Chapter 3 - Unsupervised learning

- January 2020
- DOI: 10.1016/B978-0-12-811842-9.00003-0
- In book: Advantages and Pitfalls of Pattern Recognition
- Publisher: Elsevier

- Authors: Horst Langer, Susanna Falsaperla, Conny Hammer

Abstract

Unsupervised learning is based on the definition of an appropriate metrics defining the similarity of patterns. On the basis of the metrics, we form groups or clusters of patterns following various strategies. In partitioning cluster analysis, we form disjoint clusters. Being faced with data, where clusters still exhibit heterogeneities or subclusters, we may adopt the strategy of hierarchical clustering, which leads to the generation of the so-called dendrograms. In the partitioning strategy, we choose a priori the number of clusters we wish to form, whereas in the hierarchical strategy, the number of clusters depends on the resolution we want to have. Density-based clustering considers local structures of a data set. We consider a unit volume in our data space and derive the density of samples within this volume. Moving toward neighboring volumes, we verify whether the number of samples has dropped below a threshold. If this is the case, we identify a heterogeneity, otherwise we join the neighboring volumes to a common cluster. Self-Organizing Maps (SOMs) provide a way of representing multidimensional data in much lower dimensional spaces than the original data set. The process of reducing the dimensionality of vectors is essentially a data compression technique known as vector quantization. The SOM technique creates a network that stores information in a way that it maintains the topological relationships within the patterns of the data set. Each node of the network represents a number of patterns. Assigning a color code to the nodes, the representation of pattern characteristics with high-dimensional feature vectors becomes extremely effective.