Automatic topography of high-dimensional data sets by non-parametric density peak clustering. (English) Zbl 1484.62075

Summary: Data analysis in high-dimensional spaces aims at obtaining a synthetic description of a data set, revealing its main structure and its salient features. We here introduce an approach providing this description in the form of a topography of the data, namely a human-readable chart of the probability density from which the data are harvested. The approach is based on an unsupervised extension of Density Peak clustering and on a non-parametric density estimator that measures the probability density in the manifold containing the data. This allows finding automatically the number and the height of the peaks of the probability density, and the depth of the “valleys” separating them. Importantly, the density estimator provides a measure of the error, which allows distinguishing genuine density peaks from density fluctuations due to finite sampling. The approach thus provides robust and visual information about the density peaks height, their statistical reliability and their hierarchical organization, offering a conceptually powerful extension of the standard clustering partitions. We show that this framework is particularly useful in the analysis of complex data sets.

MSC:
62H30 Classification and discrimination; cluster analysis (statistical aspects)
62G07 Density estimation

Keywords:
clustering algorithm; high-dimensional data; hierarchy visualization; density-peak clustering; nonparametric density estimation

Software:
Pfam; Scikit; Intrinsic-Dimension; hdbscan; t-SNE

Full Text: DOI arXiv

References:


Pande, V.S.; Beauchamp, G.R.; Bowman, Everything you wanted to know about markov state models but were afraid to ask, Methods 52 (2010) 99-105. Protein Folding.


Sittel, F.; Stock, G., Robust density-based clustering to identify metastable conformational states of proteins, J. Chem. Theory
Sormani, G.; Rodriguez, A.; Laio, A., Explicit characterization of the free-energy landscape of a protein in the space of all its 


Torgerson, W. S., Multidimensional scaling: I. Theory and method, Psychometrika, 17, 401-419 (1952) · Zbl 0049.37603

Vinh, N. X.; Epps, J.; Bailey, J., Information theoretic measures for clusterings comparison: variants, properties, normalization 

Wang, Y.; Pang, W.; Zhou, Y., Density propagation based adaptive multi-density clustering algorithm, Plos One, 13, Article 


This reference list is based on information provided by the publisher or from digital mathematics libraries. Its items are heuristically 
matched to zbMATH identifiers and may contain data conversion errors. It attempts to reflect the references listed in the original 
paper as accurately as possible without claiming the completeness or perfect precision of the matching.