

# Data Mining: Concepts and Techniques (2nd edition)

Jiawei Han and Micheline Kamber  
Morgan Kaufmann Publishers, 2006

## Bibliographic Notes for Chapter 7 Cluster Analysis

Clustering has been studied extensively for more than 40 years and across many disciplines due to its broad applications. Most books on pattern classification and machine learning contain chapters on cluster analysis or unsupervised learning. Several textbooks are dedicated to the methods of cluster analysis, including Hartigan [Har75], Jain and Dubes [JD88], Kaufman and Rousseeuw [KR90], and Arabie, Hubert, and De Sorte [AHS96]. There are also many survey articles on different aspects of clustering methods. Recent ones include Jain, Murty, and Flynn [JMF99] and Parsons, Haque, and Liu [PHL04].

Methods for combining variables of different types into a single dissimilarity matrix were introduced by Kaufman and Rousseeuw [KR90].

For partitioning methods, the  $k$ -means algorithm was first introduced by Lloyd [Llo57] and then MacQueen [Mac67]. The  $k$ -medoids algorithms of PAM and CLARA were proposed by Kaufman and Rousseeuw [KR90]. The  $k$ -modes (for clustering categorical data) and  $k$ -prototypes (for clustering hybrid data) algorithms were proposed by Huang [Hua98]. The  $k$ -modes clustering algorithm was also proposed independently by Chaturvedi, Green, and Carroll [CGC94, CGC01].

The CLARANS algorithm was proposed by Ng and Han [NH94]. Ester, Kriegel, and Xu [EKX95] proposed techniques for further improvement of the performance of CLARANS using efficient spatial access methods, such as  $R^*$ -tree and focusing techniques. A  $k$ -means-based scalable clustering algorithm was proposed by Bradley, Fayyad, and Reina [BFR98].

An early survey of agglomerative hierarchical clustering algorithms was conducted by Day and Edelsbrunner [DE84]. Agglomerative hierarchical clustering, such as AGNES, and divisive hierarchical clustering, such as DIANA, were introduced by Kaufman and Rousseeuw [KR90]. An interesting direction for improving the clustering quality of hierarchical clustering methods is to integrate hierarchical clustering with distance-based iterative relocation or other nonhierarchical clustering methods. For example, BIRCH, by Zhang, Ramakrishnan, and Livny [ZRL96], first performs hierarchical clustering with a CF-tree before applying other techniques. Hierarchical clustering can also be performed by sophisticated linkage analysis, transformation, or nearest-neighbor analysis, such as CURE by Guha, Rastogi, and Shim [GRS98], ROCK (for clustering categorical attributes) by Guha, Rastogi, and Shim [GRS99], and Chameleon by Karypis, Han, and Kumar [KHK99].

For density-based clustering methods, DBSCAN was proposed by Ester, Kriegel, Sander, and Xu [EKXS96]. Ankerst, Breunig, Kriegel, and Sander [ABKS99] developed OPTICS, a cluster-ordering method that facilitates density-based clustering without worrying about parameter specification. The DENCLUE algorithm, based on a set of density distribution functions, was proposed by Hinneburg and Keim [HK98].

A grid-based multiresolution approach called STING, which collects statistical information in grid cells, was proposed by Wang, Yang, and Muntz [WYM97]. WaveCluster, developed by Sheikholeslami, Chatterjee, and Zhang [SCZ98], is a multiresolution clustering approach that transforms the original feature space by wavelet transform.

For model-based clustering, the EM (Expectation-Maximization) algorithm was developed by Dempster, Laird, and Rubin [DLR77]. AutoClass is a Bayesian statistics-based method for model-based clustering by Cheeseman and Stutz [CS96] that uses a variant of the EM algorithm. There are many other extensions and applications of EM, such as Lauritzen [Lau95]. For a set of seminal papers on conceptual clustering, see Shavlik and Dietterich [SD90]. Conceptual clustering was first introduced by Michalski and Stepp [MS83]. Other examples of the conceptual clustering approach include COBWEB by Fisher [Fis87], and CLASSIT by Gennari, Langley, and Fisher [GLF89].

Studies of the neural network approach [He99] include SOM (self-organizing feature maps) by Kohonen [Koh82, Koh89], by Carpenter and Grossberg [Ce91], and by Kohonen, Kaski, Lagus, et al. [KKL<sup>+</sup>00], and competitive learning by Rumelhart and Zipser [RZ85].

Scalable methods for clustering categorical data were studied by Gibson, Kleinberg, and Raghavan [GKR98], Guha, Rastogi, and Shim [GRS99], and Ganti, Gehrke, and Ramakrishnan [GGR99]. There are also many other clustering paradigms. For example, fuzzy clustering methods are discussed in Kaufman and Rousseeuw [KR90], Bezdek [Bez81], and Bezdek and Pal [BP92].

For high-dimensional clustering, an Apriori-based dimension-growth subspace clustering algorithm called CLIQUE was proposed by Agrawal, Gehrke, Gunopulos, and Raghavan [AGGR98]. It integrates density-based and grid-based clustering methods. A sampling-based, dimension-reduction subspace clustering algorithm called PROCLUS, and its extension, ORCLUS, were proposed by Aggarwal, Procopiuc, Wolf, et al. [APW<sup>+</sup>99] and by Aggarwal and Yu [AY00], respectively. An entropy-based subspace clustering algorithm for mining numerical data, called ENCLUS, was proposed by Cheng, Fu, and Zhang [CFZ99]. For a frequent pattern-based approach to handling high-dimensional data, Beil, Ester and Xu [BEX02] proposed a method for frequent term-based text clustering. Wang, Wang, Yang, and Yu proposed pCluster, a pattern similarity-based clustering method [WWYY02].

Recent studies have proceeded to clustering stream data, as in Babcock, Babu, Datar, et al. [BBD<sup>+</sup>02]. A  $k$ -median-based data stream clustering algorithm was proposed by Guha, Mishra, Motwani, and O'Callaghan [GMMO00], and by O'Callaghan, Mishra, Meyerson, et al. [OMM<sup>+</sup>02]. A method for clustering evolving data streams was proposed by Aggarwal, Han, Wang, and Yu [AHWY03]. A framework for projected clustering of high-dimensional data streams was proposed by Aggarwal, Han, Wang, and Yu [AHWY04].

A framework for constraint-based clustering based on user-specified constraints was built by Tung, Han, Lakshmanan, and Ng [THLN01]. An efficient method for constraint-based spatial clustering in the existence of physical obstacle constraints was proposed by Tung, Hou, and Han [THH01]. The quality of unsupervised clustering can be significantly improved using supervision in the form of pairwise constraints (i.e., pairs of instances labeled as belonging to the same or different clustering). Such a process is considered semi-supervised clustering. A probabilistic framework for semi-supervised clustering was proposed by Basu, Bilenko, and Mooney [BBM04]. A CLTree method that transforms the clustering problem into a classification problem and then uses decision tree induction for cluster analysis was proposed by Liu, Xia, and Yu [LXY01].

Outlier detection and analysis can be categorized into four approaches: the statistical approach, the distance-based approach, the density-based local outlier detection, and the deviation-based approach. The statistical approach and discordancy tests are described in Barnett and Lewis [BL94]. Distance-based outlier detection is described in Knorr and Ng [KN97, KN98]. The detection of density-based local outliers was proposed by Breunig, Kriegel, Ng, and Sander [BKNS00]. Outlier detection for high-dimensional data is studied by Aggarwal and Yu [AY01]. The sequential problem approach to deviation-based outlier detection was introduced in Arning, Agrawal, and Raghavan [AAR96]. Sarawagi, Agrawal, and Megiddo [SAM98] introduced a discovery-driven method for identifying exceptions in large multidimensional data using OLAP data cubes. Jagadish, Koudas, and Muthukrishnan [JKM99] introduced an efficient method for mining deviants in time-series databases.

# Bibliography

- [AAR96] A. Arning, R. Agrawal, and P. Raghavan. A linear method for deviation detection in large databases. In *Proc. 1996 Int. Conf. Data Mining and Knowledge Discovery (KDD'96)*, pages 164–169, Portland, OR, Aug. 1996.
- [ABKS99] M. Ankerst, M. Breunig, H.-P. Kriegel, and J. Sander. OPTICS: Ordering points to identify the clustering structure. In *Proc. 1999 ACM-SIGMOD Int. Conf. Management of Data (SIGMOD'99)*, pages 49–60, Philadelphia, PA, June 1999.
- [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.
- [AHS96] P. Arabie, L. J. Hubert, and G. De Soete. *Clustering and Classification*. World Scientific, 1996.
- [AHWY03] C. C. Aggarwal, J. Han, J. Wang, and P. S. Yu. A framework for clustering evolving data streams. In *Proc. 2003 Int. Conf. Very Large Data Bases (VLDB'03)*, pages 81–92, Berlin, Germany, Sept. 2003.
- [AHWY04] C. Aggarwal, J. Han, J. Wang, and P. S. Yu. A framework for projected clustering of high dimensional data streams. In *Proc. 2004 Int. Conf. Very Large Data Bases (VLDB'04)*, pages 852–863, Toronto, Canada, Aug. 2004.
- [APW<sup>+</sup>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.
- [AY00] C. C. Aggarwal and P. S. Yu. Finding generalized projected clusters in high dimensional spaces. In *Proc. 2000 ACM-SIGMOD Int. Conf. Management of Data (SIGMOD'00)*, pages 70–81, Dallas, TX, May 2000.
- [AY01] C. C. Aggarwal and P. S. Yu. Outlier detection for high dimensional data. In *Proc. 2001 ACM-SIGMOD Int. Conf. Management of Data (SIGMOD'01)*, pages 37–46, Santa Barbara, CA, May 2001.
- [BBD<sup>+</sup>02] B. Babcock, S. Babu, M. Datar, R. Motwani, and J. Widom. Models and issues in data stream systems. In *Proc. 2002 ACM Symp. Principles of Database Systems (PODS'02)*, pages 1–16, Madison, WI, June 2002.
- [BBM04] S. Basu, M. Bilenko, and R. J. Mooney. A probabilistic framework for semi-supervised clustering. In *Proc. 2004 ACM SIGKDD Int. Conf. Knowledge Discovery in Databases (KDD'04)*, pages 59–68, Seattle, WA, Aug. 2004.
- [BEX02] F. Beil, M. Ester, and X. Xu. Frequent term-based text clustering. In *Proc. 2002 ACM SIGKDD Int. Conf. Knowledge Discovery in Databases (KDD'02)*, pages 436–442, Edmonton, Canada, July 2002.
- [Bez81] J. C. Bezdek. *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum Press, 1981.

- [BFR98] P. Bradley, U. Fayyad, and C. Reina. Scaling clustering algorithms to large databases. In *Proc. 1998 Int. Conf. Knowledge Discovery and Data Mining (KDD'98)*, pages 9–15, New York, NY, Aug. 1998.
- [BKNS00] M. M. Breunig, H.-P. Kriegel, R. Ng, and J. Sander. LOF: Identifying density-based local outliers. In *Proc. 2000 ACM-SIGMOD Int. Conf. Management of Data (SIGMOD'00)*, pages 93–104, Dallas, TX, May 2000.
- [BL94] V. Barnett and T. Lewis. *Outliers in Statistical Data*. John Wiley & Sons, 1994.
- [BP92] J. C. Bezdek and S. K. Pal. *Fuzzy Models for Pattern Recognition: Methods That Search for Structures in Data*. IEEE Press, 1992.
- [Ce91] G. A. Carpenter and S. Grossberg (eds.). *Pattern Recognition by Self-Organizing Neural Networks*. MIT Press, 1991.
- [CFZ99] C.-H. Cheng, A. W. Fu, and Y. Zhang. Entropy-based subspace clustering for mining numerical data. In *Proc. 1999 Int. Conf. Knowledge Discovery and Data Mining (KDD'99)*, pages 84–93, San Diego, CA, Aug. 1999.
- [CGC94] A. Chaturvedi, P. Green, and J. Carroll. K-means, k-medians and k-modes: Special cases of partitioning multiway data. In *The Classification Society of North America (CSNA) Meeting Presentation*, Houston, TX, 1994.
- [CGC01] A. Chaturvedi, P. Green, and J. Carroll. K-modes clustering. *J. Classification*, 18:35–55, 2001.
- [CS96] P. Cheeseman and J. Stutz. Bayesian classification (AutoClass): Theory and results. In U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining*, pages 153–180. AAAI/MIT Press, 1996.
- [DE84] W. H. E. Day and H. Edelsbrunner. Efficient algorithms for agglomerative hierarchical clustering methods. *J. Classification*, 1:7–24, 1984.
- [DLR77] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Statistical Society*, 39:1–38, 1977.
- [EKSX96] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases. In *Proc. 1996 Int. Conf. Knowledge Discovery and Data Mining (KDD'96)*, pages 226–231, Portland, OR, Aug. 1996.
- [EKX95] M. Ester, H.-P. Kriegel, and X. Xu. Knowledge discovery in large spatial databases: Focusing techniques for efficient class identification. In *Proc. 1995 Int. Symp. Large Spatial Databases (SSD'95)*, pages 67–82, Portland, ME, Aug. 1995.
- [Fis87] D. Fisher. Improving inference through conceptual clustering. In *Proc. 1987 Nat. Conf. Artificial Intelligence (AAAI'87)*, pages 461–465, Seattle, WA, July 1987.
- [GGR99] V. Ganti, J. E. Gehrke, and R. Ramakrishnan. CACTUS—clustering categorical data using summaries. In *Proc. 1999 Int. Conf. Knowledge Discovery and Data Mining (KDD'99)*, pages 73–83, San Diego, CA, 1999.
- [GKR98] D. Gibson, J. Kleinberg, and P. Raghavan. Inferring web communities from link topology. In *Proc. 9th ACM Conf. Hypertext and Hypermedia*, pages 225–234, Pittsburgh, PA, June 1998.
- [GLF89] J. Gennari, P. Langley, and D. Fisher. Models of incremental concept formation. *Artificial Intelligence*, 40:11–61, 1989.
- [GMMO00] S. Guha, N. Mishra, R. Motwani, and L. O'Callaghan. Clustering data streams. In *Proc. 2000 Symp. Foundations of Computer Science (FOCS'00)*, pages 359–366, Redondo Beach, CA, 2000.

- [GRS98] S. Guha, R. Rastogi, and K. Shim. Cure: An efficient clustering algorithm for large databases. In *Proc. 1998 ACM-SIGMOD Int. Conf. Management of Data (SIGMOD'98)*, pages 73–84, Seattle, WA, June 1998.
- [GRS99] S. Guha, R. Rastogi, and K. Shim. ROCK: A robust clustering algorithm for categorical attributes. In *Proc. 1999 Int. Conf. Data Engineering (ICDE'99)*, pages 512–521, Sydney, Australia, Mar. 1999.
- [Har75] J. A. Hartigan. *Clustering Algorithms*. John Wiley & Sons, 1975.
- [He99] G. E. Hinton and T. J. Sejnowski (eds.). *Unsupervised Learning: Foundation of Neural Network Computation*. MIT Press, 1999.
- [HK98] A. Hinneburg and D. A. Keim. An efficient approach to clustering in large multimedia databases with noise. In *Proc. 1998 Int. Conf. Knowledge Discovery and Data Mining (KDD'98)*, pages 58–65, New York, NY, Aug. 1998.
- [Hua98] Z. Huang. Extensions to the  $k$ -means algorithm for clustering large data sets with categorical values. *Data Mining and Knowledge Discovery*, 2:283–304, 1998.
- [JD88] A. K. Jain and R. C. Dubes. *Algorithms for Clustering Data*. Prentice Hall, 1988.
- [JKM99] H. V. Jagadish, N. Koudas, and S. Muthukrishnan:. Mining deviants in a time series database. In *Proc. 1999 Int. Conf. Very Large Data Bases (VLDB'99)*, pages 102–113, Edinburgh, UK, Sept. 1999.
- [JMF99] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: A survey. *ACM Comput. Surv.*, 31:264–323, 1999.
- [KHK99] G. Karypis, E.-H. Han, and V. Kumar. CHAMELEON: A hierarchical clustering algorithm using dynamic modeling. *COMPUTER*, 32:68–75, 1999.
- [KKL<sup>+</sup>00] T. Kohonen, S. Kaski, K. Lagus, J. Solojärvi, A. Paatero, and A. Saarela. Self-organization of massive document collection. *IEEE Trans. Neural Networks*, 11:574–585, 2000.
- [KN97] E. Knorr and R. Ng. A unified notion of outliers: Properties and computation. In *Proc. 1997 Int. Conf. Knowledge Discovery and Data Mining (KDD'97)*, pages 219–222, Newport Beach, CA, Aug. 1997.
- [KN98] E. Knorr and R. Ng. Algorithms for mining distance-based outliers in large datasets. In *Proc. 1998 Int. Conf. Very Large Data Bases (VLDB'98)*, pages 392–403, New York, NY, Aug. 1998.
- [Koh82] T. Kohonen. Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43:59–69, 1982.
- [Koh89] T. Kohonen. *Self-Organization and Associative Memory* (3rd ed.). Springer-Verlag, 1989.
- [KR90] L. Kaufman and P. J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley & Sons, 1990.
- [Lau95] S. L. Lauritzen. The EM algorithm for graphical association models with missing data. *Computational Statistics and Data Analysis*, 19:191–201, 1995.
- [Llo57] S. P. Lloyd. *Least Squares Quantization in PCM*. IEEE Trans. Information Theory, 28:128–137, 1982, (original version: Technical Report, Bell Labs), 1957.
- [LXY01] B. Liu, Y. Xia, and P. S. Yu. Clustering through decision tree construction. In *Proc. 2000 ACM CIKM Int. Conf. Information and Knowledge Management (CIKM'00)*, pages 20–29, McLean, VA, Nov. 2001.
- [Mac67] J. MacQueen. Some methods for classification and analysis of multivariate observations. *Proc. 5th Berkeley Symp. Math. Stat. Prob.*, 1:281–297, 1967.

- [MS83] R. S. Michalski and R. E. Stepp. Learning from observation: Conceptual clustering. In R. S. Michalski, J. G. Carbonell, , and T. M. Mitchell, editors, *Machine Learning: An Artificial Intelligence Approach, Vol. 1*. Morgan Kaufmann, 1983.
- [NH94] R. Ng and J. Han. Efficient and effective clustering method for spatial data mining. In *Proc. 1994 Int. Conf. Very Large Data Bases (VLDB'94)*, pages 144–155, Santiago, Chile, Sept. 1994.
- [OMM<sup>+</sup>02] L. O'Callaghan, A. Meyerson, R. Motwani, N. Mishra, and S. Guha. Streaming-data algorithms for high-quality clustering. In *Proc. 2002 Int. Conf. Data Engineering (ICDE'02)*, pages 685–696, San Francisco, CA, Apr. 2002.
- [PHL04] L. Parsons, E. Haque, and H. Liu. Subspace clustering for high dimensional data: A review. *SIGKDD Explorations*, 6:90–105, 2004.
- [RZ85] D. E. Rumelhart and D. Zipser. Feature discovery by competitive learning. *Cognitive Science*, 9:75–112, 1985.
- [SAM98] S. Sarawagi, R. Agrawal, and N. Megiddo. Discovery-driven exploration of OLAP data cubes. In *Proc. Int. Conf. of Extending Database Technology (EDBT'98)*, pages 168–182, Valencia, Spain, Mar. 1998.
- [SCZ98] G. Sheikholeslami, S. Chatterjee, and A. Zhang. WaveCluster: A multi-resolution clustering approach for very large spatial databases. In *Proc. 1998 Int. Conf. Very Large Data Bases (VLDB'98)*, pages 428–439, New York, NY, Aug. 1998.
- [SD90] J. W. Shavlik and T. G. Dietterich. *Readings in Machine Learning*. Morgan Kaufmann, 1990.
- [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.
- [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.
- [WYM97] W. Wang, J. Yang, and R. Muntz. STING: A statistical information grid approach to spatial data mining. In *Proc. 1997 Int. Conf. Very Large Data Bases (VLDB'97)*, pages 186–195, Athens, Greece, Aug. 1997.
- [ZRL96] T. Zhang, R. Ramakrishnan, and M. Livny. BIRCH: an efficient data clustering method for very large databases. In *Proc. 1996 ACM-SIGMOD Int. Conf. Management of Data (SIGMOD'96)*, pages 103–114, Montreal, Canada, June 1996.