Inlierness, Outlierness, Hubness and Discriminability: an Extreme-Value-Theoretic Foundation

Michael E. HOULE

The Problem

To date, no unifying theory of data mining has been proposed.

- Many ad-hoc techniques have been designed for individual problems, such as classification or clustering.
- Solutions involve much invention and reinvention, with few guidelines.

A theoretical framework that ties together different fundamental machine learning and data mining tasks (including indexing, clustering, classification, data discriminability, subspace methods, etc.) could help the discipline, and serve as a basis for future investigation.

A First Step toward a Solution

Theoretical framework based on the statistical theory of extreme values.

- Formally unites the notions of similarity, data density, data discriminability, intrinsic dimensionality, local inlierness (cluster membership), outlierness, and hubness.
- Similar distances viewed as events of extremely small probability.
- Extreme Value Theory describes how such events are statistically distributed.
- Statistical model applies even when the nature of the data is unknown.

The Curse of Dimensionality

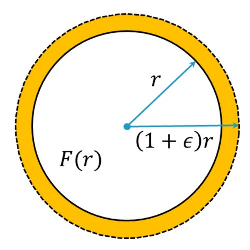
As the number of object features (data dimensionality) rises:

- Similarity values concentrate around their expected values.
- Items become less and less distinguishable.
- Data analysis based similarity (e.g. clustering and classification) becomes ineffective.

Some sets have higher *intrinsic dimensionality* (ID) than others.

- Intuitively, the minimum number of dimensions or features with which the data can be represented with minimal distortion.
- Many formalizations have been proposed (such as the Hausdorff dimension, in 1918!).

Discriminability of Distances



- Let $\mathbf{X}$ be an absolutely continuous random distance variable with c.d.f. $F_\mathbf{X}$ and p.d.f $f_\mathbf{X}$.
- Discriminability of the distance measure can be regarded as a ratio between two quantities as the distance expands infinitessimally:
- (1) the relative increase in distance, and
- (2) in probability measure.

The indiscriminability of $\mathbf{X}$ at distance r is:

$$\mathrm{InDiscr}_\mathbf{X}(r) = \lim_{\epsilon\to 0^+}\left(\frac{F_\mathbf{X}((1+\epsilon)r) - F_\mathbf{X}(r)}{\epsilon \cdot F_\mathbf{X}(r)}\right)$$

Local ID

When $F_\mathbf{X}(r) > 0$, the local intrinsic dimensionality of $\mathbf{X}$ at distance r is defined as:

$$\mathrm{IntrDim}_\mathbf{X}(r) = \lim_{\epsilon\to 0^+}\left(\frac{\ln F_\mathbf{X}((1+\epsilon)r) - \ln F_\mathbf{X}(r)}{\ln(1+\epsilon)}\right)$$

This definition is an extension for continuous distance distributions of the Expansion Dimension.

Theorem (Equivalence of ID and Indiscriminability)

Let $\mathbf{X}$ be an absolutely continuous random distance variable. If $F_\mathbf{X}$ is both positive and differentiable at r, then

$$\mathrm{IntrDim}_\mathbf{X}(r) = \mathrm{InDiscr}_\mathbf{X}(r) = \frac{r \cdot f_\mathbf{X}(r)}{F_\mathbf{X}(r)} =: \mathrm{ID}_{F_\mathbf{X}}(r).$$

ID and Scalability

Implications for Big Data:

- Data mining is greatly concerned with what happens in neighborhoods of data (clustering, classification, outlier detection, . . .).
- As the number of objects increases, the k-nearest neighbor (k-NN) distance tends to 0.
- Indiscriminability of neighborhood distances, and ID of k-NN query result, tend to $\mathrm{ID}_{F_\mathbf{X}}(0)$.

Limit effect characterizes the complexity of data.

Theorem (ID Representation Formula)

Let $\mathbf{X}$ be an absolutely continuous random distance variable such that $F_\mathbf{X}(r) > 0$ whenever $r > 0$. Then for any $r, w \in (0, z)$,

$$F_\mathbf{X}(r) = F_\mathbf{X}(w)\cdot\left(\frac{r}{w}\right)^{\mathrm{ID}_{F_\mathbf{X}}(0)}\cdot G_{F_\mathbf{X},0,w}(r),\ \text{where}$$

$$G_{F_\mathbf{X},0,w}(r) := \exp\left(\int_r^w \frac{\mathrm{ID}_{F_\mathbf{X}}(0) - \mathrm{ID}_{F_\mathbf{X}}(t)}{t}\,\mathrm{d}t\right).$$

Furthermore, for any fixed $0 < c < 1$, we have

$$\lim_{\substack{w\to 0^+\\ cw\leqslant r\leqslant w}} G_{F_\mathbf{X},0,w}(r) = 1.$$

Extreme Value Theory

- Profound importance in risk analysis, economics, civil engineering, operations research, material sciences, geophysics, ...
- Here, adapted for the lower tails of distance distributions.
- One of the three fundamental pillars of Extreme Value Theory, Karamata Characterization Theorem (1930): $F_\mathbf{X}(x) = x^{\gamma_\mathbf{X}}\ell_\mathbf{X}(1/x)$ for some constant $\gamma_\mathbf{X}$, where

$$\ell_\mathbf{X}(1/x) = \exp\left(\eta_\mathbf{X}(1/x) + \int_x^w \frac{\varepsilon_\mathbf{X}(1/t)}{t}\,\mathrm{d}t\right).$$

ID and Extreme Value Theory

- ID Representation is a more precise formulation of the Karamata Characterization, with:

$$\gamma_\mathbf{X} = \mathrm{ID}_{F_\mathbf{X}}(0)\,;$$
$$\eta_\mathbf{X}(1/x) = \ln F_\mathbf{X}(w) - \mathrm{ID}_{F_\mathbf{X}}(0)\ln w\,;$$
$$\varepsilon_\mathbf{X}(1/t) = \mathrm{ID}_{F_\mathbf{X}}(0) - \mathrm{ID}_{F_\mathbf{X}}(t)\,.$$

- $\mathrm{ID}_{F_\mathbf{X}}(0)$ is the well-studied EVT index $\gamma_\mathbf{X}$.
- Connections also exist between ID and Hausdorff dimension, and ID and the hubness phenomenon in data.

ID and Inlierness / Outlierness

If $\mathrm{ID}_{F_\mathbf{X}}(r) < \mathrm{ID}_{F_\mathbf{X}}(0)$ within neighborhood $0 < r < \epsilon$ of some point $\mathbf{p}$, then:

- The growth rate at distance r is less than that which would be expected within a uniform distribution of dimension $\mathrm{ID}_{F_\mathbf{X}}(0)$.
- The drop in indiscriminability (rise in discriminability) indicates a decrease in local density as the distance from $\mathbf{p}$ increases.
- The relationship between $\mathbf{p}$ and its neighborhood is therefore that of an *inlier*.

If instead $\mathrm{ID}_{F_\mathbf{X}}(r) > \mathrm{ID}_{F_\mathbf{X}}(0)$, $\mathbf{p}$ is an *outlier*.

2nd-Order ID

- Inlierness / outlierness is determined by the sign of $\mathrm{ID}'_{F_\mathbf{X}}(r)$ as $r \to 0^+$.
- Strength is obtained by normalizing $\mathrm{ID}'_{F_\mathbf{X}}(r)$ for distance and intrinsic dimensionality:

$$\mathrm{ID}_{\mathrm{ID}_{F_\mathbf{X}}}(r) = \frac{r\cdot \mathrm{ID}'_{F_\mathbf{X}}(r)}{\mathrm{ID}_{F_\mathbf{X}}(r)} = \mathrm{ID}_{F'_\mathbf{X}}(r) + 1 - \mathrm{ID}_{F_\mathbf{X}}(r),$$

- However, $\mathrm{ID}_{\mathrm{ID}_{F_\mathbf{X}}}(0) = 0$ always . . . need the growth rate $\mathrm{ID}_{|\mathrm{ID}_{\mathrm{ID}_{F_\mathbf{X}}}|}(0)$ of $|\mathrm{ID}_{\mathrm{ID}_{F_\mathbf{X}}}(r)|$ instead.

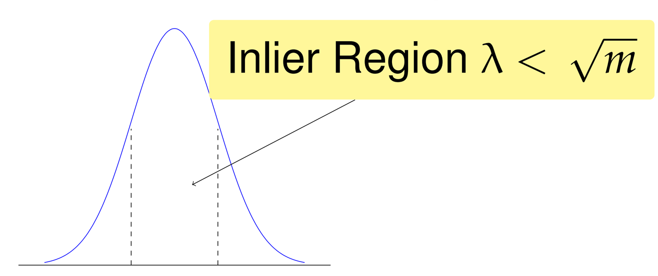
Example: Distances to a Gaussian

- Vector of normally distributed random variables with means μ_i and variances σ_i^2.
- Distance from $\mathbf{0}$ to a point $\mathbf{X} = (X_1, X_2, \ldots, X_m)$, defined as

$$Z = \sqrt{\sum_{i=1}^{m}\frac{X_i^2}{\sigma_i^2}},\ \text{and}\ \lambda = \sqrt{\sum_{i=1}^{m}\frac{\mu_i^2}{\sigma_i^2}}$$

is the normalized distance from $\mathbf{0}$ to the Gaussian mean.

- $\mathrm{ID}_{F_Z}(0) = m$ and $\mathrm{ID}_{|\mathrm{ID}_{\mathrm{ID}_{F_Z}}|}(0) = 2$ whenever $\lambda \neq \sqrt{m}$. Also, as r tends to 0, $\mathrm{ID}_{\mathrm{ID}_{F_Z}}(r) > 0$ when $\lambda > \sqrt{m}$ (tail region $\to$ outliers), and < 0 when $0 \leqslant \lambda < \sqrt{m}$ (central region $\to$ inliers).
- Normalization works! Independent of λ & m.



References

[1] M. E. Houle. “Inlierness, Outlierness, Hubness and Discriminability: an Extreme-Value-Theoretic Foundation”, NII Technical Report NII-2015-002E.

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

NII

Michael E. Houle
Address Line
☎: 03-4212-2538 FAX: 03-4212-2120 ✉: meh@nii.ac.jp