

# Preface

One of the most long-standing problems afflicting machine learning techniques is dataset dimensionality. Owing to the evolution of technologies for acquiring and creating information, however, this issue has recently become ubiquitous. In many applications to real-world problems, we deal with data with anywhere from a few dozen to many thousands of dimensions. Such high-dimensional data spaces are often encountered in areas such as medicine or biology, where DNA microarray technology and next-generation sequencing can produce a large number of measurements at once; the clustering of text documents, where, if a word-frequency vector is used, the number of dimensions equals the size of the dictionary; and many others, including data integration and management, and social network analysis. In all these cases, the dimensionality of data makes learning problems hardly tractable.

In particular, dimensionality is a highly critical factor for the clustering task. The following problems need to be addressed for clustering high-dimensional data:

- When the dimensionality is high, the volume of the space increases so fast that the available data become sparse, and we cannot find reliable clusters, as clusters are data aggregations (curse of dimensionality).
- The concept of distance becomes less precise as the number of dimensions grows, since the distance between any two points in a given dataset converges (concentration effects).
- Different clusters might be found in different subspaces, thus a global filtering of attributes is not sufficient (local feature relevance problem).
- Given a large number of attributes, it is likely that some attributes are correlated. Hence, clusters might exist in arbitrarily oriented affine subspaces.
- High-dimensional data could likely include irrelevant features, which may obscure the effect of the relevant ones.

This volume is the outcome of work done during the International Workshop on Clustering High-Dimensional Data, held at Istituto Italiano per gli Studi Filosofici, Palazzo Serra di Cassano, in Naples (Italy) on May 15, 2012, where speakers were subsequently invited to submit a paper related to their presentation.

The papers collected here aim to present an updated view of many different approaches toward clustering high-dimensional data, and can be divided by topic into three groups.

The first group introduces the general subject and issues of high-dimensional data clustering. Chapter 1 provides a general introduction, while Chapter 2 explores some properties of high-dimensional data that make it difficult to detect and even to define clusters.

The second group of chapters presents examples of techniques used to find and investigate clusters in high dimensionality. Chapter 3 focuses on an approach to *sub-space clustering*; Chapter 4 presents a selection of dimensionality-independent

methods for comparing clusterings; and Chapter 5 deals with clustering high-dimensional time series.

The third group deals with the most common approach to tackling dimensionality problems, namely, dimensionality reduction and its application in clustering. Chapter 6 introduces the topic of *intrinsic dimensionality estimation*, and Chapter 7 presents a specific technique for intrinsic dimensionality estimation. Chapter 8 compares four dimensionality reduction methods for binary data, while the last contribution, Chapter 9, focuses on dimensionality reduction by feature selection using rough-fuzzy techniques.

July 2015

Francesco Masulli  
Alfredo Petrosino  
Stefano Rovetta

Clustering High--Dimensional Data

First International Workshop, CHDD 2012, Naples, Italy,

May 15, 2012, Revised Selected Papers

Masulli, F.; Petrosino, A.; Rovetta, S. (Eds.)

2015, IX, 149 p. 41 illus. in color., Softcover

ISBN: 978-3-662-48576-7