Comparing Classifiers in Standardized Partition Spaces using Experimental Design[1]

Ursula Garczarek and Claus Weihs

Department of Statistics, University of Dortmund, 44221 Dortmund, Germany

Abstract. We propose a standardized partition space (SPS) that offers a unifying framework for the comparison of a wide variety of classification rules. Using SPS, one can define measures for the performance of classifiers w.r.t. goodness concepts beyond the expected rate of correct classifications of the objects of interest. These measures are comparable for rules from so different techniques as support vector machines, neural networks, discriminant analysis, and many more. In particular, we are interested in assessing the reliability of classification rules when used for proceeding interpretation of the relationship between the values of predictors and the membership in classes.

We will demonstrate the high potential of SPS for the comparison of classification methods in a simulation study to analyse the following problem:

Given a medium number of predictors, (10-20), and a potentially complex relation between classes and predictors, one would expect flexible classification methods like support vector machines or neural networks to do better than simple methods like e.g. the linear discriminant analysis or cart. Nevertheless, one often observes on real data sets, that the simple procedures do pretty well. Our assumption is, that simple methods are more robust against instability, and that the effect of instability superposes the effect of complexity of the relation. By instability we mean the deviation from the assumption that the collected data is some independent and identically distributed sample from some joint distribution of predictors and classes.

We analyse this problem with a simulation study using experimental design.

Keywords. Classification, Comparative Studies, Experimental Design

1 Introduction

2 Standardized Partition Spaces

2.1 Argmax rules

The method of SPS is applicable to all classification methods that finally decide for a certain class $c, c = 1, ..., G$ following some argmax rule. Argmax rules are based on a transformation $\vec{\mathbf{m}}(x) := (\mathbf{m}(x,1), ..., \mathbf{m}(x,G))' \in \mathbf{M}$ of predictor values $x \in \mathbf{X}$ from predictor space $\mathbf{X}$ into some G-dimensional space of real numbers $\mathbf{M} \subseteq \mathbb{R}^G$. The vector $\vec{\mathbf{m}}(x)$ is interpreted as a

[1]This work has been supported by the Deutsche Forschungsgemeinschaft, Sonderforschungsbereich 475.

vector of membership values in the classes. Argmax rules assign an object with predictor values $x \in \mathbf{X}$ into the class with highest membership:

$$\mathbf{cl}(x, \vec{\mathbf{m}}) \quad = \quad \arg \max_{c=1,\ldots,G} \mathbf{m}(x, c).$$

We call $\max_{c=1,\ldots,G} \mathbf{m}(x, c)$ 'assigment values' and the space of all observations that get assigned to the same class 'assigment area' of that class. Together, the assigment areas for all classes form the classifier's partition of the predictor space.

For example, all Bayes optimal classifiers are argmax classifiers.

2.2 Optimal Features

Our idea is motivated by the attempt to make any argmax rule comparable to the 'true' or 'best' Bayes optimal classifier that has complete knowledge ($=: \tau$) about the distribution of predictors and classes on the population of objects. The Bayes optimal classifier assigns objects into the class with highest conditional probability:

$$\mathbf{cl}(x, \vec{\mathbf{p}}_\tau) \quad = \quad \arg \max_{c=1,\ldots,G} \mathbf{p}(c \mid x, \tau), \quad x \in \mathbf{X}.$$

This strategy minimizes the true expected error probability. Membership values of Bayes optimal classifiers all lie in the interval $[0, 1]$ and sum up to 1. We denote this space of membership vectors by $\mathbf{M}^s \subset [0, 1]^G$.

All available information in the predictors about the membership of objects in classes is coded in the membership values of the Bayes optimal classifier. Thus, Fukunaga (1990) calls them "optimal features".

If we had these optimal features, we could use them to answer questions of interest about the interplay of predictors and class membership. In real situations, they are unknown, and estimated membership functions of argmax classifiers may be used as surrogates. Therefore, we are interested reliability of classification rules in this respect.

2.3 Scaling

In a first step, we want membership values to be directly comparable in size. We can not use their raw membership values. One obvious reason is that they may lie on various scales. They neither have to be non-negative nor add up to 1, as the optimal features do. Of course, we can standardize them ad-hoc into $\mathbf{M}^s$ by some monotone transformation without changing the final assignment into classes. But this might lead to patterns more influenced by the standardization procedure than by the rule's classification behaviour.

Less obvious, but equally important, even maximum membership values of argmax rules based on learnt conditional class probabilities are not a reliable measure for the membership of objects in the assigned classes, and thus not a reliable measure for the correctness of the rule's decision. They give information about the rule's performance from its own perspective only, whereas for comparisons, we would prefer a more objective view.

The details of the process of scaling are published in Sondhauß and Weihs (2001) . It is based on approximations of the empirical distribution of assignment values estimated on some test data within assignment areas with the beta-distribution.

All scaled membership vectors $\vec{m}^s(x) := (m^s(x,1),...,m^s(x,G))'$ of any argmax rule lie in the same space $\mathbf{M}^s$, therefore we call this method the method of standardized partition spaces. Scaled membership vectors have the following properties:

- Scaled membership values are directly comparable in size.
- The average assignment value into each class approximates the the correctness rate of that assignment on the test set.
- Scaled membership vectors of observations reflect as much as possible the position of the original membership vectors among each other within assignment areas.

2.4 Measures for the quality of scaled membership values

Following Hand's (1997) specifications for quality characteristics of conditional class probabilities, high quality membership estimations are characterizes as follows: A high assessment (relative to the assessed membership in other classes) in the assigned class should be justified (accuracy), the relative sizes of membership in classes should reflect 'true' conditional class probabilities (precision), and membership values of objects in the different classes should be well-separated (non-resemblance).

Average precision on the test set is used for the scaling of membership values, and thus precision is no longer a quantity for comparison.

The measure of accuracy is based on the Euclidean distances between scaled membership vectors $\vec{\mathbf{m}}^s(x)$ and the vector representing the corresponding *true* class corner $\vec{e}(c(x))$ for the examples (x, c_x) in the test set $\mathbf{T}$. We standardize the mean of these distances such that a measure of 1 is achieved if all vectors lie in the correct corners, and zero if they

Analogously, the measure of the ability to separate is based on the Euclidean distances between scaled membership vectors $\vec{\mathbf{m}}^s(x)$ and the vector representing the corresponding *assigned* class corner $\vec{e}(\mathbf{cl}(x, \vec{\mathbf{m}}(x)))$. We standardize the mean of these distances in the same way as above.

2.5 Potential use of scaled membership values

Once we can trust the quantitative assessment of an observations membership in classes according to some rule, we can use them as surrogates for the optimal features. This makes all kinds of analyses possible for interpretation. e.g.:

- For up to four classes, one can **visualize scaled membership vectors** of various classifier on the same data in a barycentric coordinate system to explore differences in the classifiers' patterns.
- Observations with highest membership can be used as **prototypes** for objects in certain classes

- For each predictor variable, one can **plot membership values** versus the rank of the **predictor** variable for some data to explore their connection in a scale-independent manner.

In this paper we will present the membership-predictor plots.

3 Experiment

We demonstrate the use of standardized partitions spaces in analyzing the astonishing phenomenon that simple classification methods do pretty well on real data sets though their underlying premises about the true relation between predictors and classes can not reasonably be assumed to hold. We will implement a screening experiment to detect the main influencing factors for the performance of various classifiers.

In general, influencing factors for the goodness of any classification methods are data characteristics like the number of classes (we fix as three), the number of predictor variables (we fix as 12), the number of training objects as such (we vary), the number of training objects in classes (we use balanced design only) the number of missing values (we ignore this factor at this stage), and the form of the joint distribution of classes and predictors on objects.

Our main interest is on the influence of the form of the joint distribution of classes and predictors on objects.

The form determines the shape of the optimal partition. The joint faces of partitions between the classes $g = 1, ...G$ is given implicitly by the equations

$$\mathbf{f}_{g,h} : \mathbf{X} \rightarrow \mathbf{X}_{g,h} \subset \mathbf{X}, \mathbf{p}(g|\vec{x}, \tau) = \mathbf{p}(h|\vec{x}, \tau).$$

We view these implicit functions as random variables $Y_{g,h} := f_{g,h}(\vec{X})$.

For a direct systematic scanning of this space of implicit functions via simulations, we would have to fix various functions $f_{g,h}(\vec{X})$ of increasing complexity, and determine from there possible joint distributions of C and $\vec{X}$.

This is rather complicated, and moreover, the connection of this to something one can potentially know about the problem at hand, or see by exploration of the data, is difficult to understand. Thus we do not want to model the complexity of the relation via these implicit functions, but via the conditional distributions $P(\vec{X}|c, \tau)$, $c = 1, ..., G$.

One aspect of the joint distribution is the dependency structure between predictors. Often this makes the learning results less stable. Some classification methods, like e.g. the naive bayes, even assume independence of variables. In all cases, the analysis of the relevance of predictors for the detection of classes is obscured by this inner dependency.

Concerning the shape of the bivariate distributions between one predictor and the class, we define easiness from the perspective of the classifier from a linear discriminant analysis: easy are linear functions f_g, h, which we know can be generated from multivariate normal distributions with equal covariance matrices. These are the assumptions, a linear discriminant analysis is based on. In that case, $Y_{g,h}$ - as a sum of normally distributed variables - is also normally distributed.

To see the effect of instability on the different types of classification methods, we model instability by deflected observations accounting for three factors: the percentage of deflected

Table 1: Definition of Expectation of Predictors in Groups in the non-dependent case

V	G1	G2	G3	relevant for	V	G1	G2	G3	relevant for
V1	0	-1.64	1.64	G2 vs G3	V7	0	-1	1	G2 vs G3
V2	1.64	0	-1.64	G1 vs G3	V8	1	0	-1	G1 vs G3
V3	-1.64	1.64	0	G1 vs G2	V9	-1	1	0	G1 vs G2
V4	0	1.64	1.64	G1 vs rest	V10	0	0	0	Annoyance
V5	1.64	0	1.64	G2 vs rest	V11	0	0	0	Annoyance
V6	1.64	1.64	0	G3 vs rest	V12	0	0	0	Annoyance

observations, the percentage of relevant variables in which the deflection takes place and the direction of the deflection.

Deflection only takes place on the test data. Thus the "true" generating process for the test data differs from that of the training data.

3.1 Classification Methods

We compared a set of classification methods, that are quite different in the assumptions they make about any underlying generating process of the data, and that are famous in different communities and for different tasks: the classifiers from linear and quadratic discriminant analysis 'LDA' 'QDA' from statistics, the naïve Bayes 'NB' (famous for good performance in text classification), a Neural Network 'NN' (a tool for non-linear function approximation), the strikingly simple k-Nearest neighbor classifier 'k-NN', and some decision tree classifier 'rPart' (mainly famous in the machine learning community). In Sondhauß and Weihs (2001) you find a more detailed description of the implementation of these classification methods.

3.2 Quality Characteristics

The target values in our experiment are correctness rate (CR), accuracy (Ac) and ability to separate (AS).The more overlapping the true distribution are the more difficult is the problem as such. Thus we use as target values not the goodness criteria as such but their relation ratios (rCR, rAc, rAS) to the best that can be achieved: the values of these criteria for the optimal bayes classifier.

3.3 Predictors

We also want to demonstrate the use of SPS for analysing the relevance of variables for the different classes. For that purpose we generate our predictor variables such that we have a clear concept for the relevance at least in the non-dependent case.

All univariate distributions have variance 1. They only differ in expectation, kurtosis and skewness. Table gives the defined expectations, which are either zero, one, or the upper or lower 5%-quantile of the standard normal distribution, $u_{.05} \doteq -1.64$ and $u_{.95} \doteq 1.64$.

We fix this expectations for the normal and the non-normal case. The resulting expectations, variances an covariances in the dependent case can be calculated.

3.4 Factors

We define the low (level=-1) and the high level (level=1) for our experimental design by looking at the easiness of the learning task.

The more training data we have, the more information we have to learn classification rules, thus we consider the number of examples in the training set as influencing factor **TO**. We set $N = 1000$ as low (difficulty) level and $N = 100$ as high (difficulty) level.

To model the deviance of the true distribution from the normal distribution, we use the Johnson System (Johnson, 1949), where random variables can be generated such that they have pre-defined first four central moments. Low and high skewness values (0.1^2 and 1.15^2), and low and high kurtosos values (2.7 and 5.0) are selected such that all combinations are valid, and that a wide range of distributions in the skewness-kurtosis-plane is spanned.

On the low level of dependency **Dp**, we generate independent predictor variables $X_k, k = 1, ..., K$. The high level of dependence is constructed by calculating "inverted" variables: $\tilde{X}_k := \sum_{i=1}^{K} X_i - X_k$, $k = 1, ..., K$.

We do want deflection to be the deviance from the ordinary, thus the chosen high level of 40% for the percentage of deflected observations **DO** assures that more than a half of the observations is "ordinary". We define the low level as 10%. For the low level of the percentage of deflected variables **DV** only one of the nine relevant predictor variables is deflected, on the high level, all but one.

The direction of deflection **DD** of an observation is either determined by a shift of the value in each affected predictor variable towards its mean in the true group of the observation, or away from it towards the nearest true mean of this variable in another group.

3.5 Experimental design

To do the screening for detect the relevant factors for the relative performance of the analyzed set of classifiers, we use a standard Plackett-Burman design for seven factors.

4 Results

In Table 2 we present the values of those coefficients of the approximated linear response functions that are significant to a five percent level, the measure of the fit of the linear model based on all factors R^2, and based only on the significant factors $R^2_{5\%}$.

The classifiers differ strongly in the factors that most heavily influence their relative performance: NB, k-NN, and rPart react strongly negative on the dependency, whereas QDA, and LDA react both negative on kurtosis, most others react positively, only NN does not react at all on kurtosis. The reaction on the deflection is all in all not very strong (relatively to the best that could have been learnt), only LDA is reacting positively!

The astonishing result that LDA also reacts positively on skewness, may be turn out to be the result of interactions. Especially for LDA, QDA, and NN, the low R^2-values indicate that models with interactions should be fitted.

Table 2: Descriptions of Approximated Response Functions

Meth.	Crit.	Intc.	TO	K	S	DP	DO	DV	DD	R^2	$R^2_{5\%}$
LDA	rCR	.9997	−.0189	−.0071	.0062				.0072	.89	.83
	rAc	.9973	−.0503	−.0158	.0173				.0175	.87	.81
	rAS	.999	−.0347	−.013	.0117				.0142	.89	.82
QDA	rCR	.9963	−.0583							.90	.82
	rAc	.9917	−.1334	−.0238						.91	.86
	rAS	.9933	−.0965	−.0192			.0165			.91	.88
rPart	rCR	.929	−.0966	.0479		−.1171				.91	.90
	rAc	.779	−.166	.1005		−.237				.91	.90
	rAS	.8873	−.1342	.0688		−.1806				.92	.91
NB	rCR	.999	−.1498	.1241		−.2816				.97	.96
	rAc	.9973	−.2308	.1665		−.5473	.0733			.97	.97
	rAS	.9983	−.2406	.1926		−.4381				.97	.96
k-NN	rCR	.9973	−.0586	.0286	.0167	−.0528	.0129			.98	.98
	rAc	.9927	−.1345	.0933		−.1213				.93	.91
	rAS	.9967	−.0983	.0426	.0319	−.0868	.0251			.98	.98
NN	rCR	.976	−.0384				.0139			.73	.65
	rAc	.9247	−.0894	.0396				−.0401	−.0586	.74	.72
	rAS	.9657	−.06				.0212			.74	.67

In Figure 1 you see that the influence of the third variable for the separation between group 1 and 3 can easily be read off the membership-predictor plot for plan No. 1, where all factors are set on their low values. Because of the scaling, the high level of the assignment values reflect the fact that these assignments can be trusted. Due to the black-box-nature of neural networks this information could typically not be read off that easily.

References

[Fuk90] Keinosuke Fukunaga. *Introduction to Statistical Pattern Recognitionen.* Academic Press, New York, 2nd edition, 1990.

[Han97] David J. Hand. *Construction and Assessment of Classification Rules.* John Wiley & Sons, Chichester, 1997.

[Joh49] N. L. Johnson. Bivariate distributions based on simple translation systems. *Biometrika*, 36:149–176, 1949.

[SW01] Ursula M. Sondhauss and Claus Weihs. Standardizing the comparison of partitions. Technical report, SFB 475, University of Dortmund, 2001. 31/01.

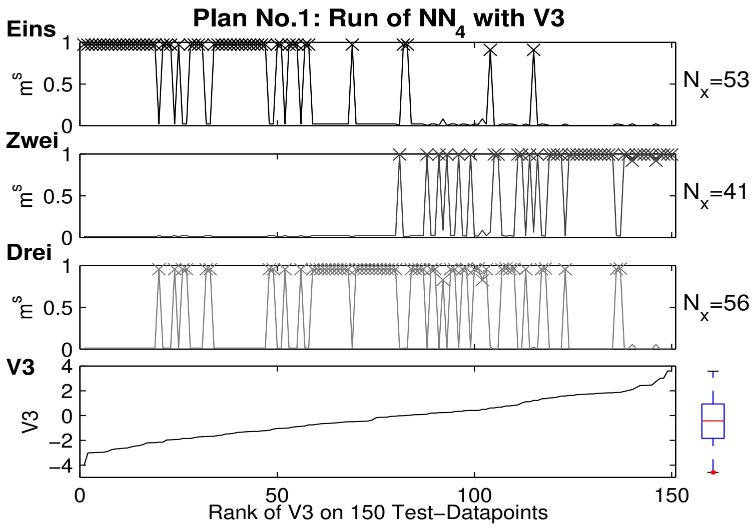

Figure 1: Example of a membership-predictor plot. Scaled membership values in all three groups are plotted versus the rank of V3 within a subsample of 150 points of the test set. Crosses mark the assignment values. The forth plot displays the run of the actual value of V3, and the shape of the marginal distribution of V3 is visualized in the boxplot at the bottom on the right side.