Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

How fuzzy partitions stabilise feature selection: a variance–fuzziness analysis of profile-based filter scores

Sisir Kumar Rajbongshi

sisir@gauhati.ac.in

Pandit Deendayal Upadhyaya Adarsha Mahavidyalaya (affiliated to Gauhati University)

N. Sathiyaseelan

Achariya College of Engineering Technology (An Autonomous Institution)

Shahidul Islam Khan

Pandit Deendayal Upadhyaya Adarsha Mahavidyalaya (affiliated to Gauhati University)

Research Article

Keywords:

Posted Date: November 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11748380/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

How fuzzy partitions stabilise feature selection: a variance–fuzziness analysis of profile-based filter scores

Sisir Kumar Rajbongshi[1,*], N. Sathiyaseelan[2] and Shahidul Islam Khan[3]

[1]Department of Computer Science, Pandit Deendayal Upadhyaya Adarsha Mahavidyalaya (affiliated to Gauhati University), Amjonga, Goalpara 783124, Assam, India

[2]Department of Mathematics, Achariya College of Engineering Technology (An Autonomous Institution), Puducherry 605110, India

[3]Department of Mathematics, Pandit Deendayal Upadhyaya Adarsha Mahavidyalaya (affiliated to Gauhati University), Amjonga, Goalpara 783124, Assam, India

E-mail: sisir@gauhati.ac.in (S.K.R.); rakhakshik786@gmail.com (S.I.K.)

*Corresponding author: sisir@gauhati.ac.in ORCID (S.K.R.): 0000-0003-1874-6882

Abstract

Numeric features are often described by fuzzy linguistic terms such as Low, Medium and High, and filter scores built from the class distributions over these terms are popular in interpretable feature selection. Since fuzzification only smooths the data, it cannot increase the class information in a feature, so it is unclear why such scores should select features better than their crisp counterparts. We show that the benefit is a variance effect. For any strong fuzzy partition, the total variance of the term memberships equals that of a crisp partition with the same term masses minus the mean quadratic fuzziness of the partition. Consequently, irrelevant features obtain smaller chance scores: for the fuzzy soft discrimination score (FSDS), which equals twice a total variation distance, we give a bias bound, an exact large-sample limit and a concentration inequality; for fuzzy mutual information, the chance level falls below the classical value $(T-1)/(2N)$ by an amount fixed by the fuzziness, which yields a fuzzy Miller–Madow correction. The same analysis predicts that fuzziness hurts when class differences are narrowly localised. Simulations confirmed every prediction: a triangular partition lowered the chance score by 20% for FSDS and 34% for mutual information. On 20 public benchmark datasets, in a comparison of ten filters including fuzzy-rough dependency, Kolmogorov–Smirnov and ReliefF, the fuzzy versions of both scores were more stable than their crisp versions in small samples (Kuncheva index 0.498 vs 0.409 for FSDS, 0.519 vs 0.443 for mutual information) and admitted fewer noise features, with unchanged classification accuracy.

Keywords: Fuzzy partition; Feature selection; Fuzzy mutual information; Total variation distance; Stability; Bias–variance trade-off

1 Introduction

Filter methods, which score each feature separately, remain a standard first step in building classifiers from tabular data because they are fast, independent of the model and easy to explain [1, 2]. In interpretable and medical applications, numeric features are often described by linguistic terms such as Low, Medium and High, represented by fuzzy sets [3, 4], and a feature is judged by how differently the classes are distributed over its terms. Fuzzy and fuzzy-rough measures of this kind

have been proposed in many forms [5–8]. A common, mostly informal, justification is that graded memberships make such scores more robust than the same scores computed on crisp intervals.

This justification raises a basic question. Fuzzification replaces each value by a vector of membership degrees, which smooths the data and can only lose information about the class (Proposition 1). Why, then, should a fuzzy score ever select features better than its crisp counterpart, and when does fuzziness hurt? To our knowledge, no existing analysis answers this in a form that predicts how large the benefit is.

We answer it for profile-based scores, that is, scores computed from the class-conditional distributions of a feature over its linguistic terms. Our running example is the fuzzy soft discrimination score (FSDS), the L_1 distance between these distributions. The answer is that fuzziness acts as a variance-reduction device: it does not add signal, but it lowers the sampling noise of the class profiles by an exact and interpretable amount. The contributions are as follows.

1. *A variance–fuzziness identity* (Theorem 1). The total variance of the term memberships equals that of a crisp partition with the same term masses minus the mean quadratic fuzziness $\bar{F}$ of the partition.

2. *Consequences for chance scores* (Theorems 2 and 3). The expected score of an irrelevant feature has a smaller upper bound under a fuzzy partition, and for large samples the expected chance score itself is smaller, with an explicit limit. The same holds for fuzzy mutual information, whose chance level falls below the classical χ^2 value $(T-1)/(2N)$ by an amount fixed by the fuzziness; this yields a fuzzy Miller–Madow correction.

3. *A variance–discrimination trade-off* (Section 4.4), which predicts that fuzziness helps for smooth class differences and hurts for narrowly localised ones, and a debiased score with a data-driven degree of fuzziness.

4. *Evidence.* Simulations confirm every theoretical prediction. On 20 public benchmark datasets we compare ten filters, including fuzzy mutual information, fuzzy-rough dependency, the Kolmogorov–Smirnov statistic and ReliefF, with three classifiers.

FSDS itself is not new; its interpretation as a total variation distance was noted, without proofs or estimation theory, in our earlier applied study [9]. The results of Section 4 and the experiments are new.

Section 2 reviews related work. Section 3 defines the setting and states what the score measures. Section 4 contains the main results. Sections 5 and 6 report the simulation and benchmark studies, Section 7 discusses limitations, and Section 8 concludes. Standard proofs are in Appendix A.

2 Related work

Fuzzy and fuzzy-rough feature selection. Fuzzy-rough methods evaluate features through fuzzy dependency degrees based on fuzzy lower approximations [5, 6], and fuzzy information measures extend entropy and mutual information to fuzzy partitions [7, 8]. These methods are motivated

by the graded nature of real measurements, but their behaviour as estimators, in particular their variance and chance level in small samples, has received little attention. Our analysis addresses exactly this point and applies to both L_1-type and information-type scores built from fuzzy profiles.

Fuzzy partitions and fuzziness. Strong (Ruspini) partitions are the standard way of defining linguistic terms on a numeric domain [10], and triangular memberships are a common choice [11]. Measures of fuzziness quantify how far a fuzzy set is from being crisp [12, 13]. Casalino et al. [14] showed that in fuzzy rule-based classifiers on strong partitions, fuzziness mainly affects the confidence of decisions rather than the decisions themselves. Their analysis concerns classification; ours concerns the estimation of a feature score and identifies a different effect of fuzziness, variance reduction.

Discretisation and divergence-based filters. Supervised and unsupervised discretisation of continuous features is well studied [15, 16]. Distances between class-conditional distributions, including the total variation distance, are classical measures of separability with known links to the Bayes error [17, 18], and the bias of plug-in information estimates is classical as well [19, 20]. FSDS connects these strands: it is a total variation distance computed after a fuzzy discretisation, and the fuzziness of that discretisation governs the precision of its estimate.

Stability of feature selection. Selected feature sets can change substantially between samples, and stability indices such as those of Kuncheva [21] and Nogueira et al. [22] quantify this. Our theory explains why fuzzy partitions improve stability relative to crisp ones.

3 Setting and the score

Let $Y \in \{0, 1\}$ be a binary class and X a real-valued feature with class-conditional laws P^1, P^0 and prior $\pi = \Pr(Y = 1)$. A training sample contains n_1 independent observations from P^1 and n_0 from P^0; the class counts are fixed, $N = n_1 + n_0$ and $h = 1/n_1 + 1/n_0$. For a signed measure ν, $\|\nu\|$ is its total variation norm, and $\mathrm{TV}(P, Q) = \frac{1}{2}\|P - Q\|$.

Definition 1 (strong fuzzy partition)**.** Measurable functions $\mu_1, \ldots, \mu_T : \mathbb{R} \to [0, 1]$ form a strong fuzzy partition if $\sum_{t=1}^{T} \mu_t(x) = 1$ for every x [10].

The Low/Medium/High partition with centres $c_1 < c_2 < c_3$ is the main example. To vary the degree of fuzziness continuously we use the family $\mu^{(s)}$, $s \in [0, 1]$: between adjacent centres $a < b$ the membership passes linearly from the lower to the upper term over an interval of width $s(b - a)$ centred at $(a + b)/2$. Thus $s = 0$ is the crisp partition cut at the midpoints and $s = 1$ the triangular partition. The partition is assumed to be fixed, that is, chosen independently of the sample on which a score is computed (see Section 7).

Definition 2 (class profiles and the score)**.** For a class c and a term t let $M^c(t) = \mathrm{E}[\mu_t(X) \mid Y = c]$ and let $\hat{M}^c(t)$ be the average of $\mu_t(X_i)$ over the training observations of class c. The fuzzy soft

discrimination score (FSDS) and its estimate are

$$\Phi = \sum_{t=1}^{T} |M^1(t) - M^0(t)| \in [0, 2], \qquad \hat{\Phi} = \sum_{t=1}^{T} |\hat{M}^1(t) - \hat{M}^0(t)|.$$

For a prior π we also write $\Phi_\pi = \sum_t |\pi M^1(t) - (1-\pi)M^0(t)|$, so that $\Phi_{1/2} = \Phi/2$.

Because the memberships sum to one, M^c and $\hat{M}^c$ are probability vectors on the terms; $\hat{M}^c(t)$ is the relative sigma-count of the fuzzy set "class-c observations described by term t". Viewing the terms as parameters, the map $t \mapsto \mu_t(x_i)$ is a fuzzy soft set over the observations [23], which is the origin of the name [9]. For a crisp partition, $\hat{\Phi}$ is the L_1 distance between class histograms.

Definition 3 (term masses, Gini impurity, mean quadratic fuzziness)**.** For a law of X let $m_t = \mathrm{E}\mu_t(X)$, $\sigma_t^2 = \operatorname{Var} \mu_t(X)$,

$$G = 1 - \sum_t m_t^2, \qquad \bar{F} = \mathrm{E}\Big[\sum_t \mu_t(X)(1 - \mu_t(X))\Big] = \mathrm{E}\Big[1 - \sum_t \mu_t(X)^2\Big].$$

$\bar{F}$ is the expected quadratic fuzziness index of the partition at X in the sense of Kaufmann [12]; $\bar{F} = 0$ if and only if the partition is crisp almost surely. Under P^c we write G^c, $\bar{F}^c$ and $V^c = G^c - \bar{F}^c$.

The first result records what Φ measures. All three parts follow from standard arguments [17, 18] (proofs in Appendix A); we state them because they fix the meaning of the score.

Proposition 1 (meaning of the score)**.** *(a) $K(t \mid x) = \mu_t(x)$ is a Markov kernel, $M^c = P^c K$, and $\Phi = 2\,\mathrm{TV}(M^1, M^0)$.*

(b) Fuzzification cannot create discrimination: $\Phi_\pi \le \|\pi P^1 - (1-\pi)P^0\|$; in particular $\Phi \le 2\,\mathrm{TV}(P^1, P^0)$.

(c) The Bayes error ε_B of any classifier observing X satisfies $\varepsilon_B \le \frac{1}{2}(1 - \Phi_\pi)$; with equal priors, $\varepsilon_B \le \frac{1}{2} - \frac{\Phi}{4}$.

Part (b) poses the question this paper answers. Fuzzification only smooths the data, so a fuzzy partition cannot carry more class information than the raw feature, or than a fine crisp partition. If fuzzy scores nevertheless select features better, the reason must lie in estimation, not in the population score.

4 Fuzziness as variance reduction

4.1 The variance–fuzziness identity

Theorem 1 (variance–fuzziness identity)**.** *For every term t, $\sigma_t^2 = m_t(1-m_t) - \mathrm{E}[\mu_t(X)(1-\mu_t(X))]$, and*

$$\sum_{t=1}^{T} \sigma_t^2 = G - \bar{F}.$$

If X has an atomless law, a crisp interval partition with the same term masses exists and has total variance G; the fuzzy partition reduces it by exactly $\bar{F}$, and the reduction is zero if and only if the partition is crisp almost surely.

Proof. Since $0 \le \mu_t \le 1$, $\mu_t^2 = \mu_t - \mu_t(1-\mu_t)$. Taking expectations, $\mathrm{E}\mu_t^2 = m_t - \mathrm{E}[\mu_t(1-\mu_t)]$ and $\sigma_t^2 = \mathrm{E}\mu_t^2 - m_t^2 = m_t(1-m_t) - \mathrm{E}[\mu_t(1-\mu_t)]$. Summing over t and using $\sum_t m_t = 1$ gives $\sum_t m_t(1-m_t) = G$ and $\sum_t \mathrm{E}[\mu_t(1-\mu_t)] = \bar{F}$. For a crisp partition each $\mu_t(X)$ is an indicator with mean m_t, so its variance is $m_t(1-m_t)$; for an atomless law such a partition with prescribed masses is obtained by cutting the line at the corresponding quantiles. Finally $\bar{F} = 0$ if and only if $\mu_t(X)(1-\mu_t(X)) = 0$ almost surely for all t. □

The identity is elementary; its value lies in what it controls. A crisp partition places each observation entirely in one term, so term counts are multinomial and have the largest variance possible for the given masses. A fuzzy partition spreads each observation over neighbouring terms, and the spread, measured by $\bar{F}$, is subtracted exactly. Theorem 1 concerns one observation; since $\hat{M}^c(t)$ averages n_c independent memberships, the sampling variance of an estimated profile is $\sum_t \mathrm{Var}\,\hat{M}^c(t) = (G^c - \bar{F}^c)/n_c$, against G^c/n_c for the crisp partition with the same masses. The rest of this section shows that this single quantity governs how profile-based scores behave on irrelevant features.

4.2 Chance scores of irrelevant features

Ranking errors in filter methods occur when an irrelevant feature obtains a large score by chance. The next two results quantify these chance scores.

Theorem 2 (bias and chance scores)**.**

$$\Phi \;\le\; \mathrm{E}\hat{\Phi} \;\le\; \Phi + \sqrt{T\Big(\frac{V^1}{n_1} + \frac{V^0}{n_0}\Big)}.$$

In particular, for an irrelevant feature ($P^1 = P^0$), $\Phi = 0$ and $\mathrm{E}\hat{\Phi} \le \sqrt{T\,h\,(G - \bar{F})}$, whereas the same bound for the crisp partition with the same term masses is $\sqrt{T\,h\,G}$; their ratio is $\sqrt{1 - \bar{F}/G}$.

Proof. Let $D_t = \hat{M}^1(t) - \hat{M}^0(t)$, so that $\mathrm{E}D_t = M^1(t) - M^0(t)$ and $\hat{\Phi} = \sum_t |D_t|$. Jensen's inequality gives $\mathrm{E}|D_t| \ge |\mathrm{E}D_t|$, hence the lower bound. For the upper bound, $|D_t| \le |\mathrm{E}D_t| + |D_t - \mathrm{E}D_t|$ and $\mathrm{E}|D_t - \mathrm{E}D_t| \le \sqrt{\mathrm{Var}\,D_t}$, so by the Cauchy–Schwarz inequality $\mathrm{E}\hat{\Phi} \le \Phi + \sqrt{T\sum_t \mathrm{Var}\,D_t}$. By independence, $\mathrm{Var}\,D_t = \mathrm{Var}_1\mu_t/n_1 + \mathrm{Var}_0\mu_t/n_0$, and Theorem 1 under each class gives $\sum_t \mathrm{Var}\,D_t = V^1/n_1 + V^0/n_0$. □

Theorem 2 compares upper bounds. The next result shows that, for large samples, the chance scores themselves are smaller, and that the same mechanism acts on information-based scores. For the latter we use the plug-in mutual information $\hat{I}$ between the class and the (fuzzy) term, computed from the sigma-count table $\hat{p}(t,c) = \frac{1}{N}\sum_{i:\,Y_i=c}\mu_t(X_i)$; with a crisp partition, $\hat{I}$ is the usual mutual

information of the discretised feature, and with a fuzzy one it is a sigma-count (fuzzy) mutual information, in the spirit of fuzzy information measures [7]. We call it fuzzy MI below.

Theorem 3 (asymptotic chance scores)**.** *Let the feature be irrelevant ($P^1 = P^0$, with term masses m_t and variances σ_t^2), let the partition be fixed, and let $n_1, n_0 \to \infty$ with $n_1/N \to p_1 \in (0,1)$.*

(a) $h^{-1/2}\,\mathrm{E}\hat{\Phi} \to \sqrt{2/\pi}\,\sum_t \sigma_t$.

(b) If $m_t > 0$ for all t, then $N\hat{I}$ converges in distribution to a random variable with mean

$$\frac{1}{2}\sum_t \frac{\sigma_t^2}{m_t} = \frac{1}{2}\Big[(T-1) - \sum_t \frac{\mathrm{E}[\mu_t(X)(1-\mu_t(X))]}{m_t}\Big].$$

In both cases the crisp partition with the same term masses gives $\sigma_t^2 = m_t(1-m_t)$, that is, the limits $\sqrt{2/\pi}\sum_t \sqrt{m_t(1-m_t)}$ and $(T-1)/2$; the fuzzy limits are smaller term by term whenever $\mathrm{E}[\mu_t(1-\mu_t)] > 0$.

Proof. (a) D_t is a difference of two independent sample means of the bounded variable $\mu_t(X)$, with mean 0 and variance $\sigma_t^2 h$. If $\sigma_t = 0$, $D_t = 0$ almost surely. Otherwise, by the central limit theorem $D_t/(\sigma_t\sqrt{h})$ converges in distribution to a standard normal variable; its second moment equals 1 for all sample sizes, so $|D_t|/(\sigma_t\sqrt{h})$ is uniformly integrable and $h^{-1/2}\mathrm{E}|D_t| \to \sigma_t\sqrt{2/\pi}$. Sum over t.

(b) Let $q_c = n_c/N$ and $\hat{m} = q_1\hat{M}^1 + q_0\hat{M}^0$, the estimated term masses. The plug-in mutual information is $\hat{I} = \sum_c q_c\,\mathrm{KL}(\hat{M}^c \,\|\, \hat{m})$, and $\hat{M}^1 - \hat{m} = q_0 D$, $\hat{M}^0 - \hat{m} = -q_1 D$. For probability vectors P, Q with Q in a neighbourhood of m, a second-order Taylor expansion gives $\mathrm{KL}(P\|Q) = \frac{1}{2}\sum_t (P_t - Q_t)^2/Q_t + R$ with $|R| \le C\|P-Q\|^3$. Hence

$$N\hat{I} = \tfrac{1}{2}\,q_1 q_0 \sum_t \frac{(\sqrt{N}D_t)^2}{\hat{m}_t} + O_p(N^{-1/2}).$$

By the multivariate central limit theorem $\sqrt{N}D$ converges to a centred Gaussian vector W with $\mathrm{Var}\,W_t = \sigma_t^2/(p_1 p_0)$, because $Nh \to 1/(p_1 p_0)$; and $\hat{m} \to m$ almost surely. By Slutsky's lemma and the continuous mapping theorem, $N\hat{I}$ converges in distribution to $Q = \frac{1}{2}p_1 p_0 \sum_t W_t^2/m_t$, whose mean is $\frac{1}{2}\sum_t \sigma_t^2/m_t$. The second expression follows from Theorem 1 term by term. □

For a crisp partition, part (b) is the classical result that $2N\hat{I}$ is asymptotically χ^2_{T-1} under independence, the basis of the Miller–Madow bias correction [19, 20]. Part (b) shows that a fuzzy partition lowers this chance level by $\frac{1}{2}\sum_t \mathrm{E}[\mu_t(1-\mu_t)]/m_t$, and it gives the corresponding fuzzy correction $\hat{I} - \frac{1}{2N}\sum_t \hat{\sigma}_t^2/\hat{m}_t$. Theorems 2 and 3 together show that the variance–fuzziness identity is not specific to FSDS: it controls the chance level of any score that is a smooth or Lipschitz function of the class profiles.

4.3 Concentration

Proposition 2 (concentration)**.** *For every* $\varepsilon > 0$, $\Pr(|\hat{\Phi} - \mathrm{E}\hat{\Phi}| \geq \varepsilon) \leq 2\exp\{-\varepsilon^2/(2h)\}$. *Consequently, if* $n_1 = n_0 = n$, *then with probability at least* $1 - \delta$,

$$|\hat{\Phi} - \Phi| \leq \sqrt{\frac{T(V^1 + V^0)}{n}} + 2\sqrt{\frac{\ln(2/\delta)}{n}}.$$

The proof (Appendix A) uses McDiarmid's inequality [24]. The bound is distribution-free and therefore conservative, but it shows that $\hat{\Phi}$ is a consistent estimate of Φ at the rate $n^{-1/2}$ for every strong fuzzy partition.

4.4 The variance–discrimination trade-off

Increasing the overlap of the partition has two effects. By Proposition 1(b), smoothing can only reduce the population score (signal). By Theorems 1–3, it reduces the sampling variance of the profiles and the chance scores of irrelevant features by amounts governed by $\bar{F}$ (noise). Fuzziness therefore helps ranking when the reduction of noise outweighs the loss of signal. This is expected when class differences are spread smoothly over the range of the feature, and not expected when the difference is concentrated in a region narrower than the overlap of the terms. Section 5 confirms both cases.

4.5 A debiased and adaptive score

Because the best degree of fuzziness depends on the unknown kind of class difference, we also consider choosing it from the data. Theorem 3(a) suggests the estimated chance level $\hat{b}_0 = \sqrt{2/\pi}\,\sum_t \sqrt{\hat{v}_t h}$, where $\hat{v}_t$ is the pooled sample variance of $\mu_t(X)$, and the standardised score

$$Z = \frac{\hat{\Phi} - \hat{b}_0}{\hat{s}_0}, \qquad \hat{s}_0 = \sqrt{\Big(1 - \frac{2}{\pi}\Big)\sum_t \hat{v}_t h}.$$

The scale $\hat{s}_0$ ignores the dependence between terms induced by $\sum_t \mu_t = 1$, so it is a normaliser rather than an exact standard deviation. Under irrelevance (4,000 replications, $n = 30$ and 300 per class, crisp and triangular partitions), Z had a simulated mean within ± 0.015 of zero and a standard deviation between 1.20 and 1.22, so Z is correctly centred but slightly over-dispersed; we use it only to put features on a common scale. The adaptive score of a feature is $\max_s Z(s)$ over the grid $s \in \{0, 0.25, 0.5, 0.75, 1\}$. Taking a maximum introduces a selection effect, but the effect applies equally to all features.

5 Simulation study

Design. All simulations used $T = 3$ terms with centres $-1, 0, 1$ and the overlap grid $s \in \{0, 0.25, 0.5, 0.75, 1\}$. Population quantities were computed from 2×10^6 draws of $N(0, 1)$. Chance

Table 1: Verification of Theorems 1–3 ($T = 3$, centres $-1, 0, 1$, $N(0,1)$ data). Columns 2–4: population quantities; chance FSDS: mean of $\hat{\Phi}$ for an irrelevant feature; $N\hat{I}$: mean of N times the binned mutual information for an irrelevant feature (4,000 replications).

				Chance FSDS, $n = 30$			$N\hat{I}$, $n = 300$	
Overlap s	$\sum_t \sigma_t^2$	$G - \bar{F}$	$\bar{F}$	Simulated	Bound (Thm 2)	Limit (Thm 3)	Simulated	Limit (Thm 3)
0 (crisp)	0.6629	0.6629	0.0000	0.294	0.364	0.290	1.006	1.000
0.25	0.6045	0.6045	0.0586	0.274	0.348	0.277	0.917	0.914
0.5	0.5468	0.5468	0.1167	0.267	0.331	0.264	0.825	0.829
0.75	0.4899	0.4899	0.1741	0.251	0.313	0.250	0.745	0.745
1 (triangular)	0.4344	0.4344	0.2304	0.236	0.295	0.235	0.667	0.661

scores were estimated for an irrelevant $N(0,1)$ feature with $n = 30, 100, 300$ and 1000 observations per class and 4,000 replications. Proposition 1 was checked by numerical integration for Gaussian mean shifts $\delta \in \{0.25, 0.5, 1, 2\}$. In the ranking experiment, each replication contained one relevant feature and 20 irrelevant $N(0,1)$ features, and we recorded whether the relevant feature obtained the highest score (ties broken at random). Four kinds of class difference were used: mean shift, $N(0.3,1)$ against $N(0,1)$; scale change, $N(0,1.4^2)$ against $N(0,1)$; skewed shift, $e^{0.25+0.6Z} - 1.2$ against $e^{0.6Z} - 1.2$ with $Z \sim N(0,1)$; narrow bump, a mixture in which 15% of the positive class is drawn from $N(0.75, 0.05^2)$ and the rest from $N(0,1)$, against $N(0,1)$. Sample sizes were 30, 100 and 300 per class with 1,500 replications, and all scores were computed on the same replications. Besides FSDS on each partition and the adaptive score, we included binned mutual information on the crisp ($s = 0$) and triangular ($s = 1$) partitions, the ANOVA F statistic and the two-sample Kolmogorov–Smirnov (KS) statistic [25].

Verification of the theory. Table 1 and Fig. 1 confirm every prediction. The identity of Theorem 1 held to machine precision at every overlap level: the total variance fell from 0.663 for the crisp partition to 0.434 for the triangular one, while $\bar{F}$ rose from 0 to 0.230 and G stayed between 0.663 and 0.665. The bound of Theorem 2 held in all 20 settings. The limit of Theorem 3(a) was accurate already at $n = 30$ (0.290 and 0.235 predicted for the crisp and triangular partitions, 0.294 and 0.236 simulated). The chance score of FSDS was about 20% lower for the triangular partition at every sample size, and the chance level of mutual information was 34% lower ($N\mathrm{E}\hat{I} = 0.667$ against 1.006 at $n = 300$, limits 0.661 and 1). Proposition 2 held in all 20 settings (empirical exceedance probabilities at most 0.0003 against the nominal 0.05), which shows that the distribution-free bound is conservative. Proposition 1(b) and (c) held for all mean shifts and both partitions.

The variance–discrimination trade-off. Fig. 2 shows how the probability of ranking the relevant feature first depends on the overlap. For the three smooth kinds of class difference it increased with s, apart from small fluctuations at $n = 30$; at $n = 100$ it rose from 0.30 to 0.41 for the mean shift, from 0.24 to 0.30 for the scale change and from 0.41 to 0.48 for the skewed shift. For the narrow bump the pattern reversed, as predicted in Section 4.4: slight fuzziness helped, but full fuzziness blurred the localised difference, and at $n = 300$ the probability fell from 0.63 at $s = 0.25$ to 0.38 at $s = 1$. Binned mutual information behaved in the same way (Table 2): the triangular

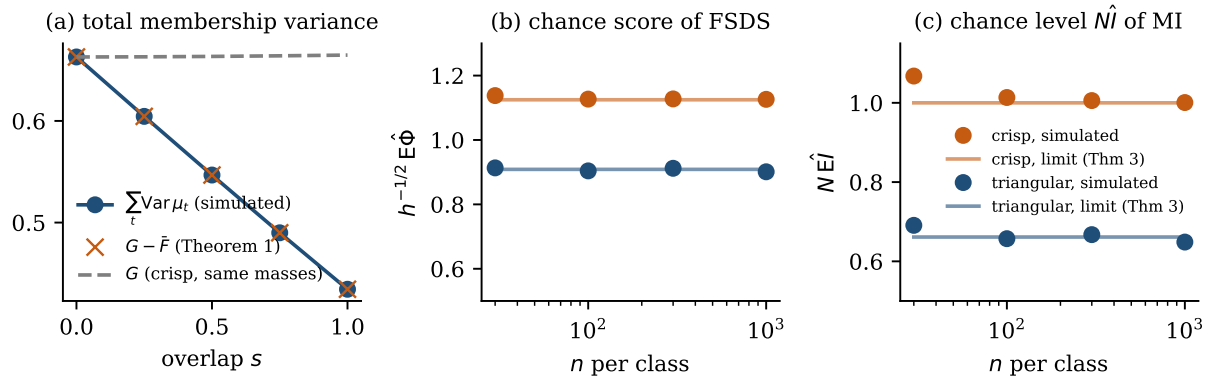

Figure 1: (a) Total variance of the term memberships against the overlap s, with the value given by Theorem 1 and the crisp reference G. (b) Scaled chance score $h^{-1/2}\mathrm{E}\hat{\Phi}$ of an irrelevant feature and the limit of Theorem 3(a). (c) Chance level $N\mathrm{E}\hat{I}$ of binned mutual information and the limit of Theorem 3(b).

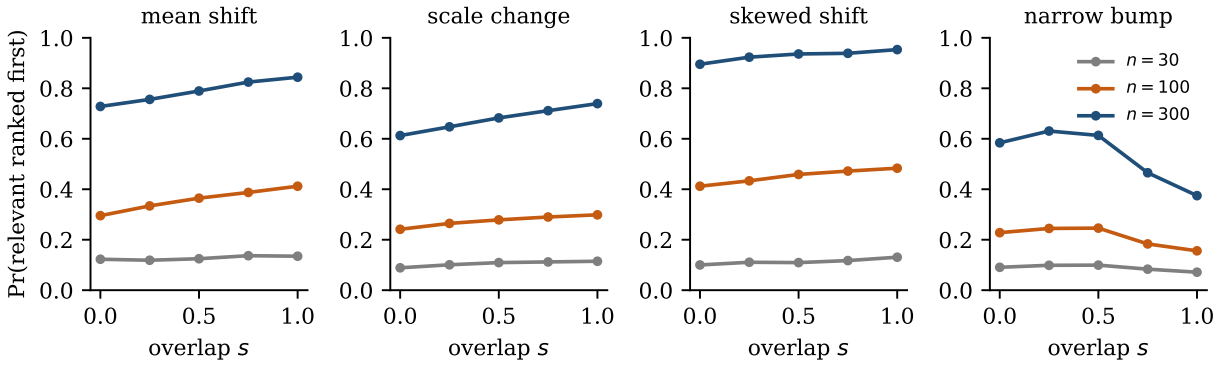

Figure 2: Probability that the relevant feature is ranked first among 21 features by FSDS, against the overlap s, for four kinds of class difference and three sample sizes per class (1,500 replications per point).

partition improved it for the three smooth differences (for example, from 0.32 to 0.42 for the mean shift at $n = 100$) and harmed it for the narrow bump.

Comparison with other filters. The adaptive score avoided the failure case: it was the best FSDS variant for the skewed shift and, for the narrow bump, matched the crisp partition instead of the much weaker triangular one, at the cost of some power for the mean and scale differences, where searching over s is unnecessary. The ANOVA F statistic, the most powerful test for location differences between normal distributions, was the best filter for the mean shift and the skewed shift, but it detects only differences in means: for the scale change it ranked the relevant feature first in about 5% of replications, the chance level, whereas FSDS did so in 30% at $n = 100$ and 74% at $n = 300$. The KS statistic was competitive for location differences and the best filter in Table 2 for the narrow bump at $n = 300$, but clearly weaker than FSDS for the scale change (0.18 against 0.30

Table 2: Probability that the relevant feature is ranked first among 21 features (1,500 replications per setting; chance level $1/21 \approx 0.048$). Crisp: $s = 0$; FSDS and fuzzy MI: triangular partition, $s = 1$.

Class difference	n	Crisp FSDS	FSDS	Adaptive	Crisp MI	Fuzzy MI	ANOVA	KS
Mean shift	30	0.123	0.135	0.125	0.121	0.141	0.173	0.135
	100	0.295	0.412	0.379	0.323	0.423	0.478	0.397
	300	0.728	0.844	0.816	0.761	0.863	0.904	0.822
Scale change	30	0.089	0.115	0.096	0.091	0.113	0.049	0.087
	100	0.241	0.299	0.257	0.238	0.298	0.057	0.175
	300	0.613	0.739	0.681	0.611	0.733	0.047	0.557
Skewed shift	30	0.100	0.131	0.213	0.187	0.192	0.249	0.235
	100	0.412	0.483	0.609	0.564	0.624	0.664	0.632
	300	0.895	0.953	0.975	0.962	0.983	0.991	0.977
Narrow bump	30	0.091	0.071	0.095	0.089	0.078	0.075	0.098
	100	0.228	0.156	0.219	0.216	0.155	0.101	0.217
	300	0.584	0.375	0.587	0.567	0.385	0.262	0.624

at $n = 100$). No single filter was best for every kind of class difference, which is the usual situation for filters; the point of the experiment is that, within each family of profile-based scores, the effect of fuzziness follows the theory.

6 Benchmark study

6.1 Design

Datasets. The datasets were taken from the Penn Machine Learning Benchmarks (PMLB) [26, 27]. Using the PMLB summary statistics, we selected all binary classification datasets with 150 to 5,000 instances, at least four continuous features and at most 80 features. Of the 37 datasets meeting these criteria, 15 were deprecated earlier copies of current datasets and were removed. Two further datasets, dis and hypothyroid, were excluded because their minority class (1.5% and 4.8% of cases) leaves only about three to five minority cases in a subsample of 100. Features with at least 10 distinct values were treated as numeric and retained. Table 3 lists the 20 datasets.

Methods. Ten filters were compared. Four are profile-based and use the Low/Medium/High partition with centres at the 5th, 50th and 95th percentiles of the training data: crisp FSDS ($s = 0$), FSDS ($s = 1$), the adaptive score of Section 4.5, and binned mutual information on the crisp and triangular partitions (crisp MI and fuzzy MI). The others are the fuzzy-rough dependency degree of the class on each feature [6] with the similarity $\max\{0, 1 - |x - z|/\mathrm{sd}\}$ and the Łukasiewicz implicator, the KS statistic, the ANOVA F statistic, mutual information with the nearest-neighbour estimator [28, 29] and ReliefF [30, 31] with 10 neighbours. ReliefF is multivariate; it is included as a widely used reference filter. In every dataset the top $k = \max(3, \mathrm{round}(0.25p))$ features were selected.

Table 3: Benchmark datasets: instances n, numeric features p, share of the positive class (Pos.) and number of selected features k.

Dataset	n	p	Pos.	k	Dataset	n	p	Pos.	k
backache	180	7	0.14	3	hepatitis	155	6	0.79	3
biomed	209	7	0.64	3	horse colic surgery	368	7	0.37	3
breast cancer wisconsin diagnostic	569	30	0.37	8	ionosphere	351	32	0.64	8
bupa	345	5	0.51	3	prnn crabs	200	6	0.50	3
churn	5000	17	0.14	4	profb	672	6	0.33	3
credit approval australia	690	7	0.56	3	saheart	462	8	0.35	3
glass2	163	8	0.47	3	sonar	208	60	0.47	15
heart disease cleveland	303	5	0.46	3	spambase	4601	57	0.39	14
heart disease hungarian	294	5	0.36	3	spectf	349	44	0.73	11
heart disease va long beach	200	5	0.74	3	tokyo1	959	37	0.64	9

Criteria. (1) *Stability*: in 100 stratified random subsamples, the average pairwise Kuncheva index [21] of the selected sets, in two regimes, half of the data and a small-sample regime of at most 100 cases. (2) *Spurious selections*: in each subsample every numeric feature was duplicated and its values were randomly permuted, which creates features with realistic marginal distributions but no relation to the class; the measure is the share of the top k positions taken by such noise features. (3) *Accuracy*: five repetitions of stratified 5-fold cross-validation, with feature selection inside each training fold, and the balanced accuracy on the test fold of three classifiers trained on the selected features: logistic regression (standardised features, L_2 penalty, $C = 1$), k-nearest neighbours ($k = 5$, standardised features) and Gaussian naive Bayes. Missing values were replaced by the median of the training data available to the method, and imputation, partition centres and scores were always computed from the subsample or training fold alone.

Statistical comparison. For each dataset and criterion, results were first averaged over the subsamples or folds, so the dataset is the independent unit of every test. Differences among the ten methods were tested with the Friedman test and pairs with the two-sided Wilcoxon signed-rank test [32]; as effect size we report the matched-pairs rank-biserial correlation r ($|r| = 1$ when one method is better on every dataset). The comparisons of each fuzzy score with its crisp counterpart are the ones predicted by the theory; the others are exploratory and not corrected for multiplicity.

Implementation and use of AI tools. All computations used Python 3.13 with NumPy 2.5, SciPy 1.18 and scikit-learn 1.9 [29]; every random generator was seeded, and the code is provided as supplementary material. A large language model (Claude, Anthropic) was used to assist in checking derivations, writing the simulation and analysis code and editing the text. All derivations, code and results were checked by the authors, who take full responsibility for them.

6.2 Results

Table 4 summarises the results and Fig. 3 shows the small-sample regime dataset by dataset.

Fuzzy against crisp. The predictions of Section 4 were confirmed for both profile-based scores. In the small-sample regime FSDS was more stable than crisp FSDS in 18 of the 20 datasets (mean

Table 4: Results on 20 benchmark datasets: mean over datasets, with the average rank among the ten methods in parentheses (1 = best). Stability: Kuncheva index of the top-k sets over 100 subsamples of half of the data or at most 100 cases. Noise share: share of the top k taken by permuted noise features. Balanced accuracy: 5×5-fold cross-validation with selection inside each training fold.

	Stability ↑		Noise share ↓		Balanced accuracy ↑		
Method	Half	≤ 100	Half	≤ 100	LR	kNN	NB
Crisp FSDS	0.559 (6.5)	0.409 (6.8)	0.097 (6.5)	0.145 (6.8)	0.689 (6.3)	0.709 (5.2)	0.694 (5.5)
FSDS	0.634 (4.3)	0.498 (4.1)	0.070 (4.3)	0.112 (4.2)	0.692 (6.0)	0.704 (6.2)	0.693 (5.3)
Adaptive FSDS	0.607 (5.2)	0.472 (5.0)	0.075 (4.9)	0.123 (5.1)	0.691 (5.8)	0.705 (6.1)	0.696 (5.0)
Crisp MI	0.590 (5.5)	0.443 (5.9)	0.081 (5.0)	0.124 (5.2)	0.696 (5.5)	0.711 (4.3)	0.697 (5.1)
Fuzzy MI	0.681 (3.3)	0.519 (3.5)	0.065 (3.2)	0.100 (2.7)	0.695 (5.1)	0.709 (4.9)	0.696 (4.8)
Fuzzy-rough	0.531 (6.4)	0.372 (7.2)	0.183 (7.8)	0.213 (7.8)	0.681 (6.6)	0.700 (5.8)	0.678 (7.4)
KS	0.649 (4.9)	0.483 (5.2)	0.080 (4.4)	0.113 (4.0)	0.701 (4.0)	0.711 (4.9)	0.688 (5.3)
ANOVA	0.639 (5.3)	0.488 (4.8)	0.075 (4.5)	0.119 (5.0)	0.698 (4.8)	0.704 (5.8)	0.694 (4.5)
MI (kNN)	0.427 (8.8)	0.314 (8.4)	0.162 (7.8)	0.203 (7.8)	0.698 (5.1)	0.708 (6.8)	0.686 (6.4)
ReliefF	0.614 (4.8)	0.522 (4.2)	0.116 (6.5)	0.165 (6.4)	0.703 (6.0)	0.717 (5.0)	0.683 (5.7)
Friedman p	2×10^{-6}	4×10^{-7}	2×10^{-10}	2×10^{-10}	0.16	0.24	0.09

Kuncheva index 0.498 against 0.409, $p < 0.001$, $r = 0.91$), and with half of the data in 17 (0.634 against 0.559, $p = 0.001$, $r = 0.78$). It admitted fewer noise features: 11.2% against 14.5% of the top k in the small-sample regime ($p < 0.001$, $r = 0.96$; fewer in 15 datasets, more in one) and 7.0% against 9.7% with half of the data ($p = 0.004$, $r = 0.87$). Fuzzy MI improved on crisp MI in the same way: stability 0.519 against 0.443 (15 of 20 datasets, $p < 0.001$, $r = 0.82$) and noise share 10.0% against 12.4% ($p < 0.001$, $r = 0.99$) in the small-sample regime. Classification accuracy did not change significantly ($p \geq 0.07$ for every classifier and both pairs).

Comparison with other filters. Fuzzy MI had the best average rank of all ten methods for stability and for noise share, in both regimes. This agrees with Theorem 3, which predicts a larger relative reduction of chance scores for mutual information (34% in Table 1) than for FSDS (20%). FSDS did not differ significantly from ANOVA, KS or ReliefF in stability ($p \geq 0.36$ in both regimes); it was more stable than the fuzzy-rough dependency degree in the small-sample regime ($p = 0.006$) and far more stable than nearest-neighbour MI ($p < 0.001$). It admitted fewer noise features than the fuzzy-rough degree, nearest-neighbour MI and ReliefF ($p \leq 0.011$ in both regimes). The fuzzy-rough degree and nearest-neighbour MI, which both depend on distances between individual observations, were the least stable scores and admitted the most noise. The adaptive score was slightly less stable than FSDS (small-sample regime 0.472 against 0.498, $p = 0.033$), which suggests that smooth class differences predominate in these datasets and that the search over s costs some stability when it is not needed. Balanced accuracy did not differ significantly among the ten methods for any classifier (Friedman $p = 0.16$, 0.24 and 0.09 for logistic regression, kNN and naive Bayes); the means ranged from 0.68 to 0.72.

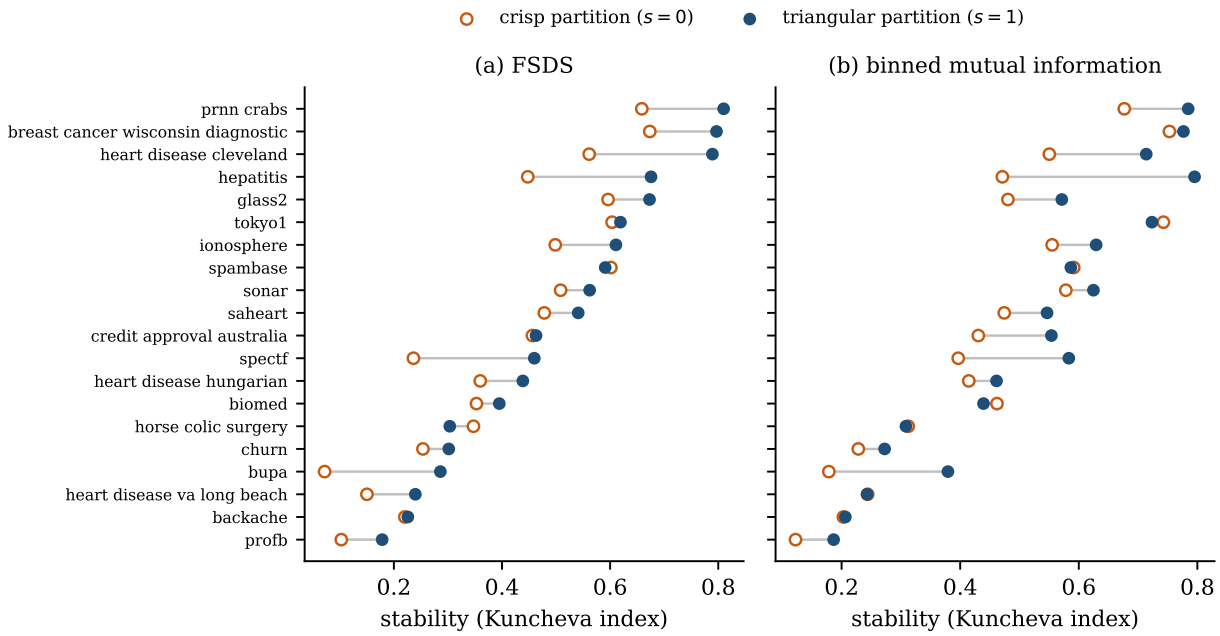

Figure 3: Stability of the selected sets in the small-sample regime (at most 100 training cases) on 20 benchmark datasets, for the crisp (open circles) and triangular (filled circles) partitions. (a) FSDS. (b) Binned mutual information. Datasets are ordered by the stability of FSDS.

7 Discussion

The analysis gives a precise answer to the question in the introduction. Fuzziness cannot increase the discrimination a feature contains (Proposition 1). What it does is reduce the variance of the term memberships, and hence the sampling variance of the class profiles, by an amount fixed by the mean quadratic fuzziness of the partition (Theorem 1). Lower variance gives irrelevant features smaller chance scores (Theorems 2 and 3), for L_1-type and information-type scores alike, and therefore fewer ranking errors. The benefit is largest in small samples, where variance dominates, and it turns into a loss when the class difference is concentrated in a region narrower than the overlap of the terms. The simulations and the benchmark results both follow this pattern.

These results also clarify what can and cannot be claimed for fuzzy feature scores. A fuzzy score is not more powerful than a well-chosen parametric statistic when that statistic's assumptions hold, and it does not by itself improve classification accuracy. Its advantages are of a different kind: it is a bounded, nonparametric measure of separation with a direct error guarantee (Proposition 1); it captures differences in spread and shape that ANOVA misses; it is more stable than the same score on crisp intervals, for a reason that can be stated and quantified; and its output is a per-term profile that a domain expert can read. Theorem 3(b) also gives practitioners a concrete tool, the fuzzy Miller–Madow correction, for comparing fuzzy mutual information across features or partitions with different chance levels.

Limitations. First, the theory assumes a fixed partition. In the experiments the centres were

estimated from the same training data as the score, which introduces a small dependence; computing the centres on a separate split restores the conditions of the theorems exactly. Second, the analysis is for one feature and two classes. Multi-class versions follow by replacing $|M^1 - M^0|$ with pairwise or one-against-rest distances, and Theorem 3(b) extends directly to C classes, but interactions between features are outside the scope of a univariate score. Third, Proposition 2 is distribution-free and therefore conservative. Fourth, the standardisation in the adaptive score is a heuristic and is slightly over-dispersed. Fifth, percentile-based centres degenerate when one value dominates a feature: in 55 of the 359 benchmark features (50 of the 57 in spambase) two of the three centres coincided, so the partition had fewer effective terms and its fuzzy and crisp versions nearly agreed, which is why spambase shows no gain in Fig. 3. Finally, the benchmark datasets are of moderate size, and the benefits of fuzziness are expected to shrink as samples grow, as Theorem 2 implies.

8 Conclusion

We have shown that the benefit of fuzzy partitions in feature scoring is a variance effect. The fuzzy soft discrimination score is a total variation distance computed through a strong fuzzy partition, and fuzzification can only lower it. The fuzziness of the partition, however, lowers the variance of the class profiles by exactly the mean quadratic fuzziness, and as a result the chance scores of irrelevant features are smaller, both for FSDS and for fuzzy mutual information, with explicit limits. Simulations confirmed every prediction, including the failure case of narrowly localised class differences, and on 20 benchmark datasets the fuzzy versions of both scores were more stable and admitted fewer noise features than their crisp versions, with unchanged classification accuracy. Because the identity holds for any strong fuzzy partition, the same reasoning applies to other scores built from class-conditional term frequencies; extending it to multivariate and fuzzy-rough measures is a natural next step.

Declarations

- **Funding.** This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

- **Competing interests.** The authors have no competing interests to declare that are relevant to the content of this article.

- **Ethics approval and consent to participate.** Not applicable. The study uses simulated data and public benchmark datasets only.

- **Consent for publication.** Not applicable.

- **Data availability.** All datasets are publicly available from the Penn Machine Learning Benchmarks repository (`https://github.com/EpistasisLab/pmlb`).

- **Code availability.** The simulation and analysis code is provided as supplementary material.
- **Author contributions.** S.K.R.: conceptualization, methodology, formal analysis, software, investigation, visualization, writing (original draft, review and editing). N.S.: validation (verification of proofs), writing (review and editing). S.I.K.: validation (verification of the mathematical results and of the reported numerical results), writing (review and editing). All authors read and approved the final manuscript.
- **Use of large language models.** The use of an AI tool in preparing this work is described in Section 6.

A Proofs of Propositions 1 and 2

Proof of Proposition 1. (a) By Definition 1, $K(\cdot \mid x)$ is a probability vector for every x, and $(P^c K)(t) = \int \mu_t \, dP^c = M^c(t)$. The identity $\Phi = 2\,\mathrm{TV}(M^1, M^0)$ is the definition of total variation on a finite set.

(b) Let $\nu = \pi P^1 - (1-\pi)P^0$ with total variation measure $|\nu|$, so $\|\nu\| = |\nu|(\mathbb{R})$. Since $\mu_t \geq 0$, $|\int \mu_t \, d\nu| \leq \int \mu_t \, d|\nu|$. Summing over t and using $\sum_t \mu_t = 1$ gives $\Phi_\pi \leq \|\nu\|$. With $\pi = \frac{1}{2}$, $\Phi/2 \leq \frac{1}{2}\|P^1 - P^0\| = \mathrm{TV}(P^1, P^0)$.

(c) For any observation Z with class-conditional laws Q^1, Q^0, the Bayes error is $\int \min\{\pi \, dQ^1, (1-\pi) \, dQ^0\} = \frac{1}{2}(1 - \|\pi Q^1 - (1-\pi)Q^0\|)$, because $\min(a,b) = \frac{1}{2}(a + b - |a-b|)$ [18]. Apply this to $Z = X$ and to a term drawn from $K(\cdot \mid X)$, whose class-conditional laws are M^1, M^0. By (b) the first norm is at least the second, which is Φ_π. □

Proof of Proposition 2. $\hat{\Phi}$ is a function of N independent observations. Replacing one observation x of class 1 by x' changes $\hat{M}^1$ by the vector $(\mu(x') - \mu(x))/n_1$, whose L_1 norm is at most $2/n_1$ because $\mu(x)$ and $\mu(x')$ are probability vectors. By the triangle inequality $\hat{\Phi}$ changes by at most $c_i = 2/n_1$, and similarly by at most $2/n_0$ for class 0. Hence $\sum_i c_i^2 = 4h$, and McDiarmid's inequality gives $\Pr(|\hat{\Phi} - \mathrm{E}\hat{\Phi}| \geq \varepsilon) \leq 2\exp(-2\varepsilon^2 / \sum_i c_i^2) = 2\exp\{-\varepsilon^2/(2h)\}$.

For the second statement let $n_1 = n_0 = n$, so $h = 2/n$, and take $\varepsilon = 2\sqrt{\ln(2/\delta)/n}$; then $|\hat{\Phi} - \mathrm{E}\hat{\Phi}| \leq \varepsilon$ with probability at least $1 - \delta$. Since $\Phi \leq \mathrm{E}\hat{\Phi}$ (Theorem 2), $\Phi - \hat{\Phi} \leq \mathrm{E}\hat{\Phi} - \hat{\Phi} \leq \varepsilon$ and $\hat{\Phi} - \Phi \leq (\hat{\Phi} - \mathrm{E}\hat{\Phi}) + (\mathrm{E}\hat{\Phi} - \Phi) \leq \varepsilon + \sqrt{T(V^1 + V^0)/n}$ by Theorem 2. □

References

[1] Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. J. Mach. Learn. Res. 3, 1157–1182 (2003)

[2] Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R.P., Tang, J., Liu, H.: Feature selection: a data perspective. ACM Comput. Surv. 50(6), 94 (2017). https://doi.org/10.1145/3136625

[3] Zadeh, L.A.: Fuzzy sets. Inf. Control 8(3), 338–353 (1965). https://doi.org/10.1016/S0019-9958(65)90241-X

[4] Hüllermeier, E.: Fuzzy sets in machine learning and data mining. Appl. Soft Comput. 11(2), 1493–1505 (2011)

[5] Jensen, R., Shen, Q.: Computational Intelligence and Feature Selection: Rough and Fuzzy Approaches. Wiley-IEEE Press, Hoboken, NJ (2008)

[6] Jensen, R., Shen, Q.: New approaches to fuzzy-rough feature selection. IEEE Trans. Fuzzy Syst. 17(4), 824–838 (2009)

[7] Hu, Q., Yu, D., Xie, Z., Liu, J.: Fuzzy probabilistic approximation spaces and their information measures. IEEE Trans. Fuzzy Syst. 14(2), 191–201 (2006)

[8] Yu, Y., Lv, M., Qian, J., Lv, J., Miao, D.: Fuzzy information gain ratio-based multi-label feature selection with label correlation. Int. J. Mach. Learn. Cybern. 15, 2737–2747 (2024). https://doi.org/10.1007/s13042-023-02060-9

[9] Rajbongshi, S.K.: Fuzzy soft sets for stroke risk feature selection: age-adjusted ranking, clinical validity and robustness to severe class imbalance. Int. J. Comput. Tech. 13(6) (2026). https://doi.org/10.5281/zenodo.24131866

[10] Ruspini, E.H.: A new approach to clustering. Inf. Control 15(1), 22–32 (1969). https://doi.org/10.1016/S0019-9958(69)90591-9

[11] Pedrycz, W.: Why triangular membership functions? Fuzzy Sets Syst. 64(1), 21–30 (1994). https://doi.org/10.1016/0165-0114(94)90003-5

[12] Kaufmann, A.: Introduction to the Theory of Fuzzy Subsets, vol. 1. Academic Press, New York (1975)

[13] De Luca, A., Termini, S.: A definition of a nonprobabilistic entropy in the setting of fuzzy sets theory. Inf. Control 20(4), 301–312 (1972). https://doi.org/10.1016/S0019-9958(72)90199-4

[14] Casalino, G., Castellano, G., Castiello, C., Mencar, C.: Effect of fuzziness in fuzzy rule-based classifiers defined by strong fuzzy partitions and winner-takes-all inference. Soft Comput. 26, 6519–6527 (2022)

[15] Fayyad, U.M., Irani, K.B.: Multi-interval discretization of continuous-valued attributes for classification learning. In: Proceedings of the 13th International Joint Conference on Artificial Intelligence, pp. 1022–1027 (1993)

[16] García, S., Luengo, J., Sáez, J.A., López, V., Herrera, F.: A survey of discretization techniques: taxonomy and empirical analysis in supervised learning. IEEE Trans. Knowl. Data Eng. 25(4), 734–750 (2013). https://doi.org/10.1109/TKDE.2012.35

[17] Cover, T.M., Thomas, J.A.: Elements of Information Theory, 2nd edn. Wiley, Hoboken, NJ (2006)

[18] Devroye, L., Györfi, L., Lugosi, G.: A Probabilistic Theory of Pattern Recognition. Springer, New York (1996)

[19] Miller, G.A.: Note on the bias of information estimates. In: Quastler, H. (ed.) Information Theory in Psychology: Problems and Methods, pp. 95–100. Free Press, Glencoe, IL (1955)

[20] Paninski, L.: Estimation of entropy and mutual information. Neural Comput. 15(6), 1191–1253 (2003). https://doi.org/10.1162/089976603321780272

[21] Kuncheva, L.I.: A stability index for feature selection. In: Proceedings of the IASTED International Conference on Artificial Intelligence and Applications, pp. 390–395, Innsbruck (2007)

[22] Nogueira, S., Sechidis, K., Brown, G.: On the stability of feature selection algorithms. J. Mach. Learn. Res. 18(174), 1–54 (2018)

[23] Maji, P.K., Biswas, R., Roy, A.R.: Fuzzy soft sets. J. Fuzzy Math. 9(3), 589–602 (2001)

[24] McDiarmid, C.: On the method of bounded differences. In: Siemons, J. (ed.) Surveys in Combinatorics. London Math. Soc. Lecture Note Ser., vol. 141, pp. 148–188. Cambridge University Press, Cambridge (1989)

[25] Massey, F.J., Jr.: The Kolmogorov–Smirnov test for goodness of fit. J. Am. Stat. Assoc. 46(253), 68–78 (1951). https://doi.org/10.1080/01621459.1951.10500769

[26] Olson, R.S., La Cava, W., Orzechowski, P., Urbanowicz, R.J., Moore, J.H.: PMLB: a large benchmark suite for machine learning evaluation and comparison. BioData Min. 10, 36 (2017). https://doi.org/10.1186/s13040-017-0154-4

[27] Romano, J.D., Le, T.T., La Cava, W., Gregg, J.T., Goldberg, D.J., Chakraborty, P., Ray, N.L., Himmelstein, D., Fu, W., Moore, J.H.: PMLB v1.0: an open-source dataset collection for benchmarking machine learning methods. Bioinformatics 38(3), 878–880 (2022). https://doi.org/10.1093/bioinformatics/btab727

[28] Ross, B.C.: Mutual information between discrete and continuous data sets. PLoS ONE 9(2), e87357 (2014). https://doi.org/10.1371/journal.pone.0087357

[29] Pedregosa, F., Varoquaux, G., Gramfort, A., et al.: Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830 (2011)

[30] Kononenko, I.: Estimating attributes: analysis and extensions of RELIEF. In: Bergadano, F., De Raedt, L. (eds.) Machine Learning: ECML-94. Lecture Notes in Computer Science, vol. 784, pp. 171–182. Springer, Berlin (1994)

[31] Robnik-Šikonja, M., Kononenko, I.: Theoretical and empirical analysis of ReliefF and RReliefF. Mach. Learn. 53, 23–69 (2003). https://doi.org/10.1023/A:1025667309714

[32] Demšar, J.: Statistical comparisons of classifiers over multiple data sets. J. Mach. Learn. Res. 7, 1–30 (2006)

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementarycodeFSDS.zip