A Geometric Theory of Feature Selection and Distance-Based Measures

Kilho Shin and Adrian Pino Angulo
University of Hyogo
Kobe, Japan
yshin@ai.u-hyogo.ac.jp

Abstract

Feature selection measures are often explained by the analogy to a rule to measure the "distance" of sets of features to the "closest" ideal sets of features. An ideal feature set is such that it can determine classes uniquely and correctly. This way of explanation was just an analogy before this paper. In this paper, we show a way to map arbitrary feature sets of datasets into a common metric space, which is indexed by a real number p with $1 \leq p \leq \infty$. Since this determines the distance between an arbitrary pair of feature sets, even if they belong to different datasets, the distance of a feature set to the closest ideal feature set can be used as a feature selection measure. Surprisingly, when $p = 1$, the measure is identical to the Bayesian risk, which is probably the feature selection measure that is used the most widely in the literature. For $1 < p \leq \infty$, the measure is novel and has significantly different properties from the Bayesian risk. We also investigate the correlation between measurements by these measures and classification accuracy through experiments. As a result, we show that our novel measures with $p > 1$ exhibit stronger correlation than the Bayesian risk.

1 Introduction

Feature selection is indeed one of the central focuses of machine learning research. In this paper, we understand the feature selection problem as follows:

> *Given a dataset, select an appropriately small number of features that faithfully determine the classes of the examples of the dataset.*

If we could find a feature subset such that the features correctly determine the classes of the examples, it could be an answer we wanted [Almuallim and Dietterich, 1994]. Such feature subsets are referred to as *reducts* in the rough set theory [Pawlak, 1991] and as *consistent feature sets* in this paper [Liu *et al.*, 1998].

A dataset, however, does not necessarily include a consistent feature subset, and in such cases, we have to be satisfied with selecting feature subsets that are sufficiently close to being consistent. In other words, we need a measure that can measure the "distance" of a feature subset to the "closest" consistent feature set. However, we only know two necessary conditions for such a measure, namely, *determinacy* and *monotonicity*: Determinacy corresponds to the identity of indiscernibles of the axioms of metrics and requires that the measurement is zero, if, and only if, the feature set is consistent; Monotonicity, on the other hand, requires that the measurement of an arbitrary feature set is no greater than the measurement of its subset. When a measure has the determinacy and monotonicity properties, we also say that it is *determinant* and *monotonous*. With a determinant and monotonous measure, a feature selection algorithm attempts to find feature sets whose measurements are close to zero.

Generally, the feature selection measures proposed in the literature are categorized into two groups. One consists of measures that simply evaluate dependence between two features [Mengle and Goharian, 2009; Bolón-Canedo *et al.*, 2011; Foithong *et al.*, 2012; Suzuki and Sugiyama, 2013; Wang, 2015]. The framework of *mRMR* [Peng *et al.*, 2005] takes advantage of measures of this group and aims to select a feature set that maximizes the difference of the *relevance* from the *redundancy* of the feature set: The relevance is the collective dependence of the feature set to class labels, while the redundancy is the the internal mutual dependence of the features of the feature set. On the other hand, the other consists of determinant and monotonous measures [Liu *et al.*, 1998; Shin and Xu, 2009; Arauzo-Azofra *et al.*, 2008], and 14 of the 17 feature selection measures studied in [Molina *et al.*, 2002] are known to belong to this group [Shin *et al.*, 2011]. In this paper, we are interested in the latter group.

Example 1 *Bayesian risk, also known as inconsistency rate [Liu* et al.*, 1998], is defined as follows.*

$$\mu_{\mathsf{br}}{}^{D}(S) = \sum_{\boldsymbol{x} \in \Omega_S} \Big(\boldsymbol{p}(S = \boldsymbol{x}) - \max_{c \in \Omega_{\mathrm{C}}} \boldsymbol{p}(S = \boldsymbol{x}, \mathrm{C} = c) \Big).$$

S is a feature subset of a dataset D, and C is the random variable to represent classes. Ω_S and Ω_{C} are the sample spaces of S and C, and $\boldsymbol{p}$ is the empirical probability distribution of D. This measure is, indeed, determinant and monotonous.

Example 2 *Zhao et al. [Zhao and Liu, 2007] proposed* Interact*, a feature selection algorithm that leverages the Bayes*

risk μ_{br} to evaluate the closeness of feature sets to the state of being consistent. Therefore, Interact selects feature sets with small μ_{br} measurements.

Example 3 *Shin et al. [Shin and Xu, 2009] also proposed the conditional entropy defined by*

$$\mu_{\mathsf{ce}}{}^{D}(S) = \sum_{\boldsymbol{x}\in\Omega_S}\sum_{c\in\Omega_{\mathrm{C}}} -\boldsymbol{p}(S=\boldsymbol{x},\mathrm{C}=c)\log\frac{\boldsymbol{p}(S=\boldsymbol{x},\mathrm{C}=c)}{\boldsymbol{p}(S=\boldsymbol{x})}$$

as a determinant and monotonous measure. On the other hand, the well-known feature selection algorithm that selects the top n features X with respect to the mutual information $I(X;\mathrm{C})$ actually selects those feature sets that minimize

$$\mu^{D}(\{X_1,\ldots,X_n\}) = \sum_{i=1}^{n}\mu_{\mathsf{ce}}{}^{D}(\{X_i\}).$$

By using μ^D instead of $\mu_{\mathsf{ce}}{}^D$, this algorithm is fast but is not very accurate, since μ^D is not determinant.

When a measure represents the distance of a feature set to the closest consistent feature set in some metric space, we call it *distance-based*. About distance-based measures, we only knew two necessary conditions, that is, determinacy and monotonicity, and hence, the distance-based measure was only an analogy to explain the desirable natures of feature selection measures. This paper changes this.

In this paper, we present a way to identify a feature subset of a dataset uniquely with a point in a common metric space. The metric space is parameterized by a real number $p \in [1,\infty]$, and hence, we have infinitely many such metric spaces. In the metric space, consistent feature sets form a closed subspace, and hence, we can determine the distance of an arbitrary feature set to the closest consistent feature set. This is nothing other than a distance-based measure. A different values of p determines a different metric function. For $p = 1$, we show that the measure is identical to the well-known Bayesian risk. On the other hand, for $p > 1$, the measure is novel and unknown in the literature. Also, through experiments, we show that measurements by our novel distance-based measures correlate with classification accuracy.

2 Formulating feature selection measures

In this paper, we do not discriminate between features and random variables. Furthermore, for a dataset D, we use the following notations.

- $\overline{\mathcal{F}}_D$ is the entire set of features of D. $\overline{\mathcal{F}}_D$ is finite and includes the feature C_D that represents classes. Also, we denote $\overline{\mathcal{F}}_D \setminus \{\mathrm{C}_D\}$ by $\mathcal{F}_D$.
- $\Omega_{\overline{\mathcal{F}}_D} = \prod_{X\in\overline{\mathcal{F}}_D}\Omega_X$ is the entire sample space, where Ω_X is the sample space of an individual feature X. For a feature set S, Ω_S means the Cartesian product $\prod_{X\in S}\Omega_X$, which is the sample space of S.
- $\boldsymbol{p}_D : \Omega_{\overline{\mathcal{F}}_D} \to \mathbb{Q}$ is the empirical probability distribution.
- We also denote D by the triplet $(\overline{\mathcal{F}}_D, \Omega_{\overline{\mathcal{F}}_D}, \boldsymbol{p}_D)$.
- For $S \subseteqq \mathcal{F}_D$, $D|_S$ is the dataset derived from D by eliminating the features in $\mathcal{F}_D \setminus S$.
- For $S \subseteqq \mathcal{F}_D$, $D^{\wedge S}$ is the dataset derived from D by replacing the features of S with a single feature $\wedge S$ such that $\Omega_{\wedge S} = \Omega_S$. The values for S of an example in D are replaced with the vector consisting of the values.

Example 4 *For the dataset D determined by (a) below, we have $\overline{\mathcal{F}}_D = \{B,R,G,\mathrm{C}_D\}$ and $\Omega_{\overline{\mathcal{F}}_D} = \{A,B,O,AB\}\times\{M,N,C,A\}\times\{M,F\}\times\{p,n\}$. When $S=\{R,G\}$, (b) and (c) determine $D\,|_S$ and $D^{\wedge S}$.*

(a) D

B	R	G	C_D
A	M	M	p
A	M	M	n
B	N	M	p
O	C	F	n
AB	A	M	n

(b) $D|_S$

R	G	C_D
M	M	p
M	M	n
N	M	p
C	F	n
A	M	n

(c) $D^{\wedge S}$

B	$\wedge S$	C_D
A	(M, M)	p
A	(M, M)	n
B	(N, M)	p
O	(C, F)	n
AB	(A, M)	n

On the other hand, a feature selection measure μ is formulated as a family of μ^D indexed by datasets D, and each μ^D is a real-valued function defined over the power set $\mathfrak{P}(\mathcal{F}_D)$ of $\mathcal{F}_D$. In this paper, we assume that a feature selection measure supports the following three requirements, which all of the 19 measures surveyed in [Shin *et al.*, 2011] support as well.

Requirement 1 (Naming invariance)
Measurements by the measure are invariant to renaming of features and feature values. This also implies that the measure assumes that every feature is categorical.

Requirement 2 (Localization invariance)
If $T \subseteqq S$, $\mu^D(T) = \mu^{D|_S}(T)$ holds.

Requirement 3 (Feature aggregation invariance)
If $T \supseteqq S$, $\mu^D(T) = \mu^{D^{\wedge S}}((T\setminus S)\cup\{\wedge S\})$ holds.

Example 5 *Bayesian risk μ_{br} has all of these invariance properties: $\mu_{\mathsf{br}}{}^D(\mathcal{F}_D) = \mu_{\mathsf{br}}{}^E(\mathcal{F}_E) = \frac{1}{5}$ indicates the naming invariance; For $S=\{R,G\}$, $\mu_{\mathsf{br}}{}^D(\{G\})$ and $\mu_{\mathsf{br}}{}^{D|_S}(\{G\})$ are identical to $\frac{2}{5}$, and $\mu_{\mathsf{br}}{}^{D^{\wedge S}}(\{B,\wedge S\}) = \frac{1}{5}$ holds These indicate the localization and feature aggregation invariance properties.*

				E
F_1	F_2	F_3	C_D
1	1	1	+1
1	1	1	-1
2	2	1	+1
3	3	2	-1
4	4	1	-1

In this paper, we will identify (D,S) and (E,T), pairs of a dataset and a feature subset, if $\mu^D(S) = \mu^E(T)$ holds for any measure μ that supports Requirements 1, 2 and 3. For this, we introduce the notion of *coherent* mappings and then prove Proposition 1.

Definition 1 *A mapping $\varphi : \mathcal{F}_D \to \mathcal{F}_E$ is* coherent *between D and E, if, and only if, there exist injections $v_Y : \Omega_{\varphi^{-1}(Y)} \to \Omega_Y$ for $Y \in \mathrm{Im}(\varphi)$ and an injection $v_{\mathrm{C}_E} : \Omega_{\mathrm{C}_D} \to \Omega_{\mathrm{C}_E}$ such that*

$$\boldsymbol{p}_D(\boldsymbol{x}) = \sum_{\boldsymbol{y}:\boldsymbol{y}|_{\mathrm{Im}(\varphi)\cup\{\mathrm{C}_E\}}=v_\pi(\boldsymbol{x})}\boldsymbol{p}_E(\boldsymbol{y}),$$

where $v_\pi = \prod_{Y\in\mathrm{Im}(\varphi)\cup\{\mathrm{C}_E\}} v_Y : \Omega_{\overline{\mathcal{F}}_D} \to \Omega_{\mathrm{Im}(\varphi)\cup\{\mathrm{C}_E\}}$.

Example 6 *We consider the datasets D of Example 4 and E below. When we define φ by $\varphi(B) = F_1$ and $\varphi(R) = \varphi(G) = F_2$, φ is coherent. To show this, we determine v_{F_1} and v_{F_2} as below and v_{C_E} by $v_{\mathrm{C}_E}(p) = 1$ and $v_{\mathrm{C}_E}(n) = 2$. For example, $\boldsymbol{p}_D(A, M, M, p) = \frac{1}{5}$ and $\boldsymbol{p}_E(1,1,1,1) + \boldsymbol{p}_E(1,1,3,1) = \frac{1}{10} + \frac{1}{10} = \frac{1}{5}$ hold.*

E

F_1	F_2	F_3	C_E	F_1	F_2	F_3	C_E
1	*1*	*1*	*1*	*2*	*3*	*4*	*1*
1	*1*	*2*	*2*	*3*	*6*	*2*	*2*
1	*1*	*3*	*1*	*3*	*6*	*3*	*2*
1	*1*	*3*	*2*	*4*	*7*	*4*	*2*
2	*3*	*1*	*1*	*4*	*7*	*4*	*2*

v_{F_1}			v_{F_2}					
A	→	*1*	*(M, M)*	→	*1*	*(C, M)*	→	*5*
B	→	*2*	*(M, F)*	→	*2*	*(C, F)*	→	*6*
O	→	*3*	*(N, M)*	→	*3*	*(A, M)*	→	*7*
AB	→	*4*	*(N, F)*	→	*4*	*(A, F)*	→	*8*

Proposition 1 *If a feature selection measure μ supports Requirement 1 to 3, $\mu^E(S) = \mu^D(\varphi^{-1}(S))$ holds for an arbitrary coherent mapping $\varphi : \mathcal{F}_D \to \mathcal{F}_E$ between D and E and an arbitrary $S \subseteqq \mathrm{Im}(\varphi)$.*

Proof. We let $\varphi = \underline{\varphi} \circ \overline{\varphi}$ such that $\overline{\varphi} : \mathcal{F}_D \to \mathrm{Im}(\varphi)$ and $\underline{\varphi} : \mathrm{Im}(\varphi) \to \mathcal{F}_E$. Since $\overline{\varphi}$ yields a sequence of feature aggregations that convert D into $E|_{\mathrm{Im}(\varphi)}$ up to renaming of features and feature values, $\mu^E(S) = \mu^{E|_{\mathrm{Im}(\varphi)}}(S) = \mu^D(\varphi^{-1}(S))$ follows from Requirement 1 to 3. ∎

Example 7 *For the same D, E and φ as Example 6, $\mu_{\mathsf{br}}{}^E(\{F_1, F_2\}) = \mu_{\mathsf{br}}{}^D(\{B, R, G\}) = \mu_{\mathsf{br}}{}^E(\{F_1\}) = \mu_{\mathsf{br}}{}^D(\{B\}) = \mu_{\mathsf{br}}{}^E(\{F_2\}) = \mu_{\mathsf{br}}{}^D(\{R, G\}) = \frac{1}{5}$ holds.*

3 Projecting a feature set of a dataset into a subspace $\mathcal{P}_{\mathbb{R}}^{+\langle\infty\rangle}$ of the ℓ^p space ($p \in [1, \infty]$)

We let $n \in \mathbb{N} \cup \{\infty\}$. We introduce the space $\mathcal{P}_0{}^{\langle n\rangle}$ into which feature sets of datasets are projected. We start with preparation. Let $\mathbb{K}$ be either the rational number field $\mathbb{Q}$ or the real number field $\mathbb{R}$. We define $\mathcal{P}_{\mathbb{K}}{}^{\langle\infty\rangle}$ and $\mathcal{P}_{\mathbb{K}}^{+\langle\infty\rangle}$ as follows, which are sets of functions $\boldsymbol{p}$ such that $\boldsymbol{p} : \mathbb{N}^2 \to \mathbb{K}$.

Definition 2 *We define $\mathcal{P}_{\mathbb{K}}{}^{\langle\infty\rangle}$ and $\mathcal{P}_{\mathbb{K}}^{+\langle\infty\rangle}$ by:*

$$\mathcal{P}_{\mathbb{K}}{}^{\langle\infty\rangle} = \Big\{\boldsymbol{p}\Big|\boldsymbol{p}(i,j) \geq 0, \sum_{i=1}^{\infty}\sum_{j=1}^{\infty}\boldsymbol{p}(i,j) = 1\Big\};$$

$$\mathcal{P}_{\mathbb{K}}^{+\langle\infty\rangle} = \Big\{\boldsymbol{p}\Big|\boldsymbol{p}(i,j) \geq 0, \sum_{i=1}^{\infty}\sum_{j=1}^{\infty}\boldsymbol{p}(i,j) \leq 1\Big\}.$$

We can identify $\mathbb{N}^2$ with $\mathbb{N}$ (for example, $f : (i,j) \mapsto \frac{(i+j-1)(i+j-2)}{2} + i$ is bijective), and therefore, $\mathcal{P}_{\mathbb{R}}{}^{\langle\infty\rangle}$ and $\mathcal{P}_{\mathbb{R}}^{+\langle\infty\rangle}$ can be viewed as subspaces of ℓ^p for $1 \leq p \leq \infty$. If we focus on datasets that include at most n classes, we can use $\mathcal{P}_{\mathbb{K}}{}^{\langle n\rangle}$ and $\mathcal{P}_{\mathbb{K}}^{+\langle n\rangle}$ instead.

Definition 3 *We define $\mathcal{P}_{\mathbb{K}}{}^{\langle n\rangle}$ and $\mathcal{P}_{\mathbb{K}}^{+\langle n\rangle}$ by:*

$$\mathcal{P}_{\mathbb{K}}{}^{\langle n\rangle} = \big\{\boldsymbol{p} \in \mathcal{P}_{\mathbb{K}}{}^{\langle\infty\rangle}\big|j > n \Rightarrow \boldsymbol{p}(i,j) = 0\big\};$$

$$\mathcal{P}_{\mathbb{K}}^{+\langle n\rangle} = \big\{\boldsymbol{p} \in \mathcal{P}_{\mathbb{K}}^{+\langle\infty\rangle}\big|j > n \Rightarrow \boldsymbol{p}(i,j) = 0\big\};$$

$\mathcal{P}_{\mathbb{R}}{}^{\langle n\rangle}$ and $\mathcal{P}_{\mathbb{R}}^{+\langle n\rangle}$ are metric spaces with the metric derived from the norm $\|\cdot\|_p$ of ℓ^p. $\mathcal{P}_{\mathbb{R}}{}^{\langle n\rangle}$ is closed in ℓ^p only when $p = 1$, while $\mathcal{P}_{\mathbb{R}}^{+\langle n\rangle}$ is the closure of $\mathcal{P}_{\mathbb{R}}{}^{\langle n\rangle}$ in ℓ^p for $1 < p \leq \infty$. Therefore, $\mathcal{P}_{\mathbb{R}}{}^{\langle n\rangle}$ in ℓ^1 and $\mathcal{P}_{\mathbb{R}}^{+\langle n\rangle}$ in ℓ^p for $1 < p \leq \infty$ are complete, since ℓ^p is a Banach space. Although $\mathcal{P}_{\mathbb{R}}{}^{\langle n\rangle}$ and $\mathcal{P}_{\mathbb{R}}^{+\langle n\rangle}$ are bounded, none of them is compact.

We define $\mathcal{P}_0{}^{\langle n\rangle}$, a subspace of $\mathcal{P}_{\mathbb{R}}{}^{\langle n\rangle}$, as follows.

Definition 4 *We let $\mathrm{supp}(\boldsymbol{p})$ be the smallest $\Omega_U \times \Omega_{\mathrm{C}}$ such that $\Omega_U \subseteqq \mathbb{N}$, $\Omega_{\mathrm{C}} \subseteqq \mathbb{N}$ and $\Omega_U \times \Omega_{\mathrm{C}} \supseteqq \mathrm{Supp}(\boldsymbol{p}) = \{(i,j) \mid \boldsymbol{p}(i,j) > 0\}$. We define $\mathcal{P}_0{}^{\langle\infty\rangle}$ and $\mathcal{P}_0{}^{\langle n\rangle}$ for $n \in \mathbb{N}$ by:*

$$\mathcal{P}_0{}^{\langle\infty\rangle} = \{\boldsymbol{p} \in \mathcal{P}_{\mathbb{Q}}{}^{\langle\infty\rangle} \mid |\mathrm{supp}(\boldsymbol{p})| < \infty\};$$

$$\mathcal{P}_0{}^{\langle n\rangle} = \{\boldsymbol{p} \in \mathcal{P}_0{}^{\langle\infty\rangle} \mid j > n \Rightarrow \boldsymbol{p}(i,j) = 0\}.$$

Example 8 *When we define $\boldsymbol{p}$ by $\boldsymbol{p}(i,i) = (1-r)r^{i-1}$ for $r \in (0,1)$, $\boldsymbol{p} \in \mathcal{P}_{\mathbb{R}}{}^{\langle\infty\rangle}$ holds, because $\sum_{i=1}^{\infty}(1-r)r^{i-1} = (1-r)\lim_{n\to\infty}\frac{1-r^n}{1-r} = 1$. Furthermore, if $r \in \mathbb{Q}$, $\boldsymbol{p} \in \mathcal{P}_{\mathbb{Q}}{}^{\langle\infty\rangle}$ holds. For this $\boldsymbol{p}$, we have $\Omega_U = \mathbb{N}$, $\Omega_{\mathrm{C}} = \mathbb{N}$ and $\mathrm{supp}(\boldsymbol{p}) = \{(i,i) \mid i \in \mathbb{N}\}$. Hence, $\boldsymbol{p} \notin \mathcal{P}_0{}^{\langle\infty\rangle}$.*

For $n \in \mathbb{N} \cup \{\infty\}$, $\mathcal{P}_0{}^{\langle n\rangle}$ turns out to be dense in $\mathcal{P}_{\mathbb{R}}{}^{\langle n\rangle}$. For $\boldsymbol{p} \in \mathcal{P}_0{}^{\langle n\rangle}$, $D_{\boldsymbol{p}} = (\{U, \mathrm{C}\}, \mathbb{N} \times \mathbb{N}, \boldsymbol{p})$ is not a dataset, because $\mathbb{N} \times \mathbb{N}$ is infinite. Nevertheless, we can think of a coherent mapping $\varphi : \mathcal{F}_D \to \{U\}$ between a dataset D and $D_{\boldsymbol{p}}$. Theorem 1 shows how to project a feature set S of a dataset D into $\mathcal{P}_0{}^{\langle n\rangle}$.

Theorem 1 *We let D be a dataset with n classes and $S \subseteqq \mathcal{F}_D$. There exists $\boldsymbol{p} \in \mathcal{P}_0{}^{\langle n\rangle}$ such that $\varphi : S \to \{U\}$ is coherent between $D|_S$ and $D_{\boldsymbol{p}}$. Conversely, for arbitrary $\boldsymbol{p} \in \mathcal{P}_0{}^{\langle n\rangle}$, there exists a dataset D with n classes such that $\varphi : \mathcal{F}_D \to \{U\}$ is coherent between D and $D_{\boldsymbol{p}}$.*

Proof. Let $v_1 : \Omega_S \to \mathbb{N}$ and $v_2 : \Omega_{\mathrm{C}_D} \to \{1, \ldots, n\}$ be injective. It suffices to define $\boldsymbol{p} \in \mathcal{P}_0{}^{\langle n\rangle}$ by $\boldsymbol{p}(i,j) = \boldsymbol{p}_{D|_S}\left(v_1^{-1}(i), v_2^{-1}(j)\right)$ if $(i,j) \in \mathrm{Im}(v_1 \times v_2)$ and $\boldsymbol{p}(i,j) = 0$ otherwise. To show the converse, we have only to let $D = (\{U, C\}, \mathrm{supp}(\boldsymbol{p}), \boldsymbol{p})$. ∎

Example 9 *To project $D|_S$ of Example 4 into $\mathcal{P}_0{}^{\langle 2\rangle}$, we determine $\boldsymbol{p} \in \mathcal{P}_0{}^{\langle 2\rangle}$ by $\boldsymbol{p}(1,1) = \boldsymbol{p}(1,2) = \boldsymbol{p}(3,1) = \boldsymbol{p}(6,2) = \boldsymbol{p}(7,4) = \frac{1}{5}$ and $\boldsymbol{p}(i,j) = 0$ otherwise. When we determine v_U and v_{C} identical to v_{F_2} and v_{C_E} of Example 6, $\varphi : S \to \{U\}$ is coherent between $D|_D$ and $D_{\boldsymbol{p}}$, and hence, $D|_D$ is projected to $\boldsymbol{p}$.*

Example 10 *A single $D|_S$ can be projected to infinitely many points in $\mathcal{P}_0{}^{\langle n\rangle}$. For example, we let $D|_S$, $\boldsymbol{p}$ and v_U be the same as Example 9. For an arbitrary bijection $\pi : \mathbb{N} \to \mathbb{N}$, we define $\boldsymbol{p}_\pi$ by $\boldsymbol{p}_\pi(\pi(i), j) = \boldsymbol{p}(i,j)$. Apparently, $D|_S$ projects to $\boldsymbol{p}_\pi$ with $\pi \circ v_U$.*

Example 11 *More than one datasets are projected to a single point in ${\mathcal{P}_0}^{\langle n\rangle}$. For example, $D|_S$ of Example 4 and $E|_{\{F_2\}}$ of Example 6 are both projected to any of $\boldsymbol{p}_\pi$ of Example 10. To be precise, a single point in ${\mathcal{P}_0}^{\langle n\rangle}$ has infinitely many datasets that project to the point.*

Based on Proposition 1 and Theorem 1, Theorem 2 asserts that a feature selection measure can be uniquely viewed as a real-valued function defined over ${\mathcal{P}_0}^{\langle n\rangle}$.

Theorem 2 *For a feature selection measure μ, $\hat{\mu} : {\mathcal{P}_0}^{\langle\infty\rangle} \to \mathbb{R}$ uniquely exists and $\hat{\mu}(\boldsymbol{p}) = \mu^D(S)$ holds for arbitrary D, S and $\boldsymbol{p}$ such that $\varphi : S \to \{U\}$ is coherent between $D|_S$ and $D_{\boldsymbol{p}}$.*

Proof. We let $D^*_{\boldsymbol{p}}$ be $(\{U, C\}, \mathrm{supp}(\boldsymbol{p}), \boldsymbol{p})$ for $\boldsymbol{p} \in {\mathcal{P}_0}^{\langle\infty\rangle}$. We determine $\hat{\mu}(\boldsymbol{p})$ by $\mu^{D^*_{\boldsymbol{p}}}(\{U\})$. For a dataset D such that $\varphi : S \to \{U\}$ is coherent between $D|_S$ and $D_{\boldsymbol{p}}$ with $v_U : \Omega_S \to \mathbb{N}$ and $v_C : \Omega_{\mathrm{C}_D} \to \mathbb{N}$, we let $D^{\#}_{\boldsymbol{p}} = (\{U, \mathrm{C}\}, \mathrm{Im}(v_U) \times \mathrm{Im}(v_C), \boldsymbol{p})$. $\varphi : S \to \{U\}$ is coherent between $D|_S$ and $D^{\#}_{\boldsymbol{p}}$, and the inclusion $\mathrm{supp}(\boldsymbol{p}) \subseteqq \mathrm{Im}(v_U) \times \mathrm{Im}(v_C)$ yields coherence between $D^*_{\boldsymbol{p}}$ and $D^{\#}_{\boldsymbol{p}}$. Therefore, $\mu^D(S) = \mu^{D^{\#}_{\boldsymbol{p}}}(\{U\}) = \mu^{D^*_{\boldsymbol{p}}}(\{U\}) = \hat{\mu}(\boldsymbol{p})$. ∎

Example 12 *Datasets that are projected to the same point in ${\mathcal{P}_0}^{\langle n\rangle}$ have the same measurement for an arbitrary feature selection measure μ. For example, as seen in Example 11, $D|_S$ of Example 4 and $E|_{\{F_2\}}$ of Example 6 are both projected to the same point ${\mathcal{P}_0}^{\langle 2\rangle}$. On the other hand, ${\mu_{\mathsf{br}}}^D(S) = {\mu_{\mathsf{br}}}^E(\{F_2\}) = \frac{1}{5}$ holds as seen in Example 7. Therefore, by letting $\hat{\mu}(\boldsymbol{p}) = \mu^D(S)$ for $D|_S$ that projects to $\boldsymbol{p}$, $\hat{\mu}$ is well defined.*

The problem of the projection determined by Theorem 1 consists in the fact that a single pair (D, S) is mapped to an infinite number of different points in ${\mathcal{P}_0}^{\langle n\rangle}$ (Example 10). In the next section, we will solve this problem.

4 Defining a quotient space $\mathcal{Q}_{\mathbb{R}}^{+\langle\infty\rangle}$ of $\mathcal{P}_{\mathbb{R}}^{+\langle\infty\rangle}$

Our idea to solve the problem is to introduce an equivalence relation that unifies points of ${\mathcal{P}_0}^{\langle n\rangle}$ that are the images of the same single pair (D, S) and then use the resulting quotient space. The relation $\sim$ defined below is an equivalence relation. We assume $n \in \mathbb{N} \cup \{\infty\}$.

Definition 5 *Let $\boldsymbol{p}$ and $\boldsymbol{q}$ be in $\mathcal{P}_{\mathbb{R}}^{+\langle\infty\rangle}$. We define $\boldsymbol{p} \sim \boldsymbol{q}$, if, and only if, there exist bijections $v_U : \mathbb{N} \to \mathbb{N}$ and $v_C : \mathbb{N} \to \mathbb{N}$ such that $\boldsymbol{p}(i, j) = \boldsymbol{q}(v_U(i), v_C(j))$.*

Example 13 *$\boldsymbol{p}_\pi \sim \boldsymbol{p}$ holds for $\boldsymbol{p}_\pi$ and π of Example 10.*

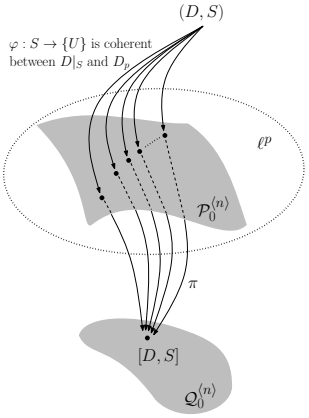
Definition 6 *We let $\mathcal{Q}_{\mathbb{R}}^{+\langle\infty\rangle} = \mathcal{P}_{\mathbb{R}}^{+\langle\infty\rangle}/\!\sim$. Moreover, for the canonical projection $\pi : \mathcal{P}_{\mathbb{R}}^{+\langle\infty\rangle} \to \mathcal{Q}_{\mathbb{R}}^{+\langle\infty\rangle}$, we let $\mathcal{Q}_{\mathbb{R}}^{+\langle n\rangle} = \pi(\mathcal{P}_{\mathbb{R}}^{+\langle n\rangle})$, $\mathcal{Q}_{\mathbb{R}}{}^{\langle n\rangle} = \pi(\mathcal{P}_{\mathbb{R}}{}^{\langle n\rangle})$ and $\mathcal{Q}_0{}^{\langle n\rangle} = \pi({\mathcal{P}_0}^{\langle n\rangle})$.*

Proposition 2 is easy to see but plays a crucial rule to prove Theorem 3 and 4.

Proposition 2 *For $\boldsymbol{p}$ and $\boldsymbol{q}$ in ${\mathcal{P}_0}^{\langle\infty\rangle}$, the following are equivalent.*

1. *$\boldsymbol{p} \sim \boldsymbol{q}$.*
2. *There exists (D, S) such that $\varphi : S \to \{U\}$ is coherent between $D|_S$ and $D_{\boldsymbol{p}}$ and between $D|_S$ and $D_{\boldsymbol{q}}$.*
3. *$\varphi : S \to \{U\}$ is coherent between $D|_S$ and $D_{\boldsymbol{p}}$, if, and only if, it is coherent between $D|_S$ and $D_{\boldsymbol{q}}$.*

Theorem 3 and 4 follow from Theorem 1, Theorem 2 and Proposition 2. In the following, $[\boldsymbol{p}]$ denotes $\pi(\boldsymbol{p})$, and $[D, S]$ denotes the image of (D, S) in $\mathcal{Q}_0{}^{\langle\infty\rangle}$.



Theorem 3 *We let D be a dataset with n classes and $S \subseteqq \mathcal{F}_D$. There uniquely exists $[\boldsymbol{p}] \in \mathcal{Q}_0{}^{\langle n\rangle}$ such that $\varphi : S \to \{U\}$ is coherent between $D|_S$ and $D_{\boldsymbol{p}}$. Conversely, for arbitrary $[\boldsymbol{p}] \in \mathcal{Q}_0{}^{\langle n\rangle}$, there exists a dataset D with n classes such that $\varphi : \mathcal{F}_D \to \{U\}$ is coherent between D and $D_{\boldsymbol{p}}$.*

Theorem 4 *For a feature selection measure μ, there uniquely exists $[\hat{\mu}] : \mathcal{Q}_0{}^{\langle\infty\rangle} \to \mathbb{R}$ such that $[\hat{\mu}]([D, S]) = \mu^D(S)$ holds for any dataset D and $S \subseteqq \mathcal{F}_D$.*

5 Introducing a metric d_p into $\mathcal{Q}_{\mathbb{R}}^{+\langle\infty\rangle}$

Although a quotient of a metric space is not always a metric space, we can derive a metric into $\mathcal{Q}_{\mathbb{R}}^{+\langle\infty\rangle}$ from ℓ^p as follows.

Definition 7 *For $[\boldsymbol{p}]$ and $[\boldsymbol{q}]$ in $\mathcal{Q}_{\mathbb{R}}^{+\langle\infty\rangle}$, we define $d_p([\boldsymbol{p}], [\boldsymbol{q}]) = \inf\{\|\boldsymbol{p}' - \boldsymbol{q}'\|_p \mid \boldsymbol{p} \sim \boldsymbol{p}', \boldsymbol{q} \sim \boldsymbol{q}'\}$.*

Theorem 5 *For $1 \leq p \leq \infty$, d_p is a metric over $\mathcal{Q}_{\mathbb{R}}^{+\langle\infty\rangle}$.*

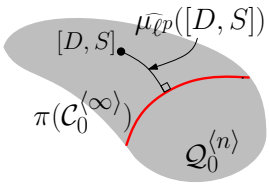
Proof. We only prove the triangle inequality $d_p([\boldsymbol{p}], [\boldsymbol{r}]) \leq d_p([\boldsymbol{p}], [\boldsymbol{q}]) + d_p([\boldsymbol{q}], [\boldsymbol{r}])$ taking advantage of Lemma 1.

Lemma 1 *For $\boldsymbol{p} \sim \boldsymbol{p}'$ and $\boldsymbol{q}$ in $\mathcal{P}_{\mathbb{R}}^{+\langle\infty\rangle}$, there exists $\boldsymbol{q}' \sim \boldsymbol{q}$ such that $\|\boldsymbol{p} - \boldsymbol{q}\|_p = \|\boldsymbol{p}' - \boldsymbol{q}'\|_p$.*

For $\varepsilon > 0$, we assume $\|\boldsymbol{p} - \boldsymbol{q}\|_p - d_p([\boldsymbol{p}], [\boldsymbol{q}]) < \frac{\varepsilon}{2}$ and $\|\boldsymbol{q}' - \boldsymbol{r}\|_p - d_p([\boldsymbol{q}], [\boldsymbol{r}]) < \frac{\varepsilon}{2}$ for $\boldsymbol{q} \sim \boldsymbol{q}'$. By Lemma 1, we have $\|\boldsymbol{q} - \boldsymbol{r}'\|_p = \|\boldsymbol{q}' - \boldsymbol{r}\|_p$ and $d_p([\boldsymbol{p}], [\boldsymbol{r}]) \leq \|\boldsymbol{p} - \boldsymbol{r}'\|_p \leq \|\boldsymbol{p} - \boldsymbol{q}\|_p + \|\boldsymbol{q} - \boldsymbol{r}'\|_p < d_p([\boldsymbol{p}], [\boldsymbol{q}]) + d_p([\boldsymbol{q}], [\boldsymbol{r}]) + \varepsilon$. ∎

6 Deriving a measure $\widehat{\mu_{\ell^p}}$ from d_p

We see that the minimum of d_p distance from $[D, S]$ to a consistent feature set determines a feature selection measure. We first define *consistent points* of $\mathcal{P}_{\mathbb{R}}^{+\langle\infty\rangle}$.



Definition 8 *$\boldsymbol{p} \in \mathcal{P}_{\mathbb{R}}^{+\langle\infty\rangle}$ is* consistent, *if, and only if, there exists $\bar{j} : \mathbb{N} \to \mathbb{N}$ such that $\boldsymbol{p}(i, j) = 0$ if $j \neq \bar{j}(i)$.*

$\mathcal{C}_{\mathbb{R}}^{+\langle\infty\rangle}$ *denotes the entire set of consistent* $\boldsymbol{p}$*, and we let* $\mathcal{C}_{\mathbb{R}}^{+\langle n\rangle} = \mathcal{C}_{\mathbb{R}}^{+\langle\infty\rangle} \cap \mathcal{P}_{\mathbb{R}}^{+\langle n\rangle}$*,* $\mathcal{C}_{\mathbb{R}}{}^{\langle n\rangle} = \mathcal{C}_{\mathbb{R}}^{+\langle\infty\rangle} \cap \mathcal{P}_{\mathbb{R}}{}^{\langle n\rangle}$ *and* $\mathcal{C}_0{}^{\langle n\rangle} = \mathcal{C}_{\mathbb{R}}^{+\langle\infty\rangle} \cap \mathcal{P}_0{}^{\langle n\rangle}$ *for* $n \in \mathbb{N} \cup \{\infty\}$*.*

Definition 9 *For* $\overline{j} : \mathbb{N} \to \mathbb{N}$ *such that* $\overline{j}(i) \in \operatorname{argmax}\{\boldsymbol{p}(i,j) \mid j \in \mathbb{N}\}$ *and* $\boldsymbol{p} \in \mathcal{P}_0{}^{\langle\infty\rangle}$*, we define:*

For $1 \le p < \infty$*,*

$$\widehat{\mu_{\ell^p}}(\boldsymbol{p}) = \sqrt[p]{\sum_{i\in\mathbb{N}} \left(\sum_{j\in\mathbb{N}} \boldsymbol{p}(i,j)^p - \boldsymbol{p}(i,\overline{j}(i))^p \right)};$$

For $p = \infty$*,*

$$\widehat{\mu_{\ell^\infty}}(\boldsymbol{p}) = \max\left\{ \boldsymbol{p}(i,j) \,|\, (i,j) \in \mathbb{N}^2, (i,j) \neq (i,\overline{j}(i)) \right\}.$$

Surprisingly, $\widehat{\mu_{\ell^1}}$ is identical to the Bayesian risk.

Proposition 3 *If* $\varphi : S \to U$ *is coherent between* $D|_S$ *and* $D_{\boldsymbol{p}}$*,* $\mu_{\mathsf{br}}{}^D(S) = \widehat{\mu_{\ell^1}}(\boldsymbol{p})$ *holds.*

Proof. We assume that $v_U : \Omega_S \to \mathbb{N}$ and $v_{\mathrm{C}} : \Omega_{\mathrm{C}} \to \mathbb{N}$ yield the coherence between $D|_S$ and $D_{\boldsymbol{p}}$.

$$\begin{aligned}
&\mu_{\mathsf{br}}{}^D(S) \\
&= \sum_{\boldsymbol{x}\in\Omega_S} \left(\boldsymbol{p}_{D|_S}(S = \boldsymbol{x}) - \max_{c\in\Omega_{\mathrm{C}}} \boldsymbol{p}_{D|_S}(S = \boldsymbol{x}, \mathrm{C} = c) \right) \\
&= \sum_{\boldsymbol{x}\in\Omega_S} \left(\sum_{c\in\Omega_{\mathrm{C}}} \boldsymbol{p}(v_U(\boldsymbol{x}), v_{\mathrm{C}}(c)) - \max_{c\in\Omega_{\mathrm{C}}} \boldsymbol{p}(v_U(\boldsymbol{x}), v_C(c)) \right) \\
&= \widehat{\mu_{\ell^1}}(\boldsymbol{p}). \blacksquare
\end{aligned}$$

Theorem 6 *For* $\boldsymbol{p} \in \mathcal{P}_0{}^{\langle\infty\rangle}$*, the following holds.*

1. $\widehat{\mu_{\ell^1}}(\boldsymbol{p}) = \frac{1}{2}\min\{d_1([\boldsymbol{p}],[\boldsymbol{q}]) \mid \boldsymbol{q} \in \mathcal{C}_0{}^{\langle\infty\rangle}\}$.
2. *For* $1 < p \le \infty$*,* $\widehat{\mu_{\ell^p}}(\boldsymbol{p}) = \inf\{d_p([\boldsymbol{p}],[\boldsymbol{q}]) \mid \boldsymbol{q} \in \mathcal{C}_0{}^{\langle\infty\rangle}\} = \min\{d_p([\boldsymbol{p}],[\boldsymbol{q}]) \mid \boldsymbol{q} \in \mathcal{C}_{\mathbb{R}}^{+\langle\infty\rangle} \wedge |\mathrm{supp}(\boldsymbol{q})| < \infty\}$.

Proof. Here, we prove only 1. For $\boldsymbol{q} \in \mathcal{C}_0{}^{\langle\infty\rangle}$, we assume $j : \mathbb{N} \to \mathbb{N}$ such that $\boldsymbol{q}(i,j) = 0$ if $j \neq j(i)$. Also, we let $\overline{j} : \mathbb{N} \to \mathbb{N}$ satisfy $\overline{j}(i) \in \operatorname{argmax}\{\boldsymbol{p}(i,j) \mid j \in \mathbb{N}\}$. We first fix $i \in \mathbb{N}$.

$$\begin{aligned}
&\sum_{j\in\mathbb{N}} |\boldsymbol{p}(i,j) - \boldsymbol{q}(i,j)| \\
&\quad = |\boldsymbol{p}(i,j(i)) - \boldsymbol{q}(i,j(i))| + \sum_{j\in\mathbb{N}\setminus\{j(i)\}} \boldsymbol{p}(i,j) \\
&\quad \ge |\boldsymbol{p}(i,j(i)) - \boldsymbol{q}(i,j(i))| + \sum_{j\in\mathbb{N}} \boldsymbol{p}(i,j) - \max_{j\in\mathbb{N}} \boldsymbol{p}(i,j).
\end{aligned}$$

Then, we sum up the both sides across all $i \in \mathbb{N}$.

$$\begin{aligned}
\|\boldsymbol{p} - \boldsymbol{q}\|_1 &\ge \sum_{i\in\mathbb{N}} |\boldsymbol{p}(i,j(i)) - \boldsymbol{q}(i,j(i))| \\
&\quad + \sum_{i\in\mathbb{N}} \left(\sum_{j\in\mathbb{N}} \boldsymbol{p}(i,j) - \max\{\boldsymbol{p}(i,j) \mid j \in \mathbb{N}\} \right) \\
&\ge \sum_{i\in\mathbb{N}} \boldsymbol{q}(i,j(i)) - \sum_{i\in\mathbb{N}} \boldsymbol{p}(i,j(i)) + \widehat{\mu_{\ell^1}}(\boldsymbol{p}) \ge 2\widehat{\mu_{\ell^1}}(\boldsymbol{p}).
\end{aligned}$$

On the other hand, for $\boldsymbol{q} \in \mathcal{C}_0{}^{\langle\infty\rangle}$ such that $\boldsymbol{q}(i,j) = \sum_{j\in\mathbb{N}} \boldsymbol{p}(i,j)$ if $j = \overline{j}(i)$ and 0 otherwise, we have

$$\|\boldsymbol{p} - \boldsymbol{q}\|_1 = 2\left(1 - \sum_{i\in\mathbb{N}} \max\{\boldsymbol{p}(i,j) \mid j \in \mathbb{N}\}\right).$$

As $\boldsymbol{q}' \in \mathcal{C}_{\mathbb{R}}{}^{\langle\infty\rangle}$ if $\boldsymbol{q}' \sim \boldsymbol{q}$, Lemma 1 implies the assertion. $\blacksquare$

Determinacy of $\widehat{\mu_{\ell^p}}$ immediately follows from Theorem 6. Also, it is easy to show monotonicity of $\widehat{\mu_{\ell^p}}$. Furthermore, $\widehat{\mu_{\ell^p}}$ turns out continuous under d_q for arbitrary $1 \le p, q \le \infty$.

7 Contrasting d_1 and d_p with $p > 1$

Proposition 3 asserts that $\widehat{\mu_{\ell^1}}$ is identical to the Bayesian risk, which is well known in the literature. On the other hand, $\widehat{\mu_{\ell^p}}$ with $p \in (1,\infty]$ is a novel measure unknown in the literature. Also, as a metric function, d_1 and d_p with $p \in (1,\infty]$ are significantly different as stated below without proof.

- $\mathcal{Q}_{\mathbb{R}}{}^{\langle\infty\rangle}$ is complete under d_1, while $\mathcal{Q}_{\mathbb{R}}^{+\langle\infty\rangle}$ is complete under d_p for $p \in (1,\infty]$.
- $\mathcal{Q}_{\mathbb{R}}{}^{\langle n\rangle}$ is *not* compact under d_1, while $\mathcal{Q}_{\mathbb{R}}^{+\langle n\rangle}$ is compact under d_p for $p \in (1,\infty]$.
- $\mathcal{Q}_0{}^{\langle n\rangle}$ is dense in $\mathcal{Q}_{\mathbb{R}}{}^{\langle n\rangle}$ under d_1, while it is dense in $\mathcal{Q}_{\mathbb{R}}^{+\langle n\rangle}$ under d_p for $p \in (1,\infty]$.

Moreover, in Section 8, we will see that $\widehat{\mu_{\ell^1}}$ and $\widehat{\mu_{\ell^p}}$ for $p \in (1,\infty]$ have different properties in terms of correlation between their measurements and classification accuracy. Although we cannot explain the reason for this difference, we imagine that it is related with the aforementioned difference between d_1 and d_p with $p \in (1,\infty]$ as a metric function.

8 Correlation between measurements by $\widehat{\mu_{\ell^p}}$ and classification accuracy

We evidently expect that there exists strong correlation between measurements by a good measure and classification accuracy. From this point of view, we ran experiments with $\widehat{\mu_{\ell^p}}$ for $p = 1, 2, \ldots, 5$. In particular, we focus on comparison between $p = 1$ and $p > 1$: $\widehat{\mu_{\ell^1}}$ is identical to the Bayesian risk and is the feature selection measure that is used the most widely in the literature; On the other hand, the other measures are novel and introduced in this paper for the first time.

Table 1: Datasets

NAME	#FEAT.	#EXAM.	#CLASSES
ARRHYTHMIA	279	452	13
AUDIOLOGY	69	226	24
MFEAT-FACTOR	216	2000	10
MFEAT-FOURIER	76	2000	10
MFEAT-KARHUNEN	64	2000	10
MFEAT-PIXEL	240	2000	10
MFEAT-ZERNIKE	47	2000	10
MUSK	166	476	2
OPTIDIGITS	64	5620	10
SONAR	60	208	2
SPAMBASE	57	4601	2
SPECTROMETER	100	531	48

8.1 Datasets

Table 1 shows the datasets that we use in our experiments as well as their important attributes, namely, the numbers of features, examples and classes. All of the datasets are obtained from the UCI repository of machine learning databases [Blake and Merz, 1998].

8.2 Methods

The following are the steps of our experiments.

Sampling feature sets

For each dataset of Table 1, 60 feature sets are selected at random. The sizes of the selected feature sets also vary at random. In total, we obtain $60 \times 12 = 720$ pairs of a feature set and a dataset.

Localizing datasets

For each pair of a feature set and a dataset, we localize the dataset by eliminating all of the features that do not belong to the relevant feature set. Consequently, we obtain 720 localized datasets.

Measuring accuracy of classification

We apply three classifiers, namely, Naïve Bayes, C4.5 and SVM, to each of the localized datasets in the manner of 10-fold cross validation and record the averaged AUC-ROC scores, which we use as classification accuracy scores.

8.3 Results

Figure 1, 2 and 3 display scatter plots of the results of our experiments with Naïve Bayes, C4.5 and SVM, respectively. The x-axis of each chart represents measurements by a measure out of (a) $\widehat{\mu_{\ell^1}}$, (b) $\widehat{\mu_{\ell^2}}$, (c) $\widehat{\mu_{\ell^3}}$, (d) $\widehat{\mu_{\ell^4}}$ and (e) $\widehat{\mu_{\ell^5}}$, while the y-axis does the averaged AUC-ROC scores. From the charts, we have the impression that $\widehat{\mu_{\ell^p}}$ with $p > 1$ shows stronger negative correlation with classification accuracy than $\widehat{\mu_{\ell^1}}$. We look into this more closely in the next subsection.

8.4 Analysis on correlation

To compare $\widehat{\mu_{\ell^1}}$ and $\widehat{\mu_{\ell^p}}$ with $p > 1$ in terms of correlation with classification accuracy, we introduce a function $P_t(x)$ as an index. We let N_x be the number of plots whose $\widehat{\mu_{\ell^p}}$ distances fall within $[x, x + 0.01)$ and $N_{t,x}$ be the number of plots whose AUC-ROC scores exceed t and $\widehat{\mu_{\ell^p}}$ distances fall within $[x, x+0.01)$. Then, we define $P_t(x)$ by $P_t(x) = \frac{N_{t,x}}{N_x}$. Intuitively, $P_t(x)$ approximates the probability of the case that the classification accuracy exceeds t when the measurement by the relevant measure is x.

Figure 4, 5 and 6 show the curves of $P_t(x)$ for Naïve Bayes, C4.5 and SVM. The value of t varies in $\{0.95, 0.90, 0.85, 0.80\}$. Since $P_t(x) \geq P_{t'}(x)$ holds for $t < t'$, the curve for t is located above the curve for t'. We observe two properties from the charts.

- The curves for $\widehat{\mu_{\ell^p}}$ for $p > 1$ are akin to one another in shape, while the curves for $\widehat{\mu_{\ell^1}}$ appear significantly different from the other measures.
- The curves for $\widehat{\mu_{\ell^p}}$ with $p > 1$ exhibits clearer negative correlation between measurements by the measure and classification accuracy than the curves for $\widehat{\mu_{\ell^1}}$.

In fact, the table below presents the correlation coefficients between the measurements x and the values of $P_t(x)$. The presented values also support the observation stated above.

$t =$	0.95	0.90	0.85	0.80
NAÏVE BAYES				
$\widehat{\mu_{\ell^1}}$	-0.42	-0.51	-0.60	-0.52
$\widehat{\mu_{\ell^2}}$	-0.58	-0.79	-0.97	-0.85
$\widehat{\mu_{\ell^3}}$	-0.56	-0.69	-0.76	-0.79
$\widehat{\mu_{\ell^4}}$	-0.54	-0.71	-0.89	-0.89
$\widehat{\mu_{\ell^5}}$	-0.54	-0.73	-0.91	-0.89
C4.5				
$\widehat{\mu_{\ell^1}}$	-0.17	-0.41	-0.34	-0.38
$\widehat{\mu_{\ell^2}}$	-0.41	-0.47	-0.79	-0.83
$\widehat{\mu_{\ell^3}}$	-0.52	-0.29	-0.72	-0.69
$\widehat{\mu_{\ell^4}}$	-0.52	-0.38	-0.77	-0.81
$\widehat{\mu_{\ell^5}}$	-0.52	-0.44	-0.82	-0.84
SVM				
$\widehat{\mu_{\ell^1}}$	-0.52	-0.62	-0.55	-0.23
$\widehat{\mu_{\ell^2}}$	-0.52	-0.54	-0.65	-0.80
$\widehat{\mu_{\ell^4}}$	-0.55	-0.52	-0.63	-0.75
$\widehat{\mu_{\ell^4}}$	-0.52	-0.52	-0.59	-0.72
$\widehat{\mu_{\ell^5}}$	-0.52	-0.52	-0.56	-0.71

9 Conclusion

We have introduced a family of feature selection measures $\{\widehat{\mu_{\ell^p}} \mid p \in [1, \infty]\}$ that are derived from the distance functions of some metric spaces. Although $\widehat{\mu_{\ell^1}}$ turns out to be identical to the Bayesian risk, $\widehat{\mu_{\ell^p}}$ for $p \in (1, \infty]$ are novel. Through experiments, we have shown that $\widehat{\mu_{\ell^p}}$ with $p > 1$ exhibits clearer negative correlation between its measurements and classification accuracy than $\widehat{\mu_{\ell^1}}$ does.

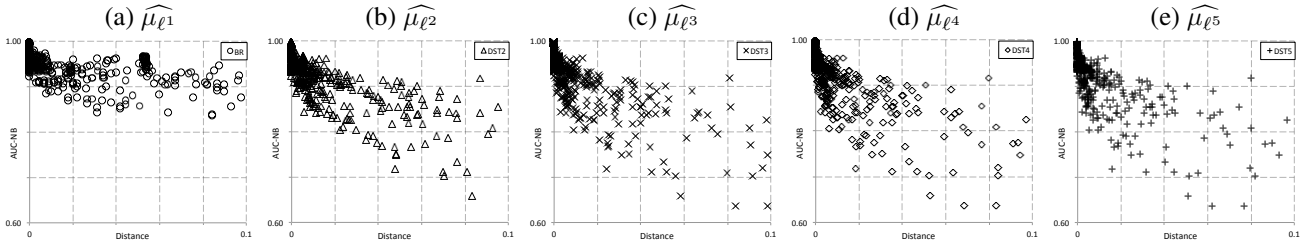


Figure 1: Scatter plots of the experimental results (Naïve Bayes)

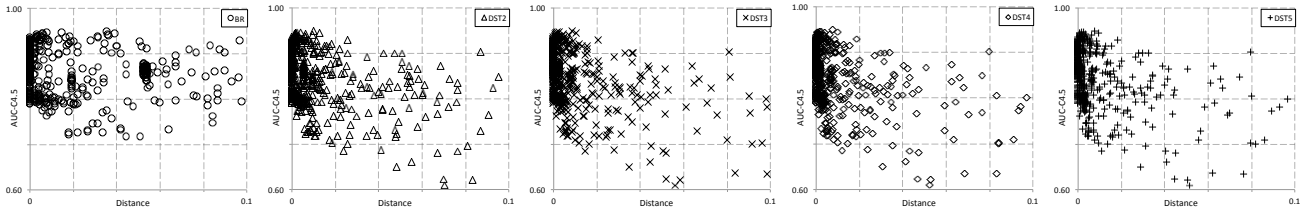


Figure 2: Scatter plots of the experimental results (C4.5)

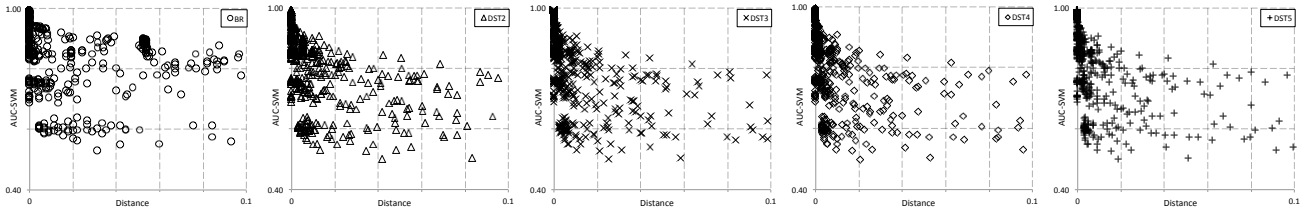


Figure 3: Scatter plots of the experimental results (SVM)

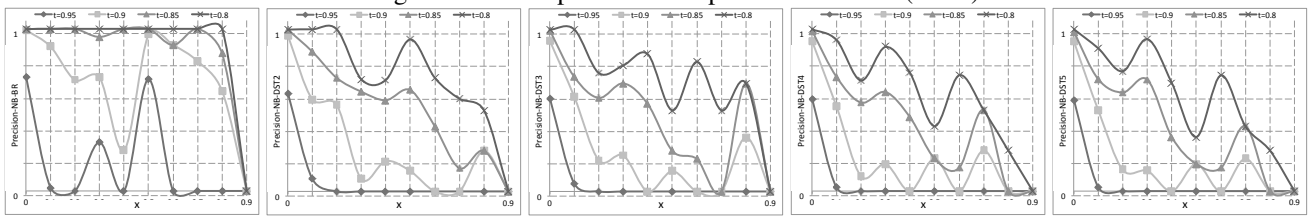


Figure 4: The curves of $P_t(x)$ for $t \in \{0.95, 0.90, 0.85, 0.80\}$ (Naïve Bayes)

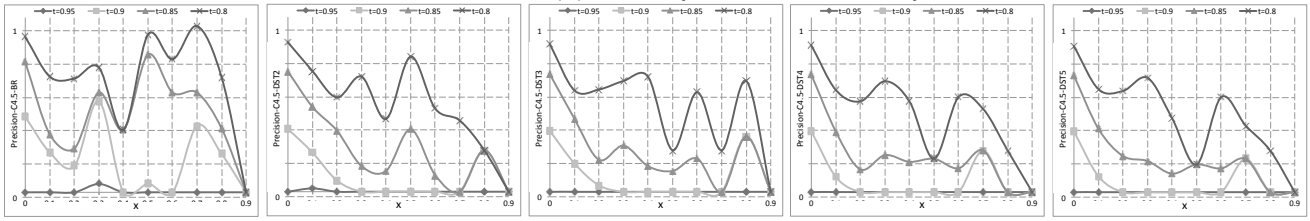


Figure 5: The curves of $P_t(x)$ for $t \in \{0.95, 0.90, 0.85, 0.80\}$ (C4.5)

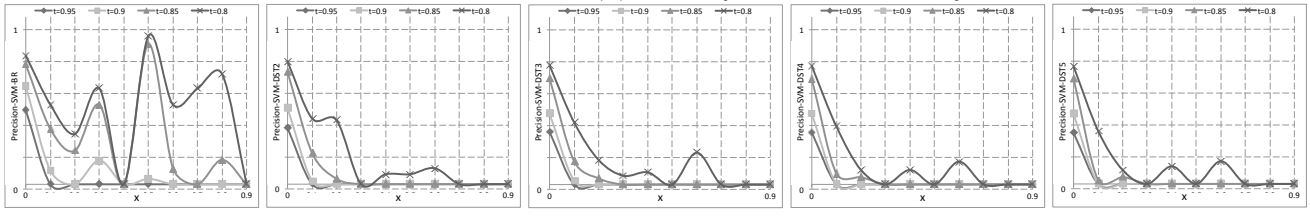


Figure 6: The curves of $P_t(x)$ for $t \in \{0.95, 0.90, 0.85, 0.80\}$ (SVM)

References

[Almuallim and Dietterich, 1994] H. Almuallim and T. G. Dietterich. Learning boolean concepts in the presence of many irrelevant features. *Artificial Intelligence*, 69(1 - 2), 1994.

[Arauzo-Azofra *et al.*, 2008] A. Arauzo-Azofra, J. M. Benitez, and J. L Castro. Consistency measures for feature selection. *Journal of Intelligent Information Systems*, 30(3):273–292, November 2008.

[Blake and Merz, 1998] C. S. Blake and C. J. Merz. UCI repository of machine learning databases. Technical report, University of California, Irvine, 1998.

[Bolón-Canedo *et al.*, 2011] V. Bolón-Canedo, S. Seth, and N. Sánchez-Maro no. Statistical dependence measure for feature selection in microarray datasets. In *ESANN 2011*, pages 23–28, 2011.

[Foithong *et al.*, 2012] S. Foithong, O. Pinngern, and B. Attachoo. Feature subset selection wrapper based on mutual information and rough sets. *Expert Systems with Applications*, 39(1):574–584, 2012.

[Liu *et al.*, 1998] H. Liu, H. Motoda, and M. Dash. A monotonic measure for optimal feature selection. In *Proceedings of European Conference on Machine Learning*, 1998.

[Mengle and Goharian, 2009] S.S.R. Mengle and N. Goharian. Ambiguity measure feature-selection algorithm. *Journal of the American Society for Information Science and Technology*, 60(5):1037–1050, 2009.

[Molina *et al.*, 2002] L. Molina, L. Belanche, and A. Nebot. Feature selection algorithms: A survey and experimental evaluation. In *Proceedings of IEEE International Conference on Data Mining*, pages 306–313, 2002.

[Pawlak, 1991] Z. Pawlak. *Rough Sets, Theoretical aspects of reasoning about data*. Kluwer Academic Publishers, 1991.

[Peng *et al.*, 2005] H. Peng, F. Long, and C. Ding. Feature selection based on mutual information: Criteria of max-dependency, max-relevance and min-redundancy. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 27(8), August 2005.

[Shin and Xu, 2009] K. Shin and X.M. Xu. Consistency-based feature selection. In *13th International Conferece on Knowledge-Based and Intelligent Information & Engineering System*, 2009.

[Shin *et al.*, 2011] K. Shin, D. Fernandes, and S. Miyazaki. Consistency measures for feature selection: A formal definition, relative sensitivity comparison, and a fast algorithm. In *22nd International Joint Conference on Artificial Intelligence*, pages 1491–1497, 2011.

[Suzuki and Sugiyama, 2013] T. Suzuki and M. Sugiyama. Sufficient dimension reduction via squared-loss mutual information estimation. *Neural Computation*, 25(3):725–758, 2013.

[Wang, 2015] L. Wang. Feature selection algorithm based on conditional dynamic mutual information. *International Journal on Smart Sensing and Intelligent Systems*, 8(1), 2015.

[Zhao and Liu, 2007] Z. Zhao and H. Liu. Searching for interacting features. In *Proceedings of International Joint Conference on Artificial Intelligence*, pages 1156 – 1161, 2007.