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 [RNI09].

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. [APW$^+$99].

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 Widom [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

[AGGR98] R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan. Automatic subspace clustering of high dimensional data for data mining applications. In *Proc. 1998 ACM-SIGMOD Int. Conf. Management of Data (SIGMOD'98)*, pages 94–105, Seattle, WA, June 1998.

[Alp11] E. Alpaydin. *Introduction to Machine Learning (2nd ed.).* MIT Press, 2011.

[APW^{+}99] C. C. Aggarwal, C. Procopiuc, J. Wolf, P. S. Yu, and J.-S. Park. Fast algorithms for projected clustering. In *Proc. 1999 ACM-SIGMOD Int. Conf. Management of Data (SIGMOD'99)*, pages 61–72, Philadelphia, PA, June 1999.

[ARV09] S. Arora, S. Rao, and U. Vazirani. Expander flows, geometric embeddings and graph partitioning. *J. ACM*, 56:5:1–5:37, 2009.

[Bez81] J. C. Bezdek. *Pattern Recognition with Fuzzy Objective Function Algorithms.* Plenum Press, 1981.

[BGRS99] K. S. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft. When is "nearest neighbor" meaningful? In *Proc. 1999 Int. Conf. Database Theory (ICDT'99)*, pages 217–235, Jerusalem, Israel, Jan. 1999.

[Bis06] C. M. Bishop. *Pattern Recognition and Machine Learning.* Springer, 2006.

[CC00] Y. Cheng and G. Church. Biclustering of expression data. In *Proc. 2000 Int. Conf. Intelligent Systems for Molecular Biology (ISMB'00)*, pages 93–103, La Jolla, CA, Aug. 2000.

[DH73] W. E. Donath and A. J. Hoffman. Lower bounds for the partitioning of graphs. *IBM J. Research and Development*, 17:420–425, 1973.

[DLR77] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Statistical Society, Series B*, 39:1–38, 1977.

[DR05] I. Davidson and S. S. Ravi. Clustering with constraints: Feasibility issues and the k-means algorithm. In *Proc. 2005 SIAM Int. Conf. Data Mining (SDM'05)*, Newport Beach, CA, Apr. 2005.

[DWB06] I. Davidson, K. L. Wagstaff, and S. Basu. Measuring constraint-set utility for partitional clustering algorithms. In *Proc. 10th European Conf. Principles and Practice of Knowledge Discovery in Databases (PKDD'06)*, Berlin, Germany, Sept. 2006.

[Fie73] M. Fiedler. Algebraic connectivity of graphs. *Czechoslovak Mathematical J.*, 23:298–305, 1973.

[FR02] C. Fraley and A. E. Raftery. Model-based clustering, discriminant analysis, and density estimation. *J. American Statistical Association*, 97:611–631, 2002.

[Har72] J. Hartigan. Direct clustering of a data matrix. *J. American Stat. Assoc.*, 67:123–129, 1972.

[HKKR99] F. Höppner, F. Klawonn, R. Kruse, and T. Runkler. *Fuzzy Cluster Analysis: Methods for Classification, Data Analysis and Image Recognition.* Wiley, 1999.

[JW02] G. Jeh and J. Widom. SimRank: a measure of structural-context similarity. In *Proc. 2002 ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD'02)*, pages 538–543, Edmonton, Canada, July 2002.

[KKZ09] H.-P. Kriegel, P. Kroeger, and A. Zimek. Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering. *ACM Trans. Knowledge Discovery from Data (TKDD)*, 3:1–58, 2009.

[Lux07] U. Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17:395–416, 2007.

[Mar09] S. Marsland. *Machine Learning: An Algorithmic Perspective.* Chapman and Hall/CRC, 2009.

[MB88] G. J. McLachlan and K. E. Basford. *Mixture Models: Inference and Applications to Clustering.* John Wiley & Sons, 1988.

[Mir98] B. Mirkin. Mathematical classification and clustering. *J. of Global Optimization*, 12:105–108, 1998.

[MO04] S. C. Madeira and A. L. Oliveira. Biclustering algorithms for biological data analysis: A survey. *IEEE/ACM Trans. Computational Biology and Bioinformatics*, 1:24–45, 2004.

[NJW01] A. Y. Ng, M. I. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems 14, MIT Press*, pages 849–856, 2001.

[PZC$^+$03] J. Pei, X. Zhang, M. Cho, H. Wang, and P. S. Yu. Maple: A fast algorithm for maximal pattern-based clustering. In *Proc. 2003 Int. Conf. Data Mining (ICDM'03)*, pages 259–266, Melbourne, FL, Dec. 2003.

[RNI09] M. Radovanović, A. Nanopoulos, and M. Ivanović. Nearest neighbors in high-dimensional data: the emergence and influence of hubs. In *Proc. 2009 Int. Conf. Machine Learning (ICML'09)*, pages 865–872, Montreal, Quebec, Canada, June 2009.

[Sch07] S. E. Schaeffer. Graph clustering. *Computer Science Review*, 1:27–64, 2007.

[THH01] A.K.H. Tung, J. Hou, and J. Han. Spatial clustering in the presence of obstacles. In *Proc. 2001 Int. Conf. Data Engineering (ICDE'01)*, pages 359–367, Heidelberg, Germany, April 2001.

[THLN01] A.K.H. Tung, J. Han, L.V.S. Lakshmanan, and R. T. Ng. Constraint-based clustering in large databases. In *Proc. 2001 Int. Conf. Database Theory (ICDT'01)*, pages 405–419, London, UK, Jan. 2001.

[TSS04] A. Tanay, R. Sharan, and R. Shamir. Biclustering algorithms: A survey. In *Handbook of Computational Molecular Biology, Chapman & Hall*, pages 26:1–17, 2004.

[WCRS01] K. Wagstaff, C. Cardie, S. Rogers, and S. Schrödl. Constrained k-means clustering with background knowledge. In *Proc. 2001 Int. Conf. Machine Learning (ICML'01)*, pages 577–584, Williamstown, MA, June 2001.

[Wu83] C. F. J. Wu. On the convergence properties of the EM algorithm. *Ann. Statistics*, 11:95–103, 1983.

[WWYY02] H. Wang, W. Wang, J. Yang, and P. S. Yu. Clustering by pattern similarity in large data sets. In *Proc. 2002 ACM-SIGMOD Int. Conf. Management of Data (SIGMOD'02)*, pages 418–427, Madison, WI, June 2002.

[XYFS07] X. Xu, N. Yuruk, Z. Feng, and T. A. J. Schweiger. SCAN: A structural clustering algorithm for networks. In *Proc. 2007 ACM SIGKDD Int. Conf. Knowledge Discovery in Databases (KDD'07)*, San Jose, CA, Aug. 2007.