Combining Mental Fit and Data Fit for Classification Rule Selection

C.Weihs and U. M.Sondhauss
Fachbereich Statistik and SFB 475
Universitaet Dortmund, D-44221 Dortmund, Germany

May 18, 2001

Abstract

Mental fit of classification rules is lately introduced to judge the adequacy of such rules for human understanding. This paper first discusses the various criteria introduced in relation to mental fit in the literature. Based on this, the paper derives a general criterion for the interpretability of partitions generated by classification rules. We introduce interpretability as a combination of mental fit and data fit, or more specifically, as a combination of comprehensibility and reliability of a partition. We introduce so-called prototypes to improve comprehensibility, and the so-called reliability of such prototypes as a measure of data fit.

Key words: mental fit, data fit, interpretability, classification rules, partitions, scaling, prototypes, reliability

1 Introduction

In the days of data mining, the number of competing classification techniques is growing steadily. Thus, it is a worthy goal to rate the classification rules from a wide range of techniques. An ideal measure for the selection of the best classification rule from candidates should combine two aspects, namely mental fit and data fit.

There are various possibilities to define data fit, which is often also called consistency or accuracy (cp. e.g. Hand (1997), p. 100). In the following we will think about data fit as predictive power measured by misclassification rate.

In contrast, mental fit has much more recently come into play. Before we will derive a criterion for classification rule selection which combines mental fit and data fit we will first discuss the different concepts introduced in the literature to characterize mental fit. The main 'definitions' are:

- Classifiers should constitute explicit concepts meaningful to humans and evaluable directly in mind (Feng and Michie (1994)).

[1] This work has been supported by the Deutsche Forschungsgemeinschaft, Sonderforschungsbereich 475. The paper was presented at the 25th Annual Conference of the Gesellschaft fuer Klassifikation (GfKl) March 14-16, 2002, Universitaet Muenchen, Germany

- The most important elements of mental fit are coverage, simplicity, and explainability (v. den Eijkel (1999)).

In the literature, comparisons of the mental fit of classification rules are often either too general or too method-specific. Surely, what is ideally needed is a unique accepted general formalization of mental fit which is always relevant and measurable. Typically, however, ratings of mental fit are inevitably method, project and customer dependent. For some techniques it is nevertheless argued that they outperform all others with respect to certain aspects of mental fit as interpretability or comprehensibility in general, e.g. for production rules (Weiss and Kulikowski (1991)), decision trees (Michie et al. (1994)), and fuzzy systems (Bodenhofer and Bauer (1999)). Often however, such statements are related to specific concepts for mental fit. Examples for such 'local measures' of mental fit used with certain methods are 'low tree complexity' for decision trees (Breiman et al. (1984)), and 'small number of involved original features in discriminant coordinates' for discriminant analysis (Weihs (1992, 1993)).

In this paper we will first discuss various aspects of mental fit (section 2), and will then propose a general strategy to formalize mental fit locally w.r.t. certain representations. In that spirit, we introduce a new simple very general rule selection method for partitions, a typical way of interpreting classification rules. To this purpose we will standardize the representation of partitions and define performance criteria that can be calculated for a wide variety of techniques.

2 Formalizations of Mental Fit

2.1 Conceptual Partitioning

"The main advantage that rule induction offers for decision-making problems is what is sometimes called a mental fit to the problem" (v. den Eijkel (1999)). In contrast, it can be argued that many statistical and neural net techniques partition the feature space into one region per class by using some kind of discriminant function difficult to understand in form and content. Moreover, often the sole goal is to optimize the (predictive) accuracy of the classification.

Rule induction classification techniques like CART (Breiman et al. (1984)), however, partition the feature space into multiple regions per class, where each region is associated with one and only one class *and has a simple shape.* This way, the partitioning of feature space can be regarded as generating 'explicit knowledge' describing the data. Rule induction is thus concerned with finding explicit reasons for partitions in that the rules are derived as simple conditions of original features. This is often called conceptual partitioning or concept description of rules.

Here is an example for conceptual partitioning. In business cycle phase prediction (see e.g. Weihs et al. (1999)), we got the following partitions characterizing classification rules. By linear discriminant analysis one region per class was produced. In Figure 1, in addition, the course of projected observations into the 2D-discriminant space of a full business cycle not used for learning is indicated by connected arrows. *Solid* arrows represent correct classifications, *broken* ones misclassified examples.

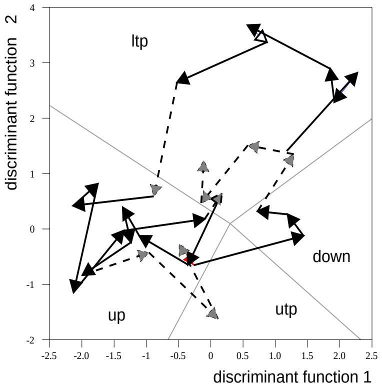

Figure 1: Linear Discriminant Analysis - one region per class

The problem with this partition is twofold. The axes are not interpretable, since they are arbitrary linear combinations of the original features, and the partition is even not easy to be described in discriminant coordinates since the borders of the regions are not parallel to the axes.

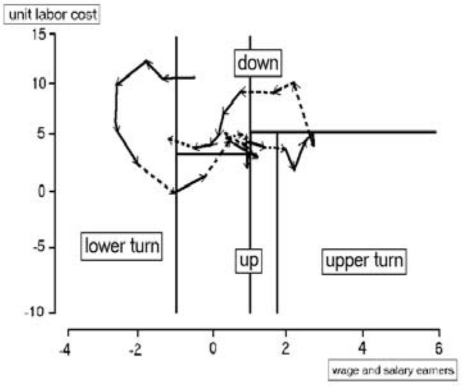

Figure 2: CART - multiple regions per class

Both problems do not exist for CART. In Figure 2 a corresponding partition is shown for two factors only. CART delivers multiple regions per class. Here, two regions are found for the phases UPswing and DOWNswing of the business cycle. Since the region borders are parallel to the axes representing original features, interpretation is easy. The illustration obviously indicates a classification problem near 0% increase of WAGE AND SALARY EARNERS and a simultaneous 4% increase of UNIT LABOR COSTS.

2.2 Simplicity

A second advantage of rule induction techniques comes from the partitioning process as well. The simplicity of rules, i.e. the low dimensionality of conditions, often intrinsically leads to dimension reduction by feature selection (v. den Eijkel (1999)). E.g., in decision trees with nearly the same data fit, a system of classification rules is chosen with low tree complexity. Here, pruning leads to a low number of features together with a low number of regions (Breiman et al. (1984)).

Corresponding simplicity criteria were developed for other methods too. For linear discriminant analysis, e.g., the discriminant functions may be simplified by minimizing the number of original features involved, retaining the discriminative power (Weihs (1992, 1993)). For neural networks a similar criterion is proposed. Network pruning aims at removing redundant links and nodes without increasing the error rate. A smaller number of nodes and links left in the network after pruning provides rules that are more concise and simple in describing the classification function (Lu et al. (1995)). A forefather of all these methods is the idea of finding classification rules with minimum description length (Rissanen (1978)). Here, the idea is obvious, the shorter the rule description, the easier to understand the rule, i.e. the bigger mental fit.

In Figure 3 you see an example for simplicity induction. The efficacy of a drug for impaired brain functions in geriatric patients is measured by a class score EFF at 4 levels determined by medical doctors, as well as by five efficacy measures (DBCRS, DIADLP, DSCAG, DSKTR, DZVTT). Linear discriminant analysis shows that one dimension can nearly perfectly describe the variation in the data (cp. Figure 3). This first discriminant function is of the following form: CD1 = 0.26*DBCRS - 0.10*DIADLP + 0.07*DSCAG + 0.02*DSKTR + 0.01*DZVTT.

Simplification of this function can be obtained by regressing the corresponding discriminant scores on the original predictors. Greedy forward selection first chooses DBCRS with $R^2 = 0.71$, then DSCAG leading to $R^2 = 0.85$, and at last DSKTR so that $R^2 = 0.95$. Linear discriminant analysis based on these predictors leads to a nearly equivalent simplified version of the first discriminant function, where only three predictors are involved.

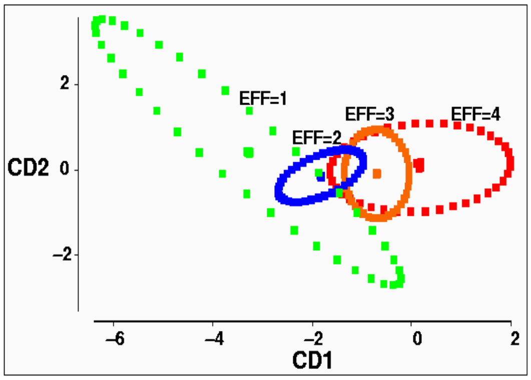

Figure 3: Linear Discriminant Analysis - confidence regions for classes

2.3 Coverage

Another term related to mental fit is coverage, often also called generality. Coverage of a partition is defined to be the higher the more powerful the regions in a partition are, i.e. the more instances (observations, examples) the regions contain. A related concept is the height of (data) density (or data non-sparseness) in the regions. Obviously, coverage influences simplicity.

2.4 Comprehensibility (Explainability)

This is the first really subjective measure in that comprehensibility is (obviously) user dependent, since the rule should be understandable to the user. Possible comprehensibility criteria for a rule formulated in certain coordinates are:

- Original coordinates are preferred to latent ones,
- small number of coordinates to many coordinates,
- explicit rules to trees,
- trees to functions,
- conditions formulated as qualifications ('high', 'low') over thresholds '$(<, >)$',
- conjunctions over conjunctions plus disjunctions.
- non-disjoint rules over disjoint rules, or vice versa, dependent on the application.

Other criteria might be important in specific applications.

2.5 Interestingness

A side step and at the same time an extension to mental fit is the term interestingness. There are two kinds of formalizations: objective and subjective ones.

2.5.1 Objective interestingness measures

For an association rule $Y \to X$, $p(X|Y)$ is called confidence of the rule. Note that for classification $X = C =$ 'one of the classes' in association rules. The larger the confidence $p(C|Y)$, the more interesting the rule. Wang et al. (1998) propose a 'onesided' variant of the J-measure of Smyth and Goodman (1992) for interestingness, namely:

$$\begin{aligned} \mathbf{J_1}(C;Y) &= \mathbf{p}(Y) \times \left[\mathbf{p}(C|Y)\mathbf{log_2}\left(\frac{\mathbf{p}(C|Y)}{\mathbf{p}(C)}\right)\right] \\ &= \text{generality of rule} \times [\text{discrimination power of rule}]. \end{aligned}$$

A second interestingness measure is

$$\mathbf{J_2}(C;Y) = \mathbf{p}(Y)\,(\mathbf{p}(C|Y)) - \mathbf{miniconf})$$

where **miniconf** is a specified minimum confidence. Obviously, $\mathbf{J_2}(C;Y)$ is the 'surplus' of the association in the rule $Y \to C$ relative to the association level specified to be the minimum confidence.

The larger the value of $\mathbf{J_1}$, $\mathbf{J_2}$, the more interesting the rule. $\mathbf{J_1} < 0$ corresponds to negative association. $\mathbf{J_2} < 0$ corresponds to an association level below minimum confidence.

2.5.2 Subjective interestingness measures

Objective measures of interestingness, although useful in many respects, usually do not capture all the complexities of the rule discovery process. Therefore, subjective measures of interestingness are needed to define interestingness of a rule (Silberschatz and Tuzhilin (1996)). Subjective measures do not depend only on the structure of a rule and on the data used in the discovery process, but also on the user's beliefs who examines the rule. Two major reasons why a rule is interesting from the subjective (user-oriented) point of view may be:

- unexpectedness - a rule is interesting when it is 'surprising' to the user,
- actionability - a rule is interesting if the user can act on it to his advantage.

Since actionability appears to be hard to be formalized, we will concentrate on unexpectedness and its relation to beliefs following Silberschatz and Tuzhilin (1996). They distinguish two kinds of beliefs: Hard beliefs are constraints that should not be changed with new evidence. Soft beliefs are beliefs the user is willing to change with new evidence.

One possibility of belief adjustment is using the Bayesian approach. In this case, the 'degree of a belief α' is defined as $\mathbf{P}(\alpha|\xi)$, given some previous evidence ξ, supporting that belief. With new evidence E, the update of the degree of belief in α, $\mathbf{P}(\alpha|E;\xi)$, is then obtained by using Bayes rule.

Interestingness of rules is then defined as follows. A rule is interesting relative to some belief system if it 'affects' this system, and the more it 'affects' it, the more interesting the rule is. If a rule contradicts the set of hard beliefs of the user then this rule is always interesting to the user. In case of soft beliefs, interestingness of rule $\mathbf{r}$ relative to a (soft) belief system $\mathbf{B}$ and previous evidence ξ is defined by

$$\mathbf{I}(\mathbf{r};\mathbf{B};\xi) = \sum_{\alpha_i \in \mathbf{B}} w_i \left|\mathbf{P}(\alpha_i|r;\xi) - \mathbf{P}(\alpha_i|\xi)\right|,$$

where w_i is the importance weight for the soft belief α_i in the belief system $\mathbf{B}$, and $\sum_{\alpha_i \in \mathbf{B}} w_i = 1$. This definition of interestingness measures of how much degrees of believe are changed as a result of a new rule $\mathbf{r}$.

Interestingness appears to be an extension to mental fit since mental fit is obviously a prerequisite to subjective interestingness, especially to the decision whether a hard belief is contradicted.

2.6 List of important aspects of mental fit

In summary, objective measures of mental fit are:

- Completeness, not yet mentioned, but obvious: the instance space should be described completely;
- Coverage: rules should be as powerful as possible;
- Simplicity: rules should concisely describe reality, the less variables, the better.

Subjective measures are:

- Comprehensibility: rules should be understandable to the user;
- Interestingness: the more unexpected the rule is in relation to a belief system, the better.

3 A general approach to rule selection

In what follows we would like to concentrate on the combination of mental fit and data fit and define what we call interpretability as

Mental Fit plus Data Fit,

More special, we will discuss interpretability of classification rules as

Comprehensibility plus Reliability,

since unreliable statements are not interpretable, even if they would be very much comprehensible.

Regarding comprehensibility, our aim is to develop what we call a local measure w.r.t. certain representations of the result of classification procedures. For those users who decide that such a representation is the preferred type, the goal is to select a rule that has biggest mental fit and data fit w.r.t. this representation.

Regarding reliability, we distinguish two cases. If we interpret the rule as such, without deducing additional entities from it, it simply gains its reliability from the correctness rate. This is the case in so-called "conventional" partitions that we treat in subsection 3.1. When rules with acceptable data fit, are too complex to be comprehensible, though, we have to do further transformations to get comprehensible information from them. This is the case in the **standardized** partitions that we introduce in subsection 3.2 where we use rules to derive prototypes for interpretation. Naturally, this process to deduce also has to be reliable.

We see mainly two typical ways rules are interpreted: the use of partitions to describe classes and to deduce a measure of importance of predictors to detect main influences. Here we want to formalize mental fit and data fit for partitions.

3.1 Interpretability of conventional partitions

Our minimum requirements for interpretability of classification rules are:

- Rules should be formulated in 2 or 3 maximum 7 dimensions, more dimensions are not cognitively manageable;

- dimensions should relate to interpretable features;
- rules should have acceptable prediction accuracy.

Ranking of classification rules according to the above definition of interpretability should be done as follows:

- the lower the dimension of the rule, the better;
- in case of equal dimension, choose the rule with the lowest error rate.

3.2 Standardized partitions

What should one do, however, if these requirements cannot be fulfilled? In what follows we will standardize partitions so that we can apply the procedure also in case of such a failure.

There is only one condition a classification method is subjected to so that any of it's rules can be used to generate a standardized partition. The final decision $\mathbf{cl} : \mathbf{X} \rightarrow 1, ..., G$ has to be an argmax rule based on (rule-specific) transformations of observations:

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

For all classifiers that fulfill this condition, including e.g. all probabilistic classifiers, support vector machines, and neural networks, we can formulate partitions in a coordinate system, where axes have a common interpretation. Let

$$m(x,c) \quad := \quad \text{strength of membership of } x \text{ in } c.$$

We denote the space of corresponding membership vectors $\vec{m}(x) := (m(x,1)$ $, ..., m(x,G))$ by $\mathcal{M} \subset \mathbb{R}^G$. For probabilistic classifiers, membership values are monotone transformations of estimated class probabilities. In that case, we use

$$m(x,c) \quad := \quad \hat{p}(c|x), x \in \mathbf{X}, c = 1, ..., G.$$

Thus, these membership values all lie in the interval $[0,1]$ and sum up to one. We denote this (future standardized) space of membership vectors by $\mathcal{M}^s \subset [0,1]^G$ which will be our space for standardized partitions in future. For $G = 3$ or $G = 4$ membership vectors in $\mathcal{M}^s$ can be visualized in a barycentric coordinate system called **simplex**, see e.g. Figure 4.

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 4 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 4 separate regions for observations that get assigned to the same class. We call these regions **assignment areas** *of the corresponding classes. Dashed borders within these regions separate observations that differ* *in the class with second highest (estimated) class probability that we call* **preference areas***.*

The closer the marker of an observation 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, only that the Bayes-QDA classifier seem 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. A comparison of the correctness rates in the different regions, however, reveals that the Bayes-QDA classifier performs worse in the assignment to any class.

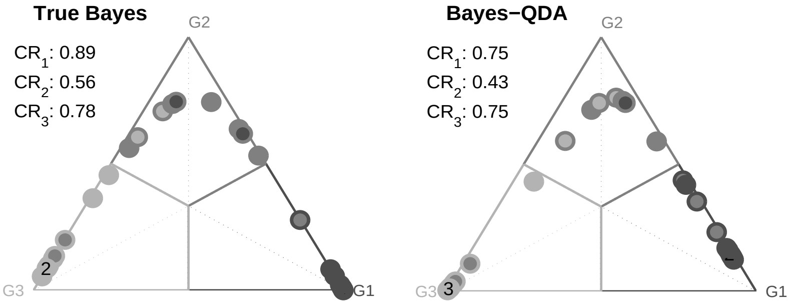

Figure 4: Simplexes representing the behaviour of the True-Bayes and the Bayes-QDA classifiers on the test set. $\mathbf{CR}_1$–$\mathbf{CR}_3$ denote the correctness rates for the assignments to the corresponding classes G_1–G_3. Solid borders separate assignment areas of classes, dashed borders preference areas within classes. The true class defines the inner color of markers, the assigned class the color of the outer circle.

In contrast to probabilistic classifiers, membership values generated by support vector machines or neural networks can be any real number. In order to compare membership vectors of different methods, we scale them into the space $\mathcal{M}^s$. We base our scaling of membership vectors on the comparison of average **confidence** with actual **competence** on some test set **T** that are defined as follows:

- A probabilistic classifier's **confidence** in a single decision $\mathbf{CF}(x)$ is reflected by its **assignment value** $m_{\mathbf{cl(x)}}(x)$:

$$\begin{aligned} \mathbf{CF}(x) &= \max_{c=1,\dots,G} m(x,c) = m(x,\mathbf{cl}(x)) \\ =: m_{\mathbf{cl}}(x) &= \hat{p}(\mathbf{cl}(x)|x),\ x \in \mathbf{X}. \end{aligned}$$

- Any classifier's **competence** to assign observations to class c is reflected by the probability for a correct assignment to this class for any random $X \in \mathbf{X}$:

$$\mathbf{P}_{C,X}(C\!=\!c \mid \mathbf{cl}(X)\!=\!c),\ c = 1, ..., G.$$

The details of the process of scaling is published in Sondhauss and Weihs (2001) and is based on approximations of the empirical distribution of assignment values within assignment areas with the beta-distribution.

The thus scaled membership vectors have the following properties:

- Scaled membership values are directly comparable in size.
- The average confidence of observations equals the actual competence on the test set.
- The empirical distribution of scaled membership vectors reflects the distribution of the original membership vectors corrected for the information in the test set.
- Scaled membership vectors of observations reflect as much as possible the position of the original membership vectors among each other within areas.

We then support interpretation by prototyping of reliable rules, where **reliability** relates to high ability to separate in the standardized partition space, and a **prototype** is the most typical instance (observation, example) in terms of scaled membership vectors.

3.2.1 Prototype

An observation where the rule has a justified high confidence in its decision, as well as no clear preference for the membership in any of the other classes, obviously has properties - from the perspective of the rule - that are quite specific for the assigned class. This is our motivation to define a prototype to be the correctly assigned observation the scaled membership vector of which is nearest to the class corner - denoted as $\vec{e}(c)$ - using euclidean distance:

$$x^*_{c,\mathbf{T}} \quad := \quad \arg \min_{x \in \mathbf{x}_{c,\mathbf{T}}} \left\| \vec{e}(c) - \vec{m^s}(x) \right\|, \tag{i}$$

where $\mathbf{x}_{c,\mathbf{T}}$ consists of all observations in the test set $\mathbf{T}$ in the assignment area of $c, c = 1, ..., G$. The prototype $\mathbf{x}_{c,\mathbf{T}}$ is also the observation the unscaled membership vector of which is nearest to the corner.

3.2.2 Reliability

As stated above, for the reliability of deduced entities from classifiers, the process of their derivation has to be checked. The selection of a prototype is based on the measure of typicalness (i) in terms of euclidean distance of scaled membership vectors to class corners. This means here, we need a definition of the reliability of the interpretation of the prototypes in terms of the reliability of the rule's induced measure of typicalness.

If scaled membership vectors reflect the important features of the observations that help to discriminate classes, we say the selected prototype is **reliable for interpretation**. We check reliability by the typicalness of the centers of scaled membership vectors within assignment areas. According to our scaling process these centers reflect the actual competence of the classifier to assign correctly. If these centers are not typical for the assignment areas then obviously (scaled) membership vectors do not reflect the important features for the discrimination of observations in the assignment areas.

For the combination of the typicalness of the centers in all areas, we use their mean, weighted by the size of the groups. We additionally standardize the measure, such that:

- A value of one corresponds to the highest possible typicalness, where all centers lie in the assigned corners. This happens, when there is no misclassification on the test set.
- A value of zero occurs, when all vectors are equal and lie in the barycenter of the simplex.
- Negative values indicate that the rule could be improved by an interchange of the assignment of two assignment areas.

The definition is as follows:

$$\mathbf{AS_T} \ := \ \frac{\sqrt{\frac{G-1}{G}} - \frac{1}{N_{\mathbf{T}}} \sum_{x \in \mathbf{T}} \| \vec{e}(\mathbf{cl}(x)) - \vec{\mathbf{p}}_{\mathbf{T}}(\vec{n} \mid \mathbf{cl}(x) = c) \|}{\sqrt{\frac{G-1}{G}}}, \qquad \text{(ii)}$$

where $\vec{n}$ denotes the vector $(1, 2, ..., G)$ of all classes. As the center of the assignment values reflects the actual competence of the rule on the test set, we can conclude that the rule does a good job in discriminating the classes on basis of its membership vectors, if $\mathbf{AS_T}$ is high and the centers lie near to the corresponding correct corner of the simplex. In other words, $\mathbf{AS_T}$ is also a measure for another known goodness aspect of classifiers, we call it **ability to separate**, which is the antonym of what Hand (1997) terms **resemblance**. The ability to separate is a characteristic of the classifier which should not be mistaken for **separability**, which is a characteristic of the problem that tells us how different the "true" probabilities of belonging to each class are.

Continuation of the Example *Using the same data as before, we demonstrate the scaling process for the Bayes-QDA classifier. In the simplex on the left hand-side of Figure 5 you see the original, and on the right hand-side the scaled membership vectors. For each area in each simplex we present mean confidences.*

At first, observing that $\mathbf{CF}_1$–$\mathbf{CF}_3$ *are bigger than the actual competence of the Bayes-QDA classifier (*$\mathbf{CR}_1$–$\mathbf{CR}_3$ *in Figure 4) reveals that the classifier is over-confident in its assignment to each class. Consequently, scaled membership vectors correct that by being moved away from the class corners. By construction* $\mathbf{CF}^s_1$–$\mathbf{CF}^s_3$ *of the scaled membership vectors are almost equal to* $\mathbf{CR}_1$–$\mathbf{CR}_3$*. Only in class* G_3 *the approximations in the scaling process do not result in equal* $\mathbf{CF}^s_3$ *and* $\mathbf{CR}_3$ *up to the second decimal point.*

Two observations that are assigned to class G_2 *even get moved out of their region: one into the region of its true class, but another one into a wrong region.*

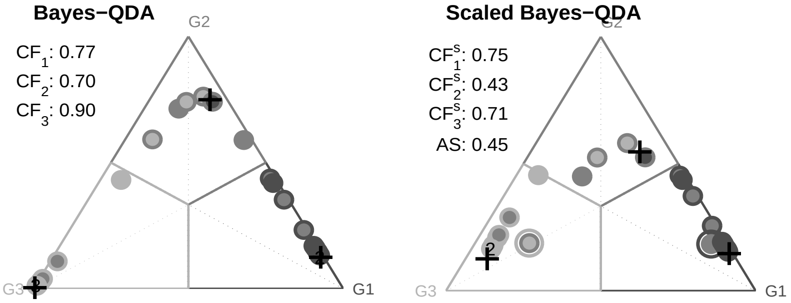

Figure 5: Simplexes illustrating the scaling process for the Bayes-QDA classifier on the test set. Mean confidences $\mathbf{CF}_1$–$\mathbf{CF}_3$ and $\mathbf{CF}_1^s$–$\mathbf{CF}_3^s$ in each area are given, based on the original and the scaled assignment values respectively. Prototypes are represented as '+' and **AS** is the actual ability to separate of the classifier on the test set. An additional second outer circle of a marker occurs two times: these observations get assigned to class G_2, but scaling moves them into the assignment areas of classes G_1 and G_3.

The prototypes for classes G_1, G_2, and G_3 are 0.1, 7.6, and 22.0, only differing in class G_2 from the prototypes of the True-Bayes classifier 0.1, 5.6, and 22.0. Note that our definition of prototypes has a connection to the mode of estimated conditional distributions of observations given the classes of probabilistic classifiers: for two classes with equal a-priori probabilities prototypes are the observations nearest to that modes.

Whether or not the size of 0.45 in the actual ability to separate $AS_{\mathbf{T}}$ of the Bayes-QDA classifier makes the interpretation reliable, is, as usual, dependent on the problem and thus, can only be evaluated in comparison with other classifiers. Clearly, in comparison with the actual ability to separate of 0.59 of the best classifier - the True-Bayes - we would prefer the prototypes of the True-Bayes.

3.2.3 Interpretability of Standardized Partitions

The use of prototypes for the description of a collection of observations is quite common, both in statistics and in machine learning. In statistics arithmetic mean, median, and mode are the most basic features that are reported when describing subpopulations of any kind. In machine learning, case-based reasoners claim that what people store for future problem solving are examples rather than rules, from which a high mental fit of examples may be induced (c.p. Aamodt and Plaza (1994)).

This justifies our approach in general, since we deduced from a rule with a conventional partition with acceptable data fit but without acceptable mental fit an entity that has a high mental fit, namely prototypes. Thus, we improved mental fit by deriving prototypes, and we defined a corresponding new measure of reliability that goes beyond the correctness rate in conventional partitions that takes care of the process of derivation.

For the rating of interpretability of these prototypes from different classifiers, their comprehensibility plays no role, as they do not differ in that respect. They only differ in their reliability which can be, as we argued above, measured in the same manner as the rule's ability to separate classes. That is, we put the interpretability of classifiers in standardized partitions down to another known goodness aspect, the method-related ability to separate.

4 Conclusion

This paper developed a general criterion for the interpretability of partitions generated by classification rules. Based on a discussion of mental fit criteria from the literature, we introduced interpretability as a combination of mental fit and data fit, or more specifically, as a combination of comprehensibility and reliability of a partition. For cases where the partition as such is not comprehensible we developed a standardized partition which is used to derive so-called prototypes to improve comprehensibility. The reliability of such prototypes is then used as a measure of data fit.

REFERENCES

AAMODT, A., and PLAZA, E (1994): Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches *AI Communications, 7(1):39-59.*

BODENHOFER, U., and BAUER, P. (1999): Towards an Axiomatic Treatment of "Interpretability", in Proceedings of the 6th International Conference on Soft Computing (IIZUKA2000), Iizuka, Japan, 334-339.

BREIMAN, L., FRIEDMAN, J. H., OLSHEN, R. A., and STONE, C. J. (1984): Classification and Regression Trees, Wadsworth, Belmont.

EIJKEL, VAN DEN G. (1999): Rule Induction, in Berthold, Hand (Eds.): Intelligent Data Analysis: An Introduction, Springer, Berlin, 195-216.

FENG, C., MICHIE, D. (1994): Machine Learning of Rules and Trees, in Michie, Spielgelhalter, and Taylor (Eds.): Machine Learning, Neural and Statistical Classification, Ellis Horwood, New York, 50-83.

HAND, D. J. (1997): Construction and Assessment of Classification Rules, Wiley, Chichester

LU, H., SETIONO, R., and LIU, H. (1995): NeuroRule: A Connectionist Approach to Data Mining, in Proceedings of the 1st VLDB Conference Zuerich, Switzerland, 1995

MICHIE, D., SPIEGELHALTER, D. J., and TAYLOR, C. C. (1994): Conclusions in Michie, Spielgelhalter, and Taylor (Eds.): Machine Learning, Neural and Statistical Classification, Ellis Horwood, New York, 213-227.

RISSANEN, J. (1978): Modeling by Shortest Data Description. *Automatica 14, 465-471.*

SONDHAUSS, U. and WEIHS, C. (2001): Standardizing the Comparison of Partitions, submitted to Unwin, Wilhelm, and Hofmann (Eds.): Special issue of the Journal of Computational Statistics for the Proceedings of the International Symposium on Data Mining and Statistics, November 20-21, 2000, University of Augsburg, Germany

SILBERSCHATZ, A., and TUZHILIN, A. (1996): What makes patterns interesting in knowledge discovery systems. *IEEE Trans. on Knowledge and Data Engineering, 8,970–974.*

SMYTH, P. and GOODMAN, R. (1992): An information theoretic approach to rule induction from databases. *IEEE Transactions on Knowledge and Data Engineering, 4, 301-316*

WANG, K., TAY, W., and LIU, B. (1998): An interestingness-based interval merger for numeric association rules, in Proceedings of the International Conference on Knowledge Discovery and Data Mining, August 1998, New York City, AAAI, 121-128.

WEIHS, C. (1992): Vorhersagefaehigkeit multivariater linearer Methoden: Simulation und Grafik, in Enke, Goelles, Haux and Wernecke (Eds.): Methoden und Werkzeuge fuer die exploratorische Datenanalyse in den Biowissenschaften, Fischer, Stuttgart, 111-127.

WEIHS, C. (1993): Multivariate Exploratory Data Analysis and Graphics: A Tutorial. *Journal of Chemometrics 7, 305-340.*

WEIHS, C., ROEHL, M. C., and THEIS, W. (1999): Multivariate classification of business phases, Technical Report 26/1999, SFB 475, Universitaet Dortmund.

WEISS, S. M. and KULIKOWSKI, C. A. (1991): Computer Systems that Learn; Morgan Kaufmann, San Francisco, 114, 168/9.