# Bibliography

Achlioptas, D. and McSherry, F. (2001). Fast computation of low rank matrix approximations. In *STOC* '01: *Proceedings of the Thirty-Third Annual ACM Symposium on Theory of Computing*, pages 611–618, New York, NY, USA. ACM Press. (Cited on p. 269)

Adolfsson, A., Ackerman, M., and Brownstein, N. C. (2019). To cluster, or not to cluster: An analysis of clusterability methods. *Pattern Recognition*, 88:13–26. (Cited on pp. 12, 295)

Agarwal, P. and Mustafa, N. (2004). $k$-means projective clustering. In *Proceedings of the Twenty-Third ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (PODS)*, pages 155–165, Paris, France. ACM Press. (Cited on pp. 154, 259)

Aggarwal, C., Han, J., Wang, J., and Yu, P. (2003). A framework for clustering evolving data streams. In *Proceedings of* 29*th International Conference on Very Large Data Bases*, pages 81–92, Berlin, Germany. Morgan Kaufmann. (Cited on p. 272)

Aggarwal, C., Han, J., Wang, J., and Yu, P. (2004). A framework for projected clustering of high dimensional data streams. In *Proceedings of the Thirtieth International Conference on Very Large Data Bases*, pages 852–863, Toronto, Canada. Morgan Kaufmann. (Cited on p. 272)

Aggarwal, C., Wolf, J., Yu, P., Procopiuc, C., and J. S. Park (1999). Fast algorithms for projected clustering. In *Proceedings of the* 1999 *ACM SIGMOD International Conference on Management of Data*, pages 61–72. New York, NY, USA. ACM Press. (Cited on pp. 69, 222, 224)

Aggarwal, C. and Yu, P. (2000). Finding generalized projected clusters in high dimensional spaces. In Chen, W., Naughton, J. F., and Bernstein, P. A., editors, *Proceedings of the* 2000 *ACM SIGMOD International Conference on Management of Data, May* 16–18, 2000, *Dallas, Texas*, Volume 29, pages 70–81. New York, NY, USA. ACM Press. (Cited on pp. 222, 226)

Aggarwal, C. and Yu, P. (2002). Redefining clustering for high-dimensional applications. *IEEE Transactions on Knowledge and Data Engineering*, 14(2):210–225. (Cited on p. 259)

Aggarwal, C. C. and Reddy, C. K., editors (2013). *Data Clustering: Algorithms and Applications*. CRC Press, Boca Raton, FL, USA. (Cited on p. 14)

Aggarwal, C. C. and Zhai, C., editors (2012). *Mining Text Data*. Springer, New York, NY, USA. (Cited on p. 22)

Aghabozorgi, S., Shirkhorshidi, A. S., and Wah, T. Y. (2015). Time-series clustering – a decade review. *Information Systems*, 53:16–38. (Cited on p. 11)

Agrawal, R., Faloutsos, C., and Swami, A. (1993). Efficient similarity search in sequence databases. In *FODO* '93: *Proceedings of the* 4*th International Conference on Foundations of Data Organization and Algorithms*, volume 730 of *Lecture Notes In Computer Science*, pages 69–84. Berlin, Germany. Springer-Verlag. (Cited on p. 81)

Agrawal, R., Gehrke, J., Gunopulos, D., and Raghavan, P. (1998). Automatic subspace clustering of high dimensional data for data mining applications. In *SIGMOD Record ACM Special Interest Group on Management of Data*, pages 94–105, New York, NY, USA. ACM Press. (Cited on pp. 221, 222, 226)

Agrawal, R., Lin, K., Sawhney, H., and Shim, K. (1995). Fast similarity search in the presence of noise, scaling, and translation in time-series databases. In *VLDB* '95: *Proceedings of the* 21*st International Conference on Very Large Data Bases*, pages 490–501, Palo Alto, CA, USA. Morgan Kaufmann. (Cited on pp. 84, 85)

Ahmad, A. and Khan, S. S. (2019). Survey of state-of-the-art mixed data clustering algorithms. *IEEE Access*, 7:31883–31902. (Cited on p. 12)

Aho, A., Hopcroft, J., and Ullman, J. (1974). *The Design and Analysis of Computer Algorithms*. Addison-Wesley Series in Computer Science and Information Processing. Reading, MA; Don Mills, Ontario, Canada. Addison-Wesley Publishing Company. (Cited on p. 224)

Al-Sultan, K. (1995). A tabu search approach to the clustering problem. *Pattern Recognition*, 28(9):1443–1451. (Cited on pp. 171, 172)

Al-Sultan, K. and Fedjki, C. (1997). A tabu search-based algorithm for the fuzzy clustering problem. *Pattern Recognition*, 30(12):2023–2030. (Cited on p. 173)

Alon, N., Matias, Y., and Szegedy, M. (1996). The space complexity of approximating the frequency moments. In *STOC* '96: *Proceedings of the Twenty-Eighth Annual ACM Symposium on Theory of Computing*, pages 20–29, New York, NY, USA. ACM Press. (Cited on p. 269)

Alpert, C. and Yao, S. (1995). Spectral partitioning: The more eigenvectors, the better. In *DAC* '95: *Proceedings of the* 32*nd ACM/IEEE Conference on Design Automation*, pages 195–200, New York, NY, USA. ACM Press. (Cited on p. 189)

Alsabti, K., Ranka, S., and Singh, V. (1998). An efficient $k$-means clustering algorithm. In *Proceedings of IPPS/SPDP Workshop on High Performance Data Mining*, Orlando, Florida. (Cited on pp. 150, 153)

Amir, A., Kashi, R., Netanyahu, N., Keim, D., and Wawryniuk, M. (2003). Analyzing high-dimensional data by subspace validity. In *Third IEEE International Conference on Data Mining,* 2003. *ICDM* 2003, pages 473–476, Los Alamitos, CA, USA. IEEE Computer Society. (Cited on p. 259)

Anderberg, M. (1973). *Cluster Analysis for Applications*. New York, USA. Academic Press. (Cited on pp. 5, 12, 25, 27, 28, 29, 39, 49, 65, 73, 95, 96, 99, 128, 149, 150, 161)

Andrade, G., Ramos, G., Madeira, D., Sachetto, R., Ferreira, R., and Rocha, L. (2013). G-DBSCAN: A GPU accelerated algorithm for density-based clustering. *Procedia Computer Science*, 18:369–378. (Cited on p. 261)

Andrews, D. (1972). Plots of high-dimensional data. *Biometrics*, 28(1):125–136. (Cited on p. 62)

Andrews, H. and Patterson, C. (1976a). Singular value decomposition image coding. *IEEE Transactions on Communications*, 4:425–432. (Cited on p. 45)

Andrews, H. and Patterson, C. (1976b). Singular value decompositions and digital image processing. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 24:26–53. (Cited on p. 45)

Andrienko, G. and Andrienko, N. (2004). Parallel coordinates for exploring properties of subsets. In *CMV* '04: *Proceedings of the Second International Conference on Coordinated & Multiple Views in Exploratory Visualization (CMV*'04*)*, pages 93–104, Washington, DC, USA. IEEE Computer Society. (Cited on p. 58)

Augustson, J. and Minker, J. (1970). An analysis of some graph theoretical cluster techniques. *Journal of the Association for Computing Machinery*, 17(4):571–588. (Cited on p. 189)

Azoff, E. (1994). *Neural Network Time Series Forecasting of Financial Markets*. Chichester, UK. Wiley & Sons. (Cited on p. 24)

Babcock, B., Datar, M., and Motwani, R. (2002). Sampling from a moving window over streaming data. In *SODA* '02: *Proceedings of the Thirteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 633–634, Philadelphia, PA, USA. Society for Industrial and Applied Mathematics. (Cited on p. 269)

Babcock, B., Datar, M., Motwani, R., and O'Callaghan, L. (2003). Maintaining variance and $k$-medians over data stream windows. In *PODS* '03: *Proceedings of the Twenty-Second ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, pages 234–243, New York, NY, USA. ACM Press. (Cited on p. 272)

Babu, G. and Murty, M. (1993). A near-optimal initial seed value selection in $k$-means algorithm using a genetic algorithm. *Pattern Recognition Letters*, 14(10):763–769. (Cited on pp. 150, 169)

Bagnall, A. and Janacek, G. (2004). Clustering time series from ARMA models with clipped data. In *KDD* '04: *Proceedings of the* 2004 *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 49–58, New York, NY, USA. ACM Press. (Cited on pp. 267, 268, 269)

Bagnall, A., Janakec, G., and Zhang, M. (2003). Clustering time series from mixture polynomial models with discretised data. Technical report CMP-C03-17, School of Computing Sciences, University of East Anglia, Norwich, England. (Cited on p. 268)

Bahmani, B., Moseley, B., Vattani, A., Kumar, R., and Vassilvitskii, S. (2012). Scalable k-means++. *Proceedings of the VLDB Endowment*, 5(7):622–633. (Cited on p. 261)

Baker, F. and Hubert, L. (1976). A graph-theoretic approach to goodness-of-fit in complete-link hierarchical clustering. *Journal of the American Statistical Association*, 71(356):870–878. (Cited on p. 189)

Bandyopadhyay, S. and Maulik, U. (2002). An evolutionary technique based on $k$-means algorithm for optimal clustering in $r^n$. *Information Sciences*, 146(1-4):221–237. (Cited on p. 168)

Bandyopadhyay, S., Murthy, C., and Pal, S. (1995). Pattern classification with genetic algorithms. *Pattern Recognition Letters*, 16(8):801–808. (Cited on p. 168)

Banfield, C. (1976). Statistical algorithms: Algorithm AS 102: Ultrametric distances for a single linkage dendrogram. *Applied Statistics*, 25(3):313–315. (Cited on p. 105)

Banfield, J. and Raftery, A. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics*, 49(3):803–821. (Cited on pp. 207, 208, 209, 220)

Bar-Joseph, Z., Gerber, G., Gifford, D., Jaakkola, T., and Simon, I. (2002). A new approach to analyzing gene expression time series data. In *RECOMB* '02: *Proceedings of the Sixth Annual International Conference on Computational Biology*, pages 39–48, New York, NY, USA. ACM Press. (Cited on p. 268)

Baraldi, A. and Blonda, P. (1999a). A survey of fuzzy clustering algorithms for pattern recognition. I. *IEEE Transactions on Systems, Man and Cybernetics, Part B*, 29(6):778–785. (Cited on p. 11)

Baraldi, A. and Blonda, P. (1999b). A survey of fuzzy clustering algorithms for pattern recognition. II. *IEEE Transactions on Systems, Man and Cybernetics, Part B*, 29(6):786–801. (Cited on p. 11)

Barbará, D. (2002). Requirements for clustering data streams. *ACM SIGKDD Explorations Newsletter*, 3(2):23–27. (Cited on pp. 269, 276)

Barbará, D., Li, Y., and Couto, J. (2002). COOLCAT: An entropy-based algorithm for categorical clustering. In *Proceedings of the Eleventh International Conference on Information and Knowledge Management*, pages 582–589, McLean, Virginia, USA. ACM Press. (Cited on pp. 218, 295)

Barber, C., Dobkin, D., and Huhdanpaa, H. (1996). The quickhull algorithm for convex hulls. *ACM Transactions on Mathematical Software*, 22(4):469–483. (Cited on pp. 138, 283)

Batagelj, V. (1981). Note on ultrametric hierarchical clustering algorithms. *Psychometrika*, 46(3):351–352. (Cited on p. 110)

Baulieu, F. (1989). A classification of presence/absence based dissimilarity coefficients. *Journal of Classification*, 6:233–246. (Cited on pp. 73, 100)

Baulieu, F. (1997). Two variant axiom systems for presence/absence based dissimilarity coefficients. *Journal of Classification*, 14:159–170. (Cited on p. 73)

Bay, S. and Pazzani, M. (1999). Detecting change in categorical data: Mining contrast sets. In Chaudhuri, S. and Madigan, D., editors, *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 302–306, New York, NY, USA. ACM Press. (Cited on p. 219)

Belacel, N., Hansen, P., and Mladenović, N. (2002). Fuzzy $J$-means: A new heuristic for fuzzy clustering. *Pattern Recognition*, 35(10):2193–2200. (Cited on p. 175)

Bell, D., McErlean, F., and Stewart, P. (1990). Application of simulated annealing to clustering tuples in databases. *Journal of American Society for Information Science and Technology*, 42(2):98–110. (Cited on p. 184)

Bellman, R., Kalaba, R., and Zadeh, L. (1966). Abstraction and pattern classification. *Journal of Mathematical Analysis and Applications*, 2:581–586. (Cited on p. 139)

Bensmail, H., Celeux, G., Raftery, A., and Robert, C. (1997). Inference in model-based cluster analysis. *Statistics and Computing*, 7(1):1–10. (Cited on p. 220)

Bentley, J. (1975). Multidimensional binary search trees used for associative searching. *Communications of the ACM*, 18(9):509–517. (Cited on pp. 351, 352)

Bentley, J. (1980). Multidimensional divide-and-conquer. *Communications of the ACM*, 23(4):214–229. (Cited on p. 351)

Bentley, J. and Friedman, J. (1979). Data structures for range searching. *ACM Computing Surveys (CSUR)*, 11(4):397–409. (Cited on p. 351)

Berndt, D. and Clifford, J. (1994). Using dynamic time warping to find patterns in time series. In *AAAI-94 Workshop on Knowledge Discovery in Databases*, pages 229–248, Menlo Park, CA, USA. AAAI Press. (Cited on p. 83)

Berry, M. and Linoff, G. (2000). *Mastering Data Mining*. New York, NY, USA. John Wiley & Sons. (Cited on p. 4)

Beyer, K., Goldstein, J., Ramakrishnan, R., and Shaft, U. (1999). When is "nearest neighbor" meaningful? In Beeri, C. and Buneman, P., editors, *Proceedings of 7th International Conference on Database Theory*, volume 1540 of *Lecture Notes in Computer Science*, pages 217–235. Berlin, Germany. Springer. (Cited on p. 221)

Beygelzimer, A., Perng, C., and Ma, S. (2001). Fast ordering of large categorical datasets for better visualization. In *KDD* '01: *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 239–244, New York, NY, USA. ACM Press. (Cited on p. 61)

Bezdek, J. (1974a). Cluster validity with fuzzy sets. *Journal of Cybernetics*, 3(3):58–72. (Cited on p. 291)

Bezdek, J. (1974b). *Fuzzy mathematics in pattern classification*. PhD thesis, Cornell University, Ithaca, NY. (Cited on p. 142)

Bezdek, J. (1980). A convergence theorem for the fuzzy ISODATA clustering algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2(1):1–8. (Cited on pp. 167, 254)

Bezdek, J. (1981). *Pattern Recognition with Fuzzy Objective Function Algorithms*. Norwell, MA, USA. Kluwer Academic Publishers. (Cited on pp. 141, 142, 145, 291)

Bezdek, J., Hathaway, R., Sabin, M., and Tucker, W. (1992). Convergence theory for fuzzy $c$-means: Counterexamples and repairs. In Bezdek, J. and Pal, S., editors, *Fuzzy Models for Pattern Recognition: Methods that Search for Approximate Structures in Data*, pages 138–142, Piscataway, NJ, USA. IEEE Computer Society. (Cited on p. 145)

Bickel, P. J. and Doksum, K. A. (1981). An analysis of transformations revisited. *Journal of the American Statistical Association*, 76(374):296–311. (Cited on p. 44)

Bijnen, E. (1973). *Cluster Analysis: Survey and Evaluation of Techniques*. Tilburg, The Netherlands. Tilburg University Press. (Cited on p. 12)

Binder, D. (1978). Bayesian cluster analysis. *Biometrika*, 65(1):31–38. (Cited on p. 207)

Blake, C. and Merz, C. (1998). UCI repository of machine learning databases. `http://www.ics.uci.edu/~mlearn/MLRepository.html`. (Cited on p. 87)

Bobisud, H. and Bobisud, L. (1972). A metric for classification. *Taxon*, 21:607–613. (Cited on p. 104)

Bobrowski, L. and Bezdek, J. (1991). $c$-means clustering with the $l_1$ and $l_\infty$ norms. *IEEE Transactions on Systems, Man and Cybernetics*, 21(3):545–554. (Cited on pp. 145, 148)

Bock, H. (1985). On some significance tests in cluster analysis. *Journal of Classification*, 2:77–108. (Cited on p. 290)

Bock, H. (1989). Probabilistic aspects in cluster analysis. In Opitz, O., editor, *Conceptual and Numerical Analysis of Data*, pages 12–44, Augsburg, FRG. Springer-Verlag. (Cited on p. 6)

Bock, H. (1996). Probabilistic models in cluster analysis. *Computational Statistics and Data Analysis*, 23(1):5–28. (Cited on pp. 207, 290, 295)

Bolla, M. (2013). *Spectral Clustering and Biclustering: Learning Large Graphs and Contingency Tables*. Hoboken, NJ, USA. Wiley & Sons. (Cited on p. 14)

Bollobás, B., Das, G., Gunopulos, D., and Mannila, H. (1997). Time-series similarity problems and well-separated geometric sets. In *SCG* '97: *Proceedings of the Thirteenth Annual Symposium on Computational Geometry*, pages 454–456. New York, NY, USA. ACM Press. (Cited on pp. 24, 80)

Borg, I. and Groenen, P. (1997). *Modern Multidimensional Scaling*. Springer Series in Statistics. Berlin, Germany. Springer. (Cited on p. 52)

Borodin, A., Ostrovsky, R., and Rabani, Y. (1999). Subquadratic approximation algorithms for clustering problems in high dimensional spaces. In *STOC* '99: *Proceedings of the Thirty-First Annual ACM Symposium on Theory of Computing*, pages 435–444, New York, NY, USA. ACM Press. (Cited on p. 270)

Bottou, L. and Bengio, Y. (1995). Convergence properties of the $k$-means algorithms. In *Advances in Neural Information Processing Systems* 7, pages 585–592. Cambridge, MA, USA. MIT Press. (Cited on p. 150)

Boutsidis, C., Zouzias, A., and Drineas, P. (2010). Random projections for k-means clustering. In *Proceedings of the* 23*rd International Conference on Neural Information Processing Systems*, NIPS'10, pages 298–306, Red Hook, NY, USA. Curran Associates. (Cited on p. 261)

Bouveyron, C. and Brunet-Saumard, C. (2014). Model-based clustering of high-dimensional data: A review. *Computational Statistics & Data Analysis*, 71:52–78. (Cited on p. 11)

Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 26(2):211–252. (Cited on p. 44)

Boyles, R. (1983). On the convergence of the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 45(1):47–50. (Cited on p. 215)

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, 52:345–370. (Cited on p. 155)

Bozkaya, T., Yazdani, N., and Özsoyoğlu, M. (1997). Matching and indexing sequences of different lengths. In *CIKM* '97: *Proceedings of the Sixth International Conference on Information and Knowledge Management*, pages 128–135, New York, NY, USA. ACM Press. (Cited on p. 84)

Bradley, P. and Fayyad, U. (1998). Refining initial points for $k$-means clustering. In Shavlik, J., editor, *Proceedings of the Fifteenth International Conference on Machine Learning*, pages 91–99, San Francisco, CA, USA. Morgan Kaufmann. (Cited on pp. 150, 241)

Bradley, P., Fayyad, U., and Reina, C. (1998). Scaling clustering algorithms to large databases. In *Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining*, pages 9–15. Menlo Park, CA, USA. AAAI Press. (Cited on p. 272)

Bradley, P. and Mangasarian, O. (2000). $k$-plane clustering. *Journal of Global Optimization*, 16(1):23–32. (Cited on p. 166)

Brazma, A. and Vilo, J. (2000). Gene expression data analysis. *Biochemical Societies*, 480:17–24. (Cited on p. 330)

Broder, A., Garcia-Pueyo, L., Josifovski, V., Vassilvitskii, S., and Venkatesan, S. (2014). Scalable $k$-means by ranked retrieval. In *Proceedings of the 7th ACM International Conference on Web Search and Data Mining*, WSDM '14, pages 233–242. New York, NY, USA. ACM Press. (Cited on p. 262)

Broder, A., Glassman, S., Manasse, M., and Zweig, G. (1997). Syntactic clustering of the web. In *Selected Papers from the Sixth International Conference on World Wide Web*, pages 1157–1166, Amsterdam, The Netherlands. Elsevier Science Publishers. (Cited on p. 273)

Broder, A. Z., Carmel, D., Herscovici, M., Soffer, A., and Zien, J. (2003). Efficient query evaluation using a two-level retrieval process. In *Proceedings of the Twelfth International Conference on Information and Knowledge Management*, CIKM '03, pages 426–434, New York, NY, USA. ACM Press. (Cited on p. 263)

Buhmann, J. (2003). Data clustering and learning. In Arbib, M., editor, *The Handbook of Brain Theory and Neural Networks*, pages 308–312, Cambridge, MA, USA. The MIT Press. (Cited on p. 8)

Buló, S. R. and Pelillo, M. (2017). Dominant-set clustering: A review. *European Journal of Operational Research*, 262(1):1–13. (Cited on p. 12)

Calinski, T. and Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics*, 3:1–27. (Cited on p. 286)

Cao, Y. and Wu, J. (2002). Projective ART for clustering data sets in high dimensional spaces. *Neural Networks*, 15(1):105–120. (Cited on pp. 222, 238)

Caraça-Valente, J. and López-Chavarrías, I. (2000). Discovering similar patterns in time series. In *KDD* '00: *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 497–505, New York, NY, USA. ACM Press. (Cited on pp. 80, 267)

Carmichael, J., George, J., and Julius, R. (1968). Finding natural clusters. *Systematic Zoology*, 17(2):144–150. (Cited on p. 6)

Carpenter, G. and Grossberg, S. (1987a). ART2: Self-organization of stable category recognition codes for analog input patterns. *Applied Optics*, 26:4919–4930. (Cited on p. 238)

Carpenter, G. and Grossberg, S. (1987b). A massively parallel architecture for a self-organizing neural pattern recognition machine. *Computer Vision, Graphics and Image Processing*, 37:54–115. (Cited on p. 238)

Carpenter, G. and Grossberg, S. (1990). ART3: Hierarchical search using chemical transmitters in self-organizing pattern recognition architectures. *Neural Networks*, 3:129–152. (Cited on p. 238)

Carroll, J. and Arabie, P. (1980). Multidimensional scaling. *Annual Review of Psychology*, 31:607–649. (Cited on p. 52)

Cattell, R. (1949). $r_p$ and other coefficients of pattern similarity. *Psychometrika*, 14(4):279–298. (Cited on pp. 92, 100)

Celeux, G. and Govaert, G. (1995). Gaussian parsimonious clustering models. *Pattern Recognition*, 28(5):781–793. (Cited on pp. 209, 210)

Cezkanowski, J. (1909). Zur differentialdiagnose der neandertalgruppe. *Korrespondenz-Blatt deutsch. Ges. Anthropol. Ethnol. Urgesch*, 40:44–47. (Cited on p. 72)

Chakrabarti, K., Keogh, E., Mehrotra, S., and Pazzani, M. (2002). Locally adaptive dimensionality reduction for indexing large time series databases. *ACM Transactions on Database Systems*, 27(2):188–228. (Cited on p. 80)

Chan, K. and Fu, W. (1999). Efficient time series matching by wavelets. In *ICDE* '99: *Proceedings of the* 15*th International Conference on Data Engineering*, pages 126–133, Washington, DC, USA. IEEE Computer Society. (Cited on p. 86)

Chan, P., Schlag, M., and Zien, J. (1994). Spectral $k$-way ratio-cut partitioning and clustering. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 13(9):1088–1096. (Cited on p. 189)

Chang, C. and Ding, Z. (2004). Categorical data visualization and clustering using subjective factors. In Kambayashi, Y., Mohania, M., and Wöß, W., editors, 6*th International Conference on Data Warehousing and Knowledge Discovery*, volume 3181, pages 229–238, Zaragoza, Spain. Heidelberg. Springer-Verlag. (Cited on p. 61)

Chang, C. and Ding, Z. (2005). Categorical data visualization and clustering using subjective factors. *Data & Knowledge Engineering*, 53(3):243–262. (Cited on pp. 60, 61)

Chang, J. and Jin, D. (2002). A new cell-based clustering method for large, high-dimensional data in data mining applications. In *SAC* '02: *Proceedings of the* 2002 *ACM Symposium on Applied Computing*, pages 503–507. New York, NY, USA. ACM Press. (Cited on pp. 198, 259)

Charikar, M., Chaudhuri, S., Motwani, R., and Narasayya, V. (2000). Towards estimation error guarantees for distinct values. In *PODS* '00: *Proceedings of the Nineteenth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, pages 268–279, New York, NY, USA. ACM Press. (Cited on p. 269)

Charikar, M. and Guha, S. (1999). Improved combinatorial algorithms for the facility location and $k$-median problems. In *FOCS* '99: *Proceedings of the* 40*th Annual Symposium on Foundations of Computer Science*, pages 378–388, Washington, DC, USA. IEEE Computer Society. (Cited on p. 271)

Charikar, M., O'Callaghan, L., and Panigrahy, R. (2003). Better streaming algorithms for clustering problems. In *STOC* '03: *Proceedings of the Thirty-Fifth Annual ACM Symposium on Theory of Computing*, pages 30–39, New York, NY, USA. ACM Press. (Cited on p. 272)

Chaturvedi, A., Green, P., and Carroll, J. (2001). $k$-modes clustering. *Journal of Classification*, 18(1):35–55. (Cited on pp. 162, 179)

Chen, H. and Meer, P. (2005). Robust fusion of uncertain information. *IEEE Transactions on Systems, Man and Cybernetics, Part B*, 35(3):578–586. (Cited on p. 159)

Chen, J., Ching, R., and Lin, Y. (2004). An extended study of the $k$-means algorithm for data clustering and its applications. *Journal of the Operational Research Society*, 55(9):976–987. (Cited on p. 153)

Chen, N., Chen, A., and Zhou, L. (2002). An incremental grid density-based clustering algorithm. *Journal of Software*, 13(1):1–7. (Cited on p. 198)

Cheng, C., Fu, A., and Zhang, Y. (1999). Entropy-based subspace clustering for mining numerical data. In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 84–93. New York, NY, USA. ACM Press. (Cited on pp. 222, 230)

Cheng, C., Lee, W., and Wong, K. (2002). A genetic algorithm-based clustering approach for database partitioning. *IEEE Transactions on Systems, Man and Cybernetics, Part C*, 32(3):215–230. (Cited on p. 168)

Cheng, Y. (1995). Mean shift, mode seeking, and clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(8):790–799. (Cited on pp. 157, 158, 159, 160, 250, 255)

Chernoff, H. (1973). The use of faces to represent points in $k$-dimensional space graphically. *Journal of the American Statistical Association*, 68(342):361–368. (Cited on p. 62)

Chi, J. T., Chi, E. C., and Baraniuk, R. G. (2016). $k$-POD: A method for $k$-means clustering of missing data. *The American Statistician*, 70(1):91–99. (Cited on p. 11)

Chiou, Y. and Lan, L. (2001). Genetic clustering algorithms. *European Journal of Operational Research*, 135(2):413–427. (Cited on p. 169)

Chiu, B., Keogh, E., and Lonardi, S. (2003). Probabilistic discovery of time series motifs. In *KDD* '03: *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 493–498, New York, NY, USA. ACM Press. (Cited on pp. 80, 267)

Cho, R., Campbell, M., Winzeler, E., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T., Gabrielian, A., Landsman, D., Lockhart, D. J., and Davis, R. (1998). A genome-wide transcriptional analysis of the mitotic cell cycle. *Molecular Cell*, 2(1):65–73. (Cited on p. 333)

Chrétien, S. and Hero III, A. (1998). Acceleration of the EM algorithm via proximal point iterations. In *Proceedings of the IEEE International Symposium on Information Theory*, 444, Piscataway, NJ, USA. IEEE Press. (Cited on p. 215)

Christopher, M. (1969). Cluster analysis and market segmentation. *British Journal of Marketing*, 3(2):99–102. (Cited on p. 4)

Chu, K. and Wong, M. (1999). Fast time-series searching with scaling and shifting. In *PODS* '99: *Proceedings of the Eighteenth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, pages 237–248, New York, NY, USA. ACM Press. (Cited on p. 84)

Clatworthy, J., Buick, D., Hankins, M., Weinman, J., and Horne, R. (2005). The use and reporting of cluster analysis in health psychology: A review. *British Journal of Health Psychology*, 10(3):329–358. (Cited on p. 4)

Cochran, W. and Hopkins, C. (1961). Some classification problems with multivariate qualitative data. *Biometrics*, 17(1):10–23. (Cited on p. 27)

Comaniciu, D. and Meer, P. (1999). Mean shift analysis and applications. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, Volume 2, pages 1197–1203, Los Alamitos, CA, USA. IEEE Computer Society. (Cited on p. 157)

Comaniciu, D. and Meer, P. (2002). Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5):603–619. (Cited on pp. 4, 157)

Conover, W. and Iman, R. (1981). Rank transformations as a bridge between parametric and nonparametric statistics. *The American Statistician*, 35(3):124–129. (Cited on p. 44)

Constantine, A. and Gower, J. (1978). Graphical representation of asymmetric matrices. *Applied Statistics*, 27:297–304. (Cited on p. 65)

Constantinescu, P. (1966). The classification of a set of elements with respect to a set of properties. *The Computer Journal*, 8(4):352–357. (Cited on p. 104)

Cormack, R. (1971). A review of classification. *Journal of the Royal Statistical Society. Series A (General)*, 134(3):321–367. (Cited on pp. 11, 17)

Cormen, T., Stein, C., Rivest, R., and Leiserson, C. (2001). *Introduction to Algorithms*. Cambridge, MA, USA. MIT Press. (Cited on p. 84)

Costa, I., de Carvalho, F., and de Souto, M. (2004). Comparative analysis of clustering methods for gene expression time course data. *Genetics and Molecular Biology*, 27(4):623–631. (Cited on p. 331)

Cotofrei, P. and Stoffel, K. (2002). Classification rules + time = temporal rules. In *ICCS* '02: *Proceedings of the International Conference on Computational Science-Part* I, Volume 2329 of *Lecture Notes in Computer Science*, pages 572–581, London, UK. Springer-Verlag. (Cited on p. 267)

Cowgill, M., Harvey, R., and Watson, L. (1999). A genetic algorithm approach to cluster analysis. *Computers & Mathematics with Applications*, 37(7):99–108. (Cited on pp. 168, 182)

Cox, T. and Cox, M. (1994). *Multidimensional Scaling*, Volume 59 of *Monographs on Statistics and Applied Probability*. London, UK. Chapman & Hall. (Cited on p. 52)

Cuesta-Albertos, J., Gordaliza, A., and Matrán, C. (1997). Trimmed $k$-means: An attempt to robustify quantizers. *The Annals of Statistics*, 25(2):553–576. (Cited on pp. 154, 155)

Cunningham, K. and Ogilvie, J. (1972). Evaluation of hierarchical grouping techniques: A preliminary study. *The Computer Journal*, 15(3):209–213. (Cited on p. 295)

Curtin, R. R., Edel, M., Lozhnikov, M., Mentekidis, Y., Ghaisas, S., and Zhang, S. (2018). mlpack 3: A fast, flexible machine learning library. *Journal of Open Source Software*, 3(26), 726. (Cited on p. 309)

Das, G., Gunopulos, D., and Mannila, H. (1997). Finding similar time series. In *Proceedings of the First European Symposium on Principles of Data Mining and Knowledge Discovery*, pages 88–100, Berlin, Germany. Springer-Verlag. (Cited on pp. 24, 80, 84, 85)

Das, G., Lin, K., Mannila, H., Renganathan, G., and Smyth, P. (1998). Rule discovery from time series. In *KDD* '98: *Proceedings of the* 4*th International Conference on Knowledge Discovery and Data Mining*, pages 16–22, Menlo Park, CA, USA. AAAI Press. (Cited on pp. 80, 267)

Das, S., Abraham, A., and Konar, A. (2009). *Metaheuristic Clustering*. New York, NY, USA. Springer. (Cited on p. 14)

Dasgupta, A. and Raftery, A. (1998). Detecting features in spatial point processes with clutter via model-based clustering. *Journal of the American Statistical Association*, 93(441):294–302. (Cited on pp. 208, 217)

Dasgupta, S. (2000). Experiments with random projection. In *Proceedings of the* 16*th Conference on Uncertainty in Artificial Intelligence*, UAI '00, pages 143–151, San Francisco, CA, USA. Morgan Kaufman. (Cited on p. 261)

Dash, M., Liu, H., and Xu, X. (2001). '1+1>2': Merging distance and density based clustering. In *Proceedings of the Seventh International Conference on Database Systems for Advanced Applications,* 2001, Hong Kong, China, pages 32–39, Los Alamitos, CA, USA. IEEE Computer Society. (Cited on pp. 200, 201)

Datar, M., Gionis, A., Indyk, P., and Motwani, R. (2002). Maintaining stream statistics over sliding windows: (extended abstract). In *SODA* '02: *Proceedings of the Thirteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 635–644, Philadelphia, PA, USA. Society for Industrial and Applied Mathematics. (Cited on p. 269)

Dave, R. (1996). Validating fuzzy partitions obtained through $c$-shells clustering. *Pattern Recognition Letters*, 17(6):613–623. (Cited on p. 291)

Davies, D. and Bouldin, D. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1(2):224–227. (Cited on pp. 35, 282, 283)

Day, N. (1969). Estimating the components of a mixture of normal distributions. *Biometrika*, 56(3):463–474. (Cited on p. 207)

Day, W. and Edelsbrunner, H. (1984). Efficient algorithms for agglomerative hierarchical clustering methods. *Journal of Classification*, 1(7):7–24. (Cited on p. 138)

de Amorim, R. C. (2016). A survey on feature weighting based $k$-means algorithms. *Journal of Classification*, 33(2):210–242. (Cited on p. 12)

De Backer, S., Naud, A., and Scheunders, P. (1998). Non-linear dimensionality reduction techniques for unsupervised feature extraction. *Pattern Recognition Letters*, 19(8):711–720. (Cited on pp. 52, 53)

de Oliveira, J. V. and Pedrycz, W. (2007). *Advances in Fuzzy Clustering and its Applications*. Hoboken, NJ, USA. Wiley & Sons. (Cited on p. 13)

De Smet, F., Mathys, J., Marchal, K., Thijs, G., De Moor, B., and Moreau, Y. (2002). Adaptive quality-based clustering of gene expression profiles. *Bioinformatics*, 18(5):735–746. (Cited on pp. 332, 333)

Debregeas, A. and Hebrail, G. (1998). Interactive interpretation of Kohonen maps applied to curves. In *KDD* '98: *Proceedings of the* 4*th International Conference of Knowledge Discovery and Data Mining*, pages 179–183, Menlo Park, CA, USA. AIAA Press. (Cited on p. 87)

Defays, D. (1977). An efficient algorithm for a complete link method. *The Computer Journal*, 20(4):364–366. (Cited on pp. 132, 275)

Delattre, M. and Hansen, P. (1980). Bicriterion cluster analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2(4):277–291. (Cited on p. 128)

Delgado, M., Skármeta, A., and Barberá, H. (1997). A tabu search approach to the fuzzy clustering problem. In *Proceedings of the Sixth IEEE International Conference on Fuzzy Systems,* 1997, Barcelona Spain, Volume 1, pages 125–130, Piscataway, NJ, USA. IEEE. (Cited on p. 170)

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38. (Cited on pp. 214, 215)

Demšar, J., Curk, T., Erjavec, A., Gorup, Č., Hočevar, T., Milutinovič, M., Možina, M., Polajnar, M., Toplak, M., Starič, A., Štajdohar, M., Umek, L., Žagar, L., Žbontar, J., Žitnik, M., and Zupan, B. (2013). Orange: Data mining toolbox in Python. *Journal of Machine Learning Research*, 14:2349–2353. (Cited on p. 309)

Deng, K. and Moore, A. (1995). Multiresolution instance-based learning. In *Proceedings of the Twelfth International Joint Conference on Artificial Intellingence*, pages 1233–1239, San Francisco. Morgan Kaufmann. (Cited on p. 351)

Deng, Z., Choi, K.-S., Jiang, Y., Wang, J., and Wang, S. (2016). A survey on soft subspace clustering. *Information Sciences*, 348:84–106. (Cited on p. 12)

D'haeseleer, P. (2005). How does gene expression clustering work? *Nature Biotechnology*, 23:1499–1501. (Cited on pp. 331, 332)

D'haeseleer, P., Wen, X., Fuhrman, S., and Somogyi, R. (1998). Mining the gene expression matrix: Inferring gene relationships from large scale gene expression data. In *Proceedings of the Second International Workshop on Information Processing in Cell and Tissues*, pages 203–212, New York, NY, USA. Plenum Press. (Cited on p. 333)

Dhillon, I., Modha, D., and Spangler, W. (1998). Visualizing class structure of multidimensional data. In Weisberg, S., editor, *Proceedings of the* 30*th Symposium on the Interface: Computing Science and Statistics*, Minneapolis, MN, Volume 30, pages 488–493, Fairfax Station, VA, USA. Interface Foundation of North America. (Cited on pp. 56, 57)

Dinesh, M., Gowda, K., and Ravi, T. (1997). Classification of symbolic data using fuzzy set theory. In *Proceedings of the* 1997 *First International Conference on Knowledge-Based Intelligent Electronic Systems,* 1997. *KES* '97, Volume 2, pages 383–386, Piscataway, NJ, USA. IEEE. (Cited on p. 24)

Ding, C. and He, X. (2004). $k$-means clustering via principal component analysis. In *Twenty-First International Conference on Machine Learning*. New York, NY, USA. ACM Press. (Cited on pp. 44, 45)

Domeniconi, C., Papadopoulos, D., Gunopulos, D., and Ma, S. (2004). Subspace clustering of high dimensional data. In *Proceedings of the SIAM International Conference on Data Mining*, Lake Buena Vista, Florida, pp. 517–521, Philadelphia, PA, USA. Society for Industrial and Applied Mathematics. (Cited on pp. 246, 259)

Drineas, P., Frieze, A., Kannan, R., Vempala, S., and Vinay, V. (1999). Clustering in large graphs and matrices. In *Proceedings of the Tenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 291–299, Philadelphia, PA, USA. Society for Industrial and Applied Mathematics. (Cited on pp. 46, 269)

Drineas, P., Frieze, A., Kannan, R., Vempala, S., and Vinay, V. (2004). Clustering large graphs via the singular value decomposition. *Machine Learning*, 56(1-3):9–33. (Cited on pp. 46, 261)

Dubes, R. (1987). How many clusters are best? - An experiment. *Pattern Recognition*, 20(6): 645–663. (Cited on p. 4)

DuBien, J. and Warde, W. (1979). A mathematical comparison of the members of an infinite family of agglomerative clustering algorithms. *The Canadian Journal of Statistics*, 7:29–38. (Cited on p. 91)

Dunn, J. (1974). Well separated clusters and optimal fuzzy partitions. *Journal of Cybernetic*, 4:95–104. (Cited on pp. 35, 283)

Dunn, J. (1977). Indices of partition fuzziness and the detection of clusters in large datasets. In Gupta, M., Saridis, G., and Gaines, B., editors, *Fuzzy Automata and Decision Processes*, pages 271–284, New York. North–Holland. (Cited on p. 283)

Dunn, J. C. (1973). A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. *Journal of Cybernetics*, 3(3):32–57. (Cited on pp. 167, 180, 283)

Duran, B. and Odell, P. (1974). *Cluster Analysis – A Survey*, volume 100 of *Lecture Notes in Economics and Mathematical Systems*. Berlin, Heidelberg, New York. Springer-Verlag. (Cited on pp. 12, 90, 91)

Edwards, A. and Cavalli-Sforza, L. (1965). A method for cluster analysis. *Biometrics*, 21(2):362–375. (Cited on pp. 128, 129, 137, 207)

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26. (Cited on p. 280)

Efron, B. (1982). *The Jackknife, the Bootstrap, and Other Resampling Plans*, volume 38 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Philadelphia, PA, USA. Society for Industrial and Applied Mathematics. (Cited on p. 333)

Efron, B. and Gong, G. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation. *The American Statistician*, 37(1):36–48. (Cited on p. 280)

Egan, M., Krishnamoorthy, M., and Rajan, K. (1998). FCLUST: A visualization tool for fuzzy clustering. In *SIGCSE* '98: *Proceedings of the Twenty-Ninth SIGCSE Technical Symposium on Computer Science Education*, pages 227–231, New York, NY, USA. ACM Press. (Cited on p. 63)

Eisen, M. and Brown, P. (1999). DNA arrays for analysis of gene expression. *Methods in Enzymology*, 303:179–205. (Cited on p. 329)

Eisen, M., Spellman, P., Brown, P., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America*, 95(25):14863–14868. (Cited on pp. 3, 330, 333)

Eklund, T., Back, B., Vanharanta, H., and Visa, A. (2003). Using the self-organizing map as a visualization tool in financial benchmarking. *Information Visualization*, 2(3):171–181. (Cited on p. 56)

El-Sonbaty, Y., Ismail, M., and Farouk, M. (2004). An efficient density based clustering algorithm for large databases. In 16*th IEEE International Conference on Tools with Artificial Intelligence,* 2004. *ICTAI* 2004, pages 673–677, Los Alamitos, CA, USA. IEEE Computer Society. (Cited on p. 199)

Engelman, L. and Hartigan, J. (1969). Percentage points of a test for clusters. *Journal of the American Statistical Association*, 64(328):1647–1648. (Cited on p. 34)

Estabrook, G. (1966). A mathematical model in graph theory for biological classification. *Journal of Theoretical Biology*, 12:297–310. (Cited on p. 189)

Estabrook, G. and Rogers, D. (1966). A general method of taxonomic description for a computed similarity measure. *BioScience*, 16:789–793. (Cited on pp. 75, 76)

Ester, M., Kriegel, H., Sander, J., Wimmer, M., and Xu, X. (1998). Incremental clustering for mining in a data warehousing environment. In *VLDB* '98: *Proceedings of the* 24*th International Conference on Very Large Data Bases*, pages 323–333. San Francisco, CA, USA. Morgan Kaufmann. (Cited on p. 200)

Ester, M., Kriegel, H., Sander, J., and Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In Simoudis, E., Han, J., and Fayyad, U., editors, *Second International Conference on Knowledge Discovery and Data Mining*, pages 226–231, Portland, Oregon. AAAI Press. (Cited on pp. 199, 200)

Ester, M., Kriegel, H., Sander, J., and Xu, X. (1997). Density-connected sets and their application for trend detection in spatial databases. In *Third International Conference on Knowledge Discovery and Data Mining (KDD*'97*)*, pages 10–15. Menlo Park, CA, USA. AAAI Press. (Cited on p. 200)

Estivill-Castro, V. (2002). Why so many clustering algorithms: A position paper. *ACM SIGKDD Explorations Newsletter*, 4(1):65–75. (Cited on p. 4)

Everitt, B. (1993). *Cluster Analysis*. 3rd edition, New York, NY, USA. Halsted Press. (Cited on pp. 6, 13, 65, 122, 128)

Everitt, B. and Hand, D. (1981). *Finite Mixture Distributions*. London, UK, USA. Chapman & Hall. (Cited on p. 218)

Everitt, B., Landau, S., and Leese, M. (2001). *Cluster Analysis*. Fourth edition, New York, NY, USA. Oxford University Press. (Cited on pp. 5, 207)

Everitt, B. and Nicholls, P. (1975). Visual techniques for representing multivariate data. *The Statistician*, 24(1):37–49. (Cited on p. 109)

Everitt, B. S., Landau, S., Leese, M., and Stahl, D. (2011). *Cluster Analysis*. Hoboken, NJ. Wiley & Sons. (Cited on p. 14)

Faber, V. (1994). Clustering and the continuous $k$-means algorithm. *Los Alamos Science*, 22:138–144. (Cited on pp. 30, 150)

Fahad, A., Alshatri, N., Tari, Z., Alamri, A., Zomaya, A., Khalil, I., Foufou, S., and Bouras, A. (2014). A survey of clustering algorithms for big data: Taxonomy and empirical analysis. *IEEE Transactions on Emerging Topics in Computing*, 2(3):267–279. (Cited on p. 11)

Faloutsos, C., Jagadish, H., Mendelzon, A., and Milo, T. (1997). A signature technique for similarity-based queries. In *Proceedings on Compression and Complexity of Sequences* 1997, Salerno, pages 2–20, Los Alamitos, CA, USA. IEEE Computer Society. (Cited on p. 86)

Faloutsos, C. and Lin, K. (1995). FastMap: A fast algorithm for indexing, data-mining and visualization of traditional and multimedia datasets. In *SIGMOD* '95: *Proceedings of the* 1995 *ACM SIGMOD International Conference on Management of Data*, pages 163–174, New York, NY, USA. ACM Press. (Cited on p. 62)

Faloutsos, C., Ranganathan, M., and Manolopoulos, Y. (1994). Fast subsequence matching in time-series databases. In *SIGMOD* '94: *Proceedings of the* 1994 *ACM SIGMOD International Conference on Management of Data*, pages 419–429, New York, NY, USA. ACM Press. (Cited on p. 86)

Farris, J. (1969). On the cophenetic correlation coefficient. *Systematic Zoology*, 18(3):279–285. (Cited on p. 95)

Fashing, M. and Tomasi, C. (2005). Mean shift is a bound optimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(3):471–474. (Cited on p. 159)

Feder, T. and Greene, D. (1988). Optimal algorithms for approximate clustering. In *STOC* '88: *Proceedings of the Twentieth Annual ACM Symposium on Theory of Computing*, pages 434–444, New York, NY, USA. ACM Press. (Cited on p. 270)

Feigenbaum, J., Kannan, S., Strauss, M., and Viswanathan, M. (1999). An approximate l1-difference algorithm for massive data streams. In *FOCS* '99: *Proceedings of the* 40*th Annual Symposium on Foundations of Computer Science*, page 501, Washington, DC, USA. IEEE Computer Society. (Cited on p. 269)

Felsenstein, J. (1985). Confidence limits on phylogenies: An approach using the bootstrap. *Evolution*, 39(4):783–791. (Cited on p. 290)

Fern, X. Z. and Brodley, C. E. (2003). Random projection for high dimensional data clustering: A cluster ensemble approach. In Fawcett, T. and Mishra, N., editors, *Proceedings of the* 20*th International Conference on Machine Learning (ICML-*03*)*, pages 186–193, Menlo Park, CA, USA. AAAI Press. (Cited on p. 261)

Filho, J., Treleaven, P., and Alippi, C. (1994). Genetic-algorithm programming environments. *IEEE Computer*, 27(6):28–43. (Cited on p. 167)

Filippone, M., Camastra, F., Masulli, F., and Rovetta, S. (2008). A survey of kernel and spectral methods for clustering. *Pattern Recognition*, 41(1):176–190. (Cited on p. 11)

Fisher, W. (1958). On grouping for maximum homogeneity. *Journal of the American Statistical Association*, 53(284):789–798. (Cited on pp. 9, 30, 32, 33, 34, 138)

Fitzgibbon, L., Allison, L., and Dowe, D. (2000). Minimum message length grouping of ordered data. *Lecture Notes in Artificial Intelligence*, 1968:56–70. (Cited on pp. 33, 34)

Florek, K., Lukaszewicz, J., Steinhaus, H., and Zubrzycki, S. (1951). Sur la liaison et la division des points d'un ensemble fini. *Colloquium Mathematicum*, 2:282–285. (Cited on p. 110)

Fowlkes, C., Belongie, S., Chung, F., and Malik, J. (2004). Spectral grouping using the Nystrom method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(2):214–225. (Cited on p. 189)

Fraley, C. (1998). Algorithms for model-based Gaussian hierarchical clustering. *SIAM Journal on Scientific Computing*, 20(1):270–281. (Cited on pp. 212, 213, 214, 220)

Fraley, C. and Raftery, A. (1998). How many clusters? which clustering method? answers via model-based cluster analysis. *The Computer Journal*, 41(8):578–588. (Cited on pp. 4, 207, 215, 217)

Fraley, C. and Raftery, A. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97(458):611–631. (Cited on pp. 207, 208, 217)

Frank, R. and Green, P. (1968). Numerical taxonomy in marketing analysis: A review article. *Journal of Marketing Research*, 5(1):83–94. (Cited on p. 4)

Friedman, H. and Rubin, J. (1967). On some invariant criteria for grouping data. *Journal of the American Statistical Association*, 62(320):1159–1178. (Cited on pp. 9, 36)

Friedman, J. and Rafsky, L. (1981). Graphics for the multivariate two-sample problem. *Journal of the American Statistical Association*, 76(374):277–287. (Cited on p. 109)

Frieze, A., Kannan, R., and Vempala, S. (1998). Fast Monte-Carlo algorithms for finding low-rank approximations. In *FOCS* '98: *Proceedings of the* 39*th Annual Symposium on Foundations of Computer Science*, pages 370–378, Washington, DC, USA. IEEE Computer Society. (Cited on p. 269)

Frieze, A., Kannan, R., and Vempala, S. (2004). Fast Monte-Carlo algorithms for finding low-rank approximations. *Journal of the ACM*, 51(6):1025–1041. (Cited on p. 269)

Fua, Y., Ward, M., and Rundensteiner, E. (1999). Hierarchical parallel coordinates for exploration of large datasets. In *VIS* '99: *Proceedings of the Conference on Visualization* '99, pages 43–50, Los Alamitos, CA, USA. IEEE Computer Society. (Cited on p. 58)

Fujikawa, Y. and Ho, T. (2002). Cluster-based algorithms for dealing with missing values. In Cheng, M.-S., Yu, P. S., and Liu, B., editors, *Advances in Knowledge Discovery and Data Mining,* 6*th Pacific-Asia Conference, PAKDD* 2002*, Taipei, Taiwan, May* 6–8, 2002, *Proceedings*, Volume 2336 of *Lecture Notes in Computer Science*, pages 549–554. Berlin, Germany. Springer-Verlag. (Cited on pp. 9, 10, 288)

Fukunaga, K. (1990). *Introduction to Statistical Pattern Recognition*. 2nd edition. Computer Science and Scientific Computing. San Diego, CA, USA. Academic Press. (Cited on pp. 13, 47, 49)

Fukunaga, K. and Hostetler, L. (1975). The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Transactions on Information Theory*, 21(1):32–40. (Cited on p. 157)

Fukuyama, Y. and Sugeno, M. (1989). A new method of choosing the number of clusters for the fuzzy $c$-means method. In *Proceedings of* 5*th Fuzzy Systems Symposium*, pages 247–250. (Cited on p. 291)

Gaber, M., Zaslavsky, A., and Krishnaswamy, S. (2005). Mining data streams: A review. *ACM SIGMOD Record*, 34(2):18–26. (Cited on pp. 11, 269)

Gaede, V. and Günther, O. (1998). Multidimensional access methods. *ACM Computing Surveys (CSUR)*, 30(2):170–231. (Cited on pp. 351, 352)

Gan, G. (2003). *Subspace clustering for high dimensional categorical data*. Master's thesis, Department of Mathematics and Statistics, York University, Toronto, Canada. (Cited on pp. 240, 244)

Gan, G. (2007). *Subspace clustering based on fuzzy models and mean shifts*. PhD thesis, Department of Mathematics and Statistics, York University, Toronto, ON, Canada. (Cited on pp. 245, 249, 256)

Gan, G. (2011). *Data Clustering in C++: An Object-Oriented Approach*. Data Mining and Knowledge Discovery Series. Boca Raton, FL, USA. Chapman & Hall/CRC Press. (Cited on pp. xix, 14, 309, 311)

Gan, G. and Chen, K. (2016). A soft subspace clustering algorithm with log-transformed distances. *Big Data and Information Analytics*, 1(1):93–109. (Cited on p. 49)

Gan, G. and Huang, J. (2017). A data mining framework for valuing large portfolios of variable annuities. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1467–1475, New York, NY, USA. ACM Press. (Cited on pp. 262, 348)

Gan, G., Lan, Q., and Sima, S. (2016). Scalable clustering by truncated fuzzy $c$-means. *Big Data and Information Analytics*, 1(2/3):247–259. (Cited on pp. 264, 265, 342)

Gan, G. and Valdez, E. A. (2016). An empirical comparison of some experimental designs for the valuation of large variable annuity portfolios. *Dependence Modeling*, 4(1):382–400. (Cited on pp. 342, 348)

Gan, G. and Valdez, E. A. (2017). Valuation of large variable annuity portfolios: Monte Carlo simulation and synthetic datasets. *Dependence Modeling*, 5:354–374. (Cited on pp. 343, 344)

Gan, G. and Valdez, E. A. (2019). *Metamodeling for Variable Annuities*. Boca Raton, FL. Chapman & Hall/CRC Press. (Cited on p. 348)

Gan, G. and Valdez, E. A. (2020). Data clustering with actuarial applications. *North American Actuarial Journal*, 24(2):168–186. (Cited on p. 348)

Gan, G. and Wu, J. (2004). Subspace clustering for high dimensional categorical data. *ACM SIGKDD Explorations Newsletter*, 6(2):87–94. (Cited on p. 239)

Gan, G., Wu, J., and Yang, Z. (2006). A fuzzy subspace algorithm for clustering high dimensional data. In Li, X., Zaiane, O., and Li, Z., editors, *Lecture Notes in Artificial Intelligence*, Volume 4093, pages 271–278, Berlin, Germany. Springer-Verlag. (Cited on p. 245)

Gan, G., Yang, Z., and Wu, J. (2005). A genetic $k$-modes algorithm for clustering categorical data. In Li, X., Wang, S., and Dong, Z., editors, *Proceedings on Advanced Data Mining and Applications: First International Conference, ADMA 2005, Wuhan, China*, volume 3584 of *Lecture Notes in Artificial Intelligence*, pages 195–202, Berlin, Germany. Springer-Verlag. (Cited on pp. 163, 169, 178)

Ganti, V., Gehrke, J., and Ramakrishnan, R. (1999). CACTUS: Clustering categorical data using summaries. In Chaudhuri, S. and Madigan, D., editors, *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 73–83, New York, NY, USA. ACM Press. (Cited on pp. 89, 186)

Ganti, V., Gehrke, J., and Ramakrishnan, R. (2000). DEMON: Mining and monitoring evolving data. In *ICDE* '00: *Proceedings of the* 16*th International Conference on Data Engineering*, pages 439–448, Washington, DC, USA. IEEE Computer Society. (Cited on p. 269)

Ganti, V., Gehrke, J., and Ramakrishnan, R. (2001). DEMON: Mining and monitoring evolving data. *IEEE Transactions on Knowledge and Data Engineering*, 13(1):50–63. (Cited on p. 269)

Garai, G. and Chaudhuri, B. (2004). A novel genetic algorithm for automatic clustering. *Pattern Recognition Letters*, 25(2):173–187. (Cited on p. 168)

García-Escudero, L. and Gordaliza, A. (1999). Robustness properties of $k$-means and trimmed $k$-means. *Journal of the American Statistical Association*, 94(447):956–969. (Cited on pp. 149, 155)

García-Escudero, L., Gordaliza, A., and Matrán, C. (1999a). Asymptotics for trimmed $k$-means and associated tolerance zones. *Journal of Statistical Planning and Inference*, 77(2):247–262. (Cited on p. 155)

García-Escudero, L., Gordaliza, A., and Matrán, C. (1999b). A central limit theorem for multivariate generalized trimmed $k$-means. *The Annuals of Statistics*, 27(3):1061–1079. (Cited on p. 155)

Gath, I. and Geva, A. (1989). Unsupervised optimal fuzzy clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(7):773–780. (Cited on pp. 142, 143)

Gavrilov, M., Anguelov, D., Indyk, P., and Motwani, R. (2000). Mining the stock market: Which measure is best? In *KDD* '00: *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 487–496, New York, NY, USA. ACM Press. (Cited on p. 87)

Ge, X. and Smyth, P. (2000). Deformable Markov model templates for time-series pattern matching. In *KDD* '00: *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 81–90, New York, NY, USA. ACM Press. (Cited on p. 86)

Gehrke, J., Ganti, V., Ramakrishnan, R., and Loh, W. (1999). BOAT—optimistic decision tree construction. In *Proceedings of the* 1999 *ACM SIGMOD international conference on Management of data*, pages 169–180. New York, NY, USA. ACM Press. (Cited on p. 237)

Gehrke, J., Ramakrishnan, R., and Ganti, V. (1998). RainForest — A framework for fast decision tree construction of large datasets. In *Proceedings of the* 24*th International Conference on Very Large Data Bases, VLDB*, pages 416–427, San Francisco, CA, USA. Morgan Kaufmann. (Cited on p. 237)

Georgescu, B., Shimshoni, I., and Meer, P. (2003). Mean shift based clustering in high dimensions: A texture classification example. In *Proceedings. Ninth IEEE International Conference on Computer Vision*, pages 456–463, Los Alamitos, CA, USA. IEEE Computer Society. (Cited on p. 159)

Geurts, P. (2001). Pattern extraction for time series classification. In *PKDD* '01: *Proceedings of the 5th European Conference on Principles of Data Mining and Knowledge Discovery*, Volume 2168 of *Lecture Notes in Computer Science,*, pages 115–127, London, UK. Springer-Verlag. (Cited on p. 87)

Gibbons, F. and Roth, F. (2002). Judging the quality of gene expression-based clustering methods using gene annotation. *Genome Research*, 12:1574–1581. (Cited on p. 331)

Gibbons, P. and Matias, Y. (1999). Synopsis data structures for massive data sets. In *SODA* '99: *Proceedings of the Tenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 909–910, Philadelphia, PA, USA. Society for Industrial and Applied Mathematics. (Cited on p. 269)

Gibson, D., Kleinberg, J., and Raghavan, P. (2000). Clustering categorical data: An approach based on dynamical systems. *The VLDB Journal*, 8(3-4):222–236. (Cited on pp. 187, 188, 295)

Gilbert, A., Guha, S., Indyk, P., Muthukrishnan, S., and Strauss, M. (2002a). Near-optimal sparse Fourier representations via sampling. In *STOC* '02: *Proceedings of the Thirty-Fourth Annual ACM Symposium on Theory of Computing*, pages 152–161, New York, NY, USA. ACM Press. (Cited on p. 269)

Gilbert, A., Kotidis, Y., Muthukrishnan, S., and Strauss, M. (2002b). How to summarize the universe: Dynamic maintenance of quantiles. In *VLDB* 2002*, Proceedings of* 28*th International Conference on Very Large Data Bases*, Hong Kong, China, pages 454–465, San Francisco, CA, USA. Morgan Kaufmann. (Cited on p. 269)

Glover, F. (1989). Tabu search-part I. *ORSA Journal on Computing*, 1(3):190–206. (Cited on p. 169)

Glover, F. (1990). Tabu search-part II. *ORSA Journal on Computing*, 2(1):4–32. (Cited on p. 169)

Glover, F., Taillard, E., and de Werra, D. (1993). A user's guide to tabu search. *Annals of Operations Research*, 41:3–28. (Cited on p. 169)

Gluck, A. and Corter, J. (1985). Information, uncertainty, and the utility of categories. In *Proceedings of the Seventh Annual Conference of the Cognitive Science Society*. Redondo Beach, CA, USA. Cognitive Science Society. (Cited on p. 295)

Goil, S., Nagesh, H., and Choudhary, A. (1999). MAFIA: Efficient and scalable subspace clustering for very large datasets. Technical report CPDC-TR-9906-010, Center for Parallel and Distributed Computing, Department of Electrical & Computer Engineering, Evanston, IL, USA. Northwestern University. (Cited on pp. 222, 234)

Goldberg, D. (1989). *Genetic Algorithms in Search, Optimization, and Machine Learning*. Reading, MA, USA. Addison-Wesley Publishing. (Cited on p. 168)

Goldin, D. and Kanellakis, P. (1995). On similarity queries for time-series data: Constraint specification and implementation. In Montanari, U. and Rossi, F., editors, *the First International Conference on Principles and Practice of Constraint Programming - CP'*95, volume 976 of *Lecture Notes in Computer Science*, pages 137–153. Berlin, Germany. Springer-Verlag. (Cited on p. 82)

Golub, T., Slonim, D., Tamayo, P., Huard, C., Mesirov, M. G. J., Coller, H., and Loh, M. (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 268(15):531–537. (Cited on p. 330)

Gonzalez, T. (1985). Clustering to minimize the maximum intercluster distance. *Theoretical Computer Science*, 38(2-3):293–306. (Cited on p. 224)

Goodall, D. (1966). A new similarity index based on probability. *Biometrics*, 22(4):882–907. (Cited on p. 100)

Goodman, L. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, 61(2):215–231. (Cited on p. 163)

Goodman, L. and Kruskal, W. (1954). Measures of association for cross classifications. *Journal of the American Statistical Association*, 49(268):732–764. (Cited on pp. 96, 97, 98)

Gordon, A. (1987). A review of hierarchical classification. *Journal of the Royal Statistical Society. Series A (General)*, 150(2):119–137. (Cited on pp. 11, 17, 105, 138)

Gordon, A. (1996). Hierarchical classification. In Arabie, P., Hubert, L., and Soete, G., editors, *Clustering and Classification*, pages 65–121, River Edge, NJ, USA. World Scientific. (Cited on pp. xv, 92, 93, 104, 106, 128, 129, 138, 289, 290)

Gordon, A. (1998). Cluster validation. In Hayashi, C., Ohsumi, N., Yajima, K., Tanaka, Y., Bock, H., and Baba, Y. editors, *Data Science, Classification, and Related Methods*, pages 22–39, Tokyo, Japan. Springer-Verlag. (Cited on p. 290)

Gordon, A. (1999). *Classification*. second edition. Boca Raton, FL, USA. Chapman & Hall/CRC. (Cited on pp. 5, 290)

Gosain, A. and Dahiya, S. (2016). Performance analysis of various fuzzy clustering algorithms: A review. *Procedia Computer Science*, 79:100–111. (Cited on p. 12)

Gotlieb, C. and Kumar, S. (1968). Semantic clustering of index terms. *Journal of the Association for Computing Machinery*, 15(4):493–513. (Cited on p. 189)

Gowda, K. and Diday, E. (1992). Symbolic clustering using a new similarity measure. *IEEE Transactions on Systems, Man and Cybernetics*, 22(2):368–378. (Cited on pp. 23, 24)

Gower, J. (1967). A comparison of some methods of cluster analysis. *Biometrics*, 23(4):623–637. (Cited on pp. 93, 122, 138)

Gower, J. (1971). A general coefficient of similarity and some of its properties. *Biometrics*, 27(4):857–874. (Cited on pp. 75, 76, 77, 163)

Gower, J. and Legendre, P. (1986). Metric and Euclidean properties of dissimilarity coefficients. *Journal of Classification*, 3:5–48. (Cited on pp. 73, 100)

Gower, J. and Ross, G. (1969). Minimum spanning trees and single linkage cluster analysis. *Applied Statistics*, 18(1):54–64. (Cited on pp. 106, 130, 131, 189)

Grabusts, P. and Borisov, A. (2002). Using grid-clustering methods in data classification. In *International Conference on Parallel Computing in Electrical Engineering,* 2002. *PARELEC* '02. *Proceedings*, Latvia, pages 425–426, Los Alamitos, CA, USA. IEEE Computer Society. (Cited on p. 191)

Green, P. and Rao, V. (1969). A note on proximity measures and cluster analysis. *Journal of Marketing Research*, 6(3):359–364. (Cited on p. 100)

Greene, W. (2003). Unsupervised hierarchical clustering via a genetic algorithm. In *The* 2003 *Congress on Evolutionary Computation,* 2003. *CEC* '03, Volume 2, pages 998–1005, Piscataway, NJ, USA. IEEE. (Cited on p. 168)

Greenwald, M. and Khanna, S. (2001). Space-efficient online computation of quantile summaries. In *SIGMOD* '01: *Proceedings of the* 2001 *ACM SIGMOD International Conference on Management of Data*, pages 58–66, New York, NY, USA. ACM Press. (Cited on p. 269)

Guha, S. and Koudas, N. (2002). Approximating a data stream for querying and estimation: Algorithms and performance evaluation. In *ICDE* '02: *Proceedings of the* 18*th International Conference on Data Engineering*, pages 567–576, Washington, DC, USA. IEEE Computer Society. (Cited on p. 269)

Guha, S., Koudas, N., and Shim, K. (2001). Data-streams and histograms. In *STOC* '01: *Proceedings of the Thirty-Third Annual ACM Symposium on Theory of Computing*, pages 471–475, New York, NY, USA. ACM Press. (Cited on p. 269)

Guha, S., Meyerson, A., Mishra, N., Motwani, R., and Callaghan, L. (2003). Clustering data streams: Theory and practice. *IEEE Transactions on Knowledge and Data Engineering*, 15(3):515–528. (Cited on pp. 269, 272)

Guha, S., Mishra, N., Motwani, R., and O'Callaghan, L. (2000a). Clustering data streams. In *FOCS* '00: *Proceedings of the* 41*st Annual Symposium on Foundations of Computer Science*, pages 359–366, Washington, DC, USA. IEEE Computer Society. (Cited on p. 270)

Guha, S., Rastogi, R., and Shim, K. (1998). CURE: an efficient clustering algorithm for large databases. In *Proceedings of the* 1998 *ACM SIGMOD International Conference on Management of Data*, pages 73–84. New York, NY, USA. ACM Press. (Cited on pp. 30, 134, 232, 261)

Guha, S., Rastogi, R., and Shim, K. (2000b). ROCK: A robust clustering algorithm for categorical attributes. *Information Systems*, 25(5):345–366. (Cited on pp. 88, 188)

Gunopulos, D. and Das, G. (2000). Time series similarity measures (tutorial PM-2). In *Tutorial Notes of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Boston, pages 243–307, New York, NY, USA. ACM Press. (Cited on pp. 24, 79, 80, 81, 82)

Gunopulos, D. and Das, G. (2001). Time series similarity measures and time series indexing. In *SIGMOD* '01: *Proceedings of the* 2001 *ACM SIGMOD International Conference on Management of Data*, page 624. New York, NY, USA. ACM Press. (Cited on pp. 80, 83, 84)

Gupta, S., Rao, K., and Bhatnagar, V. (1999). $k$-means clustering algorithm for categorical attributes. In *Proceedings of the First International Conference on Data Warehousing and Knowledge Discovery*, pages 203–208, Berlin, Germany. Springer-Verlag. (Cited on p. 164)

Haas, P., Naughton, J., Seshadri, S., and Stokes, L. (1995). Sampling-based estimation of the number of distinct values of an attribute. In *VLDB* '95: *Proceedings of the* 21*st International Conference on Very Large Data Bases*, pages 311–322, San Francisco, CA, USA. Morgan Kaufmann. (Cited on p. 269)

Halkidi, M., Batistakis, Y., and Vazirgiannis, M. (2001a). Clustering algorithms and validity measures. *Thirteenth International Conference on Scientific and Statistical Database Management*, pages 3–22, Los Alamitos, CA, USA. IEEE Computer Society. (Cited on p. 285)

Halkidi, M., Batistakis, Y., and Vazirgiannis, M. (2001b). On clustering validation techniques. *Journal of Intelligent Information Systems*, 17:107–145. (Cited on pp. 290, 291)

Halkidi, M., Batistakis, Y., and Vazirgiannis, M. (2002a). Cluster validity methods: part I. *ACM SIGMOD Record*, 31(2):40–45. (Cited on pp. 35, 277, 279, 280)

Halkidi, M., Batistakis, Y., and Vazirgiannis, M. (2002b). Clustering validity checking methods: Part II. *ACM SIGMOD Record*, 31(3):19–27. (Cited on pp. 35, 281, 282, 283, 286)

Halkidi, M. and Vazirgiannis, M. (2001). Clustering validity assessment: Finding the optimal partitioning of a dataset. *ICDM* 2001*, Proceedings of the IEEE International Conference on Data Mining*, pages 187–194, Los Alamitos, CA, USA. IEEE Computer Society. (Cited on p. 284)

Halkidi, M., Vazirgiannis, M., and Batistakis, I. (2000). Quality scheme assessment in the clustering process. *Proceedings of PKDD*, pages 265–276. Berlin, Heidelberg. Springer. (Cited on pp. 35, 283)

Hall, L., Özyurt, I., and Bezdek, J. (1999). Clustering with a genetically optimized approach. *IEEE Transactions on Evolutionary Computation*, 3(2):103–112. (Cited on p. 169)

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The WEKA data mining software: An update. *SIGKDD Explorations*, 11(1):10–18. (Cited on p. 308)

Hamerly, G. and Elkan, C. (2002). Alternatives to the $k$-means algorithm that find better clusterings. In *Proceedings of the Eleventh International Conference on Information and Knowledge Management*, pages 600–607, McLean, Virginia, USA. ACM Press. (Cited on p. 150)

Hampel, F. (1971). A general qualitative definition of robustness. *The Annals of Mathematical Statistics*, 42(6):1887–1896. (Cited on p. 155)

Handl, J., Knowles, J., and Kell, D. (2005). Computational cluster validation in post-genomic data analysis. *Bioinformatics*, 21:3201–3212. (Cited on p. 331)

Hansen, P. and Mladenović, N. (2001a). $J$-means: A new local search heuristic for minimum sum of squares clustering. *Pattern Recognition*, 34(2):405–413. (Cited on pp. 170, 173, 174)

Hansen, P. and Mladenović, N. (2001b). Variable neighborhood search: Principles and applications. *European Journal of Operational Research*, 130(3):449–467. (Cited on p. 170)

Har-Peled, S. and Mazumdar, S. (2004). On coresets for $k$-means and $k$-median clustering. In *Proceedings of the Thirty-Sixth Annual ACM Symposium on Theory of Computing*, pages 291–300. New York, NY, USA. ACM Press. (Cited on p. 153)

Har-Peled, S. and Varadarajan, K. (2002). Projective clustering in high dimensions using coresets. In *Proceedings of the Eighteenth Annual Symposium on Computational Geometry*, pages 312–318. New York, NY, USA. ACM Press. (Cited on p. 259)

Harding, E. (1967). The number of partitions of a set of $n$ points in $k$ dimensions induced by hyperplanes. In *Proceedings of the Edinburgh Mathematical Society (Series* II*)*, volume 15, pages 285–289, Edinburgh, UK. Scottish Academic Press. (Cited on p. 138)

Hartigan, J. (1967). Representation of similarity matrices by trees. *Journal of the American Statistical Association*, 62(320):1140–1158. (Cited on pp. 66, 100, 104)

Hartigan, J. (1975). *Clustering Algorithms*. Toronto, Canada. Wiley & Sons. (Cited on pp. 5, 9, 12, 147, 245)

Hartigan, J. and Mohanty, S. (1992). The runt test for multimodality. *Journal of Classification*, 9(1):63–70. (Cited on p. 290)

Hartigan, J. and Wong, M. (1979). Algorithm As 136: A $k$-means clustering algorithm. *Applied Statistics*, 28(1):100–108. (Cited on pp. 30, 109, 147, 245)

Hathaway, R. and Bezdek, J. (1984). Local convergence of the fuzzy $c$-means algorithms. *Pattern Recognition*, 19(6):477–480. (Cited on pp. 145, 167)

Haykin, S. (1999). *Neural Networks: A Comprehensive Foundation*, second edition, Upper Saddle River, NJ, USA. Prentice-Hall. (Cited on pp. 54, 55)

He, Y., Tan, H., Luo, W., Feng, S., and Fan, J. (2014). MR-DBSCAN: A scalable MapReduce-based DBSCAN algorithm for heavily skewed data. *Frontiers of Computer Science*, 8(1):83–99. (Cited on p. 262)

Hennig, C., Meila, M., Murtagh, F., and Rocci, R., editors (2016). *Handbook of Cluster Analysis*. Boca Raton, FL, USA. CRC Press. (Cited on p. 14)

Henriques, R. and Madeira, S. C. (2018). Triclustering algorithms for three-dimensional data analysis: A comprehensive survey. *ACM Computing Surveys (CSUR)*, 51(5):95. (Cited on p. 12)

Henzinger, M., Raghavan, P., and Rajagopalan, S. (1998). Computing on data streams. Technical report TR-1998-011, Digital Equipment Corp. (Cited on p. 269)

Herrero, J., Diaz-Uriarte, R., and Dopazo, J. (2003). Gene expression data preprocessing. *Bioinformatics*, 19(5):655–656. (Cited on p. 330)

Hertz, J., Krogh, A., and Palmer, R. (1991). *Introduction to the Theory of Neural Computation*. Boston, MA, USA. Addison-Wesley Longman Publishing. (Cited on p. 56)

Hetland, M. (2004). A survey of recent methods for efficient retrieval of similar time sequences. In Last, M., Kandel, A., and Bunke, H., editors, *Data Mining in Time Series Databases*, volume 57 of *Machine Perception and Artificial Intelligence*. Hackensack, NJ, USA. World Scientific. (Cited on pp. 80, 267)

Heyer, L., Kruglyak, S., and Yooseph, S. (1999). Exploring expression data: Identification and analysis of coexpressed genes. *Genome Research*, 9(11):1106–1115. (Cited on p. 333)

Hill, A., Brown, E., Whitley, M., Tucker-Kellogg, G., Hunter, C. P., and Slonim, D. (2001). Evaluation of normalization procedures for oligonucleotide array data based on spiked cRNA controls. *Genome Biology*, 2(12):55. (Cited on p. 330)

Hinneburg, A. and Keim, D. (1998). An efficient approach to clustering in large multimedia databases with noise. In *Knowledge Discovery and Data Mining*, pages 58–65, New York, NY, USA. ACM Press. (Cited on pp. 30, 203, 204)

Hipp, J., Güntzer, U., and Nakhaeizadeh, G. (2000). Algorithms for association rule mining – a general survey and comparison. *ACM SIGKDD Explorations Newsletter*, 2(1):58–64. (Cited on pp. 272, 276)

Hoare, C. (1961). Algorithm 64: Quicksort. *Communications of the ACM*, 4(7):321. (Cited on p. 32)

Hodges, K. and Wotring, J. (2000). Client typology based on functioning across domains using the CAFAS: Implications for service planning. *Journal of Behavioral Health Services and Research*, 27(3):257–270. (Cited on p. 4)

Hodson, F. (1970). Cluster analysis and archaeology: Some new developments and applications. *World Archaeology*, 1(3):299–320. (Cited on p. 138)

Hofmann, T. and Buhmann, J. (1995). Multidimensional scaling and data clustering. In Tesauro, G., Touretzky, D., and Leen, T., editors, *Advances in Neural Information Processing Systems*, Volume 7, pages 459–466. Cambridge, MA, USA. The MIT Press. (Cited on p. 62)

Holland, J. (1975). *Adaptation in Natural and Artificial Systems*. Ann Arbor, MI, USA. University of Michigan Press. (Cited on p. 167)

Holman, E. (1992). Statistical properties of large published classifications. *Journal of Classification*, 9(2):187–210. (Cited on p. 128)

Hopcroft, J. and Tarjan, R. (1973). Algorithm 447: Efficient algorithms for graph manipulation. *Commun. ACM*, 16(6):372–378. (Cited on p. 224)

Höppner, F., Klawonn, F., Kruse, R., and Runkler, T. (1999). *Fuzzy Cluster Analysis: Methods for Classification, Data Analysis and Image Recognition*. Chichester, UK. Wiley. (Cited on p. 146)

Howard, R. (1966). Classifying a population into homogeneous groups. In Lawrence, J., editor, *Operational Research and the Social Sciences*, pages 585–594, London, UK. Tavistock. (Cited on p. 174)

Hsu, C. (2006). Generalizing self-organizing map for categorical data. *IEEE Transactions on Neural Networks*, 17(2):294–304. (Cited on p. 61)

Hu, J. and Pei, J. (2018). Subspace multi-clustering: A review. *Knowledge and Information Systems*, 56(2):257–284. (Cited on p. 12)

Hua, K., Lang, S., and Lee, W. (1994). A decomposition-based simulated annealing technique for data clustering. In *Proceedings of the Thirteenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, May* 24–26, 1994, *Minneapolis, MN*, Volume 13, pages 117–128, New York, NY, USA. ACM Press. (Cited on pp. 182, 184)

Huang, Z. (1997a). Clustering large data sets with mixed numeric and categorical values. In *Knowledge Discovery and Data Mining: Techniques and Applications*, pages 21–34, Singapore. World Scientific. (Cited on p. 165)

Huang, Z. (1997b). A fast clustering algorithm to cluster very large categorical data sets in data mining. In *SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery*, Tucson, Arizona, pages 1–8, Vancouver, Canada. University of British Columbia. (Cited on pp. 71, 160)

Huang, Z. (1998). Extensions to the $k$-means algorithm for clustering large data sets with categorical values. *Data Mining and Knowledge Discovery*, 2(3):283–304. (Cited on pp. 6, 71, 148, 160, 161, 162, 165)

Huang, Z. and Ng, M. (1999). A fuzzy $k$-modes algorithm for clustering categorical data. *IEEE Transactions on Fuzzy Systems*, 7(4):446–452. (Cited on pp. 143, 144, 249)

Hubálek, Z. (1982). Coefficients of association and similarity based on binary (presence-absence) data: An evaluation. *Biological Reviews of the Cambridge Philosophical Society*, 57:669–689. (Cited on pp. 73, 100)

Hubert, L. (1974). Some applications of graph theory to clustering. *Psychometrika*, 39(3):283–309. (Cited on p. 189)

Hussein, N. (2003). A fast greedy $k$-means algorithm. Master's thesis, Computer Science, University of Amsterdam, Amsterdam, The Netherlands. (Cited on p. 178)

Ibaraki, T. and Katoh, N. (1988). *Resource Allocation Problems: Algorithmic Approaches*. Cambridge, MA, USA. MIT Press. (Cited on p. 224)

Ichino, M. (1988). General metrics for mixed features–the Cartesian space theory for pattern recognition. In *Proceedings of the* 1988 *IEEE International Conference on Systems, Man, and Cybernetics*, Volume 1, pages 494–497, Piscataway, NJ, USA. IEEE. (Cited on pp. 77, 79)

Ichino, M. and Yaguchi, H. (1994). Generalized Minkowski metrics for mixed feature-type data analysis. *IEEE Transactions on Systems, Man and Cybernetics*, 24(4):698–708. (Cited on pp. 77, 79)

Indyk, P. (1999). Sublinear time algorithms for metric space problems. In *STOC* '99: *Proceedings of the Thirty-First Annual ACM Symposium on Theory of Computing*, pages 428–434, New York, NY, USA. ACM Press. (Cited on p. 270)

Indyk, P. (2000). Stable distributions, pseudorandom generators, embeddings and data stream computation. In *FOCS* '00: *Proceedings of the* 41*st Annual Symposium on Foundations of Computer Science*, page 189, Washington, DC, USA. IEEE Computer Society. (Cited on p. 269)

Indyk, P., Koudas, N., and Muthukrishnan, S. (2000). Identifying representative trends in massive time series data sets using sketches. In *VLDB* '00: *Proceedings of the* 26*th International Conference on Very Large Data Bases*, pages 363–372, San Francisco, CA, USA. Morgan Kaufmann. (Cited on p. 87)

Inmon, W. and Linstedt, D. (2014). *Data Architecture: A Primer for the Data Scientist: Big Data, Data Warehouse and Data Vault*. Cambridge, MA. Morgan Kaufmann. (Cited on p. 19)

Inselberg, A. and Dimsdale, B. (1990). Parallel coordinates: A tool for visualizing multi-dimensional geometry. In *VIS* '90: *Proceedings of the* 1*st conference on Visualization* '90, pages 361–378, Los Alamitos, CA, USA. IEEE Computer Society. (Cited on pp. 57, 58)

Isson, J. P. (2018). *Unstructured Data Analytics: How to Improve Customer Acquisition, Customer Retention, and Fraud Detection and Prevention*. Hoboken, NJ, USA. Wiley & Sons. (Cited on pp. 19, 22)

Itoh, T., Yamaguchi, Y., Ikehata, Y., and Kajinaga, Y. (2004). Hierarchical data visualization using a fast rectangle-packing algorithm. *IEEE Transactions on Visualization and Computer Graphics*, 10(3):302–313. (Cited on p. 62)

Jacques, J. and Preda, C. (2014). Functional data clustering: A survey. *Advances in Data Analysis and Classification*, 8(3):231–255. (Cited on p. 11)

Jain, A. and Dubes, R. (1988). *Algorithms for Clustering Data*. Englewood Cliffs, New Jersey, USA. Prentice-Hall. (Cited on pp. 5, 13, 41, 43, 66, 69, 93, 103, 115, 117, 118, 122, 131, 147, 277, 278, 280, 290)

Jain, A., Duin, R., and Mao, J. (2000). Statistical pattern recognition: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1):4–37. (Cited on p. 11)

Jain, A., Murty, M., and Flynn, P. (1999). Data clustering: A review. *ACM Computing Surveys (CSUR)*, 31(3):264–323. (Cited on pp. 5, 11, 17, 69)

Jambu, M. (1978). *Classification automatique pour l'analyse de données*. Paris, France. Dunod. (Cited on pp. xv, 91, 93, 94, 128)

Jamshidian, M. and Jennrich, R. (1993). Conjugate gradient acceleration of the EM algorithm. *Journal of the American Statistical Association*, 88(421):221–228. (Cited on p. 215)

Jamshidian, M. and Jennrich, R. (1997). Acceleration of the EM algorithm by using quasi-Newton methods. *Journal of the Royal Statistical Society. Series B (Methodological)*, 59(3):569–587. (Cited on p. 215)

Januzaj, E., Kriegel, H.-P., and Pfeifle, M. (2004). DBDC: Density based distributed clustering. In Bertino, E., Christodoulakis, S., Plexousakis, D., Christophides, V., Koubarakis, M., Böhm, K., and Ferrari, E., editors, *Advances in Database Technology - EDBT* 2004: 9*th International Conference on Extending Database Technology, Heraklion, Crete, Greece, March* 14–18, 2004, pages 88–105, Berlin, Germany. Springer-Verlag. (Cited on p. 261)

Jardine, C., Jardine, N., and Sibson, R. (1967). The structure and construction of taxonomic hierarchies. *Mathematical Biosciences*, 1(2):173–179. (Cited on p. 105)

Jardine, N. (1971). A new approach to pattern recognition. *Nature*, 234:526–528. (Cited on p. 189)

Jardine, N. and Sibson, R. (1968). The construction of hierarchic and non-hierarchic classifications. *The Computer Journal*, 11(2):177–184. (Cited on pp. 9, 189)

Jeffreys, H. (1935). Some tests of significance, treated by the theory of probability. *Proceedings of the Cambridge Philosophy Society*, 31:203–222. (Cited on p. 216)

Jeffreys, H. (1961). *Theory of Probability*. third edition. Oxford, UK. Oxford University Press. (Cited on p. 216)

Jiang, D., Pei, J., and Zhang, A. (2003). DHC: A density-based hierarchical clustering method for time series gene expression data. In *Third IEEE Symposium on Bioinformatics and Bioengineering,* 2003. *Proceedings*, pages 393–400, Los Alamitos, CA, USA. IEEE Computer Society. (Cited on p. 332)

Jiang, D., Tang, C., and Zhang, A. (2004). Cluster analysis for gene expression data: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 16(11):1370–1386. (Cited on pp. 3, 11, 329, 330, 331, 333, 339)

Jiang, T. and Ma, S. (1996). Cluster analysis using genetic algorithms. In *The Third International Conference on Signal Processing,* 1996, Beijing China, Volume 2, pages 1277–1279, New York, NY, USA. IEEE Computer Society. (Cited on p. 168)

Jin, X., Lu, Y., and Shi, C. (2002). Similarity measure based on partial information of time series. In *KDD* '02: *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 544–549, New York, NY, USA. ACM Press. (Cited on p. 87)

Johansson, J., Ljung, P., Jern, M., and Cooper, M. (2006). Revealing structure in visualizations of dense 2D and 3D parallel coordinates. *Information Visualization*, 5(2):125–136. (Cited on p. 58)

Johnson, R. and Wichern, D. (1998). *Applied Multivariate Statistical Analysis*. Upper Saddle River, NJ, USA. Prentice Hall. (Cited on p. 72)

Johnson, S. (1967). Hierarchical clustering schemes. *Psychometrika*, 32(3):241–254. (Cited on pp. 105, 110, 111, 113)

Jolliffe, I. (2002). *Principal Component Analysis*. Springer Series in Statistics. Second edition. New York, NY, USA. Springer-Verlag. (Cited on pp. 44, 45, 68)

José-Garcia, A. and Gómez-Flores, W. (2016). Automatic clustering using nature-inspired metaheuristics: A survey. *Applied Soft Computing*, 41:192–213. (Cited on p. 12)

Kahveci, T., Singh, A., and Gurel, A. (2002). An efficient index structure for shift and scale invariant search of mufti-attribute time sequences. In *ICDE* '02: *Proceedings of the* 18*th International Conference on Data Engineering* 2002, page 266, San Jose, CA, USA. IEEE Computer Society. (Cited on p. 86)

Kailing, K., Kriegel, H., and Kröger, P. (2004). Density-connected subspace clustering for high-dimensional data. In *Proceedings of the SIAM International Conference on Data Mining*, Lake Buena Vista, Florida, pages 246–257, Philadelphia, PA, USA, Society for Industrial and Applied Mathematics. (Cited on p. 259)

Kalpakis, K., Gada, D., and Puttagunta, V. (2001). Distance measures for effective clustering of ARIMA time-series. In *ICDM* 2001*, Proceedings of the IEEE International Conference on Data Mining*, pages 273–280, San Jose, CA USA. IEEE Computer Society. (Cited on pp. 268, 269)

Kandogan, E. (2001). Visualizing multi-dimensional clusters, trends, and outliers using star coordinates. In *KDD* '01: *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 107–116, New York, NY, USA. ACM Press. (Cited on p. 58)

Kannan, R., Vempala, S., and Vetta, A. (2004). On clusterings: Good, bad and spectral. *Journal of the ACM*, 51(3):497–515. (Cited on p. 189)

Kantabutra, S. and Couch, A. (2000). Parallel $k$-means clustering algorithm on NOWs. *NECTEC Technical Journal*, 1(6):243–248. `http://www.nectec.or.th`. (Cited on p. 154)

Kanth, K., Agrawal, D., and Singh, A. (1998). Dimensionality reduction for similarity searching in dynamic databases. In *Proceedings of the* 1998 *ACM SIGMOD international conference on Management of data*, pages 166–176, New York, NY, USA. ACM Press. (Cited on p. 45)

Kanungo, T., Mount, D., Netanyahu, N., Piatko, C., and Wu, R. S. A. (2002). An efficient $k$-means clustering algorithm: analysis and implementation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):881–892. (Cited on p. 153)

Karypis, G., Han, E., and Kumar, V. (1999). Chameleon: Hierarchical clustering using dynamic modeling. *Computer*, 32(8):68–75. (Cited on p. 185)

Kass, R. and Raftery, A. (1995). Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795. (Cited on pp. 155, 208, 216)

Kaufman, L. and Rousseeuw, P. (1990). *Finding Groups in Data–An Introduction to Cluster Analysis*. Wiley series in probability and mathematical statistics. John Wiley & Sons, Inc., New York. (Cited on pp. 9, 65, 71, 92, 95, 106, 129, 135, 136, 150, 165, 224)

Kaufman, L. and Rousseeuw, P. J. (2005). *Finding groups in data: An introduction to cluster analysis*. Wiley, Hoboken, NJ. (Cited on p. 13)

Ke, Q. and Kanade, T. (2004). Robust subspace clustering by combined use of knnd metric and svd algorithm. In *Proceedings of the* 2004 *IEEE Computer Society Conference on Computer Vision and Pattern Recognition,* 2004. *CVPR* 2004., volume 2, pages 592–599. IEEE. (Cited on p. 259)

Keim, D. and Hinneburg, A. (1999). Optimal grid-clustering: Towards breaking the curse of dimensionality in high-dimensional clustering. In *Proceedings of the* 25*th International Conference on Very Large Data Bases(VLDB* '99*)*, pages 506–517, San Francisco. Morgan Kaufmann. (Cited on pp. 150, 192, 198)

Keim, D. and Kriegel, H. (1996). Visualization techniques for mining large databases: A comparison. *IEEE Transactions on Knowledge and Data Engineering*, 8(6):923–938. (Cited on p. 63)

Kendall, S. and Ord, J. (1990). *Time Series*. Third edition. Edward Arnold, Great Britain. (Cited on p. 24)

Keogh, E., Chu, S., and Pazzani, M. (2001). Ensemble-index: A new approach to indexing large databases. In *KDD* '01: *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 117–125. New York, NY, USA. ACM Press. (Cited on pp. 80, 267)

Keogh, E. and Kasetty, S. (2003). On the need for time series data mining benchmarks: A survey and empirical demonstration. *Data Mining and Knowledge Discovery*, 7(4):349–371. (Cited on pp. 80, 87)

Keogh, E., Lin, J., and Truppel, W. (2003). Clustering of time series subsequences is meaningless: Implications for previous and future research. In *ICDM* '03: *Proceedings of the Third IEEE International Conference on Data Mining*, pages 115–122, Washington, DC, USA. IEEE Computer Society. (Cited on pp. 267, 268)

Keogh, E. and Pazzani, M. (1998). An enhanced representation of time series which allows fast and accurate classification, clustering and relevance feedback. In Agrawal, R., Stolorz, P., and Piatetsky-Shapiro, G., editors, *Fourth International Conference on Knowledge Discovery and Data Mining (KDD*'98*)*, pages 239–241, New York, NY, USA. ACM Press. (Cited on p. 87)

Keogh, E. and Pazzani, M. (2000). Scaling up dynamic time warping for data mining applications. In *KDD* '00: *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 285–289. New York, NY, USA. ACM Press. (Cited on pp. 83, 84)

Keogh, E. and Ratanamahatana, C. (2005). Exact indexing of dynamic time warping. *Knowledge and Information Systems*, 7(3):358–386. (Cited on p. 84)

Keogh, E. and Smyth, P. (1997). A probabilistic approach to fast pattern matching in time series databases. In Heckerman, D., Mannila, H., Pregibon, D., and Uthurusamy, R., editors, *KDD* '97: *Proceedings of the* 3*rd International Conference of Knowledge Discovery and Data Mining*, pages 24–30, Newport Beach, California, USA. AAAI Press. (Cited on p. 85)

Khan, S. and Ahmad, A. (2004). Cluster center initialization algorithm for $k$-means clustering. *Pattern Recognition Letters*, 25(11):1293–1302. (Cited on p. 150)

Kim, E., Lam, J., and Han, J. (2000a). AIM: Approximate intelligent matching for time series data. In *DaWaK* 2000: *Proceedings of the Second International Conference on Data Warehousing and Knowledge Discovery*, pages 347–357, London, UK. Springer-Verlag. (Cited on p. 87)

Kim, H., Golub, G., and Park, H. (2005). Missing value estimation for DNA microarray gene expression data: Local least squares imputation. *Bioinformatics*, 21(2):187–198. (Cited on p. 330)

Kim, S., Kwon, S., and Cook, D. (2000b). Interactive visualization of hierarchical clusters using MDS and MST. *Metrika*, 51(1):39–51. (Cited on p. 62)

Klein, R. and Dubes, R. (1989). Experiments in projection and clustering by simulated annealing. *Pattern Recognition*, 22(2):213–220. (Cited on pp. 177, 184)

Klock, H. and Buhmann, J. (2000). Data visualization by multidimensional scaling: A deterministic annealing approach. *Pattern Recognition*, 33(4):651–669. (Cited on pp. 53, 62)

Kogan, J. (2007). *Introduction to Clustering Large and High-Dimensional Data*. Cambridge, UK. Cambridge University Press. (Cited on p. 13)

Kogan, J., Nicholas, C., and Teboulle, M. (2005). *Grouping Multidimensional Data: Recent Advances in Clustering*. Berlin, Germany. Springer-Verlag. (Cited on p. 13)

Kohonen, T. (1989). *Self-Organization and Associative Memory*. Third edition. New York, NY, USA. Springer-Verlag. (Cited on pp. 54, 339)

Kohonen, T. (1990). The self-organizing map. *Proceedings of the IEEE*, 78(9):1464–1480. (Cited on p. 54)

Konig, A. (2000). Interactive visualization and analysis of hierarchical neural projections for data mining. *IEEE Transactions Neural Networks*, 11(3):615–624. (Cited on p. 56)

Koren, Y. and Harel, D. (2003). A two-way visualization method for clustered data. In *KDD* '03: *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and data mining*, pages 589–594, New York, NY, USA. ACM Press. (Cited on p. 62)

Kriegel, H.-P., Kröger, P., and Zimek, A. (2009). Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering. *ACM Transactions on Knowledge Discovery from Data*, 3(1):1:1–1:58. (Cited on p. 11)

Krishna, K. and Narasimha, M. (1999). Genetic $k$-means algorithm. *IEEE Transactions on Systems, Man and Cybernetics, Part B*, 29(3):433–439. (Cited on pp. 168, 169, 175, 181)

Krishnapuram, R. and Freg, C. (1991). Fuzzy algorithms to find linear and planar clusters and their applications. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition,* 1991. *Proceedings CVPR* '91, pages 426–431, Los Alamitos, CA, USA. IEEE Computer Society. (Cited on p. 259)

Kruskal, J. (1956). On the shortest spanning subtree of a graph and the traveling salesman problem. *Proceedings of the American Mathematical Society*, 7(1):48–50. (Cited on p. 131)

Kruskal, J. (1964). Nonmetric multidimensional scaling: Numerical method. *Psychometrika*, 29(2):115–129. (Cited on p. 53)

Kruskal, J. (1971). Comments on "A nonlinear mapping for data structure analysis." *IEEE Transactions on Computers*, C-20(12):1614–1614. (Cited on p. 52)

Kruskal, J. and Landwehr, J. (1983). Icicle plots: Better displays for hierarchical clustering. *The American Statistician*, 37(2):162–168. (Cited on pp. 108, 109)

Krzanowski, W. and Lai, Y. (1988). A criterion for determining the number of groups in a data set using sum-of-squares clustering. *Biometrics*, 44(1):23–34. (Cited on pp. 36, 37)

Kubat, M. (2017). *An Introduction to Machine Learning*. second edition. New York, NY, USA. Springer-Verlag. (Cited on p. 4)

Kuiper, F. and Fisher, L. (1975). Shorter communications 391: A Monte Carlo comparison of six clustering procedures. *Biometrics*, 31(3):777–783. (Cited on p. 128)

Lance, G. and Williams, W. (1966). A generalised sorting strategy for computer classifications. *Nature*, 212:218. (Cited on p. 93)

Lance, G. and Williams, W. (1967a). A general theory of classificatory sorting strategies I. Hierarchical systems. *The Computer Journal*, 9(4):373–380. (Cited on pp. 91, 122, 128, 138)

Lance, G. and Williams, W. (1967b). A general theory of classificatory sorting strategies II. Clustering systems. *The Computer Journal*, 10(3):271–277. (Cited on p. 110)

Lance, G. and Williams, W. (1967c). Mixed-data classificatory programs I - agglomerative systems. *Australian Computer Journal*, 1(1):15–20. (Cited on p. 138)

Lanyon, S. (1985). Detecting internal inconsistencies in distance data. *Systematic Zoology*, 34(4):397–403. (Cited on p. 290)

Lee, R., Slagle, J., and Mong, C. (1976). Application of clustering to estimate missing data and improve data integrity. In *Proceedings of the* 2*nd International Conference on Software Engineering*, pages 539–544, San Francisco, California, United States. IEEE Computer Society. (Cited on p. 11)

Lee, S., Chun, S., Kim, D., Lee, J., and Chung, C. (2000). Similarity search for multidimensional data sequences. In *Proceedings of the* 16*th International Conference on Data Engineering,* 2000, pages 599–608, San Diego, CA USA. IEEE. (Cited on pp. 81, 87)

Lee, S. and Hayes, M. H. (2004). Properties of the singular value decomposition for efficient data clustering. *IEEE Signal Processing Letters*, 11(11):862–866. (Cited on p. 261)

Legendre, L. and Legendre, P. (1983). *Numerical Ecology*. New York, NY, USA. Elsevier Scientific. (Cited on pp. 5, 70, 71, 72, 74)

Legendre, P. and Rogers, D. (1972). Characters and clustering in taxonomy: A synthesis of two taximetric procedures. *Taxon*, 21:567–606. (Cited on p. 189)

Levin, M. S. (2015). Combinatorial clustering: Literature review, methods, examples. *Journal of Communications Technology and Electronics*, 60(12):1403–1428. (Cited on p. 11)

Li, C., Yu, P., and Castelli, V. (1998). MALM: A framework for mining sequence database at multiple abstraction levels. In *CIKM* '98: *Proceedings of the Seventh International Conference on Information and Knowledge Management*, pages 267–272, New York, NY, USA. ACM Press. (Cited on pp. 87, 267)

Likas, A. and Verbeek, N. V. J. (2003). The global $k$-means clustering algorithm. *Pattern Recognition*, 36(2):45–461. (Cited on p. 177)

Lin, J., Keogh, E., and Truppel, W. (2003). Clustering of streaming time series is meaningless. In *DMKD* '03: *Proceedings of the* 8*th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, pages 56–65, New York, NY, USA. ACM Press. (Cited on pp. 267, 268)

Liu, B., Xia, Y., and Yu, P. (2000). Clustering through decision tree construction. In *Proceedings of the Ninth International Conference on Information and Knowledge Management*, pages 20–29, McLean, Virginia, USA. ACM Press. (Cited on pp. 222, 237)

Liu, Y., Chen, K., Liao, X., and Zhang, W. (2004). A genetic clustering method for intrusion detection. *Pattern Recognition*, 37(5):927–942. (Cited on p. 169)

Lloyd, S. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137. (Cited on p. 150)

Lockhart, D., Dong, H., Byrne, M., Follettie, M., Chee, M. G. M., Mittmann, M., Wang, C., Kobayashi, M., and Brown, H. N. E. (1996). Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnology*, 14:1675–1680. (Cited on p. 329)

Lorr, M. (1983). *Cluster Analysis for Social Scientists*. The Jossey-Bass Social and Behavioral Science Series. San Francisco, CA, USA. Jossey-Bass. (Cited on pp. 6, 8, 13)

Lu, Y. and Huang, Y. (2005). Mining data streams using clustering. In *Proceedings of the* 2005 *International Conference on Machine Learning and Cybernetics*, Volume 4, pages 2079–2083, Red Hook, NY, USA. Curran Associates. (Cited on p. 269)

Lu, Y., Lu, S., Fotouchi, F., Deng, Y., and Brown, S. (2004a). Incremental genetic $k$-means algorithm and its application in gene expression data analysis. *BMC Bioinformatics*, 5(172):1–27. (Cited on pp. 169, 177)

Lu, Y., Lu, S., Fotouhi, F., Deng, Y., and Brown, S. (2004b). FGKA: A fast genetic $k$-means clustering algorithm. In *Proceedings of the* 2004 *ACM Symposium on Applied Computing*, pages 622–623. New York, NY, USA. ACM Press. (Cited on pp. 169, 177, 178, 179, 180)

Ma, E. and Chow, T. (2004). A new shifting grid clustering algorithm. *Pattern Recognition*, 37(3):503–514. (Cited on p. 198)

MacCuish, J. D. and MacCuish, N. E. (2010). *Clustering in Bioinformatics and Drug Discovery*. Boca Raton, FL, USA. CRC Press. (Cited on p. 14)

Macnaughton-Smith, P., Williams, W., Dale, M., and Mockett, L. (1964). Dissimilarity analysis: A new technique of hierarchical sub-division. *Nature*, 202:1034–1035. (Cited on pp. 129, 135)

Macqueen, J. (1967). Some methods for classification and analysis of multivariate observations. In LeCam, L. and Neyman, J., editors, *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, Volume 1, pages 281–297, Berkeley, CA, USA. University of California Press. (Cited on pp. 30, 136, 147, 150, 162, 208)

Maharaj, E. (2000). Clusters of time series. *Journal of Classification*, 17(2):297–314. (Cited on pp. 268, 269)

Malerba, D., Esposito, F., Gioviale, V., and Tamma, V. (2001). Comparing dissimilarity measures for symbolic data analysis. In *Proceedings of the Joint Conferences on "New Techniques and Technologies for Statistics" and "Exchange of Technology and Know-how"(ETK-NTTS'*01*)*, pages 473–481, Luxembourg. Office for Official Publications of the European Communities. (Cited on pp. 23, 24)

Mamun, A.-A., Aseltine, R., and Rajasekaran, S. (2016). Efficient record linkage algorithms using complete linkage clustering. *PLOS ONE*, 11:1–21. (Cited on p. 261)

Manku, G., Rajagopalan, S., and Lindsay, B. (1998). Approximate medians and other quantiles in one pass and with limited memory. In *SIGMOD* '98: *Proceedings of the* 1998 *ACM SIGMOD International Conference on Management of Data*, pages 426–435, New York, NY, USA. ACM Press. (Cited on p. 269)

Manku, G., Rajagopalan, S., and Lindsay, B. (1999). Random sampling techniques for space efficient online computation of order statistics of large datasets. In *SIGMOD* '99: *Proceedings of the* 1999 *ACM SIGMOD International Conference on Management of Data*, pages 251–262, New York, NY, USA. ACM Press. (Cited on p. 269)

Mántaras, R. (1991). A distance-based attribute selection measure for decision tree induction. In *Machine Learning*, Volume 6, pages 81–92, Boston. Kluwer Academic. (Cited on pp. 10, 99, 100, 288, 289)

Mao, J. and Jain, A. (1996). A self-organizing network for hyperellipsoidal clustering (hec). *IEEE Transactions on Neural Networks*, 7(1):16–29. (Cited on p. 69)

Marriott, F. (1971). Practical problems in a method of cluster analysis. *Biometrics*, 27(3):501–514. (Cited on p. 36)

Marshall, A. and Hodgson, J. (1998). DNA chips: An array of possibilities. *Nature Biotechnology*, 16:27–31. (Cited on p. 329)

Martinez, W. and Martinez, A. (2005). *Exploratory data analysis with MATLAB*. Computer Science and Data Analysis. Boca Raton, FL, USA. Chapman & Hall/CRC. (Cited on pp. 207, 208, 212, 216, 218)

Matoušek, J. (2000). On approximate geometric $k$-clustering. *Discrete and Computational Geometry*, 24(1):61–84. (Cited on p. 153)

Maulik, U. and Bandyopadhyay, S. (2000). Genetic algorithm-based clustering technique. *Pattern Recognition*, 33(9):1455–1465. (Cited on p. 168)

Maulik, U. and Bandyopadhyay, S. (2002). Performance evaluation of some clustering algorithms and validity indices. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(12):1650–1654. (Cited on p. 286)

Maulik, U., Bandyopadhyay, S., and Mukhopadhyay, A. (2011). *Multiobjective Genetic Algorithms for Clustering: Applications in Data Mining and Bioinformatics*. New York, NY, USA. Springer-Verlag. (Cited on p. 14)

McCallum, A., Nigam, K., and Ungar, L. H. (2000). Efficient clustering of high-dimensional data sets with application to reference matching. In *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD* '00, pages 169–178, New York, NY, USA. ACM Press. (Cited on p. 262)

McErlean, F., Bell, D., and McClean, S. (1990). The use of simulated annealing for clustering data in databases. *Information Systems*, 15(2):233–245. (Cited on p. 184)

McLachlan, G. and Basford, K. (1988). *Mixture Models: Inference and Applications to Clustering*, volume 84 of *STATISTICS: Textbooks and Monographs*. New York, NY, USA. Marcel Dekker. (Cited on pp. 207, 216)

McLachlan, G., Bean, R., and Peel, D. (2002). A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics*, 18(3):413–422. (Cited on p. 220)

McLachlan, G. and Krishnan, T. (1997). *The EM Algorithm and Extensions*. New York, NY, USA. Wiley & Sons. (Cited on p. 216)

McLachlan, G. and Peel, D. (2000). *Finite Mixture Models*. New York, NY, USA. Wiley & Sons. (Cited on p. 207)

McMorris, F., Meronk, D., and Neumann, D. (1983). A view of some consensus methods for trees. In Felsenstein, J., editor, *Numerical Taxonomy*, pages 122–126, Berlin, Germany. Springer-Verlag. (Cited on p. 104)

McNicholas, P. D. (2016). Model-based clustering. *Journal of Classification*, 33:331–373. (Cited on p. 12)

McQuitty, L. (1957). Elementary linkage analysis for isolating orthogonal and oblique types and typal relevancies. *Educational and Psychological Measurement*, 17:207–222. (Cited on p. 110)

McQuitty, L. (1960). Hierarchical linkage analysis for the isolation of types. *Educational and Psychological Measurement*, 20:55–67. (Cited on p. 93)

McQuitty, L. (1966). Similarity analysis by reciprocal pairs for discrete and continuous data. *Educational and Psychological Measurement*, 26:825–831. (Cited on p. 93)

McQuitty, L. (1967). Expansion of similarity analysis by reciprocal pairs for discrete and continuous data. *Educational and Psychological Measurement*, 27:253–255. (Cited on p. 93)

Meng, X. and van Dyk, D. (1997). The EM algorithm–an old folk-song sung to a fast new tune. *Journal of the Royal Statistical Society. Series B (Methodological)*, 59(3):511–567. (Cited on p. 214)

Mettu, R. and Plaxton, C. (2004). Optimal time bounds for approximate clustering. *Machine Learning*, 56(1-3):35–60. (Cited on p. 270)

Michalewicz, Z. (1992). *Genetic algorithms + data structures = evolution programs*. Berlin, Germany. Springer-Verlag. (Cited on p. 168)

Michaud, P. (1997). Clustering techniques. *Future Generation Computer Systems*, 13(2-3):135–147. (Cited on pp. 6, 164)

Miller, J. and Wegman, E. (1991). Construction of line densities for parallel coordinate plots. In *Computing and Graphics in Statistics*, pages 107–123, New York, NY, USA. Springer-Verlag. (Cited on p. 58)

Milligan, G. (1979). Ultrametric hierarchical clustering algorithms. *Psychometrika*, 44:343–346. (Cited on p. 110)

Milligan, G. (1989). A study of the beta-flexible clustering method. *Multivariate Behavioral Research*, 24:163–176. (Cited on p. 92)

Milligan, G. and Cooper, M. (1988). A study of standardization of variables in cluster analysis. *Journal of Classification*, 5:181–204. (Cited on pp. 41, 43, 44)

Min, E., Guo, X., Liu, Q., Zhang, G., Cui, J., and Long, J. (2018). A survey of clustering with deep learning: From the perspective of network architecture. *IEEE Access*, 6:39501–39514. (Cited on p. 12)

Mirkin, B. (1996). *Mathematical Classification and Clustering*. New York, NY, USA. Springer-Verlag. (Cited on p. 13)

Mirkin, B. (2005). *Clustering for Data Mining: A Data Recovery Approach*. Computer Science and Data Analysis Series. Boca Raton, FL, USA. Chapman & Hall/CRC. (Cited on p. 13)

Miyamoto, S., Ichihashi, H., and Honda, K. (2008). *Algorithms for Fuzzy Clustering: Methods in c-Means Clustering with Applications*. New York, NY, USA. Springer-Verlag. (Cited on p. 13)

Mladenović, N. and Hansen, P. (1997). Variable neighborhood search. *Computers and Operations Research*, 24(11):1097–1100. (Cited on p. 170)

Moore, A. (1990). *Efficient Memory-based Learning for Robot Control*. PhD thesis, Computer Laboratory, University of Cambridge, Cambridge, UK. (Cited on p. 351)

Moore, A. (1999). Very fast EM-based mixture model clustering using multiresolution $kd$-trees. In Kearns, M. and Cohn, D., editors, *Proceedings of the* 1998 *Conference on Advances in Neural Information Processing Systems* II, pages 543–549. Cambridge, MA, USA. MIT Press. (Cited on p. 352)

Morgan, B. (1981). Three applications of methods of cluster-analysis. *The Statistician*, 30(3):205–223. (Cited on p. 9)

Morrison, D. (1967). Measurement problems in cluster analysis. *Management Science (Series B, Managerial)*, 13(12):B775–B780. (Cited on p. 70)

Motwani, R. and Raghavan, P. (1995). *Randomized Algorithms*. New York, NY, USA. Cambridge University Press. (Cited on pp. 134, 232)

Mukhopadhyay, A., Maulik, U., and Bandyopadhyay, S. (2015). A survey of multiobjective evolutionary clustering. *ACM Computing Surveys (CSUR)*, 47(4):61. (Cited on p. 11)

Munro, J. and Paterson, M. (1980). Selection and sorting with limited storage. *Theoretical Computer Science*, 12(3):315–323. (Cited on p. 269)

Murtagh, F. (1983). A survey of recent advances in hierarchical clustering algorithms. *The Computer Journal*, 26(4):354–359. (Cited on pp. 9, 11, 17, 109, 129, 133, 138)

Murtagh, F. (1984a). Complexities of hierarchic clustering algorithms: State of the art. *Computational Statistics Quarterly*, 1(2):101–113. (Cited on p. 138)

Murtagh, F. (1984b). Counting dendrograms: A survey. *Discrete Applied Mathematics*, 7(2):191–199. (Cited on pp. 11, 104)

Murtagh, F. and Raftery, A. (1984). Fitting straight lines to points patterns. *Pattern Recognition*, 17:479–483. (Cited on p. 208)

Murthy, C. and Chowdhury, N. (1996). In search of optimal clusters using genetic algorithms. *Pattern Recognition Letters*, 17(8):825–832. (Cited on p. 168)

Nagesh, H., Goil, S., and Choudhary, A. (2000). A scalable parallel subspace clustering algorithm for massive data sets. In 2000 *International Conference on Parallel Processing (ICPP'*00), pages 477–486, Washington, DC, USA. IEEE Computer Society. (Cited on p. 234)

Nagesh, H., Goil, S., and Choudhary, A. (2001). Adaptive grids for clustering massive data sets. In *First SIAM International Conference on Data Mining*, Chicago, IL. Philadelphia, PA, USA. Society for Industrial and Applied Mathematics. (Cited on p. 198)

Naouali, S., Salem, S. B., and Chtourou, Z. (2020). Clustering categorical data: A survey. *International Journal of Information Technology & Decision Making*, 19(1):49–96. (Cited on p. 12)

Narahashi, M. and Suzuki, E. (2002). Subspace clustering based on compressibility. *Lecture Notes in Computer Science*, Volume 2534, pages 435–440, Berlin, Germany. Springer-Verlag. (Cited on p. 259)

Nascimento, M. C. and de Carvalho, A. C. (2011). Spectral methods for graph clustering – a survey. *European Journal of Operational Research*, 211(2):221–231. (Cited on p. 11)

Ng, A., Jordan, M., and Weiss, Y. (2002). On spectral clustering: Analysis and an algorithm. In Dietterich, T., Becker, S., and Ghahramani, Z., editors, *Advances in Neural Information Processing Systems*, volume 14, Cambridge, MA, USA. MIT Press. (Cited on p. 189)

Ng, M. and Wong, J. (2002). Clustering categorical data sets using tabu search techniques. *Pattern Recognition*, 35(12):2783–2790. (Cited on pp. 167, 172, 173)

Ng, R. and Han, J. (1994). Efficient and effective clustering methods for spatial data mining. In Bocca, J., Jarke, M., and Zaniolo, C., editors, 20*th International Conference on Very Large Data Bases, September* 12–15, 1994, *Santiago, Chile Proceedings*, pages 144–155, Los Altos, CA, USA. Morgan Kaufmann. (Cited on p. 30)

Ng, R. T. and Han, J. (2002). CLARANS: A method for clustering objects for spatial data mining. *IEEE Transactions on Knowledge and Data Engineering*, 14(5):1003–1016. (Cited on p. 261)

Nievergelt, J., Hinterberger, H., and Sevcik, K. (1984). The grid file: An adaptable, symmetric multikey file structure. *ACM Transactions on Database Systems*, 9(1):38–71. (Cited on p. 194)

Novikov, A. (2019). PyClustering: Data mining library. *Journal of Open Source Software*, 4(36):1230. (Cited on p. 308)

Oates, T., Firoiu, L., and Cohen, P. (2001). Using dynamic time warping to bootstrap HMM-based clustering of time series. In *Sequence Learning - Paradigms, Algorithms, and Applications*, volume 1828 of *Lecture Notes in Computer Science*, pages 35–52, London, UK. Springer-Verlag. (Cited on p. 268)

O'Callaghan, L., Mishra, N., Meyerson, A., Guha, S., and Motwani, R. (2002). Streaming-data algorithms for high-quality clustering. In *ICDE* '02: *Proceedings of the* 18*th International Conference on Data Engineering*, pages 685–694, Washington, DC, USA. IEEE Computer Society. (Cited on p. 272)

Oktar, Y. and Turkan, M. (2018). A review of sparsity-based clustering methods. *Signal Processing*, 148:20–30. (Cited on p. 12)

Olson, C. F. (1995). Parallel algorithms for hierarchical clustering. *Parallel Computing*, 21(8):1313–1325. (Cited on p. 261)

Ordonez, C. (2003). Clustering binary data streams with $k$-means. In *DMKD* '03: *Proceedings of the* 8*th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, pages 12–19, New York, NY, USA. ACM Press. (Cited on p. 272)

Orlóci, L. (1967). An agglomerative method for classification of plant communities. *Journal of Ecology*, 55:193–205. (Cited on p. 71)

Overall, J. and Klett, C. (1972). *Applied multivariate analysis*. McGraw-Hill Series in Psychology. New York, NY, USA. McGraw-Hill. (Cited on p. 41)

Pal, N. and Biswas, J. (1997). Cluster validation using graph theoretic concepts. *Pattern Recognition*, 30(6):847–857. (Cited on pp. 145, 282, 283)

Pandove, D., Goel, S., and Rani, R. (2018). Systematic review of clustering high-dimensional and large datasets. *ACM Transactions on Knowledge Discovery from Data*, 12(2):16:1–16:68. (Cited on p. 12)

Park, N. and Lee, W. (2004). Statistical grid-based clustering over data streams. *SIGMOD Record*, 33(1):32–37. (Cited on p. 198)

Park, S. and Chu, D. L. W. (1999). Fast retrieval of similar subsequences in long sequence databases. In *(KDEX* '99*) Proceedings of the* 1999 *Workshop on Knowledge and Data Engineering Exchange,* 1999, pages 60–67, Chicago, IL. Los Alamitos, CA, USA. IEEE Computer Society. (Cited on p. 87)

Park, S., Chu, W., Yoon, J., and Hsu, C. (2000). Efficient searches for similar subsequences of different lengths in sequence databases. In *ICDE* '00: *Proceedings of the* 16*th International Conference on Data Engineering,* 2000, pages 23–32, San Diego, CA, USA. IEEE Computer Society. (Cited on p. 87)

Park, S., Kim, S., and Chu, W. (2001). Segment-based approach for subsequence searches in sequence databases. In *SAC* '01: *Proceedings of the* 2001 *ACM Symposium on Applied Computing*, pages 248–252, New York, NY, USA. ACM Press. (Cited on p. 87)

Parsons, L., Haque, E., and Liu, H. (2004a). Evaluating subspace clustering algorithms. In *Workshop on Clustering High Dimensional Data and its Applications, SIAM International Conference on Data Mining (SDM* 2004*)*, pages 48–56, Philadelphia, PA, USA. Society for Industrial and Applied Mathematics. (Cited on p. 259)

Parsons, L., Haque, E., and Liu, H. (2004b). Subspace clustering for high dimensional data: a review. *SIGKDD, Newsletter of the ACM Special Interest Group on Knowledge Discovery and Data Mining*, 6(1):90–105. (Cited on pp. 11, 258)

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Édouard Duchesnay (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830. (Cited on p. 308)

Pedrycz, W. (2005). *Knowledge-Based Clustering: From Data to Information Granules*. Hoboken, NJ, USA. Wiley-Interscience. (Cited on p. 13)

Pei, J. and Yang, X. (2000). Study of clustering validity based on fuzzy similarity. In *Proceedings of the* 3*rd World Congress on Intelligent Control and Automation,* 2000, Volume 4, Hefei, China, pages 2444–2447, Piscataway, NJ, USA. IEEE. (Cited on p. 292)

Pelleg, D. and Moore, A. (1999). Accelerating exact $k$-means algorithms with geometric reasoning. In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 277–281, San Diego, CA, USA. ACM Press. (Cited on pp. 152, 153, 156, 241, 351, 352)

Pelleg, D. and Moore, A. (2000). $x$-means: Extending $k$-means with efficient estimation of the number of clusters. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pages 727–734, San Francisco, CA, USA. Morgan Kaufmann. (Cited on p. 155)

Peña, J., Lozano, J., and Larrańaga, P. (1999). An empirical comparison of four initialization methods for the $k$-means algorithm. *Pattern Recognition Letters*, 20(10):1027–1040. (Cited on pp. 149, 150)

Perng, C., Wang, H., Zhang, S., and Parker, D. (2000). Landmarks: A new model for similarity-based pattern querying in time series databases. In *ICDE* '00: *Proceedings of the* 16*th International Conference on Data Engineering*, pages 33–42, Washington, DC, USA. IEEE Computer Society. (Cited on pp. 86, 87)

Phillips, S. (2002). Acceleration of $k$-means and related clustering algorithms. In Mount, D. and Stein, C., editors, *ALENEX: International Workshop on Algorithm Engineering and Experimentation*, San Francisco, CA, Volume 2409 of Lecture Notes in Computer Science, pages 166–177, Heidelberg, Germany. Springer-Verlag. (Cited on pp. 149, 151, 152)

Podani, J. (1989). New combinatorial clustering methods. *Vegetatio*, 81:61–77. (Cited on pp. 93, 94, 128)

Pollard, D. (1981). Strong consistency of $k$-means clustering. *The Annals of Statistics*, 9(1):135–140. (Cited on p. 149)

Pollard, D. (1982). A central limit theorem for $k$-means clustering. *The Annals of Probability*, 10(4):919–926. (Cited on p. 149)

Pölzlbauer, G., Dittenbach, M., and Rauber, A. (2006). Advanced visualization of self-organizing maps with vector fields. *Neural Networks*, 19(6-7):911–922. (Cited on p. 62)

Popivanov, I. and Miller, R. (2002). Similarity search over time-series data using wavelets. In *ICDE* '02: *Proceedings of the* 18*th International Conference on Data Engineering (ICDE*'02*)*, page 212, Washington, DC, USA. IEEE Computer Society. (Cited on p. 86)

Posse, C. (2001). Hierarchical model-based clustering for large datasets. *Journal of Computational & Graphical Statistics*, 10(3):464–486. (Cited on p. 138)

Pratt, K. and Fink, E. (2002). Search for patterns in compressed time series. *International Journal of Image and Graphics*, 2(1):89–106. (Cited on p. 87)

Preparata, F. and Shamos, M. (1985). *Computational Geometry: An Introduction*. New York, NY, USA. Springer-Verlag. (Cited on p. 351)

Pries, K. H. and Dunnigan, R. (2015). *Big Data Analytics: A Practical Guide for Managers*. Boca Raton, FL, USA. CRC Press. (Cited on p. 19)

Prim, R. (1957). Shortest connection matrix network and some generalizations. *Bell System Technical Journal*, 36:1389–1401. (Cited on p. 131)

Procopiuc, C., Jones, M., Agarwal, P., and Murali, T. (2002). A Monte Carlo algorithm for fast projective clustering. In *Proceedings of the* 2002 *ACM SIGMOD International Conference on Management of Data*, pages 418–427, New York, NY, USA. ACM Press. (Cited on pp. 222, 235, 236)

Qu, Y., Wang, C., and Wang, X. (1998). Supporting fast search in time series for movement patterns in multiple scales. In *CIKM* '98: *Proceedings of the Seventh International Conference on Information and Knowledge Management*, pages 251–258, New York, NY, USA. ACM Press. (Cited on p. 87)

Quinlan, J. (1993). *C4.5: Program for machine learning*. San Mateo, CA, USA. Morgan Kaufmann. (Cited on p. 237)

Raff, E. (2017). JSAT: Java statistical analysis tool, a library for machine learning. *Journal of Machine Learning Research*, 18(23):1–5. (Cited on p. 308)

Rafiei, D. and Mendelzon, A. (1997). Similarity-based queries for time series data. In *SIGMOD* '97: *Proceedings of the* 1997 *ACM SIGMOD International Conference on Management of Data*, pages 13–25. New York, NY, USA. ACM Press. (Cited on pp. 80, 82)

Rafiei, D. and Mendelzon, A. (1998). Efficient retrieval of similar time sequences using DFT. In *FODO* '98: *Proceedings of* 5*th International Conference on Foundations of Data Organizations and Algorithms,* 1998, Kobe, Japan. Boston, MA, USA. Kluwer Academic. (Cited on p. 86)

Rajasekaran, S. (2005). Efficient parallel hierarchical clustering algorithms. *IEEE Transactions on Parallel and Distributed Systems*, 16(6):497–502. (Cited on p. 261)

Ramoni, M., Sebastiani, P., and Cohen, P. (2002). Bayesian clustering by dynamics. *Machine Learning*, 47(1):91–121. (Cited on pp. 80, 268)

Ramsay, G. (1998). DNA chips: State-of-the art. *Nature Biotechnology*, 16:40–44. (Cited on p. 329)

Rand, W. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850. (Cited on pp. 9, 287)

Ranka, S. and Sahni, S. (1991). Clustering on a hypercube multicomputer. *IEEE Transactions on Parallel and Distributed Systems*, 2(2):129–137. (Cited on p. 261)

Ray, S. and Turi, R. (1999). Determination of number of clusters in $k$-means clustering and application in colour image segmentation. In Pal, N., De, A., and Das, J., editors, *Proceedings of the* 4*th International Conference on Advances in Pattern Recognition and Digital Techniques (ICAPRDT'*99*)*, pages 137–143, Calcutta, India. Narosa Publishing House. (Cited on pp. 35, 36)

Reymond, P., Weber, H., Damond, M., and Farmer, E. (2000). Differential gene expression in response to mechanical wounding and insect feeding in arabidopsis. *The Plant Cell*, 12(5):707–720. (Cited on p. 333)

Rhee, H. and Oh, K. (1996). A performance measure for the fuzzy cluster validity. *IEEE "Soft Computing in Intelligent Systems and Information Processing." Proceedings of the* 1996 *Asian Fuzzy Systems Symposium*, pages 364–369, Piscataway, NJ, USA. IEEE. (Cited on p. 294)

Rogers, D. and Tanimoto, T. (1960). A computer program for classifying plants. *Science*, 132:1115–1118. (Cited on p. 73)

Rohlf, F. (1970). Adaptive hierarchical clustering schemes. *Systematic Zoology*, 19(1):58–82. (Cited on p. 138)

Rohlf, F. (1973). Algorithm 76: Hierarchical clustering using the minimum spanning tree. *The Computer Journal*, 16(1):93–95. (Cited on p. 132)

Rohlf, F. (1974). Algorithm 81: Dendrogram plot. *The Computer Journal*, 17(1):89–91. (Cited on p. 106)

Rohlf, F. (1982). Single link clustering algorithms. In Krishnaiah, P. and Kanal, L., editors, *Handbook of Statistics*, Volume 2, pages 267–284, Amsterdam, The Netherlands. North-Holland. (Cited on p. 111)

Rose, K., Gurewitz, E., and Fox, G. C. (1990). Statistical mechanics and phase transitions in clustering. *Physical Review Letters*, 65:945–948. (Cited on pp. 158, 159, 249, 250)

Ross, G. (1969). Algorithm AS 15: Single linkage cluster analysis. *Applied Statistics*, 18(1):106–110. (Cited on p. 106)

Rousseeuw, P. (1986). A visual display for hierarchical classification. In Diday, E., Escoufier, Y., Lebart, L., Pagès, J., and Tomassone, Y. S. R., editors, *Data analysis and Informatics* 4, pages 743–748, Amsterdam, The Netherlands. North-Holland. (Cited on p. 106)

Rousseeuw, P. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65. (Cited on p. 109)

Rummel, R. (1970). *Applied Factor Analysis*. Evanston, IL, USA. Northwestern University Press. (Cited on p. 67)

Ruspini, E. (1969). A new approach to clustering. *Information and Control*, 15:22–32. (Cited on p. 139)

Salton, G. and McGill, M. (1983). *Introduction to Modern Information Retrieval*. New York, NY, USA. McGraw-Hill. (Cited on p. 88)

Sammon, J. (1969). A nonlinear mapping for data structure analysis. *IEEE Transactions on Computing*, C18:401–409. (Cited on pp. 51, 52)

Sander, J., Ester, M., Kriegel, H., and Xu, X. (1998). Density-based clustering in spatial databases: The algorithm gdbscan and its applications. *Data Mining and Knowledge Discovery*, 2(2):169–194. (Cited on p. 200)

Sarafis, I., Trinder, P., and Zalzala, A. (2003). Towards effective subspace clustering with an evolutionary algorithm. In 2003. *CEC* '03. *The* 2003 *Congress on Evolutionary Computation*, Volume 2, pages 797–806, Piscataway, NJ, USA. IEEE. (Cited on p. 259)

SAS Institute Inc. (1983). SAS Technical report A-108, Cubic Clustering Criterion. SAS Institute. (Cited on p. 37)

Sato-Ilic, M. (2006). *Innovations in Fuzzy Clustering: Theory and Applications*. New York, NY. Springer-Verlag. (Cited on p. 13)

Saunders, J. (1980). Cluster analysis for market segmentation. *European Journal of Marketing*, 14(7):422–435. (Cited on p. 4)

Saxena, A., Prasad, M., Gupta, A., Bharill, N., Patel, O. P., Tiwari, A., Er, M. J., Ding, W., and Lin, C.-T. (2017). A review of clustering techniques and developments. *Neurocomputing*, 267:664–681. (Cited on p. 12)

Scheibler, D. and Schneider, W. (1985). Monte Carlo test of the accuracy of cluster analysis algorithms–a comparison of hierarchical and nonhierarchical methods. *Multivariate Behavioral Research*, 20:283–304. (Cited on p. 92)

Schelling, B. and Plant, C. (2019). Dataset-transformation: Improving clustering by enhancing the structure with DipScaling and DipTransformation. *Knowledge and Information Systems*, 62:457–484. (Cited on p. 49)

Schena, M., Shalon, D., Davis, R., and Brown, P. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270(5235):467–470. (Cited on p. 329)

Schikuta, E. (1996). Grid-clustering: A efficient hierarchical clustering method for very large data sets. In *Proceedings of the* 13*th International Conference on Pattern Recognition,* 1996, Volume 2, pages 101–105, Vienna Austria. Los Alamitos, CA, USA. IEEE Computer Society. (Cited on pp. 194, 195)

Schikuta, E. and Erhart, M. (1997). The BANG-clustering system: Grid-based data analysis. In Liu, X., Cohen, P., and Berthold, M., editors, *Lecture Notes in Computer Science*, Volume 1280, pages 513–524, Berlin, Germany. Springer-Verlag. (Cited on p. 195)

Schuchhardt, J., Beule, D., Malik, A., Wolski, E., Eickhoff, H., Lehrach, H., and Herzel, H. (2000). Normalization strategies for cDNA microarrays. *Nucleic Acids Research*, 28(10):e47–e47. (Cited on p. 330)

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464. (Cited on pp. 155, 208)

Scott, A. and Symons, M. (1971a). 297. Note: On the Edwards and Cavalli-Sforza method of cluster analysis. *Biometrics*, 27(1):217–219. (Cited on pp. 128, 138)

Scott, A. and Symons, M. (1971b). Clustering methods based on likelihood ratio criteria. *Biometrics*, 27(2):387–397. (Cited on pp. 36, 207)

Sedgewick, R. (1978). Implementing quicksort programs. *Communications of the ACM*, 21(10):847–857. (Cited on p. 32)

Selim, S. and Al-Sultan, K. (1991). A simulated annealing algorithm for the clustering problem. *Pattern Recognition*, 24:1003–1008. (Cited on p. 184)

Selim, S. and Ismail, M. (1984). $k$-means-type algorithms: A generalized convergence theorem and characterization of local optimality. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(1):81–87. (Cited on pp. 148, 149, 167)

Selim, S. and Ismail, M. (1986). Fuzzy $c$-means: Optimality of solutions and effective termination of the algorithm. *Pattern Recognition*, 19(6):651–663. (Cited on p. 167)

Serinko, R. and Babu, G. (1992). Weak limit theorems for univariate $k$-means clustering under a nonregular condition. *Journal of Multivariate Analysis*, 41:273–296. (Cited on p. 149)

Shadbolt, J. and Taylor, J., editors (2002). *Neural Networks and the Financial Markets*. Springer. (Cited on p. 24)

Shafer, J., Agrawal, R., and Mehta, M. (1996). SPRINT: A scalable parallel classifier for data mining. In Vijayaraman, T. M., Buchmann, A. P., Mohan, C., and L.Sarda, N., editors, *VLDB*'96*, Proceedings of the* 22*nd International Conference on Very Large Data Bases*, Mumbai (Bombay), India, pages 544–555. San Mateo, CA, USA. Morgan Kaufmann. (Cited on p. 237)

Sharan, R. and Shamir, R. (2000). CLICK: A clustering algorithm with applications to gene expression analysis. In *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology (ISMB)*, pages 307–316. Menlo Park, CA. AAAI Press. (Cited on p. 332)

Sharma, S. (1996). *Applied Multivariate Techniques*. New York, NY, USA. John Wiley & Sons. (Cited on pp. 285, 286)

Sheikholeslami, G., Chatterjee, S., and Zhang, A. (2000). WaveCluster: A wavelet-based clustering approach for spatial data in very large databases. *The VLDB Journal*, 8(3-4):289–304. (Cited on pp. 197, 198)

Shepard, R. (1962). The analysis of proximities: Multidimensional scaling with an unknown distance function I. *Psychometrica*, 27:125–140. (Cited on p. 53)

Shirkhorshidi, A., Aghabozorgi, S., Wah, T., and Herawan, T. (2014). Big data clustering: A review. In *Computational Science and Its Applications - ICCSA* 2014, volume 8583 of *Lecture Notes in Computer Science*, pages 707–720, Cham, Switzerland. Springer-Verlag. (Cited on p. 261)

Shneiderman, B. (1992). Tree visualization with tree-maps: 2-d space-filling approach. *ACM Transactions on Graphics*, 11(1):92–99. (Cited on p. 58)

Sibson, R. (1973). SLINK: An optimally efficient algorithm for the single link cluster method. *The Computer Journal*, 16(1):30–34. (Cited on pp. 105, 106, 107, 129, 130, 132, 275)

Silva, J. A., Faria, E. R., Barros, R. C., Hruschka, E. R., de Carvalho, A. C. P. L. F., and Gama, J. (2013). Data stream clustering: A survey. *ACM Computing Surveys*, 46(1):13:1–13:31. (Cited on p. 11)

Sim, K., Gopalkrishnan, V., Zimek, A., and Cong, G. (2013). A survey on enhanced subspace clustering. *Data Mining and Knowledge Discovery*, 26(2):332–397. (Cited on p. 11)

Sneath, P. (1957). The applications of computers to taxonomy. *Journal of General Microbiology*, 17:201–226. (Cited on pp. 93, 110)

Sneath, P. (1967). Some statistical problems in numerical taxonomy. *The Statistician*, 17(1):1–12. (Cited on p. 290)

Sneath, P. (1969). Evaluation of clustering methods (with discussion). In Cole, A., editor, *Numerical Taxonomy*, pages 257–271, London, UK. Academic Press. (Cited on p. 295)

Sokal, R. and Rohlf, F. (1962). The comparison of dendrograms by objective methods. *Taxon*, 11:33–40. (Cited on p. 106)

Sokal, R. and Sneath, P. (1963). *Principles of Numerical Taxonomy*. San Francisco, CA, USA. W.H. Freeman. (Cited on pp. 12, 74, 100)

Sokal, R. and Sneath, P. (1973). *Numerical taxonomy: the principles and practice of numerical classification*. San Francisco, CA, USA. W.H. Freeman. (Cited on pp. 5, 12, 100, 109)

Somorjai, R., Dolenko, B., Demko, A., Mandelzweig, M., Baumgartner, A. N. R., and Pizzi, N. (2004). Mapping high-dimensional data onto a relative distance plane–an exact method for visualizing and characterizing high-dimensional patterns. *Journal of Biomedical Informatics*, 37(5):366–379. (Cited on p. 62)

Späth, H. (1980). *Cluster Analysis Algorithms*. West Sussex, England. Ellis Horwood Limited. (Cited on pp. 129, 136, 137)

Spielmat, D. and Teng, S. (1996). Spectral partitioning works: Planar graphs and finite element meshes. In 37*th Annual Symposium on Foundations of Computer Science,* 1996. *Proceedings*, pages 96–105, Los Alamitos, CA, USA. IEEE Computer Society. (Cited on p. 189)

Sprenger, T., Brunella, R., and Gross, M. (2000). H-BLOB: A hierarchical visual clustering method using implicit surfaces. In *VIS* '00: *Proceedings of the Conference on Visualization* '00, pages 61–68, Los Alamitos, CA, USA. IEEE Computer Society. (Cited on p. 62)

Srivastava, A. N. and Sahami, M., editors (2009). *Text Mining: Classification, Clustering, and Applications*. Boca Raton, FL, USA. CRC Press. (Cited on p. 14)

StatSoft, Inc. (2005). Time series analysis. `http://www.statsoftinc.com/textbook/sttimser.html`. (Cited on p. 79)

Stute, W. and Zhu, L. (1995). Asymptotics of $k$-means clustering based on projection pursuit. *The Indian Journal Of Statistics*, 57(3):462–471. (Cited on p. 154)

Su, M. and Chou, C. (2001). A modified version of the $k$-means algorithm with a distance based on cluster symmetry. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6):674–680. (Cited on pp. 153, 154)

Su, Z., Yang, Q., Zhang, H., Xu, X., Hu, Y., and Ma, S. (2002). Correlation-based web document clustering for adaptive web interface design. *Knowledge and Information Systems*, 4(2):151–167. (Cited on p. 200)

Sung, C. and Jin, H. (2000). A tabu-search-based heuristic for clustering. *Pattern Recognition*, 33(5):849–858. (Cited on p. 169)

Tamura, S., Higuchi, S., and Tanaka, K. (1971). Pattern classification based on fuzzy relations. *IEEE Transactions on Systems, Man and Cybernetics*, 1(1):61–66. (Cited on p. 142)

Tang, C. and Zhang, A. (2002). An iterative strategy for pattern discovery in high-dimensional data sets. In *Proceedings of the Eleventh International Conference on Information and Knowledge Management*, pages 10–17, New York, NY, USA. ACM Press. (Cited on p. 332)

Tang, C., Zhang, L., Ramanathan, M., and Zhang, A. (2001). Interrelated two-way clustering: An unsupervised approach for gene expression data analysis. In *Proceedings of the* 2*nd IEEE International Symposium on Bioinformatics and Bioengineering*, pages 41–48, Washington, DC, USA. IEEE Computer Society. (Cited on p. 332)

Tango, T. (2010). *Statistical Methods for Disease Clustering*. New York, NY, USA. Springer-Verlag. (Cited on p. 14)

Tarsitano, A. (2003). A computational study of several relocation methods for $k$-means algorithms. *Pattern Recognition*, 36(12):2955–2966. (Cited on p. 150)

Tavazoie, S., Hughes, J., Campbell, M., Cho, R., and Church, G. (1999). Systematic determination of genetic network architecture. *Nature Genetics*, 22:281–285. (Cited on pp. 330, 333)

Thaper, N., Guha, S., Indyk, P., and Koudas, N. (2002). Dynamic multidimensional histograms. In *SIGMOD* '02: *Proceedings of the* 2002 *ACM SIGMOD International Conference on Management of Data*, pages 428–439, New York, NY, USA. ACM Press. (Cited on p. 269)

The Geneva Association (2013). Variable annuities - an analysis of financial stability. Available online at: `https://www.genevaassociation.org/sites/default/files/research-topics-document-type/pdf_public/ga2013-variable_annuities_0.pdf`. Accessed on Feb 17, 2020. (Cited on p. 348)

Theodoridis, S. and Koutroubas, K. (1999). *Pattern Recognition*. London, UK. Academic Press. (Cited on pp. 277, 281)

Thomasian, A., Castelli, V., and Li, C. (1998). Clustering and singular value decomposition for approximate indexing in high dimensional spaces. In *Proceedings of the Seventh International Conference on Information and Knowledge Management*, pages 201–207. New York, NY, USA. ACM Press. (Cited on p. 46)

Thorup, M. (2001). Quick $k$-median, $k$-center, and facility location for sparse graphs. In *ICALP* '01: *Proceedings of the* 28*th International Colloquium on Automata, Languages and Programming*, volume 2076 of *Lecture Notes In Computer Science*, pages 249–260, London, UK. Springer-Verlag. (Cited on p. 270)

Tibshirani, R., Walther, G., and Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society Series B (Methodological)*, 63(2):411–423. (Cited on p. 37)

Trauwaert, E. (1988). On the meaning of Dunn's partition coefficient for fuzzy clusters. *Fuzzy Sets and Systems*, 25(2):217–242. (Cited on p. 291)

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Tibshirani, T. H. R., Botstein, D., and Altman, R. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17(6):520–525. (Cited on p. 330)

Tsay, R. (2002). *Analysis of Financial Time Series*. Wiley Series in Probability and Statistics. New York, NY, USA. John Wiley & Sons. (Cited on p. 24)

Tseng, L. and Yang, S. (2000). A genetic clustering algorithm for data with non-spherical-shape clusters. *Pattern Recognition*, 33(7):1251–1259. (Cited on p. 168)

Tseng, L. and Yang, S. (2001). A genetic approach to the automatic clustering problem. *Pattern Recognition*, 34(2):415–424. (Cited on p. 168)

Tsumoto, S. (1999). Rule discovery in large time-series medical databases. In *PKDD* '99: *Proceedings of the Third European Conference on Principles of Data Mining and Knowledge Discovery*, pages 23–31, London, UK. Springer-Verlag. (Cited on pp. 80, 267)

Tubbs, J. (1989). A note on binary template matching. *Pattern Recognition*, 22(4):359–365. (Cited on p. 74)

van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605. (Cited on p. 59)

van Groenewoud, H. and Ihm, P. (1974). A cluster analysis based on graph theory. *Vegetatio*, 29:115–120. (Not cited)

van Rijsbergen, C. (1970). Algorithm 52: A fast hierarchic clustering algorithm. *The Computer Journal*, 13(3):324–326. (Not cited)

Vesanto, J. (1999). SOM-based data visualization methods. *Intelligent Data Analysis*, 3(2):111–126. (Cited on p. 62)

Vesanto, J. (2000). Neural network tool for data mining: SOM toolbox. In *Proceedings of Symposium on Tool Environments and Development Methods for Intelligent Systems (TOOLMET*2000*)*, pages 184–196, Oulu, Finland. University of Oulun. (Cited on p. 56)

Vesanto, J., Alhoniemi, E., Himberg, J., Kiviluoto, K., and Parviainen, J. (1999a). Self-organizing map for data mining in Matlab: The SOM toolbox. *Simulation News Europe*, page 54. (Cited on p. 56)

Vesanto, J., Himberg, J., Alhoniemi, E., and Parhankangas, J. (1999b). Self-organizing map in Matlab: The SOM toolbox. In *Proceedings of the Matlab DSP Conference* 1999, pages 35–40, Espoo, Finland. (Cited on p. 56)

Vlachos, M., Hadjieleftheriou, M., Gunopulos, D., and Keogh, E. (2003). Indexing multidimensional time-series with support for multiple distance measures. In *KDD* '03: *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 216–225, New York, NY, USA. ACM Press. (Cited on p. 84)

Wang, C. and Wang, X. (2000). Supporting content-based searches on time series via approximation. In *SSDBM* '00: *Proceedings of the* 12*th International Conference on Scientific and Statistical Database Management (SSDBM*'00*)*, pages 69–81, Washington, DC, USA. IEEE Computer Society. (Cited on p. 87)

Wang, H., Chu, F., Fan, W., Yu, P., and Pei, J. (2004). A fast algorithm for subspace clustering by pattern similarity. In *Proceedings of the* 16*th International Conference on Scientific and Statistical Database Management,* 2004, pages 51–60, Los Alamitos, CA, USA. IEEE Computer Society. (Cited on p. 259)

Wang, H., Wang, W., Yang, J., and Yu, P. (2002). Clustering by pattern similarity in large data sets. In *Proceedings of the* 2002 *ACM SIGMOD International Conference on Management of Data*, pages 394–405, New York, NY, USA. ACM Press. (Cited on p. 332)

Wang, K., Xu, C., and Liu, B. (1999a). Clustering transactions using large items. In *Proceedings of the Eighth International Conference on Information and Knowledge Management*, pages 483–490. New York, NY, USA. ACM Press. (Cited on pp. 23, 273)

Wang, L. and Wang, Z. (2003). CUBN: A clustering algorithm based on density and distance. In *2003 International Conference on Machine Learning and Cybernetics*, pages 108–112, Piscataway, NJ, USA. IEEE. (Cited on p. 204)

Wang, W., Yang, J., and Muntz, R. (1997). STING: A statistical information grid approach to spatial data mining. In Jarke, M., Carey, M., Dittrich, K., Lochovsky, F., and Jeusfeld, P. L. M., editors, *Twenty-Third International Conference on Very Large Data Bases*, Athens, Greece, pages 186–195, San Francisco, CA, USA. Morgan Kaufmann. (Cited on pp. 191, 198)

Wang, W., Yang, J., and Muntz, R. (1999b). STING+: An approach to active spatial data mining. In *Fifteenth International Conference on Data Engineering*, Sydney, Australia, pages 116–125, Los Alamitos, CA, USA. IEEE Computer Society. (Cited on pp. 191, 192)

Ward, Jr., J. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244. (Cited on pp. 93, 123, 208)

Ward, Jr., J. and Hook, M. (1963). Application of an hierarchical grouping procedure to a problem of grouping profiles. *Educational and Psychological Measurement*, 23(1):69–81. (Cited on p. 123)

Wegman, E. (1990). Hyperdimensional data analysis using parallel coordinates. *Journal of the American Statistical Association*, 85(411):664–675. (Cited on p. 58)

Whittaker, R. (1952). A study of summer foliage insect communities in the great Smoky Mountains. *Ecological Monographs*, 22:1–44. (Cited on p. 72)

Wierzchoń, S. and Kłopotek, M. (2018). *Modern Algorithms of Cluster Analysis*. Cham, Switzerland. Springer-Verlag. (Cited on pp. 14, 17)

Wilks, S. (1962). *Mathematical Statistics*. New York, NY, USA. John Wiley and Sons. (Cited on p. 67)

Willett, P. (1988). Recent trends in hierarchical document clustering: A critical review. *Information Processing and Management*, 24(5):577–597. (Cited on pp. 128, 129, 138)

Williams, W. and Lambert, J. (1966). Multivariate methods in plant ecology: V. Similarity analyses and information-analysis. *Journal of Ecology*, 54(2):427–445. (Cited on p. 89)

Wills, G. (1998). An interactive view for hierarchical clustering. In *INFOVIS* '98: *Proceedings of the* 1998 *IEEE Symposium on Information Visualization*, pages 26–31, Washington, DC, USA. IEEE Computer Society. (Cited on pp. 58, 59)

Wirth, M., Estabrook, G., and Rogers, D. (1966). A graph theory model for systematic biology, with an example for the oncidiinae (orchidaceae). *Systematic Zoology*, 15(1):59–69. (Cited on pp. 109, 189)

Wishart, D. (1969). 256. Note: An algorithm for hierarchical classifications. *Biometrics*, 25(1):165–170. (Cited on pp. 93, 124)

Wishart, D. (1978). Treatment of missing values in cluster analysis. In *Compstat: Proceedings in Computational Statistics*, pages 281–297, Würzburg, Germany. Physica-Verlag. (Cited on pp. 11, 164)

Wishart, D. (2002). $k$-means clustering with outlier detection, mixed variables and missing values. In Schwaiger, M. and Opitz, O., editors, *Exploratory Data Analysis in Empirical Research*, pages 216–226. Cham, Switzerland. Springer-Verlag. (Cited on pp. 69, 76, 77, 163, 243)

Witten, I. H., Frank, E., and Hall, M. A. (2011). *Data Mining: Practical Machine Learning Tools and Techniques*. Third edition. New York, NY, USA. Morgan Kaufmann. (Cited on p. 4)

Wolfe, J. (1970). Pattern clustering by multivariate mixture analysis. *Multivariate Behavioral Research*, 5:329–350. (Cited on p. 207)

Woo, K. and Lee, J. (2002). *FINDIT: A fast and intelligent subspace clustering algorithm using dimension voting*. PhD thesis, Korea Advanced Institute of Science and Technology, Department of Electrical Engineering and Computer Science. (Cited on pp. 222, 232)

Wu, C. (1983). On the convergence properties of the EM algorithm. *The Annals of Statistics*, 11(1):95–103. (Cited on p. 215)

Wu, C. (1986). Jackknife, bootstrap and other resampling methods in regression analysis. *The Annals of Statistics*, 14(4):1261–1295. (Cited on p. 333)

Wu, L., Faloutsos, C., Sycara, K., and Payne, T. (2000a). FALCON: Feedback adaptive loop for content-based retrieval. In *VLDB* '00: *Proceedings of the* 26*th International Conference on Very Large Data Bases*, Cairo, Egypt, pages 297–306, Orlando, FL, USA. Morgan Kaufmann. (Cited on p. 81)

Wu, X. and Barbará, D. (2002). Learning missing values from summary constraints. *ACM SIGKDD Explorations Newsletter*, 4(1):21–30. (Cited on p. 11)

Wu, Y., Agrawal, D., and Abbadi, A. (2000b). A comparison of DFT and DWT based similarity search in time-series databases. In *CIKM* '00: *Proceedings of the Ninth International Conference on Information and Knowledge Management*, pages 488–495, New York, NY, USA. ACM Press. (Cited on p. 86)

Wu, Z. and Leahy, R. (1993). An optimal graph theoretic approach to data clustering: Theory and its application to image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(11):1101–1113. (Cited on p. 189)

Xiao, Y. and Dunham, M. (2001). Interactive clustering for transaction data. In Volume 2114 of *Lecture Notes in Computer Science*, pages 121–130, New York, NY, USA. Springer-Verlag. (Cited on pp. 23, 88, 273, 275)

Xie, X. and Beni, G. (1991). A validity measure for fuzzy clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(8):841–847. (Cited on pp. 139, 293)

Xiong, Y. and Yeung, D. (2002). Mixtures of ARMA models for model-based time series clustering. In *ICDM* '02: *Proceedings of the* 2002 *IEEE International Conference on Data Mining (ICDM'*02*)*, pages 717–720, Washington, DC, USA. IEEE Computer Society. (Cited on pp. 268, 269)

Xu, D. and Tian, Y. (2015). A comprehensive survey of clustering algorithms. *Annals of Data Science*, 2(2):165–193. (Cited on p. 11)

Xu, H., Wang, H., and Li, C. (2002). Fuzzy tabu search method for the clustering problem. In 2002 *International Conference on Machine Learning and Cybernetics,* 2002*, Proceedings*, Volume 2, Beijing. IEEE, pages 876–880, Piscataway, NJ, USA. IEEE. (Cited on p. 170)

Xu, J., Xiong, H., Sung, S., and Kumar, V. (2003). A new clustering algorithm for transaction data via caucus. In *Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining,* 7*th PAKDD* 2003, volume 2637 of *Lecture Notes in Computer Science*, pages 551–562. Berlin, Germany. Springer-Verlag. (Cited on pp. 272, 276)

Xu, R. and Wunsch, D. C. (2010). Clustering algorithms in biomedical research: A review. *IEEE Reviews in Biomedical Engineering*, 3:120–154. (Cited on p. 11)

Xu, R. and Wunsch II, D. (2005). Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, 16(3):645–678. (Cited on p. 11)

Xu, X., Ester, M., Kriegel, H., and Sander, J. (1998). A distribution-based clustering algorithm for mining in large spatial databases. In *ICDE* '98: *Proceedings of the Fourteenth International Conference on Data Engineering*, pages 324–331. Los Alamitos, CA, USA. IEEE Computer Society. (Cited on p. 201)

Xu, X., Jäger, J., and Kriegel, H. (1999). A fast parallel clustering algorithm for large spatial databases. *Data Mining and Knowledge Discovery*, 3(3):263–290. (Cited on p. 200)

Yang, C., Duraiswami, R., DeMenthon, D., and Davis, L. (2003a). Mean-shift analysis using quasi-Newton methods. In *Proceedings of the* 2003 *International Conference on Image Processing (ICIP)*, Volume 3, pages II–447–50, Piscataway, NJ, USA. IEEE. (Cited on p. 159)

Yang, J., Wang, W., Wang, H., and Yu, P. (2002a). $\delta$-clusters: Capturing subspace correlation in a large data set. 18*th International Conference on Data Engineering,* 2002*. Proceedings*, pages 517–528, Los Alamitos, CA. IEEE Computer Society. (Cited on pp. 222, 259, 332)

Yang, J., Ward, M., and Rundensteiner, E. (2003b). Interactive hierarchical displays: A general framework for visualization and exploration of large multivariate data sets. *Computers and Graphics*, 26(2):265–283. (Cited on p. 58)

Yang, K. and Shahabi, C. (2004). A pca-based similarity measure for multivariate time series. In *MMDB* '04: *Proceedings of the* 2*nd ACM international workshop on Multimedia databases*, pages 65–74, New York, NY, USA. ACM Press. (Cited on p. 80)

Yang, M. (1993). A survey of fuzzy clustering. *Mathematical and Computer Modelling*, 18(11):1–16. (Cited on pp. 11, 141, 143)

Yang, M. and Wu, K. (2001). A new validity index for fuzzy clustering. *The* 10*th IEEE International Conference on Fuzzy Systems*, Volume 1, pages 89–92, Piscataway, NJ, USA. IEEE. (Cited on pp. 291, 293)

Yang, Y., Guan, X., and You, J. (2002b). CLOPE: A fast and effective clustering algorithm for transactional data. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 682–687, New York, NY, USA. ACM Press. (Cited on pp. 23, 273, 274)

Yang, Y. and Wang, H. (2018). Multi-view clustering: A survey. *Big Data Mining and Analytics*, 1(2):83–107. (Cited on p. 12)

Yazdani, N. and Ozsoyoglu, Z. (1996). Sequence matching of images. In *SSDBM* '96: *Proceedings of the Eighth International Conference on Scientific and Statistical Database Management*, pages 53–62, Washington, DC, USA. IEEE Computer Society. (Cited on p. 84)

Yeung, K., Fraley, C., Murua, A., Raftery, A., and Ruzzo, W. (2001). Model-based clustering and data transformations for gene expression data. *Bioinformatics*, 17(10):977–987. (Cited on pp. 24, 220)

Yeung, K., Medvedovic, M., and Bumgarner, R. (2003). Clustering gene-expression data with repeated measurements. *Genome Biology*, 4(5):G34.1–17. (Cited on p. 3)

Yeung, K. and Ruzzo, W. (2001). Principal component analysis for clustering gene expression data. *Bioinformatics*, 17(9):763–774. (Cited on p. 45)

Yi, B. and Faloutsos, C. (2000). Fast time sequence indexing for arbitrary $l_p$ norms. In Abbadi, A., Brodie, M., Chakravarthy, S., Dayal, U., Schlageter, N. K. G., and Whang, K., editors, *VLDB* '00: *Proceedings of the* 26*th International Conference on Very Large Data Bases*, pages 385–394. Orlando, FL, USA. Morgan Kaufmann. (Cited on p. 81)

Yi, B., Jagadish, H., and Faloutsos, C. (1998). Efficient retrieval of similar time sequences under time warping. In *ICDE* '98: *Proceedings of the Fourteenth International Conference on Data Engineering*, Orlando, FL USA, pages 201–208, Los Alamitos, CA, USA. IEEE Computer Society. (Cited on p. 84)

Yun, C., Chuang, K., and Chen, M. (2002). Using category-based adherence to cluster market-basket data. In *Proceedings of the* 2002 *IEEE International Conference on Data Mining (ICDM)*, pages 546–553, Los Alamitos, CA, USA. IEEE Computer Society. (Cited on pp. 272, 276)

Zadeh, L. (1965). Fuzzy sets. *Information and Control*, 8:338–353. (Cited on pp. 139, 141)

Zadeh, L. (1992). Fuzzy sets. In Bezdek, J. and Pal, S., editors, *Fuzzy Models for Pattern Recognition*, Chapter 2, pages 35–45. New York, NY, USA. IEEE. (Cited on p. 139)

Zahn, C. (1971). Graph-theoretical methods for detecting and describing Gestalt cluster. *IEEE Transactions on Computers*, C-20:68–86. (Cited on p. 189)

Zaiane, O. and Lee, C. (2002). Clustering spatial data in the presence of obstacles: A density-based approach. In *Proceedings of the International Database Engineering and Applications Symposium,* 2002, pages 214–223, Los Alamitos, CA, USA. IEEE Computer Society. (Cited on p. 200)

Zaït, M. and Messatfa, H. (1997). A comparative study of clustering methods. *Future Generation Computer Systems*, 13(2-3):149–159. (Cited on pp. 9, 287)

Zeng, G. and Dubes, R. (1985). A test for spatial randomness based on $k$-NN distances. *Pattern Recognition Letters*, 3(2):85–91. (Cited on p. 290)

Zhang, B. and Hsu, M. (2000). Scale up center-based data clustering algorithms by parallelism. Technical Report HPL-2000-6, Hewlett-Packard Laboratories. `http://www.hpl.hp.com/techreports/2000/HPL-2000-6.html`. (Cited on p. 166)

Zhang, B., Hsu, M., and Dayal, U. (1999). $k$-harmonic means - A data clustering algorithm. Technical report, Hewlett Packard Laboratories. (Cited on p. 156)

Zhang, B., Hsu, M., and Dayal, U. (2001a). $k$-harmonic means: A spatial clustering algorithm with boosting. In Roddick, J. and Hornsby, K., editors, *International Workshop on Temporal, Spatial and Spatio-Temporal Data Mining, TSDM*2000, volume 2007 of *Lecture Notes in Artificial Intelligence*, Lyon, France, pages 31–45, Berlin, Germany. Springer-Verlag. (Cited on p. 156)

Zhang, B. and Srihari, S. (2003). Properties of binary vector dissimilarity measures. Technical report, CEDAR, Department of Computer Science & Engineering, University at Buffalo,the State University of New York. `http://www.cedar.buffalo.edu/papers/publications.html`. (Cited on pp. 22, 65, 74)

Zhang, T., Ramakrishnan, R., and Livny, M. (1996). BIRCH: an efficient data clustering method for very large databases. In *Proceedings of the* 1996 *ACM SIGMOD International Conference on Management of Data*, pages 103–114, New York, NY, USA. ACM Press. (Cited on pp. 30, 134, 153, 226, 229, 261, 272)

Zhang, Y., Fu, A., Cai, C., and Heng, P. (2000b). Clustering categorical data. In *ICDE* 2000. *Proceedings of the* 16*th International Conference on Data Engineering*, page 305. Los Alamitos, CA, USA. IEEE Computer Society. (Cited on pp. 187, 188, 295)

Zhao, L., Tsujimura, Y., and Gen, M. (1996). Genetic algorithm for fuzzy clustering. In *Proceedings of IEEE International Conference on Evolutionary Computation,* 1996, Nagoya Japan, pages 716–719, Piscataway, NJ, USA. IEEE Computer Society. (Cited on pp. 169, 181)

Zhao, W., Ma, H., and He, Q. (2009). Parallel k-means clustering based on MapReduce. In *Proceedings of the* 1*st International Conference on Cloud Computing*, pages 674–679, Berlin, Germany. Springer-Verlag. (Cited on p. 262)

Zhao, Y. and Song, J. (2001). GDILC: A grid-based density-isoline clustering algorithm. In *International Conferences on Info-tech and Info-net,* 2001. *Proceedings of ICII* 2001, Volume 3, Beijing, China, pages 140–145, Piscataway, NJ, USA. IEEE Computer Society. (Cited on p. 196)

Zhong, S. and Ghosh, J. (2003a). Scalable, balanced model-based clustering. In *Proceedings of the Third SIAM International Conference on Data Mining*, San Francisco, CA, pages 71–82, Philadelphia, PA, USA. Society for Industrial and Applied Mathematics. (Cited on p. 268)

Zhong, S. and Ghosh, J. (2003b). A unified framework for model-based clustering. *The Journal of Machine Learning Research*, 4:1001–1037. (Cited on p. 220)

Zien, J., Chan, P., and Schlag, M. (1997). Hybrid spectral/iterative partitioning. In *ICCAD* '97: *Proceedings of the* 1997 *IEEE/ACM international conference on Computer-aided design*, pages 436–440, Los Alamitos, CA, USA. IEEE Computer Society. (Cited on p. 189)

Zien, J., Schlag, M., and Chan, P. (1999). Multilevel spectral hypergraph partitioning with arbitrary vertex sizes. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 18(9):1389–1399. (Cited on p. 189)

# Index

Data clustering, also known as cluster analysis, is an unsupervised process that divides a set of objects into homogeneous groups. Since the publication of the first edition of this monograph in 2007, development in the area has exploded, especially in clustering algorithms for big data and open-source software for cluster analysis. This second edition reflects these new developments.

*Data Clustering: Theory, Algorithms, and Applications, Second Edition*

- covers the basics of data clustering,
- includes a list of popular clustering algorithms, and
- provides program code that helps users implement clustering algorithms.

This book will be of interest to researchers, practitioners, and data scientists as well as undergraduate and graduate students.

**Guojun Gan** is an associate professor in the Department of Mathematics at the University of Connecticut. Before moving to academia, he worked at a life insurance company and at a hedge fund. He is a Fellow of the Society of Actuaries and his research interests fall within the interdisciplinary areas of actuarial science and data science.

**Chaoqun Ma** is a professor of management science in the College of Business Administration at Hunan University, where he served as dean from 2009 to 2019. His research interests include financial engineering, financial risk management, computational management, resource and environmental management, system optimization, and decision-making theory.

**Jianhong Wu** is a University Distinguished Research Professor in the Department of Mathematics and Statistics, the founding director of the Laboratory for Industrial and Applied Mathematics, and a senior Canada Research Chair in Industrial and Applied Mathematics at York University. He has received several prestigious awards, including the Queen Elizabeth II Diamond Jubilee Medal from the Government of Canada, and the CAIMS-Fields Industrial Mathematics Prize. His research interests include nonlinear dynamics, delay differential equations, neural networks, pattern recognition, mathematical ecology, epidemiology, and big data analytics.

---

For more information about SIAM books, journals, conferences, memberships, or activities, contact:

siam

Society for Industrial and Applied Mathematics
3600 Market Street, 6th Floor
Philadelphia, PA 19104-2688 USA
+1-215-382-9800 • Fax +1-215-386-7999
*siam@siam.org* • *www.siam.org*

**MN05**

ISBN: 978-1-611976-32-8
90000
9781611976328