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 discriminality, subspace methods, etc.) could help the discipline, and serve as a basis for future investigation.

Theoretical framework based on the statistical theory of extreme values. Formally unites the notions of similarity, data density, data discriminality, 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 X be an absolutely continuous random distance variable with c.d.f. F_X and p.d.f f_X.

Discriminability of the distance measure can be regarded as a ratio between two quantities as the distance expands infinitesimally:

1. The relative increase in distance, and
2. (in probability measure).

The indiscriminability of X at distance r is:

\[
\text{Indiscr}_X(r) = \lim_{\epsilon \to 0^+} \frac{F_X((1+\epsilon)r) - F_X(r)}{\epsilon \cdot F_X(r)}
\]

Local ID
When \( F_X(r) > 0 \), the local intrinsic dimensionality of X at distance r is defined as:

\[
\text{IntrDim}_X(r) = \lim_{\epsilon \to 0^+} \frac{\ln F_X((1+\epsilon)r) - \ln F_X(r)}{\ln(1+\epsilon)}
\]

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

Equivalence of ID and Indiscriminability
Let X be an absolutely continuous random distance variable. If F_X is both positive and differentiable at r, then

\[
\text{IntrDim}_X(r) = \text{Indiscr}_X(r) = r \cdot \left( \frac{F_X(r)}{F_X'(r)} \right) = ID_X(r).
\]

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 neighborhoods distances, and ID of k-NN query result, tend to ID_X(0).

Limit effect characterizes the complexity of data.

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):

\[
F_X(x) = x^{\gamma_X} \cdot \eta_X(1/x) \quad \text{for some constant } \gamma_X,
\]

where

\[
\eta_X(1/x) = \exp \left( \frac{\ln F_X(1/x)}{r} \cdot \int_x^\infty \frac{\ln(1/t)}{t} \, dt \right).
\]

ID and Extreme Value Theory
ID Representation is a more precise formulation of the Karamata Characterization, with:

\[
\gamma_X = ID_X(0);
\]

\[
\eta_X(1/x) = \ln F_X(w) - ID_X(0) \ln w;
\]

\[
x_X(1/t) = ID_X(0) - ID_X(t);
\]

\[
\lambda_X = ID_X(0)\text{ is the well-studied EVT index } \gamma_X.
\]

Connections also exist between ID and Hausdorff dimension, and ID and the hubness phenomenon in data.

Theorem (Equivalence of ID and Indiscriminability)

2nd-Order ID

- Inlierness / outlierness is determined by the sign of ID_X(r) as r → 0⁺.
- Strength is obtained by normalizing ID_X(r) for distance and intrinsic dimensionality:

\[
ID_X(r) - ID_X(0) = ID_X(r) + 1 - ID_X(0),
\]

- However, ID_X(0) = 0 always . . . need the growth rate ID_X(0) of ID_X(r) instead.

Example: Distances to a Gaussian

- Vector of normally distributed random variables with means \( \mu_i \) and variances \( \sigma_i^2 \).

- Distance from 0 to a point \( X = (X_1, X_2, \ldots, X_m) \), defined as

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

is the normalized distance from 0 to the Gaussian mean.

\[
ID_X(0) = m \quad \text{and } ID_{D_{\text{D}}}(0) = 2 \quad \text{whenever } \lambda < \sqrt{m}.
\]

Also, as r tends to 0, ID_X(r) → 0 when \( \lambda > \sqrt{m} \) (tail region → outliers), and < 0 when \( 0 < \lambda < \sqrt{m} \) (central region → inliers).

Normalization works! Independent of \( \lambda \) and \( m \).

References
