Fuzzy classification rules based on similarity*

Martin Holeňa[1] and David Štefka[2]

[1] Institute of Computer Science, Academy of Sciences of the Czech Republic
Pod Vodárenskou věží 2, 182 07 Prague
[2] Faculty of Nuclear Science and Physical Engineering
Czech Technical University
Trojanova 13, 120 00 Prague

Abstract. *The paper deals with the aggregation of classification rules by means of fuzzy integrals, in particular with the fuzzy measures employed in that aggregation. It points out that the kinds of fuzzy measures commonly encountered in this context do not take into account the diversity of classification rules. As a remedy, a new kind of fuzzy measures is proposed, called similarity-aware measures, and several useful properties of such measures are proven. Finally, results of extensive experiments on a number of benchmark datasets are reported, in which a particular similarity-aware measure was applied to a combination of Choquet or Sugeno integrals with three different ways of creating ensembles of classification rules. In the experiments, the new measure was compared with the traditional Sugeno λ-measure, to which it was clearly superior.*

1 Introduction

Logical formulas of specific kinds, usually called *rules*, are a traditional way of formally representing knowledge. Therefore, it is not surprising that they are also the most frequent representation of the knowledge discovered in data mining.

The most natural base for differentiating between existing rules extraction methods is the *syntax and semantics of the extracted rules* [10]. Syntactical differences between them are, however, not very deep because, principally, any rule r from a ruleset $\mathcal{R}$ has one of the forms $S_r \sim S'_r$, or $A_r \to C_r$, where S_r, S'_r, A_r and C_r are formulas of the considered logic, and $\sim$, $\to$ are symbols of the language of that logic. The difference between both forms concerns semantic properties of the symbols $\sim$ and $\to$: $S_r \sim S'_r$ is symmetric with respect to S_r, S'_r in the sense that its validity always coincides with that of $S'_r \sim S_r$ whereas $A_r \to C_r$ is not symmetric with respect to A_r, C_r in that sense. In the case of a propositional logic, $\sim$ and $\to$ are the connectives equivalence ($\equiv$) and implication, respectively, whereas in the case of a predicate logic, they are generalized quantifiers. To distinguish the formulas involved in the asymmetric case, A_r is called *antecedent* and C_r *consequent* of r.

More important is the semantic of the rules (cf. [5]), especially the difference between *rules of the Boolean logic* and *rules of a fuzzy logic*. Due to the semantics of Boolean and fuzzy formulas, the former are valid for crisp sets of objects, whereas the validity of the latter is a fuzzy set on the universe of all considered objects. Boolean rulesets are extracted more frequently, especially some specific types of them, such as *classification rulesets* [6, 9]. Those are sets of implications such that $\{A_r\}_{r\in\mathcal{R}}$ and $\{C_r\}_{r\in\mathcal{R}}$ partition the set $\mathcal{O}$ of considered objects, where $\{\cdot\}_{r\in\mathcal{R}}$ stands for the set of distinct formulas in $(\cdot)_{r\in\mathcal{R}}$. Abandoning the requirement that $\{A_r\}_{r\in\mathcal{R}}$ partitions $\mathcal{O}$ (at least in the sense of a crisp partitioning) allows to generalize those rulesets also to fuzzy antecedents [15]. For Boolean antecedents, however, this requirement entails a natural definition of the validity of a whole classification ruleset $\mathcal{R}$ for an object x. Assuming that all information about x conveyed by $\mathcal{R}$ is conveyed by the single rule r covering x (i.e., with A_r valid for x), the validity of $\mathcal{R}$ for x can be defined to coincide with the validity of $A_r \to C_r$ for that r, which in turn equals the validity of C_r for x.

It is also possible to combine several existing classification rules into a new one. Such aggregation can be either *static*, i.e., the result is the same for all inputs, or *dynamic*, where it is adapted to the currently classified input [11, 19]. In the aggregation of classification rules, we usually try to create a team of rules that are not similar. This property is called *diversity* [14]. There are many methods for building a diverse team of classifiers [2, 3, 16].

One of popular aggregation operators is the *fuzzy integral* [7, 12, 13, 17]. It aggregates the outputs of the individual classification rules with respect to a fuzzy measure. The role of fuzzy measures in the aggregation of classification rules, in particular their role with respect to the diversity of the rules, was the subject of the research reported in this paper.

The following section recalls the fuzzy integrals and fuzzy measures encountered in the aggregation of classification rules. In Section 3, which is the key section of the paper, a new fuzzy measure, called similarity-aware measure, is introduced and its theoretical properties are studied. Finally, in Section 4, results of ex-

* The research reported in this paper has been supported by the Czech Science Foundation (GA ČR) grant P202/11/1368.

tensive experiments and comparison with the traditional Sugeno λ-measure are reported.

2 Fuzzy integrals and measures in classification rules aggregation

Several definitions of a fuzzy integral exists in the literature – among them, the Choquet integral and the Sugeno integral are used most often. The role played in usual integration by additive measures (such as probability or Lebesgue measure) is in fuzzy integration played by fuzzy measures. In this section, basic concepts pertaining to different kinds of fuzzy measures will be recalled, as well as the definitions of Choquet and Sugeno integrals. Due to the intended context of aggregation of classification rules, we restrict attention to $[0, 1]$-valued functions on finite sets.

Definition 1. *A* fuzzy measure μ *on a finite set* $\mathcal{U} = \{u_1, \ldots, u_r\}$ *is a function on the power set of* $\mathcal{U}$,

$$\mu : \mathcal{P}(\mathcal{U}) \to [0, 1] \tag{1}$$

fulfilling:

1. *the boundary conditions*

$$\mu(\emptyset) = 0, \mu(\mathcal{U}) = 1 \tag{2}$$

2. *the monotonicity*

$$A \subseteq B \Rightarrow \mu(A) \leq \mu(B) \tag{3}$$

The values $\mu(u_1), \ldots, \mu(u_r)$ *are called* fuzzy densities.

Definition 2. *The* Choquet integral *of a function* $f : \mathcal{U} \to [0, 1]$, $f(u_i) = f_i$, $i = 1, \ldots, r$, *with respect to a fuzzy measure* μ *is defined as:*

$$(\mathrm{Ch}) \int f d\mu = \sum_{i=1}^{r} (f_{<i>} - f_{<i-1>}) \mu(A_{<i>}), \tag{4}$$

where $< \cdot >$ *indicates that the indices have been permuted, such that* $0 = f_{<0>} \leq f_{<1>} \leq \cdots \leq f_{<r>} \leq 1$. $A_{<i>} = \{u_{<i>}, \ldots, u_{<r>}\}$ *denotes the set of of elements of* $\mathcal{U}$ *corresponding to the* $(r - i + 1)$ *highest values of* f.

Definition 3. *The* Sugeno integral *of a function* $f : \mathcal{U} \to [0, 1]$, $f(u_i) = f_i$, $i = 1, \ldots, r$, *with respect to a fuzzy measure* μ *is defined as:*

$$(\mathrm{Su}) \int f d\mu = \max_{i=1}^{r} \min(f_{<i>}, \mu(A_{<i>})). \tag{5}$$

To define a general fuzzy measure in the discrete case, we need to define all its 2^r values, which is usually very complicated. To overcome this weakness, measures which do not need all the 2^r values have been developed [7, 17]:

Definition 4. *A fuzzy measure* μ *on* $\mathcal{U}$ *is called* symmetric *if*

$$|A| = |B| \Rightarrow \mu(A) = \mu(B) \tag{6}$$

$$\text{for } A, B \subseteq \mathcal{U}, \tag{7}$$

where $|\cdot|$ *denotes the cardinality of a set.*

Consequently, the value of a symmetric measure depends only on the cardinality of its argument. If a symmetric measure is used in Choquet integral, the integral reduces to the ordered weighted average operator [17]. However, symmetric measures assume that all elements of $\mathcal{U}$ have the same importance, thus they do not take into account the diversity of elements.

Definition 5. *Let* $\perp$ *be a t-conorm. A fuzzy measure* μ *is called* $\perp$-decomposable *if*

$$\mu(A \cup B) = \mu(A) \perp \mu(B) \quad \text{for disjoint } A, B \subseteq \mathcal{U} \tag{8}$$

Hence, $\perp$-decomposable measures need only the r fuzzy densities, whereas all the other values are computed using the formula (8). Particular cases of this kind of fuzzy measures are *additive measures*, including probabilistic measures ($\perp$ being the bounded sum), and the Sugeno λ-measure.

Definition 6. Sugeno λ-measure *[7, 17] on a finite set* $\mathcal{U} = \{u_1, \ldots, u_r\}$ *is defined*

$$\mu(A \cup B) = \mu(A) + \mu(B) + \lambda\mu(A)\mu(B), \tag{9}$$

for disjoint $A, B \in \mathcal{U}$, *and some fixed* $\lambda > -1$. *The value of* λ *is:*

a) *computed as the unique non-zero root greater than* -1 *of the equation*

$$\lambda + 1 = \prod_{i=1,\ldots,r} (1 + \lambda\mu(\{u_i\})) \tag{10}$$

if the densities do not sum up to 1;

b) $\lambda = 0$ *else.*

If the densities sum up to 1, the fuzzy measure is additive. Sugeno λ measure is a $\perp$-decomposable measure for the t-norm

$$x \perp y = \min(1, x + y + \lambda xy). \tag{11}$$

A serious weakness of any $\perp$-decomposable measure is that the fuzzy measure of a set of two (or more) classification rules is fully determined by the formula (8) for a fixed $\perp$. Therefore, if interactions between elements are to be taken into account, then they have to be incorporated directly into the fuzzy measure. That fact motivated our attempt to elaborate the concept of similarity-aware fuzzy measures.

3 Similarity-aware measures and their properties

Before introducing similarity-aware measures, let us first recall the notion of similarity [8].

Definition 7. *Let $\wedge$ be a t-norm and let $\sim: \mathcal{U} \times \mathcal{U} \to [0,1]$ be a fuzzy relation. $\sim$ is called a* similarity on $\mathcal{U}$ with respect to $\wedge$ *if the following holds for $a, b, c \in \mathcal{U}$:*

$$\sim (a,a) = 1 \text{ (reflexivity)}, \tag{12}$$

$$\sim (a,b) = \sim (b,a) \text{ (symmetry)}, \tag{13}$$

$$\sim (a,b) \wedge \sim (b,c) \leq \sim (a,c) \text{ (transitivity w.r.t. } \wedge \text{)}. \tag{14}$$

In the context of aggregation of crisp classification rules, we will work with an empirically defined relation, which, for rules ϕ_k, ϕ_l, is defined as the proportion of equal consequents on some validation set of patterns $\mathcal{V} \subset \mathcal{O}$,

$$\sim (\phi_k, \phi_l) = \frac{\sum_{x \in \mathcal{V}} I(C_{\phi_k}(x) = C_{\phi_l}(x))}{|\mathcal{V}|}. \tag{15}$$

It is easily seen that the relation (15) is a similarity with respect to the Łukasiewicz t-norm

$$\wedge_L(a,b) = \max(a+b-1, 0), \tag{16}$$

but it is not a similarity with respect to the standard (minimum, Gödel) t-norm

$$\wedge_S(a,b) = \min(a,b), \tag{17}$$

or the product t-norm

$$\wedge_P(a,b) = ab. \tag{18}$$

Fuzzy integral represents a convenient tool to work with the diversity of classification rules: As we are computing the fuzzy measure values $\mu(A_{<i>})$, we are considering a single rule $\phi_{<i>}$ at each step i, and therefore we can influence the increase of the fuzzy measure based on the similarity of $\phi_{<i>}$ to the set of rules already involved in the integration, i.e., $A_{<i+1>} = \{\phi_{<i+1>}, \ldots, \phi_{<r>}\}$. If $\phi_{<i>}$ is similar to the classifiers in $A_{<i+1>}$, the increase in the fuzzy measure should be small (since the importance of the set $A_{<i>}$ should be similar to the importance of the set $A_{<i+1>}$), and if $\phi_{<i>}$ is not similar to the classifiers in $A_{<i+1>}$, the increase of the fuzzy measure should be large. These ideas motivated the following definition:

Definition 8. *Let $\mathcal{U} = \{u_1, \ldots, u_r\}$ be a set, let $\sim$ be a similarity w.r.t. a t-norm $\wedge$, and let S be a an $r \times r$ matrix such that:*

$$S = (s_{i,j})_{i,j=1}^r \text{ with } s_{i,j} = \sim (u_i, u_j). \tag{19}$$

Let further $\kappa_i \in [0,1]$, $i = 1, \ldots, r$ denote some kind of weight (confidence, importance) of u_i, and let $[\cdot]$ denote index ordering according to κ, such that $0 \leq \kappa_{[1]} \leq \cdots \leq \kappa_{[r]} \leq 1$. Finally, let

$$\tilde{\mu}^{(S)} : \mathcal{P}(\mathcal{U}) \to [0, \infty) \tag{20}$$

be a mapping such that for $X \subseteq \mathcal{U}$,

$$\tilde{\mu}^{(S)}(X) = \sum_{i=1}^r I(u_{[i]} \in X)\kappa_{[i]}(1 - \max_{j=i+1}^r s_{[i],[j]}), \tag{21}$$

where we define $\max_{j=r+1}^r s_{[r],[j]} = 0$, and I denotes the indicator of thruth value, i.e.,

$$I(true) = 1, I(false) = 0. \tag{22}$$

Then the mapping

$$\mu^{(S)} : \mathcal{P}(\mathcal{U}) \to [0,1], \text{ defined} \tag{23}$$

$$\mu^{(S)}(X) = \frac{\tilde{\mu}^{(S)}(X)}{\tilde{\mu}^{(S)}(\mathcal{U})}, \tag{24}$$

is called a similarity-aware measure based on S.

Proposition 1. *$\mu^{(S)}$ is a fuzzy measure on $\mathcal{U}$.*

Proof. The boundary conditions follow directly from the definition of $\mu^{(S)}$. For the monotonicity, let $A \subseteq B$; then

$$\tilde{\mu}^{(S)}(A) = \sum_{i=1}^r I(u_{[i]} \in A)\kappa_{[i]}(1 - \max_{j=i+1}^r s_{[i],[j]}) \leq$$
$$\leq \sum_{i=1}^r I(u_{[i]} \in B)\kappa_{[i]}(1 - \max_{j=i+1}^r s_{[i],[j]}) =$$
$$= \tilde{\mu}^{(S)}(B), \tag{25}$$

due to $I(u_{[i]} \in A) = 1 \Rightarrow I(u_{[i]} \in B) = 1$.

Proposition 2. *For any of the 2^r subsets $X \subset \mathcal{U}$, the value $\mu(X)$ can be expressed simply as the sum of values of μ on singletons*

$$\mu^{(S)}(X) = \sum_{u_i \in X} \mu^{(S)}(u_i). \tag{26}$$

Proof. According to (21) and (23), the value of μ on the singletosn $u_i, i = 1, \ldots, r$ is

$$\mu^{(S)}(u_i) = \frac{1}{\tilde{\mu}^{(S)}(\mathcal{U})}\kappa_{[i]}(1 - \max_{j=i+1}^r s_{[i],[j]}). \tag{27}$$

Then (26) follows directly from (21).

The following propositions show that if for some i, the i-th classification rule is totally similar to some other rule in $A_{<i+1>}$, then $\mu^{(S)}$ does not increase, and if it is totally unsimilar to all classifiers in $A_{<i+1>}$, the increase in $\mu^{(S)}$ is maximal.

Proposition 3. *Let $f : \mathcal{U} \to [0,1]$, and let the matrix S in (19) fulfills*

$$s_{i,j} = 1 \text{ for } i \neq j. \tag{28}$$

Then:

1. $(\forall X \subseteq \mathcal{U})\ u_{[r]} \in X \Rightarrow \mu^{(S)} = 1$,
2. $(\forall X \subseteq \mathcal{U})\ u_{[r]} \notin X \Rightarrow \mu^{(S)} = 0$,
3. $(\mathrm{Ch}) \int f d\mu^{(S)} = (\mathrm{Su}) \int f d\mu^{(S)} = f_{[r]}$.

Proof. 1. and 2. follow directly from the fact that

$$\max_{j=i+1}^{r} s_{[i],[j]} = \begin{cases} 0 \text{ for } i = r, \\ 1 \text{ for } i < r. \end{cases} \tag{29}$$

and therefore

$$\tilde{\mu}^{(S)} = I(u_{[r]} \in X)\kappa_{[r]}. \tag{30}$$

We will prove 3. only for the Choquet integral, the case of Sugeno integral is analogous. Let $j \in \{1, \dots, r\}$ such that $< j >= [r]$; then $(\forall i > j)\ u_{[r]} \notin A_{<i>}$, and therefore $\mu^{(S)}(A_{<i>}) = 0$; $(\forall i \leq j)\ u_{[r]} \in A_{<i>}$, and therefore $\mu^{(S)} = 1$. Using this in the definition of the Choquet integral, we obtain

$$\begin{aligned} (\mathrm{Ch}) \int f d\mu^{(S)} &= \\ &= \sum_{i=1}^{r} (f_{<i>} - f_{<i-1>})\mu^{(S)}(A_{<i>}) = \\ &= \sum_{i=1}^{j} (f_{<i>} - f_{<i-1>}) = \\ &= f_{<j>} = f_{[r]}. \end{aligned} \tag{31}$$

Proposition 4. *Let $f : \mathcal{U} \to [0,1]$, and let the matrix S in (19) fulfills $s_{i,j} = 0$ for $i \neq j$. Then:*

1. $(\forall X \subseteq \mathcal{U})\ \mu^{(S)} = \frac{\sum_{i:u_{[i]} \in X} \kappa_{[i]}}{\sum_{i=1}^{r} \kappa_i}$,
2. $(\mathrm{Ch}) \int f d\mu^{(S)} \mu^{(S)} = \frac{\sum_{i=1}^{r} \kappa_i f_i}{\sum_{i=1}^{r} \kappa_i}$,
3. $(\mathrm{Su}) \int f d\mu^{(S)} = \max_{k=1}^{r}(f_{<k>}, \frac{\sum_{i=k}^{r} \kappa_{<i>}}{\sum_{i=1}^{r} \kappa_i})$.

Proof. 1. follows directly from the definition of similarity-aware measure, and 2. and 3. are applications of 1. to the definition of the Choquet/Sugeno integral.

4 Experimental testing

We have experimentally compared the performance of the proposed measure with the Sugeno λ-measure for the aggregation of classification rules by fuzzy integrals (Choquet, Sugeno). The ensembles have been created as random forests from rules obtained with classification trees [3], by bagging [2] from rules obtained with k-NN classifiers, and by the multiple feature subset method [1] from rules obtained with quadratic discriminant analysis.

In this section, we present results of comparing the measures using 10-fold crossvalidation on 5 artificial and 11 real-world datasets (the properties of the datasets are shown in Table 1). For the random forests, the number of trees was set to $r = 20$, the number of features to explore in each node varied between 2 and 5 (depending on the dimensionality of the particular dataset), the maximal size of a leaf was set to 10 (see [3] for description of the parameters). For the QDA and k-NN based ensembles, their size was set also to $r = 20$, and we used $k = 5$ as the number of neighbors for k-NN classifiers. As the weights $\kappa_1, \dots, \kappa_r$ of the classification rules, we used

$$\kappa_i(\phi) = \frac{\sum_{x \in \mathcal{V}(A_\phi)} I(C'_\phi(x) = C_\phi(x))}{|\mathcal{V}(A_\phi)|}, \tag{32}$$

where $\mathcal{V}(A_\phi) \subseteq \mathcal{V}$ is the set of validation patterns belonging to some kind of neighborhood of A_ϕ. For example, if A_ϕ concerns values of vectors in an Euclidean space, then $\mathcal{V}(A_\phi)$ is the set of k nearest neighbors under Euclidean metric of the set where the antecedent A_ϕ is valid. The number of neighbors was set to 5, 10, or 20, depending on the size of the dataset.

Table 2 shows the results of the performed comparisons. We also measured the statistical significance of the pairwse improvements (using the analysis of variance on the 5% confidence level by the Tukey-Kramer method).

We interpret the results presented in Table 2 as a confirmation of the usefulness of similarity-aware fuzzy measures proposed in Definition 8.

5 Conclusion

In this paper, we have studied the application of the fuzzy integral as an aggregation operator for classification rules in the context of their similarities. We have shown that traditionally used symmetric, or additive and other $\perp$-decomposable measures are not a good choice for combining classification rules by fuzzy integral and we have defined similarity-aware measures, which take into account both the confidence / importance and the similarities of the aggregated rules. We have shown some basic theoretical properties and special cases of the measures, including the fact that apart the singletons, the 2^r values of μ are obtained using only summation. In addition, we have experimentally compared the performance of the measures to the Sugeno λ-measure using Choquet and Sugeno fuzzy integrals on 16 benchmark datasets for 3 different ways

dataset	nr. of patterns	nr. of classes	dimension
Artificial			
clouds [4]	5000	2	2
concentric [4]	2500	2	2
gauss 3D [4]	5000	2	3
ringnorm [18]	3000	2	20
waveform [18]	5000	3	21
Real-world			
glass [18]	214	7	9
letters [18]	20000	26	16
pendigits [18]	10992	10	16
phoneme [4]	5427	2	5
pima [18]	768	2	8
poker [18]	4828	3	10
satimage [4]	6435	6	4
transfusion [18]	748	2	4
vowel [18]	990	11	10
wine [18]	178	3	13
yeast [18]	1484	4	8

Table 1. Datasets used in the experiments.

of obtaining ensembles of classification rules. The experimental comparison clearly supports our theoretical conjecture that similarity-aware measures are more suitable for the aggregation of classification rules than traditionally used additive and $\perp$-decomposable fuzzy measures.

References

1. S. D. Bay: *Nearest neighbor classification from multiple featre subsets.* Intelligent Data Analysis 3, 1999, 191–209.
2. L. Breiman: *Bagging predictors.* Machine Learning 24, 1996, 123–140.
3. L. Breiman: *Random forests.* Machine Learning 45, 2001, 5–32.
4. Machine Learning Group Catholic University of Leuven. Elena database. `http://mlg.info.ucl.ac.be/index.php?page=Elena`.
5. D. Dubois, Hüllermeier, H. Prade: *A systematic approach to the assessment of fuzzy association rules.* Data Mining and Knowledge Discovery 13, 2006, 167–192.
6. L. Geng, H. J. Hamilton: *Choosing the right lens: Finding what is interesting in data mining.* In F. Guillet and H. J. Hamilton, (Eds), Quality Measures in Data Mining, Springer Verlag, Berlin, 2007, 3–24.
7. M. Grabisch, H. T. Nguyen, E. A. Walker: *Fundamentals of uncertainty calculi with applications to fuzzy inference.* Kluwer Academic Publishers, Dordrecht, 1994.
8. P. Hájek: *Metamathematics of fuzzy logic.* Kluwer Academic Publishers, Dordrecht, 1998.
9. D. J. Hand: *Construction and assessment of classification rules.* John Wiley and Sons, New York, 1997.
10. M. Holeňa: *Measures of ruleset quality capable to represent uncertain validity.* Submitted to International Journal of Approximate Reasoning.
11. A. H. R. Ko, R. Sabourin, A. S. Britto: *From dynamic classifier selection to dynamic ensemble selection.* Pattern Recognition 41, 2008, 1718–1731.
12. L. I. Kuncheva: *Fuzzy versus nonfuzzy in combining classifiers designed by boosting.* IEEE Transactions on Fuzzy Systems 11, 2003, 729–741.
13. L. I. Kunchev: *Combining pattern classifiers: methods and algorithms.* John Wiley and Sons, New York, 2004.
14. L. I. Kuncheva C. J. Whitaker: *Measures of diversity in classifier ensembles.* Machine Learning 51, 2003, 181–207.
15. L. E. Peterson, M. A. Coleman: *Machine learning based receiver operating characteristic (ROC) curves for crisp and fuzzy classification of DNA microarrays in cancer research.* International Journal of Approximate Reasoning 47, 2008, 17–36.
16. L. Rokach: *Taxonomy for characterizing ensemble methods in classification tasks: A review and annotated bibliography.* Computational Statistics and Data Analysis 53, 2009, 4046–4072.
17. V. Torra, Y. Narukawa: *Modeling decisions: information fusion and aggregation operators.* Springer Verlag, Berlin, 2007.
18. Machine Learning Group University of California Irwine. Repository of machine learning databases. `http://www.ics.uci.edu/ mlearn/MLRepository.html`.
19. D. Štefka, M. Holeňa: *Dynamic classifier systems and their applications to random forest ensembles.* In Adaptive and Natural Computing Algorithms. Lecture Notes in Computer Science 5495, Springer Verlag, Berlin, 2009, 458–468.

dataset	Choquet integral		Sugeno integral	
	λ-measure	μ^S	λ-measure	μ^S
random forests				
clouds	12.40 ± 1.81	$\mathbf{12.25 \pm 1.85}$	12.80 ± 1.64	12.33 ± 1.47
concentric	4.32 ± 1.25	$\mathbf{2.82 \pm 1.30}$	3.24 ± 1.37	2.98 ± 1.50
gauss-3D	23.92 ± 2.97	$\mathbf{22.76 \pm 1.59}$	24.60 ± 1.36	23.28 ± 1.58
glass	21.3 ± 10.3	$\mathbf{14.1 \pm 3.5}$	24.1 ± 7.0	17.5 ± 9.4
letters	$\mathbf{7.1 \pm 0.6}$	7.3 ± 0.2	8.0 ± 0.6	7.9 ± 0.8
pendigits	3.1 ± 0.5	$\mathbf{2.7 \pm 0.5}$	3.2 ± 0.4	3.8 ± 0.7
phoneme	$\mathbf{12.4 \pm 1.2}$	13.2 ± 1.9	12.7 ± 0.8	13.3 ± 1.6
pima	26.0 ± 4.8	$\mathbf{23.8 \pm 2.0}$	25.0 ± 2.2	23.9 ± 3.6
poker	46.5 ± 3.0	$\mathbf{44.4 \pm 1.3}$	46.5 ± 1.5	45.1 ± 1.9
ringnorm	13.27 ± 2.11	*7.69 ± 2.06*	12.74 ± 2.08	***7.46 ± 1.79***
satimage	14.7 ± 1.4	$\mathbf{14.3 \pm 1.3}$	14.9 ± 0.9	14.8 ± 1.4
transfusion	4.8 ± 1.1	$\mathbf{2.3 \pm 0.7}$	4.9 ± 1.0	2.6 ± 0.7
vowel	14.5 ± 3.0	$\mathbf{13.1 \pm 3.5}$	17.0 ± 5.3	13.4 ± 3.8
waveform	18.56 ± 2.42	$\mathbf{17.93 \pm 1.89}$	18.24 ± 3.04	18.23 ± 1.58
wine	5.6 ± 6.0	$\mathbf{3.3 \pm 5.5}$	3.4 ± 4.0	6.6 ± 5.8
yeast	38.2 ± 4.1	$\mathbf{34.8 \pm 2.6}$	38.5 ± 3.7	36.3 ± 3.4
k-NN classifiers				
clouds	$\mathbf{11.93 \pm 2.29}$	12.12 ± 1.57	12.64 ± 2.48	12.96 ± 2.26
concentric	1.39 ± 0.77	1.72 ± 0.57	$\mathbf{1.30 \pm 0.80}$	1.56 ± 0.64
gauss-3D	26.71 ± 2.55	$\mathbf{26.00 \pm 2.88}$	27.68 ± 3.66	26.28 ± 2.74
glass	22.4 ± 9.8	20.7 ± 10.3	21.7 ± 11.1	$\mathbf{19.3 \pm 6.5}$
letters	17.7 ± 2.7	$\mathbf{17.6 \pm 2.9}$	19.3 ± 3.1	19.1 ± 2.7
pendigits	1.3 ± 0.8	$\mathbf{1.4 \pm 0.8}$	1.3 ± 0.5	1.3 ± 0.7
phoneme	14.6 ± 0.9	$\mathbf{14.2 \pm 2.4}$	14.4 ± 1.8	14.5 ± 1.7
pima	$\mathbf{29.1 \pm 5.1}$	30.2 ± 7.2	29.5 ± 4.4	30.3 ± 6.6
poker	45.3 ± 2.4	$\mathbf{43.5 \pm 2.3}$	47.2 ± 2.7	43.9 ± 1.4
ringnorm	36.20 ± 4.41	$\mathbf{34.28 \pm 2.59}$	33.56 ± 3.34	33.48 ± 2.94
satimage	16.5 ± 2.0	$\mathbf{15.5 \pm 1.7}$	16.8 ± 2.4	16.2 ± 2.3
transfusion	24.0 ± 4.0	$\mathbf{23.4 \pm 4.7}$	25.2 ± 3.5	24.0 ± 4.7
vowel	4.8 ± 2.2	$\mathbf{4.0 \pm 1.9}$	5.6 ± 2.1	7.0 ± 1.8
waveform	19.40 ± 2.10	$\mathbf{18.28 \pm 2.85}$	19.57 ± 2.20	19.04 ± 2.99
wine	30.0 ± 10.3	$\mathbf{28.6 \pm 14.8}$	31.2 ± 6.7	33.4 ± 16.0
yeast	41.8 ± 4.7	$\mathbf{40.5 \pm 2.6}$	42.6 ± 3.7	40.7 ± 3.6

QDA with multiple subsets				
clouds	26.00 ± 2.70	23.14 ± 2.49	25.74 ± 1.92	$\boldsymbol{\mathit{22.66 \pm 1.10}}$
concentric	4.36 ± 1.96	3.68 ± 1.68	5.72 ± 1.84	$\mathbf{3.40 \pm 0.98}$
gauss-3D	23.87 ± 1.86	$\mathbf{22.06 \pm 2.10}$	23.96 ± 2.03	22.36 ± 2.12
glass	42.3 ± 10.9	38.5 ± 12.0	43.2 ± 14.9	$\mathbf{32.4 \pm 12.5}$
letters	17.1 ± 0.7	$\boldsymbol{\mathit{14.7 \pm 0.7}}$	17.2 ± 0.7	$\mathit{14.7 \pm 0.8}$
pendigits	2.8 ± 0.5	$\mathbf{2.2 \pm 0.2}$	2.7 ± 0.5	2.7 ± 0.6
phoneme	25.4 ± 2.4	$\mathit{20.8 \pm 1.4}$	24.7 ± 1.0	$\boldsymbol{\mathit{20.2 \pm 2.2}}$
pima	27.9 ± 4.7	$\mathbf{25.5 \pm 4.2}$	28.3 ± 3.3	26.1 ± 5.1
poker	66.1 ± 2.1	$\mathit{55.1 \pm 2.3}$	66.3 ± 3.9	$\boldsymbol{\mathit{55.1 \pm 2.1}}$
ringnorm	1.94 ± 0.96	2.53 ± 1.01	$\mathbf{1.68 \pm 0.60}$	3.66 ± 1.31
satimage	17.0 ± 1.3	$\mathbf{15.7 \pm 1.1}$	17.2 ± 2.0	16.2 ± 1.3
transfusion	29.6 ± 8.6	$\mathbf{22.3 \pm 4.5}$	29.2 ± 7.1	23.4 ± 3.4
vowel	16.7 ± 5.4	$\mathbf{14.0 \pm 3.8}$	18.1 ± 4.1	15.6 ± 3.3
waveform	15.73 ± 2.07	$\mathbf{14.52 \pm 1.59}$	15.33 ± 1.72	14.54 ± 1.80
wine	1.2 ± 2.5	2.8 ± 3.9	$\mathbf{0.6 \pm 1.9}$	3.3 ± 2.9
yeast	49.0 ± 4.3	$\mathit{39.8 \pm 4.3}$	49.5 ± 4.9	$\boldsymbol{\mathit{39.1 \pm 3.8}}$

Table 2. Mean error rates ± standard deviation of the error rate [%], based on 10-fold crossvalidation. The best result for each dataset is displayed in boldface, statistically significant improvements (measured by the analysis of variance using the Tukey-Kramer method at the 5% level) are displayed in italics