Chapter 11

Advanced Cluster Analysis

11.1 Bibliographic Notes

Höppner et al. [HKKR99] provide a thorough discussion on fuzzy clustering. The fuzzy c-means algorithm (on which Example 11.7 is based) was proposed by Bezdek [Bez81]. Fraley and Raftery [FR02] give a comprehensive overview of model-based cluster analysis and probabilistic models. McLachlan and Bkasford [MB88] present a systematic introduction to mixture models and applications in cluster analysis.

Dempster, Laird, and Rubin [DLR77] are recognized as the first to introduce the EM algorithm and give it its name. However, the idea of the EM algorithm had been “proposed many times in special circumstances” before, as admitted in [DLR77]. Wu [Wu83] gives the correct analysis of the EM algorithm.

Mixture models and EM algorithms are used extensively in many data mining applications. Introductions to model-based clustering, mixture models, and EM algorithms can be found in recent textbooks on machine learning and statistical learning, such as [Bis06, Mar09, Alp11].

The increase of dimensionality has severe effects on distance functions, as indicated by Beyer et al. [BGRS99]. It also has had a dramatic impact on various techniques for classification, clustering, and semi-supervised learning [RN09].

Kriegel, Kröger, and Zimek [KKZ09] present a comprehensive survey on methods for clustering high-dimensional data. The CLIQUE algorithm was developed by Agrawal, Gehrke, Gunopulos, and Raghavan [AGGR98]. The PROCLUS algorithm was proposed by Aggawal, Procopiuc, Wolf et al. [APW99].

The technique of bi-clustering was initially proposed by Hartigan [Har72]. The term of bi-clustering was coined by Mirkin [Mir98]. Cheng and Church [CC00] introduced bi-clustering into gene expression data analysis. There are many studies on bi-clustering models and methods. The notion of δ-pCluster was introduced by Wang, Wang, Yang, and Yu [WWYY02]. For informative surveys, see Madeira and Oliveira [MO04] and Tanay, Sharan, and Shamir [TSS04]. In this chapter, we introduced the δ-cluster algorithm by Cheng...
and Church [CC00] and MaPle by Pei, Zhang, Cho, et al. [PZC+03] as examples of optimization-based methods and enumeration methods for bi-clustering, respectively.

Donath and Hoffman [DH73] and Fiedler [Fie73] pioneered spectral clustering. In this chapter, we use an algorithm proposed by Ng, Jordan, and Weiss [NJW01] as an example. For a thorough tutorial on spectral clustering, see Luxburg [Lux07].

Clustering graph and network data is an important and fast growing topic. Schaeffer [Sch07] provides a survey. The SimRank measure of similarity was developed by Jeh and Widomi [JW02]. Xu et al. [XYFS07] proposed the SCAN algorithm. Arora, Rao, and Vazirani [ARV09] discuss the sparsest cuts and approximation algorithms.

Clustering with constraints has been extensively studied. Davidson, Wagstaff, and Basu [DWB06] proposed the measures of informativeness and coherence. The COP-k-means algorithm is given by Wagstaff et al. [WCRS01]. The CVQE algorithm was proposed by Davidson and Ravi [DR05]. Tung, Han, Lakshmanan, and Ng [THLN01] presented a framework for constraint-based clustering based on user-specified constraints. An efficient method for constraint-based spatial clustering in the existence of physical obstacle constraints was proposed by Tung, Hou and Han [THH01].
Bibliography


