DEPARTMENT OF MATHEMATICS AND COMPUTER SCIENCE
UNIVERSITY OF SOUTHERN DENMARK, ODENSE

COMPUTER SCIENCE COLLOQUIUM

An Extreme-Value-Theoretic Foundation for Similarity Applications

Michael E. Houle
National Institute of Informatics
Tokyo, Japan

Monday, 28 August, 2017 at 10:15

IMADA's Seminar Room

Abstract:

For many large-scale applications in data mining, machine learning, and multimedia, fundamental operations such as similarity search, retrieval, classification, clustering, and anomaly detection generally suffer from an effect known as the 'curse of dimensionality'. As the dimensionality of the data increases, distance values tend to become less discriminative due to their increasing relative concentration about the mean of their distribution. For this reason, researchers have considered the analysis of similarity applications in terms of measures of the intrinsic dimensionality (ID) of the data sets. This presentation is concerned with a generalization of a discrete measure of ID, the expansion dimension, to the case of continuous distance distributions. This notion of the ID of a distance distribution is shown to precisely coincide with a natural notion of the indiscriminability of distances, thereby establishing a theoretically-founded relationship among probability density, the cumulative density (cumulative probability divided by distance), intrinsic dimensionality, and discriminability. The proposed indiscriminability function is shown to completely determine an extreme-value-theoretic representation of the distance distribution. From this representation, a characterization in terms of continuous ID is derived for the notions of outlierness and inlierness of data. Some implications for the analysis of feature ensembles will also be presented.

Host: Arthur Zimek