Standardizing the Comparison of Partitions [1]

Ursula Sondhauss[2] and Claus Weihs[2]

[2] Department of Statistics, University of Dortmund
Vogelpothsweg 87, 44221 Dortmund, Germany

Summary

We propose a standardized partition space that offers a unifying framework for the comparison of a wide variety of classification rules. Using standardized partition spaces, one can define measures for the performance of classifiers w.r.t. goodness concepts beyond the expected rate of correct classifications such that they are comparable for rules from so different techniques as support vector machines, neural networks, discriminant analysis, and many more. For classification problems with up to four classes, one can visualize partitions from classification rules that allow for a direct comparison of characteristic patterns of the rules. We use these visualizations to motivate measures for accuracy and non-resemblance in the sense of Hand (1997), enhanced for non-probabilistic classifiers.

1 Motivation

In the days of data mining, the number of competing classification techniques is growing steadily. Thus, it is a worthy goal to rate the goodness of classi-

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

fication rules from a wide range of techniques related to diverse theoretical backgrounds. Restricting the term *goodness* to what can be most easily formalized and measured is dissatisfactory. That is, 'goodness' of a classification rule can stand for much more than only its ability to assign objects correctly to classes. This is, however, the only aspect that is measured by the most famous performance measure, the misclassification rate. Misclassification rates do not cover the variety of demands on classification rules in practice.

In this context, Hand (1997) attaches importance on goodness concepts regarding the rule's quantitative assessment of the membership of objects in classes. This assessment typically determines the final assignment into classes: 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**).

Beyond the reliable quantitative assessment of the membership of new objects in classes, in many practical applications of classification techniques it is important that this assessment can be easily understood and is comprehensible. For that purpose one often looks at the range of values of predictors assigned to the same class. This relates to the rule's induced partitions of predictor space. This predictor space is either the original space of observed features as such, or, in cases where this space is too big for an understandable description, a space of suitably derived features. In the second case, a direct comparison of the partitions from different classifiers would only be possible, if all classification methods would deduce at least resembling entities. This is not true for the wide variety of methods that are used in data mining, e.g. discriminant analysis, neural networks, support vector machines, decision trees,...

Therefore, we will standardize the space of induced partitions such that in the standardized partition space we can compare and visualize the basic pattern of rules, and additionally we can measure performance w.r.t. goodness concepts like accuracy, precision, and non-resemblance.

2 Argmax Rules

Our method is applicable to all classification methods that finally decide for a certain class $c, c = 1, ..., G$, using an argmax rule $\mathbf{cl}$ (like, e.g., Bayes optimal classifiers) based on transformations $\vec{\mathbf{m}} \in \mathbf{M}$ of the observed predictor values from a predictor space $\mathbf{X}$ into some G-dimensional space of real numbers $\mathbf{M} \subseteq \mathbb{R}^G$:

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

The vector $\vec{\mathbf{m}}(x) := (\mathbf{m}(x,c), ..., \mathbf{m}(x,G))$ is interpreted as a vector of membership values for classes.

Our idea is motivated by the attempt to make any argmax rule comparable to the 'true' or 'best' Bayes optimal classifier. Any Bayes optimal classifier maximizes the probability that a new object will be classified correctly, given some learning set of examples $\mathbf{L} := \{(\mathbf{x_1}, \mathbf{c_1}), ..., (\mathbf{x_{N_L}}, \mathbf{c_{N_L}})\}$ and some prior knowledge ξ:

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

where $\mathbf{p}(c \mid x, \mathbf{L}, \xi), \mathbf{c} = \mathbf{1}, ..., \mathbf{G}, \mathbf{x} \in \mathbf{X}$, are learnt conditional class probabilities that are the membership values of this type of classifier, and $\vec{\mathbf{p}}_{\mathbf{L},\xi}(x)$ is the vector of these probabilities.

By τ we denote the complete knowledge about the relationship between predictors and classes that can be expressed in a probability model (including a deterministic relationship as a special case) independent of the training set $\mathbf{L}$. We call the corresponding classifier the true (or best) Bayes optimal classifier:

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

Membership values of Bayes optimal classifiers all lie in the interval $[0, 1]$ and sum up to one. We denote this space of membership vectors by $\mathbf{M}^s \subset [0, 1]^G$. In future, this will also be the space for standardized partitions. For up to four classes, the partition of a Bayes optimal classifier can be visualized in a so-called regular simplex, also known as a barycentric coordinate system. Such a diagram is well known in experimental design to represent mixtures of components, and is used e.g. by Anderson (1958) to display regions of risk for Bayes classification procedures. For the purpose of visualizing the rule's behaviour in such coordinates, we represent the conditional class probabilities $\mathbf{p}(c \mid x, \mathbf{L}, \xi), (\mathbf{x}, \mathbf{c_x}) \in \mathbf{T}$ of some test set objects that were not used for the learning of the rule.

Example *For illustration, we generated small data sets (27 observations each) for the training and the testing of a quadratic discriminant classifier with bayes-rule (Bayes-QDA). Observations come from three classes with χ^2-distributions with parameters $\nu_1 = 2$, $\nu_2 = 8$, and $\nu_3 = 16$. The simplex on the left hand-side in Figure 1 presents the vectors of true conditional class probabilities of the observations in the test set. On the right hand-side you see the estimated conditional probability vectors of the same observations of the Bayes-QDA classifier.*

Solid borders in Figure 1 separate regions for observations that get assigned *to the same class. Dashed borders within these regions separate observations that differ in the class with second highest (estimated) class probability.*

The closer the marker of an object is to the class corner the higher its (estimated) probability in that class. The layout of markers in the two simplexes look pretty much the same, with the exception that themarkers for the Bayes-QDA classifier appear to be shifted to the right so that it seems to assume G_2 to be closer in probability to G_1 than to G_3, though this is not the case as can be seen by the symmetry in the simplex of the True-Bayes classifier. Indeed, a comparison of the correctness rates in the different regions reveals that the Bayes-QDA classifier performs worse in the assignment to any class.

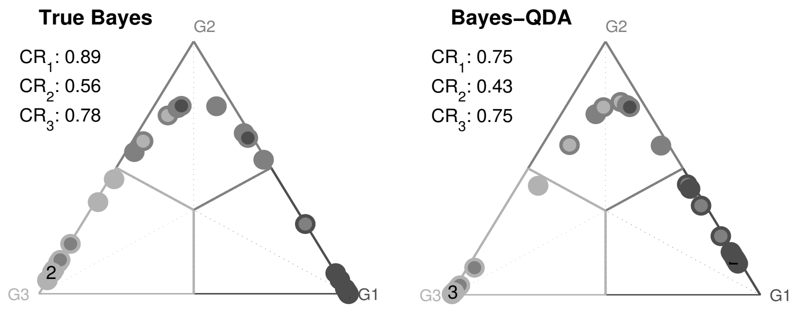

Figure 1: Simplexes representing the behaviour of the True-Bayes and the Bayes-QDA classifiers on the test set. $\mathbf{CR}_1$–$\mathbf{CR}_3$ denote the correctness rates for the assignments to the corresponding classes G_1–G_3. The true class defines the inner color of markers, the assigned class the color of the outer circle.

For obtaining comparable partitions from arbitrary argmax rules it is not appropriate to simply display their membership values. One obvious reason is that they neither have to be non-negative nor add up to one. Moreover, any ad-hoc standardization of membership values into $\mathbf{M}^s$ might lead to patterns more influenced by the standardization procedure than by the rule's classification behaviour. Even the membership values of argmax rules based on learnt conditional class probabilities are not appropriate for comparisons, because they give information about the rule's behaviour from its own perspective only, whereas for comparisons, we would prefer a more objective view. Thus, from now on we do no longer distinguish between membership values of probabilistic classifiers and non-probabilistic classifiers, assuming the latter to be 'appropriately' standardized into the space $\mathbf{M}^s$. Appropriately means here that for any $x \in \mathbf{X}$ at least the order of the membership values of classes stays the same whether it is based on the original membership values or on the standardized ones. In this respect valid transformations are, for example, subtraction of the absolute zero-point (if this is finite) or an estimator thereof and division by the sum of membership values within the membership vector of an object. For classifiers with membership values that are not on a metric scale, we recommend to use the ranks of $\mathbf{m}(x,1),...,\mathbf{m}(x,G)$ for each $x \in \mathbf{X}$ divided by the number of classes G, as the relative distance between membership values can not be interpreted.

Our aim is to scale membership vectors $\vec{\mathbf{m}} \in \mathbf{M}^s \to \vec{\mathbf{m}}^s \in \mathbf{M}^s$, such that scaled membership vectors of some test set observations reflect the rule's characteristic of classification and give a realistic impression of the rule's performance on the test set.

3 Concepts under consideration

The goodness concepts of accuracy, precision, and ability to separate we are focussing on refer to the concepts of inaccuracy, imprecision, and resemblance of Hand (1997, p. 99). There are two main differences. First, we use counterparts, i.e. high values and not to low values are desirable. Second, Hand (1997) restricts his attention to probabilistic classifiers where membership values are equal to estimated conditional class probabilities. Our aim is to generalize these concepts to be applicable for a wider range of techniques, by using scaled membership values instead of estimated conditional class probabilities.

Accuracy tells us something about the effectiveness of the rule in the assignment of objects into classes. Measures of accuracy in the literature typically typically assess whether true classes are the same as assigned classes. We call these measures **correctness** measures to distinguish them from Hands measures of accuracy that quantify the difference between a-posteriori class probabilities of an observed object (1-0) and its estimated conditional class probabilities of a probabilistic classifier. Given an accurate rule in the sense of Hand allows to interpret the size of the estimated probability in the assigned class as a reliable measure of the certainty we can have about that assignment. Of course, we can measure the accuracy of any non-probabilistic classifier simply by 'estimating' the conditional class probability of an observation in the assigned class as one that is by using a correctness measure. But this rule-dependent estimation is very crude. There is typically more information about the assessed membership of objects in classes available in non-probabilistic classifiers. With our scaled membership values we intend to propose a more sophisticated way to compare accuracy of probabilistic and non-probabilistic classifiers.

Ability to separate tells us, how well classes are distinguished, given the transformations the classifier uses to assess the membership of an object in the classes. Measures of the ability to separate are based on the diversity of the *vectors* of 'true' conditional class probabilities *among objects assigned to different classes* given their membership values. This is slightly different from Hand's concept of non-resemblance, where the diversity of the 'true' conditional class probabilities *among classes* within probability vectors given membership values is of interest. Measures of the ability to separate compare vectors, whereas measures of resemblance compare values within vectors. Both concepts are highly related though, because class probability vectors of objects that get assigned to different classes differ a lot, if the class probabilities within the vectors differ a lot, and vice-versa. The reason is that the probabilities within vectors are non-negative and add up to one. We want to obtain visualizations of scaled membership values that give a realistic impression of the rules's ability to separate on the test set.

Precision tells us, how good the classifier estimates 'true' conditional class probabilities. Measures compare the rule's membership values with 'true' conditional class probabilities. To measure precision, we obviously need knowledge about 'true' conditional class probabilities. Since our scaled membership values should reflect as precisely as possible the information in the original membership values and the rule's performance on the test set, the empirical precision will be enforced by our scaling procedure. Thus scaled membership values can not be used to assess precision, but mirror information of the rule's performance on the test set.

Note that, Hand (1997) also defines **separability** which is substantially different from the concepts above as it is a characteristic of classification *problems* and not of rules. It tells us, how different the 'true' conditional class probabilities of objects are given the observed features. Separability determines an upper bound for any rule's ability to separate. A measure for separability has to be based on the diversity of 'true' conditional class probabilities given the values of the predictors of objects.

4 Scaling

We scale membership values of the observations of a test set such that they resemble precise estimators of conditional class probabilities. So in a first step, we assess the precision of a classifier interpreting its membership values as estimators of conditional class probabilities. And then we scale these membership values towards a better precision.

For the quantification of precision as such we would need knowledge on 'true' conditional class probabilities. This is only available in simulation studies. In all other cases, we need additional consideration: A basic strategy to judge precision is to compare summary statistics of the estimated conditional class probabilities with corresponding summary statistics computed from a test set $\mathbf{T} := \{(\mathbf{x_1}, \mathbf{c_1}), ..., (\mathbf{x_N}, \mathbf{c_N})\}$. In our case, for a fair comparison, we have to use summary statistics that make the least model assumptions possible, or in other words, that is based on some minimum commonly accepted prior knowledge μ. Only assuming a finite number of classes, we know that for any fixed observation the true conditional class probabilities $\mathbf{p}(1|x,\tau), ..., \mathbf{p}(G|x,\tau)$ are parameters of a multinomial distribution. It is widely accepted to use frequencies as point estimators of such parameters, when no expert knowledge allows for a more sophisticated modeling. Thus, observed frequencies on the test set are the summary statistics of choice for our purpose.

If there is only a finite number of possible values $\{a_1, ..., a_K\}$ of x, $(x, c_x) \in \mathbf{T}$, and there are enough (up to the considerations of the analyst for a given task) observations for each of them - say $N_{\mathbf{T},\mathbf{k}}, k = 1, ..., K$ - in the test set, we can estimate:

$$\mathbf{p}\left(c \mid X = a_k, \mathbf{T}, \mu\right) = \frac{\sum_{(x,c_x) \in \mathbf{T}: x = a_k} \mathbb{I}_c(c_x)}{N_{\mathbf{T},\mathbf{k}}},$$

where $\mathbb{I}_c$ is the indicator function corresponding to class c, $c = 1, ..., G$. In most cases, however, we will need to partition objects into a small number of regions $\mathbf{R}(1), ..., \mathbf{R}(K)$, $\bigcup_{k=1,...,K} \mathbf{R}(k) = \mathbf{X} \times \{1, ..., G\}$, to obtain a reasonable estimator for the conditional class probabilities $\mathbf{p}(1|(X,C) \in \mathbf{R}(k), \tau)$,...,$\mathbf{p}(G|(X,C) \in \mathbf{R}(k), \tau)$ in that region:

$$\begin{aligned}
\mathbf{p}\left(c \mid \mathbf{R}(k), \mathbf{T}, \mu\right) &:= \mathbf{p}\left(c \mid (X,C) \in \mathbf{R}(k), \mathbf{T}, \mu\right) \\
&= \frac{\sum_{(x,c_x) \in \mathbf{R_T}(k)} \mathbb{I}_c(c_x)}{N_{\mathbf{T},\mathbf{k}}}, c = 1, ..., G,
\end{aligned}$$

where $\mathbf{R_T}(k) := \mathbf{R}(k) \cap \mathbf{T}$, $k = 1, ..., K$. We call these estimators the **region-conditional class frequencies** and denote their vector by $\vec{\mathbf{p}}_{\mathbf{T},\mu}(\mathbf{R}(k))$, $k = 1, ..., K$.

We want to define regions in the same way for all argmax rules. Also, we want objects with the same membership values to lie in the same region. Thus, for the definition of regions we use information in the membership vectors $\vec{\mathbf{m}}$, reduced to what is comparable among all argmax rules, namely the order of the classes sorted by descending membership values. We define a group to consist of all observations with the same order $\vec{\mathbf{o}} = (\mathbf{o}_1, ..., \mathbf{o}_d) : \mathbf{M}^s \to \mathbf{O}_d \subset \{1, ..., G\}^d$, of depth d, $d \leq G-1$ due to the first $d \in \mathbb{N}$ highest membership values:

$$\begin{aligned}
\mathbf{o}_1(\vec{\mathbf{m}}(x)) &= \arg\max \{\mathbf{m}(x,c), c = 1, ..., G\} \\
&= \mathbf{cl}(x, \vec{\mathbf{m}}) \\
\mathbf{o}_2(\vec{\mathbf{m}}(x)) &= \arg\max \{\mathbf{m}(x,c), c \in \{1, ..., G\} \setminus \mathbf{o}_1\} \\
&\quad ... \\
\mathbf{o}_d(\vec{\mathbf{m}}(x)) &= \arg\max \{\mathbf{m}(x,c), c \in \{1, ..., G\} \setminus \{\mathbf{o}_1, ..., \mathbf{o}_{d-1}\}\},
\end{aligned}$$

where we omitted for ease of notation the membership vector in the notation of the orders on the right hand sides of the equations. Let $\mathbf{R}(\vec{o})$ denote the region that corresponds to objects with a certain ordering $\vec{o}$ of the membership values. With a few ties in the order, we propose to randomize tied observations to the corresponding regions. With many ties (may be due to the classifiers algorithm) it is better to join regions, if possible.

For G classes and maximum depth $G-1$ we get $G!$ regions. Depending on the size of the test set, this may be still too many regions for the number of observations in these regions to be considered 'sufficiently' high. Thus, we introduce the depth as a parameter to customize for required sub-sample sizes. Within regions we then scale membership values such that their mean is approximately equal to the estimated conditional class probabilities that define the region only, that is

$$\frac{1}{N_{\mathbf{T},\vec{\mathbf{o}}}} \sum_{(x,c_x)\in\mathbf{R_T}(\vec{o})} m^s(x,c) \approx \mathbf{p}\left(c \mid \mathbf{R}(\vec{o}), \mathbf{T}, \mu\right),$$

with $c = \{\mathbf{o}_1, ..., \mathbf{o}_d\}$ and $N_{\mathbf{T},\vec{\mathbf{o}}}$ denoting the number of objects in $\mathbf{R_T}(\vec{o})$.

Scaling at depth one

The most coarse and yet still useful graining we get for depth one, where only the highest value $\mathbf{m}_{\mathbf{cl}(x,\vec{\mathbf{m}})}(x) := \mathbf{m}(x, \mathbf{cl}(x, \vec{\mathbf{m}})), (x, c_x) \in \mathbf{T}$, responsible for the class-assignment defines the region. We call $\mathbf{m}_{\mathbf{cl}(x,\vec{\mathbf{m}})}(x)$ the **assignment value** of an object $(x, c_x) \in \mathbf{T}$.

We approximate the empirical distribution $\mathbf{F_{T,c}} : \mathbf{M}^s \to [0,1]$ of the assignment values $\mathbf{m}_c(x), (x, c_x) \in \mathbf{R_T}(c)$ within each region $\mathbf{R}(c)$ by a Beta distribution $\mathbf{B}(\alpha_c, \beta_c)$, $c = 1, ..., G$. We estimate suitable parameters α_c, β_c using the method of moments (c.p. Gelman et al. (1995), p. 481) for a Beta distribution. Thus in region $\mathbf{R}(c)$ with $N_{\mathbf{T,c}}$ objects in the test set, we get:

$$\begin{aligned}
\alpha_c + \beta_c &= \frac{\overline{\mathbf{m_{T,c}}}\left(1 - \overline{\mathbf{m_{T,c}}}\right)}{\frac{1}{N_{\mathbf{T,c}}-1} \sum_{x\in\mathbf{R_T}(c)} \left(\mathbf{m}_c(x) - \overline{\mathbf{m_{T,c}}}\right)^2}, \\
\alpha_c &= (\alpha_c + \beta_c)\overline{\mathbf{m_{T,c}}}, \\
\beta_c &= (\alpha_c + \beta_c)(1 - \overline{\mathbf{m_{T,c}}}),
\end{aligned}$$

with $\overline{\mathbf{m_{T,c}}} := \frac{1}{N_{\mathbf{T,c}}} \sum_{(x,c_x)\in\mathbf{R_T}(c)} \mathbf{m}_c(x)$, denoting the arithmetic mean of the observed assignment values in the cth region $\mathbf{R}(c), c = 1, ..., G$.

Thus, the empirical probability that the assignment value $\mathbf{m}(Y)$ of any random object $(Y, C) \in \mathbf{R}(c)$ is smaller or equal to any observed assignment value $m_c(x), (x, c_x) \in \mathbf{R_T}(c)$ in that region is approximated by:

$$\mathbf{P_{T,c}}(\mathbf{m}_c(Y) \leq \mathbf{m}_c(x)) \approx \mathbf{F}_{\alpha_c,\beta_c}\left(\mathbf{m}_c(x)\right), c = 1, ..., G.$$

We want this probability to be approximately valid for scaled membership values as well, that is

$$\mathbf{P}_{\mathbf{T},\mathbf{c}}(\mathbf{m}_c^s(Y) \leq \mathbf{m}_c^s(x)) \approx \mathbf{P}_{\mathbf{T},\mathbf{c}}(\mathbf{m}_c(Y) \leq \mathbf{m}_c(x))$$

for any $(x, c_x) \in \mathbf{R}_\mathbf{T}(c)$, $c = 1, ..., G$. For their empirical distribution $\mathbf{F}^s_{\mathbf{T},\mathbf{c}} : \mathbf{M}^s \to [0,1]$, though, we require that it gives rise to an estimated Beta distribution with different parameters that are corrected for the actual behaviour of the classifier on the test set:

$$\begin{aligned} N^s_{\mathbf{T},\mathbf{c}} := \alpha_c^s + \beta_c^s \quad &:= \quad \min\left\{N_{\mathbf{T},\mathbf{c}}, \alpha_c + \beta_c\right\}, && \text{(i)} \\ \alpha_c^s \quad &:= \quad N^s_{\mathbf{T},\mathbf{c}}\ \ \mathbf{p}(c \mid \mathbf{R}_c, \mathbf{T}, \mu). && \text{(ii)} \end{aligned}$$

The scaled membership value $\mathbf{m}_c^s(x)$ can be calculated numerically from the equation

$$\mathbf{F}_{\alpha_c^s, \beta_c^s}\left(\mathbf{m}_c^s(x)\right) \quad = \quad F_{\alpha, \beta_c}\left(\mathbf{m}_c(x)\right)$$

for all $(x, c_x) \in \mathbf{R}_\mathbf{T}(c)$ and each $c = 1, ..., G$.

For the other membership values in the vector $\vec{\mathbf{m}}(x)$, we keep their ratio:

$$\begin{aligned} \mathbf{m}^s(x,g) \quad &= \quad \frac{1 - \mathbf{m}_c^s(x)}{1 - \mathbf{m}_c(x)} \mathbf{m}(x,g), \quad g = 1, .., G, g \neq c \\ \Rightarrow \frac{\mathbf{m}^s(x,g)}{\mathbf{m}^s(x,i)} \quad &= \quad \frac{\mathbf{m}(x,g)}{\mathbf{m}(x,i)}, \quad g, i = 1, ..., G, \ g, i \neq c, \end{aligned}$$

for all $(x, c_x) \in \mathbf{R}_\mathbf{T}(c)$ and each $c = 1, ..., G$.

Justification of the scaling

We use the Beta distribution for the approximations of $\mathbf{F}_{\mathbf{T},\mathbf{c}}$ and $\mathbf{F}^s_{\mathbf{T},\mathbf{c}}$, because of its flexibility and the implicit interpretation of its parameters as it is the conjugate family of the Bernoulli distribution $\mathbf{Be}(p)$. Starting with an improper prior $\mathbf{B}(0,0)$, and after α successes in N Bernoulli trails, the posterior distribution of probability p for success is $\mathbf{B}(\alpha, N - \alpha)$. Thus, α/N is an estimator of success probability p and N can be interpreted as certainty about the estimator.

With that in mind, we can view the assignment to class c, given the rule assigns to this class, as a Bernoulli trial. Assignment values then reflect the rule's estimate of its probability to decide successfully.

In the test set, we see $N_{\mathbf{T},\mathbf{c}} \ * \ \mathbf{p}(c \mid \mathbf{R}_c, \mathbf{T}, \mu)$ examples for success in $N_{\mathbf{T},\mathbf{c}}$ trials of the rule's assignment to class c. Thus, the appropriate choice for a corrected success estimator is $\frac{\alpha_c^s}{N^s_{\mathbf{T},\mathbf{c}}}$, $c = 1, ..., G$, as given in equation (ii).

An appropriate choice for the parameters $N^s_{\mathbf{T},\mathbf{c}}, c = 1, ..., G$ is less obvious. Using $N_{\mathbf{T},\mathbf{c}}$ is not appropriate, because of its interpretation as certainty of the current knowledge about parameter p of the Bernoulli distribution. This would lead to some non-intuitive behaviour of our scaling as, e.g. for $N_{\mathbf{T},\mathbf{c}} \to \infty$, all scaled assignment values of objects in that region approach p. The parameter $N^s_{\mathbf{T},\mathbf{c}}$ should rather have an interpretation as a measure of the inverse dispersion of assignment values. Because of that, no huge scaling of assignment values near to true conditional class probabilities should take place. This is an argument favoring $N^s_{\mathbf{T},\mathbf{c}} = \alpha_c + \beta_c$, but only for probabilistic classifiers. For other assignment values $\alpha_c + \beta_c$ is dependent on the ad-hoc standardization of the corresponding membership vector into $\mathbf{M}^s$, and might be misleadingly high. The definition in (i) avoids unwarranted large certainty parameters $N^s_{\mathbf{T},\mathbf{c}}$ for each $c = 1, ..., G$.

Scaling at depths $d > 1$

If regions are defined for more than the assigned class, that is $d > 1$, we intend to use the dirichlet distribution for the approximation. As this is d-dimensional there exists no inverse of the distribution function. Therefore, our idea is to perform the scaling stepwise then, starting with the membership values in $\vec{o}_1$ and continuing with those in $\vec{o}_2, ..., \vec{o}_d$. On each level, we scale as described above, but with a different way to determine parameters. They will be defined analogously to the definition of parameters when sampling from a Dirichlet distribution using the Beta distribution as described in Gelman et al. (1995, p. 482).

5 Measures

We now define measures for the rating of the performance of argmax classification rules w.r.t. accuracy and ability to separate. These measures are based on Euclidean distances between scaled membership vectors of test set observations and vectors of the corners of the simplex $\vec{e}(c)$ (see Figure 1), with components $e_g = \mathbb{I}_c(g), g = 1, ..., G, c = 1, ..., G$. This is not only useful for the understanding of the measures as such but also for a visualization of the performance of classifiers for classification problems with up to four classes.

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 one is achieved if all vectors lie in the correct corners, and zero if they all lie in the centroid of the simplex. The measure of accuracy is thus:

$$\mathbf{Ac_T} \quad := \quad \frac{\frac{G-1}{G} - \frac{1}{N} \sum_{(x,c_x) \in \mathbf{T}} \| \vec{e}(c_x) - \vec{\mathbf{m}}^s(x) \|_2}{\frac{G-1}{G}},$$

where N is the number of examples in the test set $\mathbf{T}$.

Continuation of the example. *We can now compare the behaviour of the Bayes-QDA classifier with the behaviour of some neural network (NN) classifier that originally used membership values in $\mathbb{R}$ with respect to accuracy. One can see in Figure 2 that the NN-classifier is slightly better.*

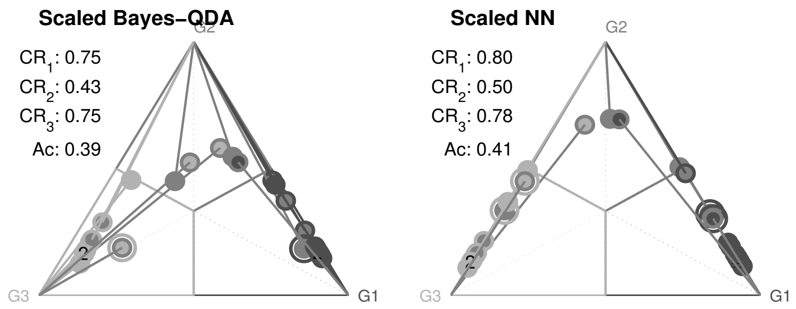

Figure 2: Simplexes representing the behaviour of the QDA-Bayes and the NN classifiers on the test set. $\mathbf{CR}_1$–$\mathbf{CR}_3$ denote the correctness rates for the assignments to the corresponding classes G_1–G_3, $\mathbf{Ac}$ the achieved value of accuracy. Lines are drawn to illustrate the Euclidean distances that determine the $\mathbf{AC}$ values.

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 $\vec{e}(\mathbf{cl}(x, \vec{\mathbf{m}}(x)))$. Note that in particular for poor classifiers the assignment of an observation based on its scaled membership values might be different from the assignment based on the original membership values, such that $\vec{e}(\mathbf{cl}(x, \vec{\mathbf{m}}^s(x))) \neq \vec{e}(\mathbf{cl}(x, \vec{\mathbf{m}}(x)))$, and that we really want to use the original assignment in our definition. We standardize the mean of this distances in the same way as above, such that our measure of the ability to separate is defined as:

$$\mathbf{AS_T} \quad := \quad \frac{\frac{G-1}{G} - \frac{1}{N}\sum_{(x,c_x)\in\mathbf{T}} \parallel \vec{e}(\mathbf{cl}(x, \vec{\mathbf{m}}(x))) - \vec{\mathbf{m}}^s(x) \parallel_2}{\frac{G-1}{G}}.$$

Continuation of the example. *Again, the NN classifier is superior to the Bayes-QDA classifier with respect to its ability to separate as you can see in Figure 3.*

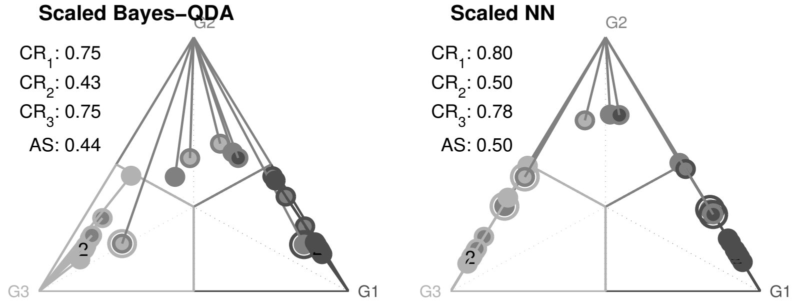

Figure 3: Simplexes representing the ability to separate of the QDA-Bayes and the NN classifiers on the test set. $\mathbf{CR}_1$–$\mathbf{CR}_3$ denote the correctness rates for the assignments to the corresponding classes G_1–G_3, **AS** the achieved value of ability to separate. Lines are drawn to illustrate the Euclidean distances that determine the **AS** values.

6 Conclusions

In this paper, we introduced standardized partitions as a mean to compare so-called argmax classification rules. These partitions build an excellent basis for the comparison of rules w.r.t. their quantitative assessment of the membership of objects in classes, in particular by means of accuracy and ability to separate.

References

[1] *Anderson, T.W. (1958). An Introduction to Multivariate Statistical Analysis. John Wiley& Sons, Canada.*

[2] *Gelman, A., Carlin, J.B., Stern H.S., and Rubin, D.B. (1995). Bayesian Data Analysis. Chapman & Hall, London.*

[3] *Hand, D.J. (1997). Construction and Assessment of Classification Rules. John Wiley & Sons, Baffins Lane, England.*