

Cluster Discovery and Visualization of Scientific Datasets

- Clustering: Grouping of objects characterized by feature vectors (or attribute vectors) into classes
- Objects in the same class are similar according to some defined similarity measure
- Pairs of objects from different classes are dissimilar under the same measure
- Questions:
- How do we find these similar classes - Techniques?
- Having found them how do we present them - Cluster Visualization ?
- How do we validate the techniques - Quality Measure ?

Some Issues - Taxonomy

- Dimensionality: low (≤ 10) vs. high dimensions (> 10)
- Size: low ($\leq 10^6$) vs. large (10^7 - 10^9)
- Stability: Streaming vs. storable
- Method of Processing: serial vs. parallel computation

Some Known Algorithms

- Extensive list of clustering algorithms have been proposed according to the domain of application
 - See survey article by Jain and Murty, 1999
 - Data Mining Book by Han and Kamber, 2001.
 - Research Group of D. Keim - <http://infovis.uni-konstanz.de/index.php?region=publications>
- Some Classification of Clustering techniques :
 - partitioning,
 - hierarchical agglomerative; hierarchical divisive
 - Density (or Grid) based clustering
- We will refer to density/grid based methods also as *Cell-Based*.

Hierarchical Clustering in M-dimensional space

- Classical hierarchical agglomerative clustering uses defined distance measures and similarity measures to construct a *dendogram*. - single-link, complete-link, average-link
- The Euclidean distance between a pair of objects, $r_p = \langle a_{p,0}, a_{p,1} \dots a_{p,M-1} \rangle$ and $r_q = \langle a_{q,0}, a_{q,1} \dots a_{q,M-1} \rangle$ is

$$\delta_2(r_q, r_p) = \sqrt{\sum_{j=0}^{M-1} (a_{p,j} - a_{q,j})^2} \quad (1)$$

- Similarity value is defined by

$$s_{p,q} = 1 - \frac{\delta_q(r_q, r_p)}{\delta_{max}} \quad (2)$$

where δ_{max} is the distance between the two farthest pair.

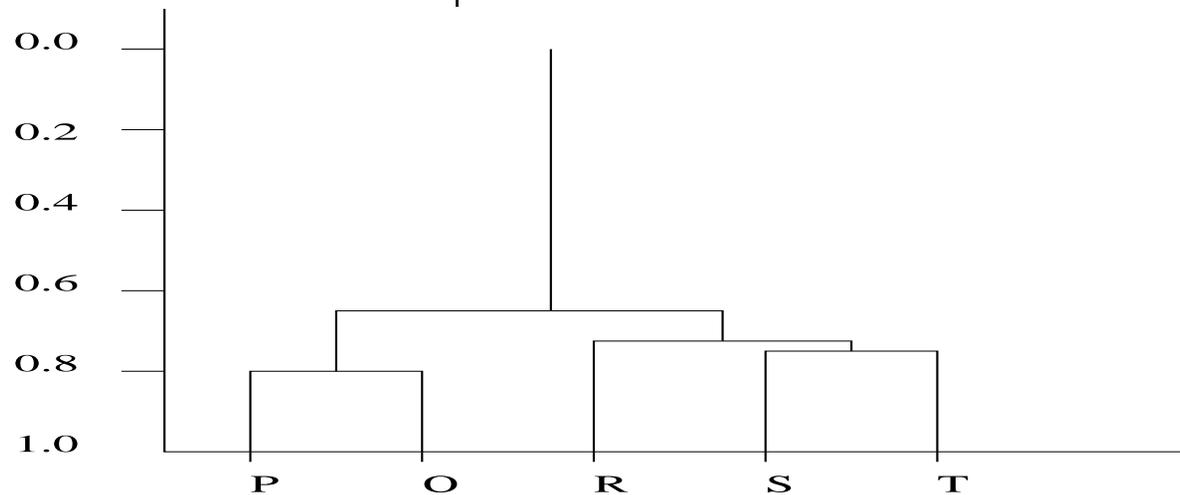
Example of Hierarchical Clustering

Object	A_0	A_1	A_2
P	0.31	17.8	3.0
Q	0.10	9.30	3.0
R	0.11	21.5	1.0
S	0.58	22.0	2.0
T	0.50	16.0	1.0

Sample Dataset

Object	P	Q	R	S	T
P	1.0				
Q	0.79	1.0			
R	0.58	0.36	1.0		
S	0.69	0.17	0.15	1.0	
T	0.61	0.34	0.72	0.75	1.0

Similarity Matrix



Single-Link Dendrogram

Other Known Algorithms

BIRCH: Derives clusters; handles noise; requires knowledge of inter-cluster distance for merging; works for small N.

CURE: Derives clusters in specified dimensions K; high running time; Used only on small data sets with $K = 2$; works by sampling.

OPTICS: For cluster analysis only; does not generate clusters; creates order of potential clusters. Nice for predicting the number of potential clusters for use in say k-Means clustering

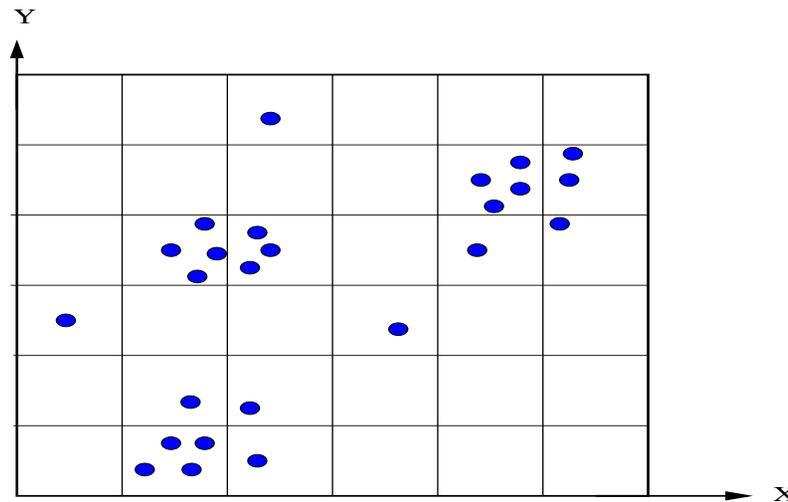
CELTYC/TDC: Derives clusters of any shapes; depends on bin resolution.

CLIQUE: Generates clusters in the form of DNF expressions; requires input of limiting dimension K; rectilinear clusters only.

Others: MAFFIA(PMAFFIA), GridGlus, STING, WaveCluster.

Grid/Density Based Clustering

- More appropriate for massively large datasets
- Takes a Spatial view of data representation
- Given a set of bounded ordered domains A_0, A_1, \dots, A_{M-1} , let $\mathcal{S}^M = A_0 \times A_1 \times \dots \times A_{M-1}$ be a *feature space*.
- Let $\mathcal{R} = \{r_0, r_1, \dots, r_{N-1}\}$ be a set of objects where $r_i = \langle a_{i,0}, a_{i,1} \dots a_{i,M-1} \rangle$, and $a_{i,j} \in A_j$.
- Each object corresponds to a point (i.e., an "image point" of the object), in the multidimensional feature space \mathcal{S}^M .



One Approach - HyCeltyc

1. Take sample of size S from original large dataset;
 - Either uniform sampling or biased sampling.
2. Apply dimensional reduction technique that preserves clusters from $M \rightarrow K$;
 - Techniques are: PCA, MDS, FastMap.
 - Points are mapped to a new feature space.
3. Cluster points in the transformed space using *fast* clustering algorithm much like connected component labeling.
4. Refine the clusters in the original M -dimensional space.
5. Using derived clusters, select the K most discriminating dimensions of these clusters.
6. Apply a linear clustering algorithm to the rest of the data points based on these K -significant dimensions.

Methods of Visualizing Cluster

- Highly problematic when we have to deal with high dimensions
- Topic discussed in detail in
 - Patrick Hoffman and George Grinstein, A survey of visualizations for high-dimensional data mining
 - D. Keim, Vis Tutorial Notes at <http://infovis.uni-konstanz.de/members/keim/PDF/Vis04Tutorial.pdf>
- Methods include
 - Scatter Plot Matrix
 - Dimensional Staking
 - Dendograms
 - Parallel Coordinates
 - Circular parallel Coordinates
- Some Viz Tools: Xgobi, GGobi and IBM's CViz

Scatter Plot Matrix

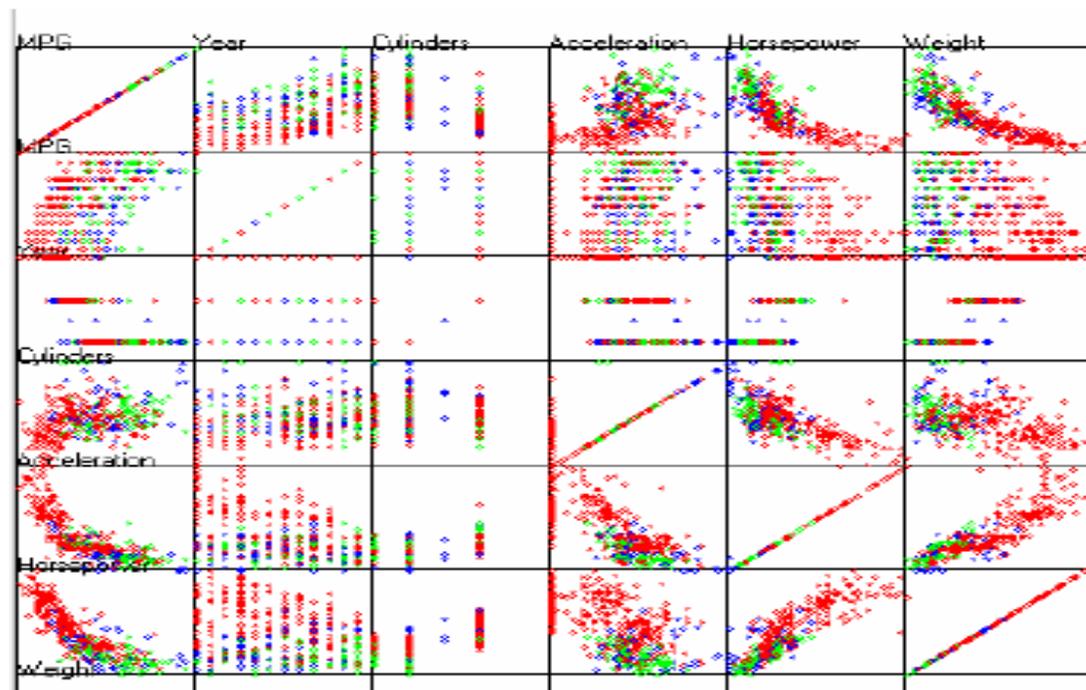


Figure 2 A Scatter Plot Matrix of the Car Dataset

Dendograms

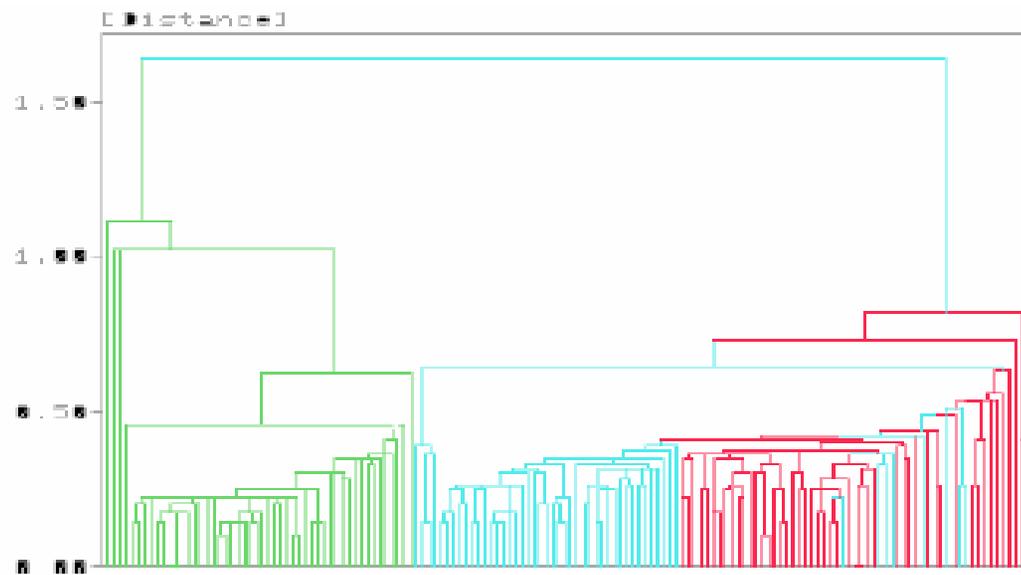


Figure 20 A Dendrogram Clustering of the Iris Flower Dataset (From the Inspect program)

Parallel Coordinates

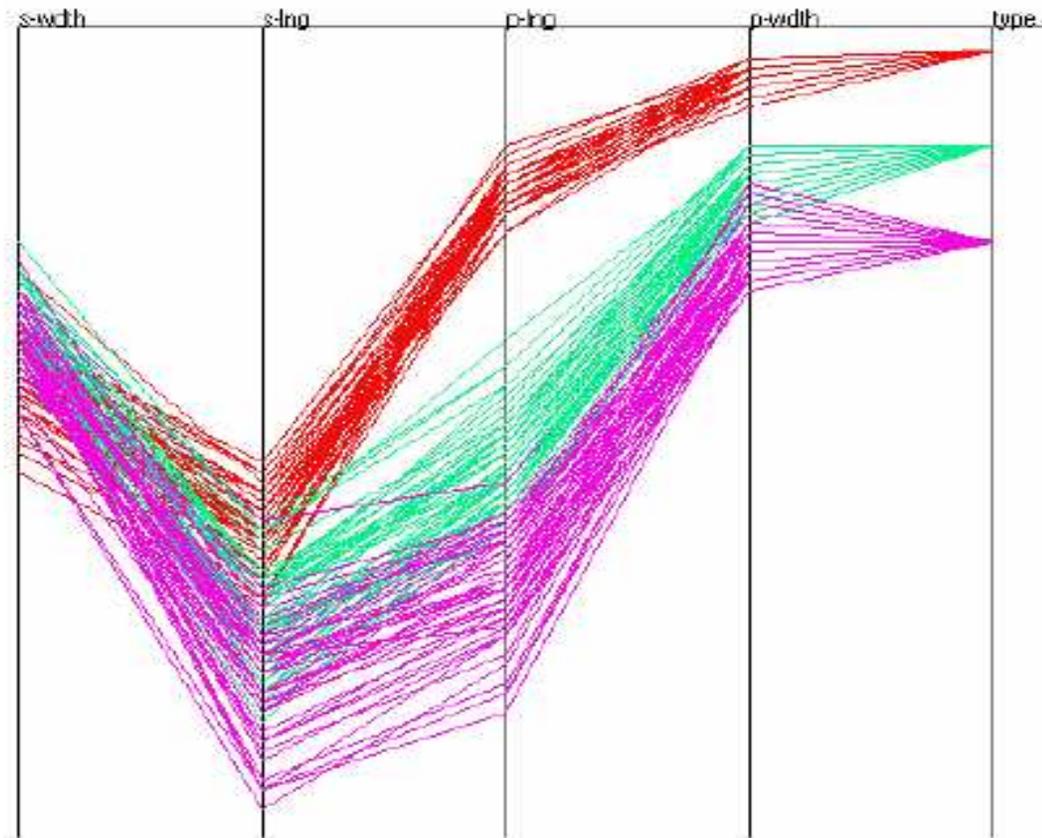


Figure 25 Parallel Coordinates of the Iris dataset (Global Normalization)

Circular Parallel Coordinates

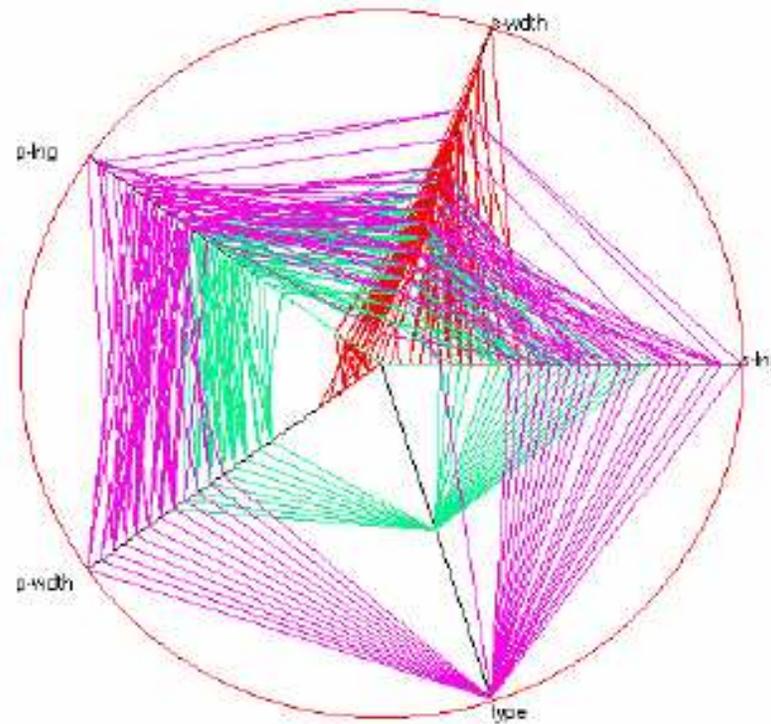


Figure 28 Circular Parallel Coordinates of the Iris Flower dataset (Local Normalization)

Some relevant references are given below:

Clustering: [2, 4, 6, 7, 8, 11, 12, 3, 1, 13, 15, 16, 17]

Visualization of Clusters: [10, 9, 5, 14]

Course Notes and Sites:

<http://infovis.uni-konstanz.de/members/keim/PDF/Vis04Tutorial.pdf>

<http://infovis.uni-konstanz.de/publications/papers/InfoVisLausanne.pdf>

<http://home.comcast.net/patrick.hoffman/viz/MIV-datamining.pdf>

http://www.cs.uml.edu/mtrutsch/research/High-Dimensional_Visualizations-KDD2001-color.pdf

<http://www.ggobi.org/> <http://www.research.att.com/areas/stat/xgobi/>

<http://www.alphaworks.ibm.com/tech/cviz>

References

- [1] R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan. Automatic subspace clustering of high dimensional data for data mining applications. In *Proc. ACM SIGMOD Int. Conf. on Management of Data*, pages 94–105, Seattle, WA, USA, 1998.
- [2] D. et al. Barbara. The new jersey data reduction report. *Bulletin of the Tech. Committee on Data Eng.*, 20(4), Dec. 1997.
- [3] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proc. 2nd Int'l. Conf. on Knowledge Discovery and Data Mining (KDD'1996)*, Portland, Oregon, 1996. AAAI Press.
- [4] C. Faloutsos and K. Lin. Fastmap: A fast algorithm for indexing, data-mining and visualization of traditional and multimedia datasets. In *Proc. ACM SIGMOD, SIGMOD RECORD*, pages 163 – 174, Jun. 1995.
- [5] G. Georges Grinstein, M. Trutschl, and U. Cvek. High-dimensional visualization. In *Proc. 7th ACM SIGKDD Int'l. Conf. on Knowledge Discovery and Data Mining, KDD'01*, New York, NY, USA, 2001. ACM Press.
- [6] J. Han and M. Kamber. *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers, San Francisco, California, 2001.
- [7] A Hinneburg and D. Keim. An efficient approach to clustering in large multimedia databases with noise. In *Proc. 4th Int. Conf. on Knowledge Discovery and Data Mining, AAAI Press*, pages 58–65, 1998.
- [8] A. Hinneburg and D. A. Keim. Optimal grid-clustering: Towards breaking the curse of dimensionality in high-dimensional clustering. In *Proc. 25th VLDB Conf.*, Edinburgh, Scotland, 1999.
- [9] P. Hoffman and G. Grinstein. Visualizations for high dimensional data mining - table visualizations, 1997.
- [10] Patrick E. Hoffman and Georges G. Grinstein. A survey of visualizations for high-dimensional data mining. pages 47–82, 2002.

- [11] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: A review. *ACM Comput. Surveys*, 31(3):264 – 323, Sept. 1999.
- [12] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. Piatko, R. Silverman, and A. Y. Wu. The analysis of a simple k-means clustering algorithm. In *Proc. 16th Symp. on Computational Geometry*, pages 100–109, Clear Water Bay Hong Kong, Jun. 2000.
- [13] G. Sheikholeslami, S. Chatterjee, and A. Zhang. Wavecluster: A multi-resolution clustering approach for very large spatial databases. In *Proc. 24th VLDB Conf.*, pages 428–439, Aug. 1998.
- [14] A. Strehl and J. Ghosh. Relationship-based clustering and visualization for high-dimensional data mining, 2002.
- [15] J. S. Vitter. An efficient algorithm for sequential random sampling. *ACM Transactions on Mathematical Software*, 13(1):58 – 67, 1987.
- [16] W. Wang, J. Yang, and R. Muntz. Sting: A statistical information grid approach to spatial data mining. In *Proc. 23rd Conf. on Very large Databases*, pages 186 – 195, Athens, Greece, 1997.
- [17] T. Zhang, R. Ramakrishnan, and M. Livny. Birch: An efficient data clustering method for very large databases. In H. V. Jagadish and I. S. Mumick, editors, *Proc. ACM SIGMOD Conf.*, pages 103–114, Montreal, Canada, Jun. 1996. ACM Press.