Unsupervised Feature Selection Using Local ID

Oussama CHELLY   Michael E. HOULE   Ken-ichi KAWARABAYASHI

デー タはどのように効率的に作るのか？

我々は、関連する機 能を保持し、冗 長性と無関係なも のを除去する特徴選 択方法を提案する。
学習アルゴリズムの効率と効果を向上するために、デー タの前処理、デー タの収集、 保存、および処理のコストを低減する目標である。
特徴選択ア ルゴリズムは、医 療データ、マルチメディア、財 務データしょりなどの多くの用途に 適する。

What is feature selection?

Feature selection consists of retaining ‘relevant’ features and removing redundant and irrelevant ones. This pre-processing of data:

- reduces the costs of data collection, storage and processing.
- reduces the effects of noise and overfitting.
- improves the efficiency and effectiveness of learning algorithms.

Applying feature selection algorithms does not require domain knowledge. Therefore, these algorithms have a very wide range of applications including medical data, multimedia, financial data, etc.

Feature Selection

- Generation of a subset of the features that realizes the best possible improvement in the performance of machine learning tasks.
- Given a matrix $X \in M_{n,m}(\mathbb{R})$ representing a set of n points in a space of dimension m
$$X = (x_1, x_2, ..., x_n) = (f_1, f_2, ..., f_m)^\top$$
where $x_i \in R^m$ are points and $f_i \in R^n$ are features.
- A feature selection algorithm consists of finding $F^* = \{f_1^*, ..., f_d^*\}$, a subset of $F = \{f_1, ..., f_m\}$ that satisfies some optimality criterion.
- Most feature optimality criteria are affected by the Curse of Dimensionality.

InDiscriminability & Intrinsic Dimensionality

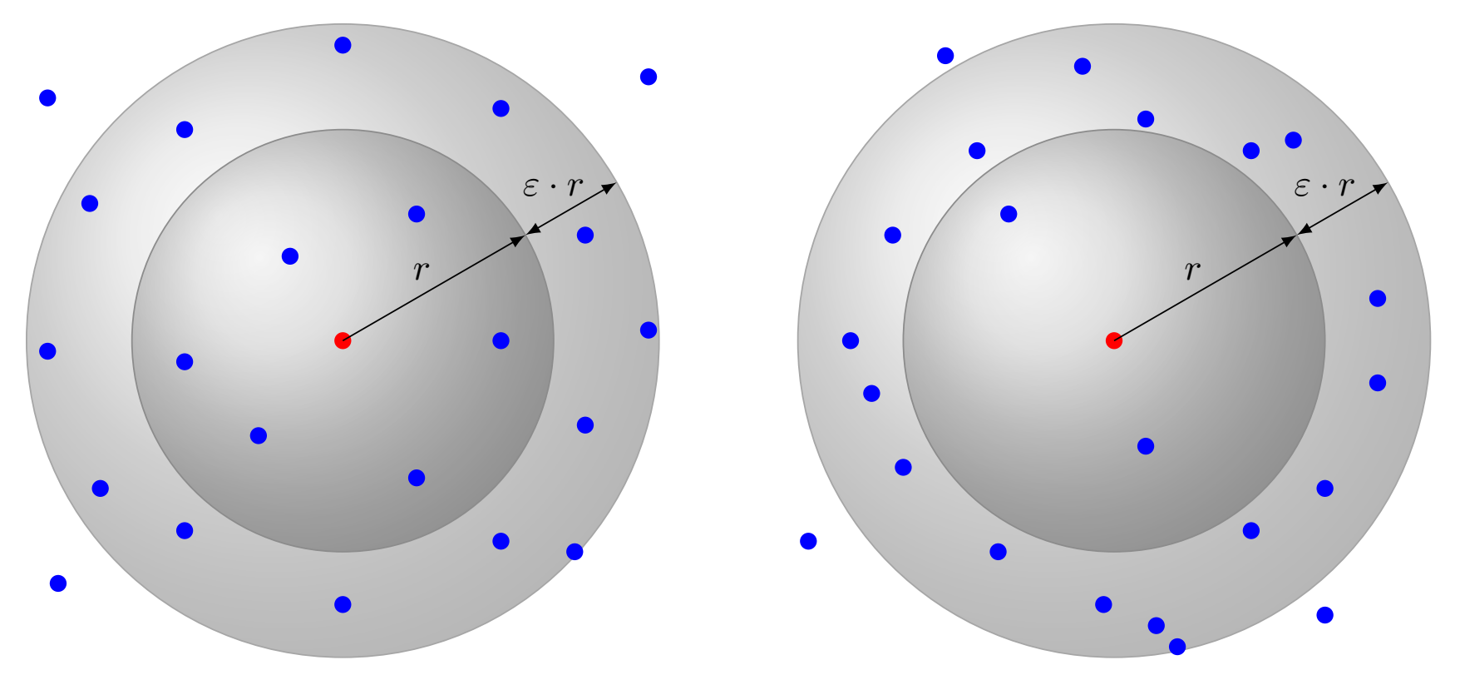


Figure: Blue points are more discriminable in the left-hand setting than in the right-hand setting.

In L_p-norm spaces, if V is a measure of volume then the representational dimension m is:

$$m = \frac{\ln V((1+\epsilon)r) - \ln V(r)}{\ln((1+\epsilon)r) - \ln r}$$

Let $\mathbf{R}$ be an absolutely continuous random distance variable with

- cumulative distribution function F_R
- probability density function Φ_R

InDiscriminability & Intrinsic Dimensionality

When $F_R(r) > 0$, the local intrinsic dimensionality of R at distance r is defined as

$$\mathrm{IntrDim}_R(x, r) = \lim_{\epsilon \to 0^+} \left(\frac{\ln F_R((1+\epsilon)r) - \ln F_R(r)}{\ln(1+\epsilon)} \right)$$

When $F_R(r) > 0$, the indiscriminability of R at distance r is given by the following limit

$$\mathrm{InDiscr}_R(\mathrm{x}, \mathrm{r}) = \lim_{\epsilon \to 0^+} \left(\frac{F_R((1+\epsilon)r) - F_R(r)}{\epsilon \cdot F_R(r)} \right)$$

With these definitions, Local Intrinsic Dimension and InDiscriminability are the same:

$$\mathrm{IntrDim}_R(x, r) = \mathrm{InDiscr}_R(x, r)$$
$$\triangleq \mathrm{ID}_R(x, r) = \frac{r \cdot \phi_R(r)}{F_R(r)}$$

Extreme Value Theory Estimation of ID

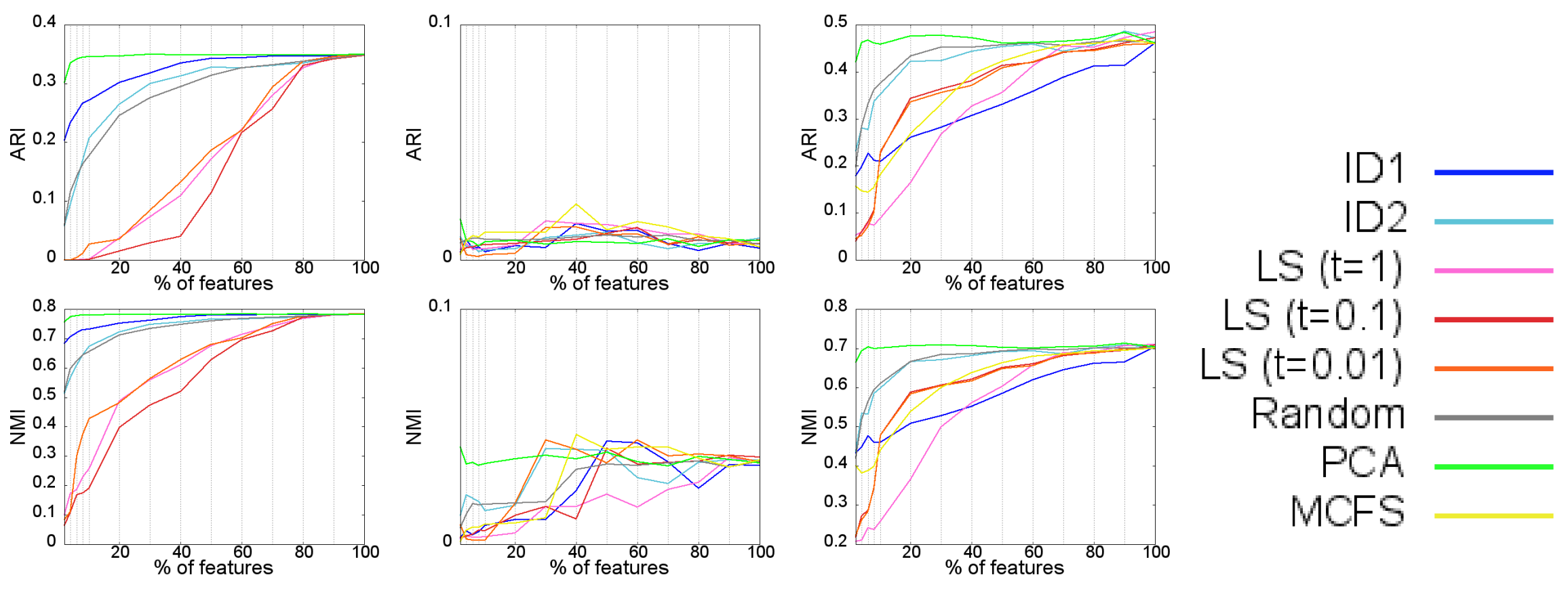
- Using Extreme Value Theory, the tail of distance distributions can be modeled as a Generalized Pareto Distribution.
- Under this assumption, maximum likelihood estimation of ID is obtained by:
$$\widehat{\mathrm{ID}}_R(x, k) = \left(-\frac{1}{k} \sum_{i=1}^{k} \ln \frac{r_i}{r_k} \right)^{-1},$$
$r_{i, i \in [1,k]}$ being the ordered distances from the reference point x to its k nearest neighbors.

Experimental Framework

Dataset	Instances	Attributes	Classes
Aloi	110250	641	1000
BCI5	31216	96	3
Gisette	7000	5000	2
Isolet	7797	617	26

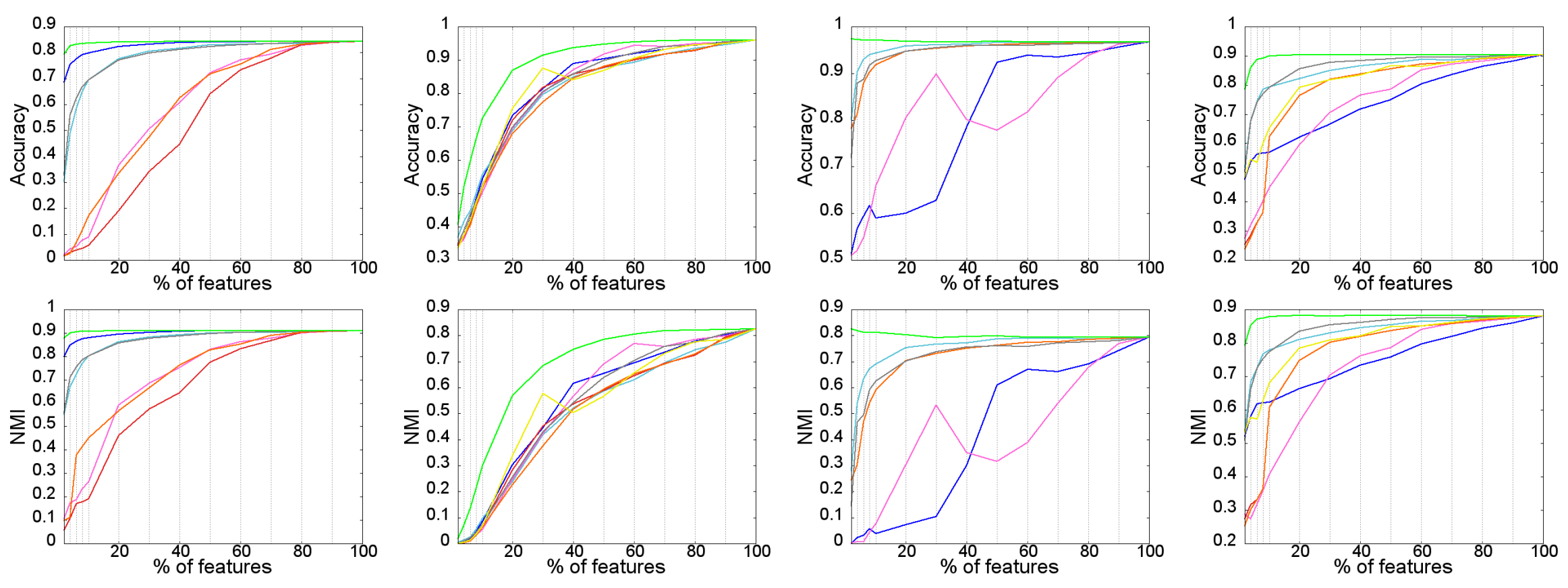
Table: Datasets used in the framework.

Experimental Results



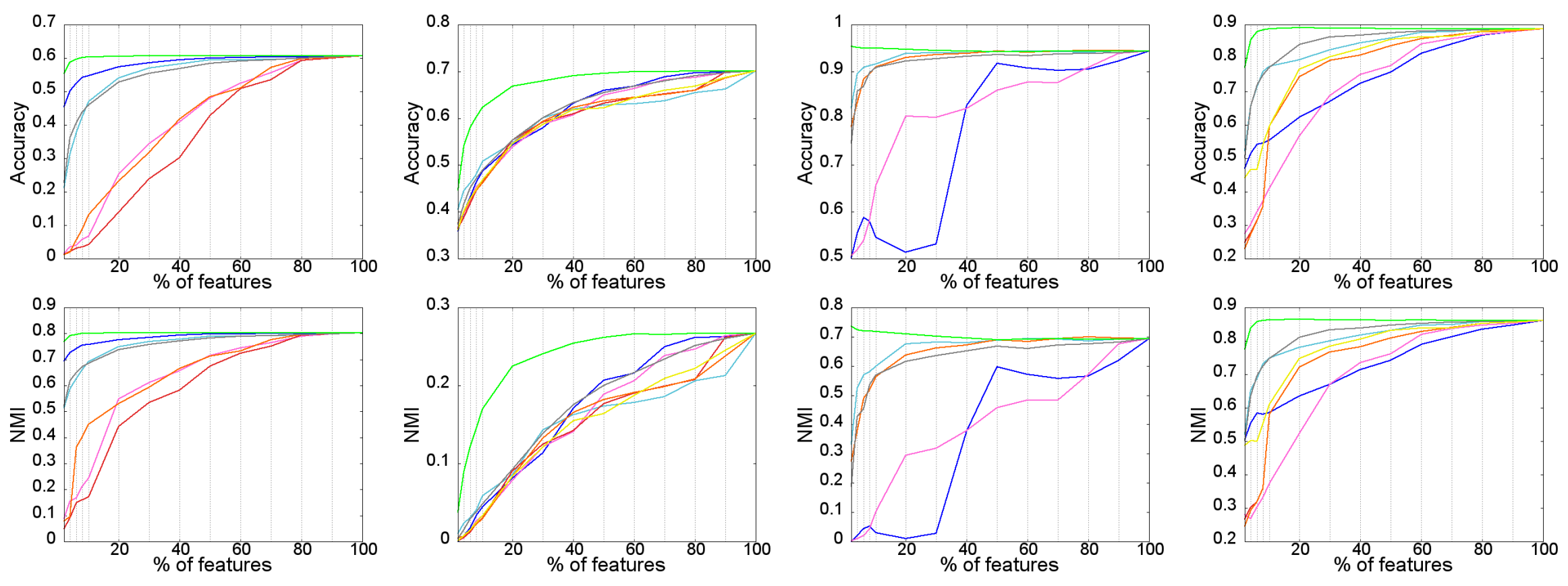
(a) Aloi (b) BCI5 (c) Isolet

Figure: ARI and NMI of K-means clustering.



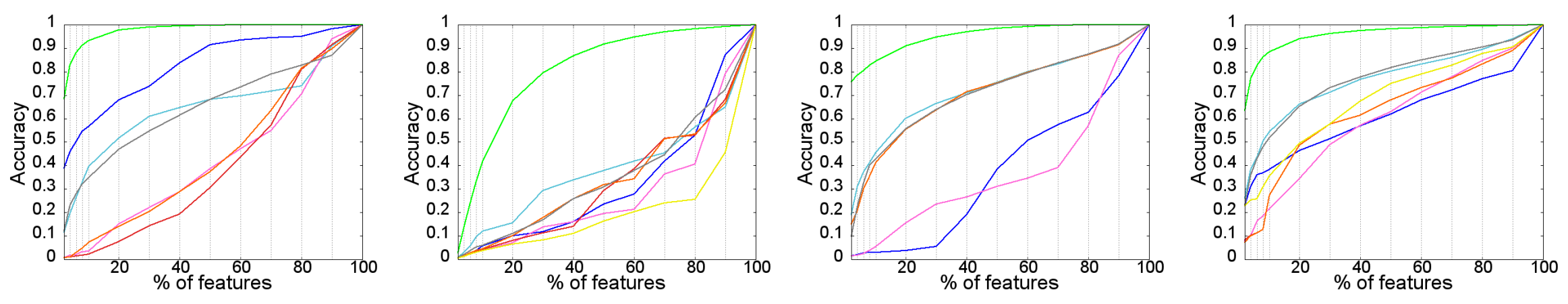
(a) Aloi (b) BCI5 (c) Gisette (d) Isolet

Figure: Accuracy and NMI of 10-NN classification.



(a) Aloi (b) BCI5 (c) Gisette (d) Isolet

Figure: Accuracy and NMI of 100-NN classification.



(a) Aloi (b) BCI5 (c) Gisette (d) Isolet

Figure: Accuracy of 100-NN indexing.

Selecting features using ID

When the representational dimension is high, ID can measure the difficulty to perform machine learning tasks. It can be engineered in order to conduct feature selection. Given a dataset $X = (x_1, x_2, ..., x_n) = (f_1, f_2, ..., f_m)^\top$ and a range k:

- Select a random subset $X^* \subset X$ of points.
- Calculate dimensionality estimates ID for each point $x \in X^*$ and for each feature f
$$\widehat{\mathrm{ID}}_f(x, k), x \in X^*, f \in \{f_1, f_2, ..., f_m\}.$$
- $F^* = \{\}$,

Univariate algorithm

Given a quantile $q \in [0, 1]$,

- Score each feature f by the q-quantile of the dimensionality estimates over the subset X^*:
$$\Theta_{\mathrm{ID}}(f, X^*) = \{\widehat{\mathrm{ID}}_f(x, k), x \in X^*\}_{(q)}.$$
- Rank features in the order of increasing scores and return:
$$F^* := \{f_1^*, f_2^*, ..., f_d^*\},$$
where $\Theta_{\mathrm{ID}}(f_i^*, X^*) < \Theta_{\mathrm{ID}}(f_j^*, X^*)$, $\forall i < j$.

Multivariate heuristic

The optimality guarantee is $(1 - 1/e)$.
A feature subset $A \subset \Omega$ is evaluated using the following score:

$$\Phi(A) = \sum_{x \in X} \phi(A, x)$$
$$\phi(A, x) = \sum_{a \in A} \alpha(\rho(a, A, x)) \cdot \beta(\mathrm{ID}_a(x)),$$

where $\alpha : \mathbb{N} \to \mathbb{R}$ is a decaying convex function, $\beta : \mathbb{R} \to \mathbb{R}$ is a decaying function, and $\rho(a, A, x)$ is the rank of the feature a in the set A in the order of increasing $\mathrm{ID}_a(x)$.

- for each point $x \in X$: rank features $f \in \Omega \setminus F^*$ by increasing $\mathrm{ID}_f(x)$, then evaluate $\phi(F^* \cup \{f\})$.
- for each $f \in \Omega \setminus F^*$, evaluate $\Phi(F^* \cup \{f\})$.
- $f^* := argmax_{f \in \Omega \setminus F^*} \Phi(F^* \cup \{f\})$.
- $F^* := F^* \cup \{f^*\}$.
- repeat until $|F^*| = d$.
- return F^*.

References

[1] M. E. Houle “Dimensionality, discriminability, density & distance distributions.”, ICDMW 2013.

[2] L. Amsaleg, O. Chelly, T. Furon, S. Girard, M. E. Houle & M. Nett “Estimating Local Intrinsic Dimensionality”, KDD2015.

Michael E. Houle　(フール マイケル)　客員教授
国立情報学研究所　〒101-8430　東京都千代田区一ツ橋2-1-2-1403
☎: 03-4212-2538　FAX: 03-4212-2120　✉: meh@nii.ac.jp