# Cluster Analysis: an Overview

Yanchang Zhao

http://www.RDataMining.com

Machine Learning 102 Workshop
S P Jain School of Global Management, Mumbai, India

30 May - 5 June 2016

# Outline

- Clustering Concepts and Techniques
- Partitioning clustering
- Hierarchical clustering
- Density-Based clustering
- Grid-Based Clustering
- Model-Based Clustering
- Fuzzy Clustering
- Subspace Clustering
- Other Clustering Techniques
- Clustering Evaluation
- Applications
- Further Readings and Online Resources

# What is Data Clustering?

- Data clustering is to partition data into groups, where the data in the same group are similar to one another and the data from different groups are dissimilar [Han and Kamber, 2000].
- To segment data into clusters so that the *intra-cluster similarity* is maximized and that the *inter-cluster similarity* is minimized.
- The groups obtained are a partition of data, which can be used for customer segmentation, document categorization, etc.

# Clustering - Membership of Objects

In terms of the membership of objects, there are two kinds of clustering.

- *Fuzzy clustering* is also known as *soft clustering*, where an object can be in more than one cluster, but with different membership degree.
- In contrast, an object in *hard clustering* can belong to one cluster only.

Generally speaking, clustering is referred to as hard clustering implicitly.

# Clustering - Type of Data

In terms of the type of data, there are

- Spatial data clustering
- Text clustering
- Multimedia clustering
- Time series clustering
- Graph clustering

# Clustering - Approaches

In terms of approaches, data clustering techniques can be classified into the following groups:

- Partitioning clustering
- Hierarchical clustering
- Density-based clustering
- Grid-based clustering
- Model-based clustering

# Clustering - Others

- Semi-supervised clustering
- Subspace clustering
- Bi-clustering
- Data stream clustering

# Features of a Good Clustering Algorithm

- Detecting clusters with various shapes and different distributions
- Finding clusters with considerably different sizes
- Working when outliers are present
- No or few parameters needed as input
- Scalable to both the size and the dimensionality of data

# Outline

Clustering Concepts and Techniques

Partitioning clustering

Hierarchical clustering

Density-Based clustering

Grid-Based Clustering

Model-Based Clustering

Fuzzy Clustering

Subspace Clustering

Other Clustering Techniques

Clustering Evaluation

Applications

Further Readings and Online Resources

# Partitioning clustering - I

- Partitioning the data into k groups first and then trying to improve the quality of clustering by moving objects from one group to another
- *k*-means [Alsabti et al., 1998, Macqueen, 1967]: randomly selects *k* objects as cluster centers and assigns other objects to the nearest cluster centers, and then improves the clustering by iteratively updating the cluster centers and reassigning the objects to the new centers.
- *k*-medoids [Huang, 1998]: a variation of *k*-means for categorical data, where the medoid (i.e., the object closest to the center), instead of the centroid, is used to represent a cluster.
- PAM and CLARA [Kaufman and Rousseeuw, 1990]
- CLARANS [Ng and Han, 1994]

# Partitioning clustering - II

- The result of partitioning clustering is dependent on the selection of initial cluster centers and it may result in a local optimum instead of a global one. (Improvement: run k-means multiple times with different initial centers and then choose the best clustering result.)
- Tends to result in sphere-shaped clusters with similar sizes
- Sentitive to outliers
- Non-trivial to choose an appropriate value for $k$

# Outline

Clustering Concepts and Techniques

Partitioning clustering

Hierarchical clustering

Density-Based clustering

Grid-Based Clustering

Model-Based Clustering

Fuzzy Clustering

Subspace Clustering

Other Clustering Techniques

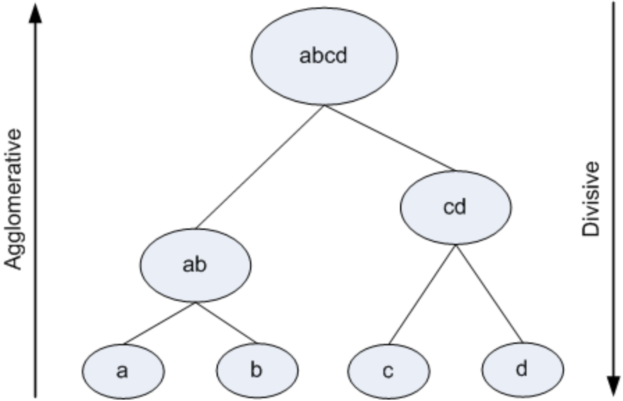
Clustering Evaluation

Applications

Further Readings and Online Resources

# Hierarchical Clustering - I

- With hierarchical clustering approach, a hierarchical decomposition of data is built in either bottom-up (agglomerative) or top-down (divisive) way.
- Generally a dendrogram is generated and a user may select to cut it at a certain level to get the clusters.



# Hierarchical Clustering - II

- With agglomerative clustering, every single object is taken as a cluster and then iteratively the two nearest clusters are merged to build bigger clusters until the expected number of clusters is obtained or when only one cluster is left.
    - AGENS [Kaufman and Rousseeuw, 1990]
- Divisive clustering works in an opposite way, which puts all objects in a single cluster and then divides the cluster into smaller and smaller ones.
    - DIANA [Kaufman and Rousseeuw, 1990]
    - BIRCH [Zhang et al., 1996]
    - CURE [Guha et al., 1998]
    - ROCK [Guha et al., 1999]
    - Chameleon [Karypis et al., 1999]

# Hierarchical Clustering - III

In hierarchical clustering, there are four different methods to measure the distance between clusters:

- *Centroid distance* is the distance between the centroids of two clusters.
- *Average distance* is the average of the distances between every pair of objects from two clusters.
- *Single-link distance*, a.k.a. *minimum distance*, is the distance between the two nearest objects from two clusters.
- *Complete-link distance*, a.k.a. *maximum distance*, is the distance between the two objects which are the farthest from each other from two clusters.

# Outline

- Clustering Concepts and Techniques
- Partitioning clustering
- Hierarchical clustering
- Density-Based clustering
- Grid-Based Clustering
- Model-Based Clustering
- Fuzzy Clustering
- Subspace Clustering
- Other Clustering Techniques
- Clustering Evaluation
- Applications
- Further Readings and Online Resources

# Density-Based Clustering - I

- The rationale of density-based clustering is that a cluster is composed of well-connected dense region, while objects in sparse areas are removed as noises.
- DBSCAN is a typical density-based clustering algorithm, which works by expanding clusters to their dense neighborhood [Ester et al., 1996].
- Although a user is not required to guess the number of clusters before clustering, he has to provide two other parameters, the radius of neighborhood and the density threshold, to run DBSCAN.

# Density-Based Clustering - II

- AGRID [Zhao and Song, 2003] (Zhao & Song, 2003) is an efficient density-based algorithm in that it uses grid to reduce the complexity of distance computation and cluster merging. By partitioning the data space into cells, only neighboring cells are taken into account when computing density and merging clusters.
- Some other density-based clustering techniques are OPTICS [Ankerst et al., 1999] and DENCLUE [Hinneburg and Keim, 1998].
- The advantage of density-based clustering is that it can filter out noise and find clusters of arbitrary shapes (as long as they are composed of connected dense regions).
- However, most density-based approaches utilize the indexing techniques for efficient neighborhood inquiry, such as R-tree and R*-tree, which do not scalable well to high-dimensional space.

# Outline

Clustering Concepts and Techniques

Partitioning clustering

Hierarchical clustering

Density-Based clustering

Grid-Based Clustering

Model-Based Clustering

Fuzzy Clustering

Subspace Clustering

Other Clustering Techniques

Clustering Evaluation

Applications

Further Readings and Online Resources

# Grid-Based Clustering - I

- Grid-based clustering works by partitioning the data space into cells with grid and then merging them to build clusters.
- Some grid-based methods, such as STING, WaveCluster and CLIQUE, use regular grids to partition data space
- Some employ adaptive or irregular grids, such as adaptive grid (Goil, Nagesh, & Choudhary, 1999) and optimal grid (Hinneburg & Keim, 1999).

# Grid-Based Clustering - II

- STING [Wang et al., 1997] is a grid-base multi-resolution clustering technique in which the spatial area is divided into rectangular cells and organized into a statistical information cell hierarchy. Statistical information, such as the count, mean, minimum, maximum, standard deviation and distribution, is stored for each cell. Thus, the statistical information of cells is captured and clustering can be performed without recourse to the individual objects.
- CLIQUE [Agrawal et al., 1998] works like APRIORI [Agrawal and Srikant, 1994]. It partitions each dimension into intervals and computes the dense units in all dimensions. Then the dense units are combined to generate the dense units in higher dimensions.

# Grid-Based Clustering - III

- WaveCluster [Sheikholeslami et al., 1998] is proposed to look at the multi-dimensional data space from a signal processing perspective. The objects are taken as a d-dimensional signal, and the high frequency parts of the signal correspond to the boundary of clusters, where the distribution of objects changes rapidly. The low frequency parts with high amplitude correspond to clusters, where data are concentrated.

# Grid-Based Clustering - IV

- The advantage of gird-based clustering is that the processing time is very fast, because it is independent on the number of objects, but dependent on the number of cells.
- Its disadvantage is that the quality of clustering is dependent on the granularity of cells and that the number of cells increases exponentially with the dimensionality of data.
- Therefore, adaptive grid and optimal grid are designed to tackle the above problems.
  - MAFIA [Goil et al., 1999] uses adaptive grids to partition a dimension depending on the distribution of data in the dimension, in contrast to partition every dimension evenly.
  - OptiGrid [Hinneburg and Keim, 1999] uses contracting projections of data to determine the optimal cutting hyper-planes for data partitioning, where the grid is arbitrary, as compared with equidistant, axis-parallel grids.

# Outline

Clustering Concepts and Techniques

Partitioning clustering

Hierarchical clustering

Density-Based clustering

Grid-Based Clustering

**Model-Based Clustering**

Fuzzy Clustering

Subspace Clustering

Other Clustering Techniques

Clustering Evaluation

Applications

Further Readings and Online Resources

# Model-Based Clustering

- Model-based clustering assumes that the data are generated by a mixture of probability distributions, and it attempts to learn statistical probability models from data, with each model representing one particular cluster [Zhong and Ghosh, 2003].
- The model type is often set as Gaussian or hidden Markov models (HMMs), and then the model structure is determined by model selection techniques and parameters are estimated using maximum likelihood algorithms.
- Algorithms
  - DMBC (Distributed Model-Based Clustering) [Kriegel et al., 2005]
  - model-based clustering based on data summarization (bEMADS and gEMADS) [Jin et al., 2005]
  - neural network approaches, such as SOM (self-organizing feature maps) [Kohonen, 1988]

# Outline

- Clustering Concepts and Techniques
- Partitioning clustering
- Hierarchical clustering
- Density-Based clustering
- Grid-Based Clustering
- Model-Based Clustering
- Fuzzy Clustering
- Subspace Clustering
- Other Clustering Techniques
- Clustering Evaluation
- Applications
- Further Readings and Online Resources

# Fuzzy Clustering - I

- Generally speaking, an object is either a member of a cluster or not a member of it. Fuzzy clustering is an extension by allowing an object to be a member of more than one cluster.
- For example, an object is the member of three clusters with membership as 0.5, 0.3 and 0.2, respectively.
- Fuzzy clustering is also known as soft clustering, as compared with hard clustering where the membership can be either 1 or 0 only.
- Algorithms: fuzzy k-means [Karayiannis, 1995] and EM (Expectation Maximization) algorithm [Dempster, 1977]

# Fuzzy Clustering - II

- EM algorithm generates fuzzy clustering in two steps.
  - The expectation (E) step calculates for each object the expectation of probabilities in each cluster.
  - The maximization (M) step computes the distribution parameters of clusters, i.e., maximizing the likelihood of the distributions given the data.
- The two steps are repeated to improve the clustering, until the improvement (say, the increase in log-likelihood) becomes negligible.
- EM algorithms may result in a local maximum instead of the global maximum.
- Similar to *k*-means, the chance to get the global maximum can be improved by running EM for multiple times with different initial guesses and then choosing the best clustering.

# Outline

- Clustering Concepts and Techniques
- Partitioning clustering
- Hierarchical clustering
- Density-Based clustering
- Grid-Based Clustering
- Model-Based Clustering
- Fuzzy Clustering
- **Subspace Clustering**
- Other Clustering Techniques
- Clustering Evaluation
- Applications
- Further Readings and Online Resources

# Subspace Clustering - I

- In some real-world applications, the dimensionality of data is in hundreds or even thousands, and more often than not, no meaningful clusters can be found in the full dimensional space.
- Subspace clustering is proposed for high dimensional data, where two clusters can be in two different subspaces and the subspaces can be of different dimensionalities.
- Algorithms:
  - CLIQUE [Agrawal et al., 1998]
  - MAFIA [Goil et al., 1999]
  - Random Projection [Fern and Brodley, 2003]
  - Projected clustering [Aggarwal et al., 1999]
  - Monte-Carlo [Procopiuc et al., 2002]
  - Projective clustering [Ng et al., 2005]
- With most techniques for subspace clustering, a cluster is defined as an axis-parallel hyper-rectangle in a subspace.

# Subspace Clustering - II

- A technique for subspace clustering proposed by [Fern and Brodley, 2003] is to find the subspaces in a random projection and ensemble way.
- The dataset is first projected into random subspaces, and then EM algorithm is used to discover clusters in the projected dataset.
- The algorithm generates several groups of clusters with the above method and then combines them into a similarity matrix, from which the final clusters are discovered with an agglomerative clustering method.

# Subspace Clustering - III

- MAFIA [Goil et al., 1999] is an efficient algorithm for subspace clustering using a density and grid based approach.
- It uses adaptive grids to partition a dimension depending on the distribution of data in the dimension.
- The bins and cells that have low density of data are pruned to reduce computation.
- The boundaries of the bins are not rigid, which improves the quality of clustering.

# Outline

- Clustering Concepts and Techniques
- Partitioning clustering
- Hierarchical clustering
- Density-Based clustering
- Grid-Based Clustering
- Model-Based Clustering
- Fuzzy Clustering
- Subspace Clustering
- Other Clustering Techniques
- Clustering Evaluation
- Applications
- Further Readings and Online Resources

# Bi-Clustering

- *Bi-clustering*, a.k.a. *co-clustering* or *two-way clustering*, is to group objects for a subset of attributes by performing simultaneous clustering of both rows and columns [Cheng, 2000, Madeira, 2004].
- It is a kind of subspace clustering.
- Bi-clustering is widely used for clustering microarray data to analyze the activities of genes under many different conditions.
- Microarray data can be viewed as a matrix, where each row represents a gene, each column stands for a condition and each entry gives the expression level of a gene under a condition.
- From microarray data, four major types of biclusters are to discover: 1) biclusters with constant values, 2) biclusters with constant values on rows or columns, 3) biclusters with coherent values, and 4) biclusters with coherent evolutions [Madeira, 2004].

## Text Clustering

- Text clustering, also referred to as document clustering, is to group documents based on the similarity in the terms used and is widely used for document categorization, information retrieval and web search engine [Beil, 2002, Zhong, 2005].
- Most algorithms for text clustering are based on a vector space model, where each document is represented by a vector of frequencies of terms.
- At first, a bag of words is collected for each document by filtering tags, stemming and pruning.
- Then the similarity of two documents is measured by how many words they share in common, and the documents can be clustered into groups with one of the clustering algorithms introduced above.
- Algorithms: Suffix Tree Clustering [Zamir and Etzioni, 1998], FTC(Frequent Term-based Clustering) and HFTC (Hierarchical FTC) [Beil, 2002]

# Data Stream Clustering

- Clustering data streams where new data arrive continuously, such as click-streams, retail transactions and stock prices [Aggarwal et al., 2003, Guha et al., 2003]
- The clusters are adjusted dynamically according to new data, and emerging clusters are also detected.
- New data can come in two ways, either as new records, such as transaction data, or as new dimensions, such as stock price data and other time series data.
- Aggarwal et al. (2003) proposed the concepts of pyramid time frame and micro-clustering to cluster evolving data streams. The statistical information of data is stored as micro-clusters, and the micro-clusters are stored at snapshots in time following a pyramidal pattern. The above are then used in an offline process to explore stream clustering over different horizons.

# Semi-Supervised Clustering

- Sometimes there is a small amount of knowledge which can be used to guide clustering, and the knowledge available is normally not enough for a supervised learning to classify the data.
- The knowledge can be either pairwise constraints, such as must-link and cannot-link, or class labels for some objects.
- Algorithms: COP-COBWEB (Constraint-Partitioning COBWEB) [Wagstaff and Cardie, 2000], CCL (Constrained Complete-Link) [Klein et al., 2002], MPC-KMeans (Metric Pairwise Constrained KMeans) [Basu et al., 2003], semi-supervised clustering with user feedback [Cohn et al., 2003], and a probabilistic model for semi-supervised clustering [Basu et al., 2004]

# Outline

Clustering Concepts and Techniques

Partitioning clustering

Hierarchical clustering

Density-Based clustering

Grid-Based Clustering

Model-Based Clustering

Fuzzy Clustering

Subspace Clustering

Other Clustering Techniques

Clustering Evaluation

Applications

Further Readings and Online Resources

# Evaluation of Clustering

The validation measures can be classified into three categories [Halkidi et al., 2001, Brun et al., 2007]:

- Internal validation
    - Based on calculating the properties of result clusters
    - *Compactness*, *Dunns validation index*, *Silhouette index* and *Huberts correlation with distance matrix*
- Relative validation
    - Based on comparisons of partitions
    - *Figure of merit* and *Stability*
- External validation
    - Based on comparing with a known true partition of data
    - *Conditional Entropy (CE)*, *Normalized Mutual Information (NMI)*, *Huberts correlation*, *Rand statistics*, *Jaccard coefficient*, and *Folkes and Mallows index*

# Outline

Clustering Concepts and Techniques

Partitioning clustering

Hierarchical clustering

Density-Based clustering

Grid-Based Clustering

Model-Based Clustering

Fuzzy Clustering

Subspace Clustering

Other Clustering Techniques

Clustering Evaluation

Applications

Further Readings and Online Resources

# Applications - I

- Customer segmentation
  To partition the general population of consumers into market segments and to better understand the relationships between different groups of consumers/potential customers, and for use in market segmentation, product positioning, new product development and selecting test markets. [1]
- Image segmentation
  Clustering can be used to divide a digital image into distinct regions for border detection or object recognition.
- Anomaly detection
  To detect anomalies or outliers with respect to clustering structure in data.

[1] https://en.wikipedia.org/wiki/Cluster_analysis

# Applications - II

- Crime analysis
  To identify areas where there are greater incidences of particular types of crime. By identifying these distinct areas or "hot spots" where a similar crime has happened over a period of time, it is possible to manage law enforcement resources more effectively.
- Educational data mining
  To identify groups of schools or students with similar properties.
- Social network analysis
  To recognize communities within large groups of people.

# Outline

Clustering Concepts and Techniques

Partitioning clustering

Hierarchical clustering

Density-Based clustering

Grid-Based Clustering

Model-Based Clustering

Fuzzy Clustering

Subspace Clustering

Other Clustering Techniques

Clustering Evaluation

Applications

Further Readings and Online Resources

# Further Readings - Clustering

- A breif overview of various apporaches for clustering

  Yanchang Zhao, et al. "Data Clustering." In Ferraggine et al. (Eds.), Handbook of Research on Innovations in Database Technologies and Applications, Feb 2009. http://yanchang.rdatamining.com/publications/Overview-of-Data-Clustering.pdf
- Cluster Analysis & Evaluation Measures

  https://en.wikipedia.org/wiki/Cluster_analysis
- Detailed reviews of algorithms for data clustering

  Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: a review. ACM Computing Surveys, 31(3), 264-323.

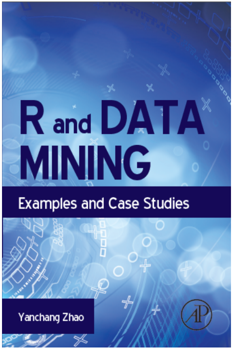
  Berkhin, P. (2002). Survey of Clustering Data Mining Techniques. Accrue Software, San Jose, CA, USA. http://citeseer.ist.psu.edu/berkhin02survey.html.
- A comprehensive textbook on data mining

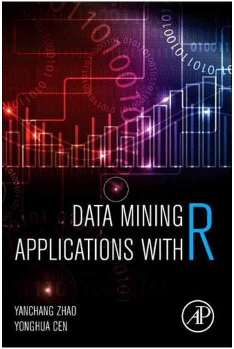
  Han, J., & Kamber, M. (2000). Data mining: concepts and techniques. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

# Online Resources

- Chapter 6 - Clustering, in book
  *R and Data Mining: Examples and Case Studies*
  `http://www.rdatamining.com/docs/RDataMining-book.pdf`
- RDataMining Reference Card
  `http://www.rdatamining.com/docs/RDataMining-reference-card.pdf`
- Free online courses and documents
  `http://www.rdatamining.com/resources/`
- RDataMining Group on LinkedIn (20,000+ members)
  `http://group.rdatamining.com`
- Twitter (2,500+ followers)
  @RDataMining

# The End





Thanks!

Email: yanchang(at)rdatamining.com

# References I

Aggarwal, C. C., Han, J., Wang, J., and Yu, P. (2003).
A framework for clustering evolving data streams.
In *29th VLDB Conference*, Berlin, Germany.
pyramidal time frame.

Aggarwal, C. C., Wolf, J. L., Yu, P. S., Procopiuc, C., and Park, J. S. (1999).
Fast algorithms for projected clustering.
In *SIGMOD '99: Proceedings of the 1999 ACM SIGMOD international conference on Management of data*, pages 61–72, New York, NY, USA. ACM Press.
PROCLUS: projected clustering.

Agrawal, R., Gehrke, J., Gunopulos, D., and Raghavan, P. (1998).
Automatic subspace clustering of high dimensional data for data mining applications.
In *SIGMOD '98: Proceedings of the 1998 ACM SIGMOD international conference on Management of data*, pages 94–105, New York, NY, USA. ACM Press.
CLIQUE.

Agrawal, R. and Srikant, R. (1994).
Fast algorithms for mining association rules in large databases.
In *Proc. of the 20th International Conference on Very Large Data Bases*, pages 487–499, Santiago, Chile.

Alsabti, K., Ranka, S., and Singh, V. (1998).
An efficient k-means clustering algorithm.
In *Proc. the First Workshop on High Performance Data Mining*, Orlando, Florida.

Ankerst, M., Breunig, M. M., Kriegel, H.-P., and Sander, J. (1999).
OPTICS: ordering points to identify the clustering structure.
In *SIGMOD '99: Proceedings of the 1999 ACM SIGMOD international conference on Management of data*, pages 49–60, New York, NY, USA. ACM Press.

# References II

Basu, S., Bilenko, M., and Mooney, R. (2003).
Comparing and unifying search-based and similarity-based approaches to semi-supervised clustering.
In *Proceedings of the ICML-2003*.

Basu, S., Bilenko, M., and Mooney, R. J. (2004).
A probabilistic framework for semi-supervised clustering.
In *KDD '04: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 59–68, New York, NY, USA. ACM.

Beil, F., E. M. . X. X. (2002).
Frequent term-based text clustering.
In *the eighth ACM SIGKDD international conference on Knowledge Discovery and Data Mining*.

Brun, M., Sima, C., Hua, J., Lowey, J., Carroll, B., Suh, E., and Dougherty, E. R. (2007).
Model-based evaluation of clustering validation measures.
*Pattern Recogn.*, 40(3):807–824.

Cheng, Y., . C. G. M. (2000).
Biclustering of expression data.
In *the eighth International Conference on Intelligent Systems for Molecular Biology*.

Cohn, D., Caruana, R., and McCallum, A. (2003).
Semi-supervised clustering with user feedback. technical report tr2003-1892.
Technical Report Technical Report TR2003-1892, Cornell University, USA.

Dempster, A., L. N. . R. D. (1977).
Maximum likelihood from incomplete data via the em algorithm.
*Journal of the Royal Statistical Society, Series B, 39(1), 1-38.*

# References III

Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. (1996).
A density-based algorithm for discovering clusters in large spatial databases with noise.
In *KDD*, pages 226–231.

Fern, X. and Brodley, C. (2003).
Random projection for high dimensional data clustering: A cluster ensemble approach.
In *The Twentieth International Conference on Machine Learning (ICML-2003)*.
random-projection.

Goil, S., Nagesh, H., and Choudhary, A. (1999).
MAFIA: Efficient and scalable subspace clustering for very large data sets.
Technical Report Technical Report CPDC-TR-9906-010, Northwestern University.

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

Guha, S., Rastogi, R., and Shim, K. (1998).
CURE: an efficient clustering algorithm for large databases.
In *SIGMOD '98: Proceedings of the 1998 ACM SIGMOD international conference on Management of data*, pages 73–84, New York, NY, USA. ACM Press.

Guha, S., Rastogi, R., and Shim, K. (1999).
ROCK: A robust clustering algorithm for categorical attributes.
In *Proceedings of the 15th International Conference on Data Engineering, 23-26 March 1999, Sydney, Austrialia*, pages 512–521. IEEE Computer Society.

Halkidi, M., Batistakis, Y., and Vazirgiannis, M. (2001).
On clustering validation techniques.
*Journal of Intelligent Information Systems*, 17(2-3):107–145.

# References IV

Han, J. and Kamber, M. (2000).
*Data Mining: Concepts and Techniques*.
Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Hinneburg, A. and Keim, D. A. (1998).
An efficient approach to clustering in large multimedia databases with noise.
In *KDD*, pages 58–65.
DENCLUE.

Hinneburg, A. and Keim, D. A. (1999).
Optimal grid-clustering: Towards breaking the curse of dimensionality in high-dimensional clustering.
In Atkinson, M. P., Orlowska, M. E., Valduriez, P., Zdonik, S. B., and Brodie, M. L., editors, *VLDB'99, Proceedings of 25th International Conference on Very Large Data Bases, September 7-10, 1999, Edinburgh, Scotland, UK*, pages 506–517. Morgan Kaufmann.
OptiGrid.

Huang, Z. (1998).
Extensions to the k-means algorithm for clustering large data sets with categorical values.
*Data Min. Knowl. Discov.*, 2(3):283–304.

Jin, H., Wong, M.-L., and Leung, K.-S. (2005).
Scalable model-based clustering for large databases based on data summarization.
*IEEE Trans. Pattern Anal. Mach. Intell.*, 27(11):1710–1719.
Member-Huidong Jin and Member-Man-Leung Wong and Senior Member-K. -S. Leung.

Karayiannis, N. B. (1995).
Generalized fuzzy k-means algorithms and their application in image compression.
In *the Society of Photo-Optical Instrumentation Engineers (SPIE) Conference*.

# References V

Karypis, G., Han, E.-H., and Kumar, V. (1999).
Chameleon: hierarchical clustering using dynamic modeling.
*Computer*, 32(8):68–75.

Kaufman, L. and Rousseeuw, P. J. (1990).
*Finding groups in data. an introduction to cluster analysis.*
Wiley Series in Probability and Mathematical Statistics. Applied Probability and Statistics, New York: Wiley, 1990.

Klein, D., Kamvar, S. D., and Manning, C. D. (2002).
From instance-level constraints to space-level constraints: Making the most of prior knowledge in data clustering.
In *ICML '02: Proceedings of the Nineteenth International Conference on Machine Learning*, pages 307–314, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Kohonen, T. (1988).
Self-organized formation of topologically correct feature maps.
*Neurocomputing: foundations of research*, pages 509–521.

Kriegel, H.-P., Kroger, P., Pryakhin, A., and Schubert, M. (2005).
Effective and efficient distributed model-based clustering.
In *ICDM '05: Proceedings of the Fifth IEEE International Conference on Data Mining*, pages 258–265, Washington, DC, USA. IEEE Computer Society.

Macqueen, J. B. (1967).
Some methods of classification and analysis of multivariate observations.
In *the Fifth Berkeley Symposium on Mathematical Statistics and Probability*.

Madeira, S. C., . O. A. L. (2004).
Biclustering algorithms for biological data analysis: A survey.
*IEEE/ACM Transactions Computational Biology and Bioinformatics, 1(1), 24-45.*

# References VI

Ng, E. K. K., chee Fu, A. W., and Wong, R. C.-W. (2005).
Projective clustering by histograms.
*IEEE Transactions on Knowledge and Data Engineering*, 17(3):369–383.

Ng, R. T. and Han, J. (1994).
Efficient and effective clustering methods for spatial data mining.
In *VLDB '94: Proceedings of the 20th International Conference on Very Large Data Bases*, pages 144–155, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Procopiuc, C. M., Jones, M., Agarwal, P. K., and Murali, T. M. (2002).
A monte carlo algorithm for fast projective clustering.
In *SIGMOD '02: Proceedings of the 2002 ACM SIGMOD international conference on Management of data*, pages 418–427, New York, NY, USA. ACM Press.

Sheikholeslami, G., Chatterjee, S., and Zhang, A. (1998).
Wavecluster: A multi-resolution clustering approach for very large spatial databases.
In *VLDB '98: Proceedings of the 24rd International Conference on Very Large Data Bases*, pages 428–439, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.
WaveCluster.

Wagstaff, K. and Cardie, C. (2000).
Clustering with instance-level constraints.
In *ICML '00: Proceedings of the Seventeenth International Conference on Machine Learning*, pages 1103–1110, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Wang, W., Yang, J., and Muntz, R. R. (1997).
STING: A statistical information grid approach to spatial data mining.
In Jarke, M., Carey, M. J., Dittrich, K. R., Lochovsky, F. H., Loucopoulos, P., and Jeusfeld, M. A., editors, *VLDB'97, Proceedings of 23rd International Conference on Very Large Data Bases, August 25-29, 1997, Athens, Greece*, pages 186–195. Morgan Kaufmann.

# References VII

Zamir, O. and Etzioni, O. (1998).
Web document clustering: a feasibility demonstration.
In *SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 46–54, New York, NY, USA. ACM.

Zhang, T., Ramakrishnan, R., and Livny, M. (1996).
BIRCH: an efficient data clustering method for very large databases.
In *SIGMOD '96: Proceedings of the 1996 ACM SIGMOD international conference on Management of data*, pages 103–114, New York, NY, USA. ACM Press.

Zhao, Y. and Song, J. (2003).
AGRID: An efficient algorithm for clustering large high-dimensional datasets.
In Whang, K.-Y., Jeon, J., Shim, K., and Srivastava, J., editors, *Proc. of the 7th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 03)*, volume 2637 of *Lecture Notes in Computer Science*, pages 271–282. Springer.
AGRID.

Zhong, S., . G. J. (2005).
Generative model-based document clustering: a comparative study.
*Knowledge and Information Systems, 8(3), 374-384.*

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