon Press Oxford, 1992.

[2] T. Bernholt and P. Fischer. The complexity of computing the MCD-estimator. *Theoretical Computer Science*, 326:383–398, 2004.

[3] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Wadsworth, Belmont, CA, 1984.

[4] K. DeJong. Learning with genetic algorithms: An overview. *Machine Learning*, 3(2/3):121–138, 1988.

[5] S. Haustein. private communication, 2004.

[6] R. Kohavi and G. H. John. The wrapper approach. In H. Liu and H. Motoda, editors, *Feature Extraction, Construction, and Selection: A Data Mining Perspective*, pages 33–50. Kluwer, 1998.

[7] D. Mladenic, J. Brank, M. Grobelnik, and N. Milic-Frayling. Feature selection using linear classifier weights: interaction with classification models. In *SIGIR 2004*, pages 234–.241, 2004.

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

[9] R.H. Myers and C. Montgomery. *Response Surface Methodology*. John Wiley & Sons, Inc., 1995.

9 Appendix

Correlation-based Feature Selection on the data set wine
(178 observations, 13 factors)

	d=4	d=7	d=10	nr. obs.
		lda		
all features	0.02	0.02	0.02	178
fs standard	163.46%	130.77%	1.92%	178
fs doptimal	361.54%	228.85%	65.38%	14
fs doptimal it	261.54%	130.77%	1.92%	160
		rpart		
all features	0.13	0.13	0.13	178
fs standard	8.65%	0.00%	0.00%	178
fs doptimal	-8.65%	-4.33%	0.00%	14
fs doptimal it	8.65%	0.00%	0.00%	160
		svmdot		
all features	0.01	0.01	0.01	178
fs standard	242.86%	145.71%	48.57%	178
fs doptimal	385.71%	48.57%	0.00%	14
fs doptimal it	437.14%	145.71%	48.57%	160
		svmrbf		
all features	0.01	0.01	0.01	178
fs standard	252.94%	52.94%	152.94%	178
fs doptimal	302.94%	50.00%	50.00%	14
fs doptimal it	302.94%	102.94%	102.94%	160

Correlation-based Feature Selection on the data set balance
(576 observations, 4 factors)

	d=1	d=2	nr. obs.
	lda		
all features	0.05	0.05	576
fs standard	547.53%	428.34%	576
fs doptimal	498.99%	337.88%	5
fs doptimal it	505.84%	347.31%	323
	rpart		
all features	0.12	0.12	576
fs standard	171.84%	126.78%	576
fs doptimal	153.37%	119.65%	5
fs doptimal it	167.55%	106.97%	323
	svmdot		
all features	0.05	0.05	576
fs standard	528.69%	425.49%	576
fs doptimal	470.24%	380.40%	5
fs doptimal it	463.78%	399.78%	323
	svmrbf		
all features	0.01	0.01	576
fs standard	2314.60%	1814.38%	576
fs doptimal	2151.20%	1702.40%	5
fs doptimal it	2277.12%	1713.29%	323

Correlation-based Feature Selection on the data set iris
(150 observations, 4 factors)

	d=1	d=2	nr. obs.
		lda	
all features	0	0	150
fs standard	0.67%	0.67%	150
fs doptimal	4.00%	0.00%	5
fs doptimal it	0.67%	0.67%	127
		rpart	
all features	0	0	150
fs standard	0.00%	0.00%	150
fs doptimal	4.00%	0.00%	5
fs doptimal it	0.00%	0.00%	127
		svmdot	
all features	0	0	150
fs standard	0.00%	2.00%	150
fs doptimal	4.00%	0.00%	5
fs doptimal it	0.00%	2.00%	127
		svmrbf	
all features	0	0	150
fs standard	0.00%	0.00%	150
fs doptimal	2.00%	0.00%	5
fs doptimal it	0.00%	0.00%	127

Correlation-based Feature Selection on the data set liver
(345 observations, 6 factors)

	d=2	d=4	d=5	nr. obs.
		lda		
all features	0.30	0.30	0.30	345
fs standard	33.21%	29.57%	8.62%	345
fs doptimal	30.31%	33.46%	19.20%	7
fs doptimal it	41.06%	32.41%	16.19%	267
		rpart		
all features	0.31	0.31	0.31	345
fs standard	30.38%	48.16%	22.74%	345
fs doptimal	18.05%	31.42%	0.79%	7
fs doptimal it	17.01%	11.37%	17.89%	267
		svmdot		
all features	0.30	0.30	0.30	345
fs standard	38.08%	43.85%	6.71%	345
fs doptimal	19.76%	17.24%	21.00%	7
fs doptimal it	33.24%	28.43%	34.48%	267
		svmrbf		
all features	0.30	0.30	0.30	345
fs standard	39.18%	33.48%	10.51%	345
fs doptimal	37.33%	11.65%	6.64%	7
fs doptimal it	33.45%	16.24%	5.59%	267

Correlation-based Feature Selection on the data set medicine
(6610 observations, 18 factors)

	d=5	d=10	d=14	nr. obs.
		lda		
all features	0.25	0.25	0.25	6610
fs standard	15.79%	2.78%	0.54%	6610
fs doptimal	12.76%	7.62%	2.36%	19
fs doptimal it	10.47%	5.14%	1.69%	1614
		rpart		
all features	0.21	0.21	0.21	6610
fs standard	31.59%	10.46%	0.50%	6610
fs doptimal	27.01%	12.39%	4.51%	19
fs doptimal it	29.15%	11.89%	1.29%	1614
		svmdot		
all features	0.26	0.26	0.26	6610
fs standard	7.18%	7.35%	5.31%	6610
fs doptimal	7.18%	5.95%	5.60%	19
fs doptimal it	7.18%	7.18%	2.16%	1614
		svmrbf		
all features	0.20	0.20	0.20	6610
fs standard	41.85%	17.92%	2.70%	6610
fs doptimal	38.30%	17.61%	10.42%	19
fs doptimal it	40.23%	17.84%	2.78%	1614

Tree-based Feature Selection on the data set business
(157 observations, 13 factors)

	d=4	d=7	d=10	nr. obs.
		lda		
all features	0.16	0.16	0.16	157
fs standard	8.16%	23.68%	7.89%	157
fs doptimal	47.89%	21.05%	24.47%	14
fs doptimal it	20.26%	27.89%	-20.53%	141
		rpart		
all features	0.25	0.25	0.25	157
fs standard	0.00%	0.00%	0.00%	157
fs doptimal	30.49%	13.07%	25.29%	14
fs doptimal it	30.49%	12.73%	15.41%	141
		svmdot		
all features	0.14	0.14	0.14	157
fs standard	9.28%	23.05%	0.30%	157
fs doptimal	93.11%	51.50%	13.47%	14
fs doptimal it	100.30%	8.38%	9.28%	141
		svmrbf		
all features	0.13	0.13	0.13	157
fs standard	24.09%	38.44%	23.75%	157
fs doptimal	90.94%	24.06%	19.38%	14
fs doptimal it	92.19%	57.19%	38.44%	141

Tree-based Feature Selection on the data set breast-cancer
(683 observations, 9 factors)

	d=2	d=5	d=8	nr. obs.
		lda		
all features	0.04	0.04	0.04	683
fs standard	59.39%	18.55%	0.00%	683
fs doptimal	166.61%	11.27%	14.89%	10
fs doptimal it	122.65%	29.72%	3.72%	467
		rpart		
all features	0.08	0.08	0.08	683
fs standard	0.00%	0.00%	0.00%	683
fs doptimal	15.83%	8.81%	0.00%	10
fs doptimal it	5.11%	0.00%	0.00%	467
		svmdot		
all features	0.03	0.03	0.03	683
fs standard	47.64%	9.57%	4.85%	683
fs doptimal	119.00%	47.78%	-4.79%	10
fs doptimal it	119.00%	14.15%	0.07%	467
		svmrbf		
all features	0.03	0.03	0.03	683
fs standard	66.77%	33.23%	0.00%	683
fs doptimal	199.60%	61.03%	5.50%	10
fs doptimal it	138.32%	61.12%	0.00%	467

Tree-based Feature Selection on the data set diabetes
(768 observations, 8 factors)

	d=2	d=4	d=7	nr. obs.
		lda		
all features	0.22	0.22	0.22	768
fs standard	14.65%	14.66%	2.92%	768
fs doptimal	36.36%	34.65%	8.22%	9
fs doptimal it	32.88%	24.68%	7.63%	495
		rpart		
all features	0.34	0.34	0.34	768
fs standard	0.00%	0.00%	0.00%	768
fs doptimal	3.86%	3.86%	0.39%	9
fs doptimal it	3.86%	3.86%	0.00%	495
		svmdot		
all features	0.23	0.23	0.23	768
fs standard	11.43%	6.28%	5.14%	768
fs doptimal	41.85%	34.92%	6.30%	9
fs doptimal it	30.43%	22.90%	8.01%	495
		svmrbf		
all features	0.25	0.25	0.25	768
fs standard	-1.62%	-2.67%	-3.70%	768
fs doptimal	24.43%	9.05%	8.46%	9
fs doptimal it	28.67%	3.16%	-1.62%	495

Tree-based Feature Selection on the data set wine
(178 observations, 13 factors)

	d=4	d=7	d=10	nr. obs.
		lda		
all features	0.02	0.02	0.02	178
fs standard	590.39%	165.38%	63.46%	178
fs doptimal	826.92%	232.69%	65.38%	14
fs doptimal it	926.92%	394.23%	167.31%	160
		rpart		
all features	0.11	0.11	0.11	178
fs standard	0.00%	0.00%	0.00%	178
fs doptimal	121.85%	101.23%	0.00%	14
fs doptimal it	52.31%	57.54%	5.23%	160
		svmdot		
all features	0.01	0.01	0.01	178
fs standard	582.86%	391.43%	245.71%	178
fs doptimal	1245.71%	297.14%	242.86%	14
fs doptimal it	1414.29%	340.00%	242.86%	160
		svmrbf		
all features	0.01	0.01	0.01	178
fs standard	658.82%	352.94%	300.00%	178
fs doptimal	958.82%	250.00%	300.00%	14
fs doptimal it	870.59%	355.88%	252.94%	160

Tree-based Feature Selection on the data set balance
(576 observations, 4 factors)

	d=1	d=2	nr. obs.
	lda		
all features	0.05	0.05	576
fs standard	541.30%	396.30%	576
fs doptimal	528.28%	405.95%	5
fs doptimal it	476.77%	347.47%	323
	rpart		
all features	0.35	0.35	576
fs standard	0.00%	0.00%	576
fs doptimal	-11.99%	-2.50%	5
fs doptimal it	-15.97%	0.00%	323
	svmdot		
all features	0.05	0.05	576
fs standard	544.86%	396.46%	576
fs doptimal	509.04%	331.78%	5
fs doptimal it	499.27%	377.15%	323
	svmrbf		
all features	0.01	0.01	576
fs standard	2414.60%	1601.31%	576
fs doptimal	2150.11%	1625.49%	5
fs doptimal it	2201.52%	1800.44%	323

Tree-based Feature Selection on the data set iris
(150 observations,4 factors)

	d=1	d=2	nr. obs.
		lda	
all features	0	0	150
fs standard	0.67%	0.67%	150
fs doptimal	9.33%	0.67%	5
fs doptimal it	2.67%	0.67%	127
		rpart	
all features	0	0	150
fs standard	0.00%	0.00%	150
fs doptimal	9.33%	2.00%	5
fs doptimal it	2.67%	0.00%	127
		svmdot	
all features	0	0	150
fs standard	0.00%	0.00%	150
fs doptimal	8.67%	2.00%	5
fs doptimal it	10.67%	2.00%	127
		svmrbf	
all features	0	0	150
fs standard	0.00%	0.00%	150
fs doptimal	7.33%	0.00%	5
fs doptimal it	14.00%	0.00%	127

Tree-based Feature Selection on the data set liver
(345 observations, 6 factors)

	d=2	d=4	d=5	nr. obs.
		lda		
all features	0.30	0.30	0.30	345
fs standard	36.06%	32.41%	32.41%	345
fs doptimal	41.72%	15.28%	15.28%	7
fs doptimal it	29.59%	38.30%	38.30%	267
		rpart		
all features	0.42	0.42	0.42	345
fs standard	0.00%	0.00%	0.00%	345
fs doptimal	0.00%	0.00%	0.00%	7
fs doptimal it	0.00%	0.00%	0.00%	267
		svmdot		
all features	0.30	0.30	0.30	345
fs standard	39.02%	18.90%	18.90%	345
fs doptimal	29.43%	34.37%	34.37%	7
fs doptimal it	37.14%	41.03%	41.03%	267
		svmrbf		
all features	0.30	0.30	0.30	345
fs standard	35.45%	19.98%	19.98%	345
fs doptimal	33.51%	24.96%	24.96%	7
fs doptimal it	37.24%	18.18%	18.18%	267

Tree-based Feature Selection on the data set medicine
(6610 observations, 18 factors)

	d=5	d=10	d=14	nr. obs.
		lda		
all features	0.25	0.25	0.25	6610
fs standard	10.22%	8.41%	6.17%	6610
fs doptimal	13.31%	10.16%	3.33%	19
fs doptimal it	11.98%	9.86%	2.06%	1614
		rpart		
all features	0.28	0.28	0.28	6610
fs standard	0.00%	0.00%	0.00%	6610
fs doptimal	0.00%	0.00%	0.00%	19
fs doptimal it	0.00%	0.00%	0.00%	1614
		svmdot		
all features	0.26 0.26	0.26	6610	
fs standard	7.18%	6.01%	6.42%	6610
fs doptimal	7.18%	7.18%	6.53%	19
fs doptimal it	7.18%	7.18%	6.01%	1614
		svmrbf		
all features	0.20	0.20	0.20	6610
fs standard	35.83%	27.95%	6.25%	6610
fs doptimal	36.53%	14.83%	7.64%	19
fs doptimal it	39.61%	20.62%	4.40%	1614