Spectral Clustering of Chemical Datasets

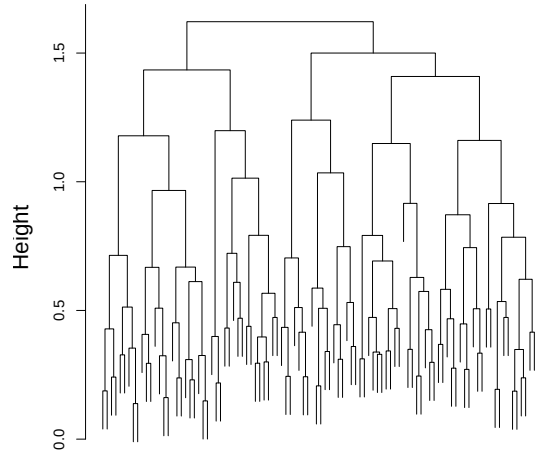
Rajarshi Guha and David J. Wild

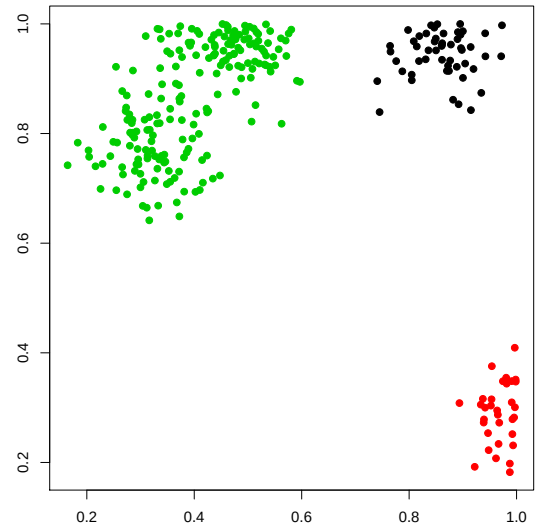
School of Informatics
Indiana University

26th March, 2007

Clustering Chemical Datasets

- Fundamental step in library design and compound selection
- Data partitioning method to focus on small, more similar subsets
  - Useful for local regression
  - Characterize multiple SAR's
- Clustering can be used as a classification technique





The Problem & Solutions

Problems

- Chemical datasets can be very large ($> 10^5$)
- The feature space can be high dimensional
- Correlated features

Some solutions

- Throw more CPU to the problem
- Use an approximation algorithm
- Avoid the distance matrix calculation

The Problem & Solutions

Problems

- Chemical datasets can be very large ($> 10^5$)
- The feature space can be high dimensional
- Correlated features

Some solutions

- Throw more CPU to the problem
- Use an approximation algorithm
- Avoid the distance matrix calculation

The Problem & Solutions

Problems

- Chemical datasets can be very large ($> 10^5$)
- The feature space can be high dimensional
- Correlated features

Some solutions

- Throw more CPU to the problem
- Use an approximation algorithm
- Avoid the distance matrix calculation

Spectral Clustering

What is it?

- For N observations, make an $N \times N$ similarity matrix
- Evaluate the eigenvectors
  - Usually the first or second
- Partition original points according to whether the corresponding element of the eigenvector is positive or negative
- Leads to a binary partition
- We can get more clusters by
  - Subdividing each partition
  - Consider multiple eigenvectors and apply k-means etc.

Caveats

- A full similarity matrix may not be a good approach
- Can consider kNN based similarity matrix (discrete)
- Exponentially decreasing similarities (continuous)

SVD Clustering

- Based on the Singular Value Decomposition
- Projects the original matrix onto a k-D subspace
- Clustering is performed in the reduced subspace
- The algorithm is a polynomial time approximation
- It has been shown that the SVD *itself* represents a clustering

- Where is it used?
  - Image segmentation
  - Web searches (Google Pagerank)

Drineas, P. et al., *Machine Learning*, **2004**, *56*, 9–33

Why is the Fast SVD Better?

- The original matrix is randomly sampled
- The sub-matrix is then decomposed
- Though the fast SVD utilizes SVD, the matrix being decomposed is significantly smaller

The Fast SVD

A $\leftarrow m \times n$ matrix
Choose $c \leq m$, $k \leq c$, P_i's
D $\leftarrow m \times m$ distance matrix
$T \leftarrow 1$
while $T < c$ **do**
  Select i from $\{1, \ldots, m\}$
  with probability P_i
  $C_i \leftarrow D_i / \sqrt{cP_i}$
  $T \leftarrow T + 1$
end while
$H \leftarrow$ top k left SV's

Drineas, P. et al., *Machine Learning*, **2004**, *56*, 9–33

Datasets & Descriptors

Datasest 1

- AMES mutagenecity
- 4337 compounds
- Categorical

Dataset 2

- Aqueous solubility
- 1236 compounds
- Continuous

Descriptors

- 166 bit MACCS fingerprints
- Constitutional, geometric and topological descriptors
- Calculations and analysis performed with MOE and R 2.2.1

Huuskonen, J., *J. Chem. Inf. Comput. Sci.*, **2000**, *40*, 773–777

Kazius, J. et al., *J. Med. Chem.*, **2005**, *48*, 312–320

Datasets & Descriptors

Datasest 1

- AMES mutagenecity
- 4337 compounds
- Categorical

Dataset 2

- Aqueous solubility
- 1236 compounds
- Continuous

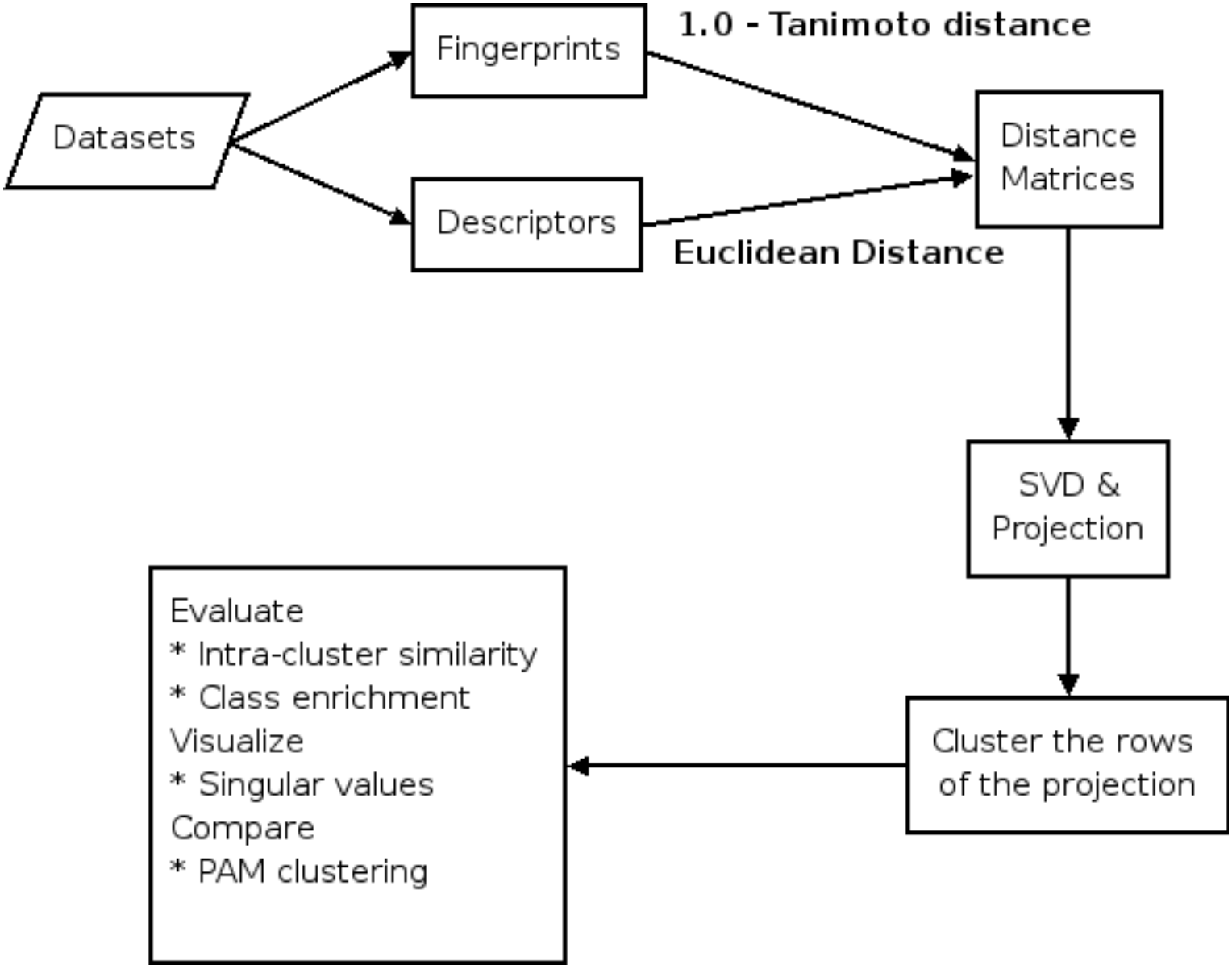
Descriptors

- 166 bit MACCS fingerprints
- Constitutional, geometric and topological descriptors
- Calculations and analysis performed with MOE and R 2.2.1

Huuskonen, J., *J. Chem. Inf. Comput. Sci.*, **2000**, *40*, 773–777

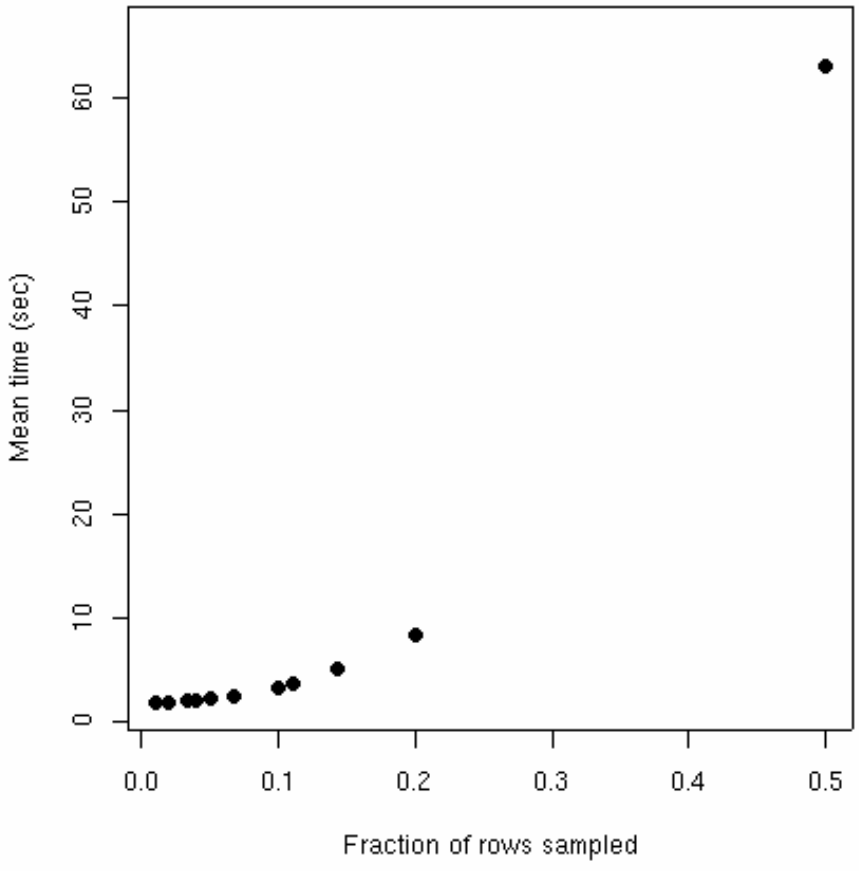
Kazius, J. et al., *J. Med. Chem.*, **2005**, *48*, 312–320

Methodology



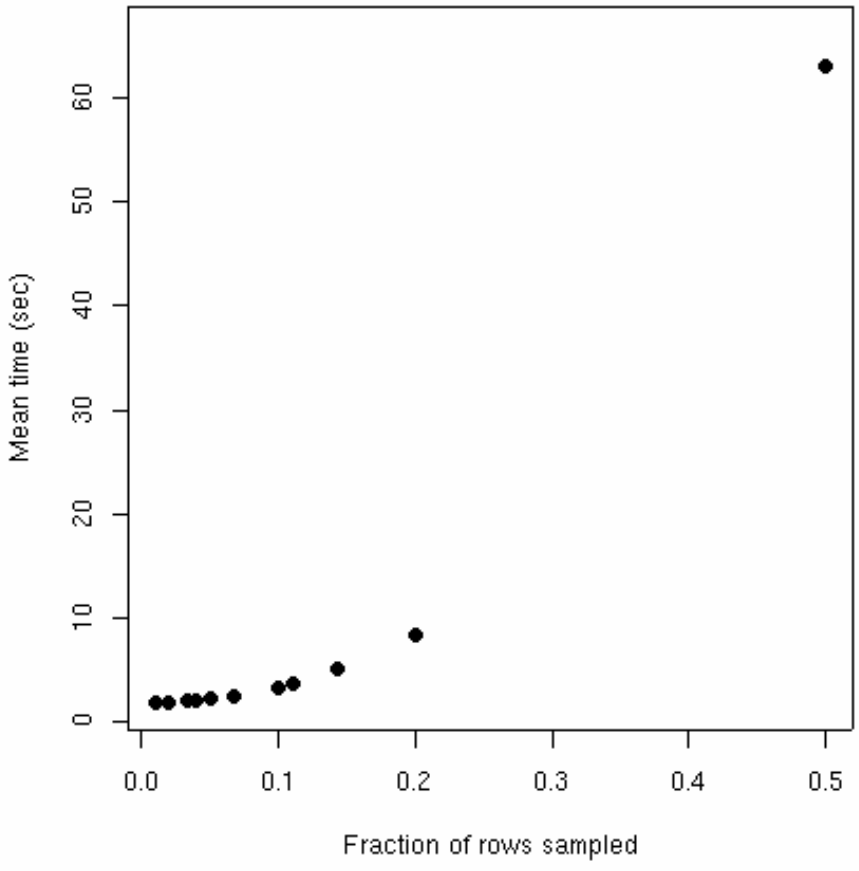
How Fast is the Fast SVD?

- Benchmarked on the Ames dataset
  - 4337×4337 matrix
- Each case was run 10 times
- No significant error till less than 10% of the rows are sampled
- Simple SVD = 368s



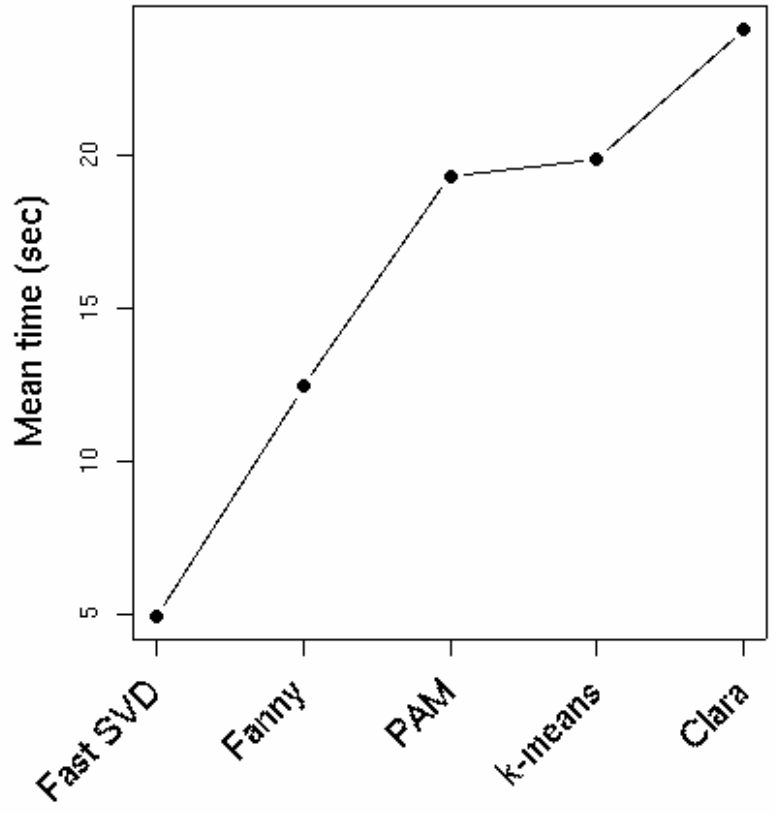
How Fast is the Fast SVD?

- Benchmarked on the Ames dataset
  - 4337×4337 matrix
- Each case was run 10 times
- No significant error till less than 10% of the rows are sampled
- Simple SVD = 368s



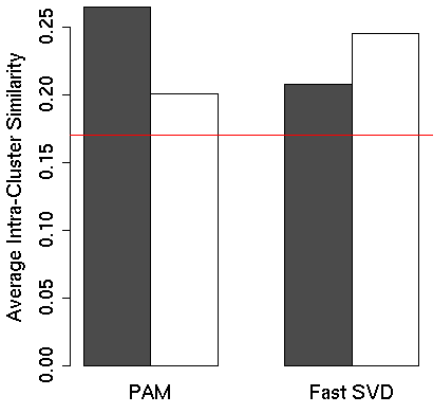
Comparison of Timings

- PAM is a more robust version of k-means
- Clara is an extension of PAM but is more suitable for large datasets as it uses a sampling process
- The fast SVD is significantly faster than all other partitioning methods

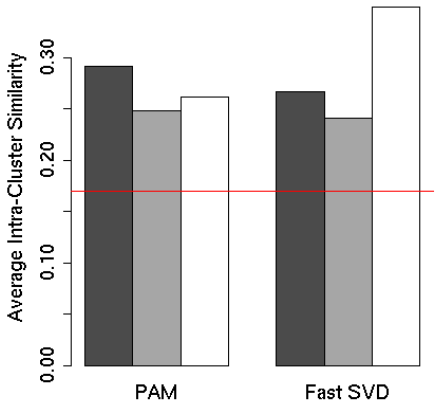


For all cases $k = 2$ and times reported are the average of ten runs

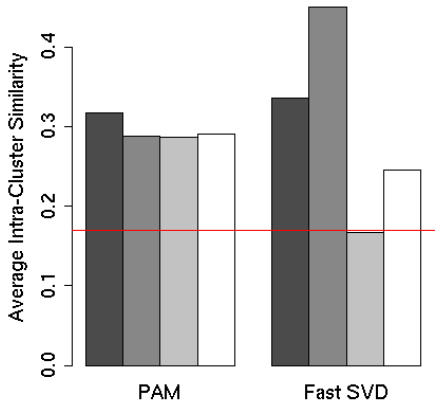
Aqueous Solubility - Class Similarity



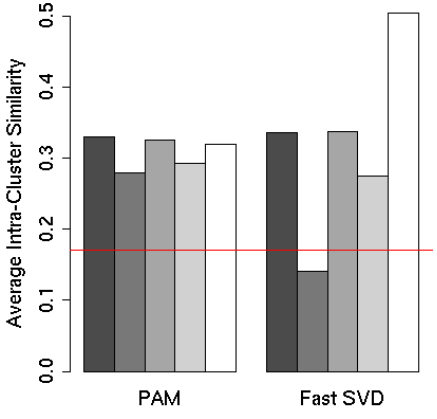
2 Cluster



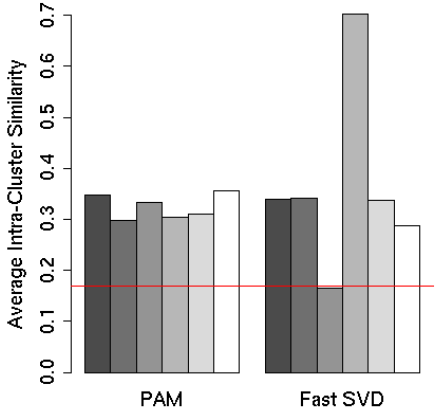
3 Cluster



4 Cluster



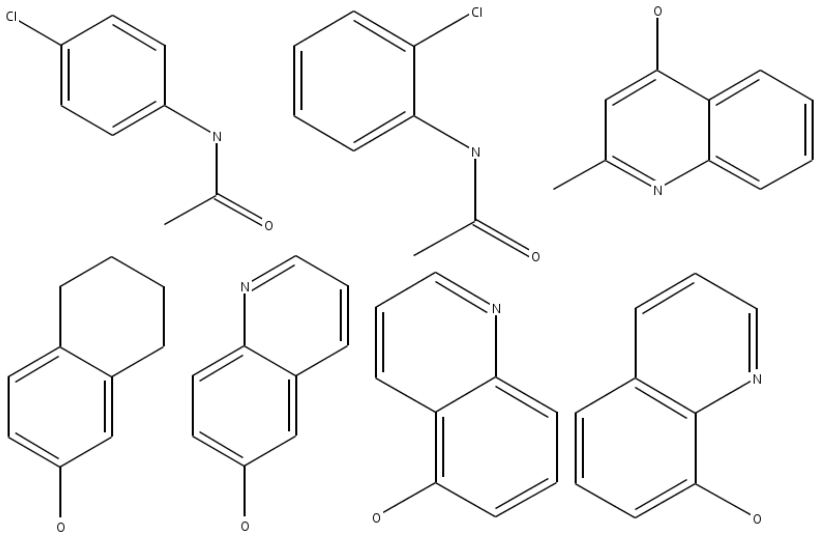
5 Cluster

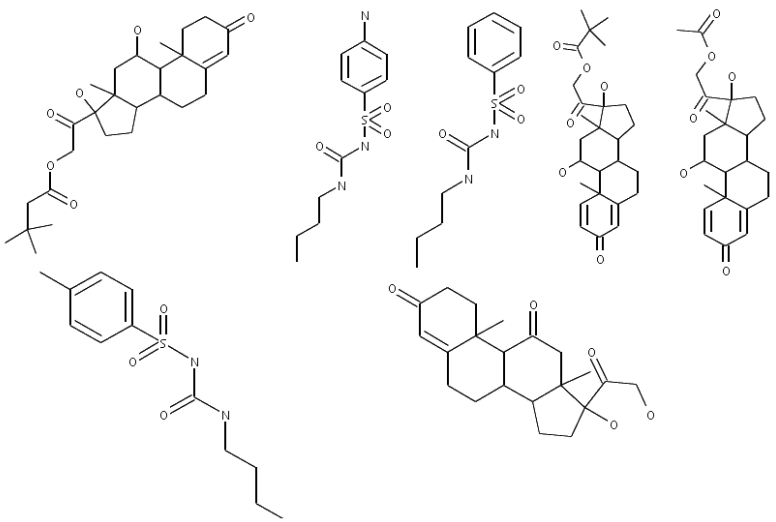


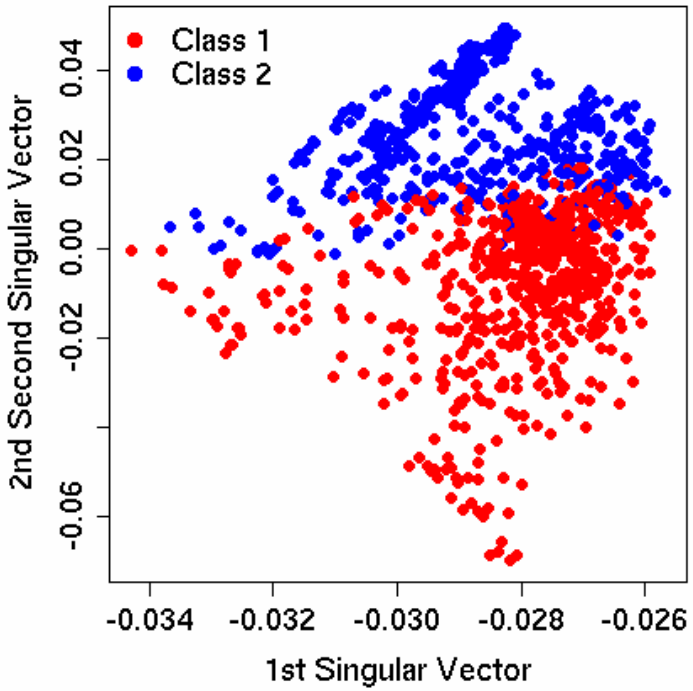
6 Cluster

- SVD favors specific clusters
- Results compare well with PAM

Aqueous Solubility - Class Members

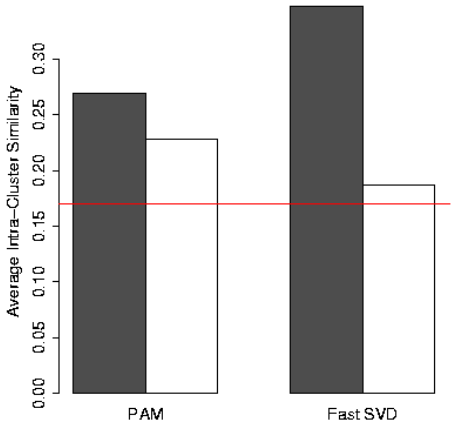




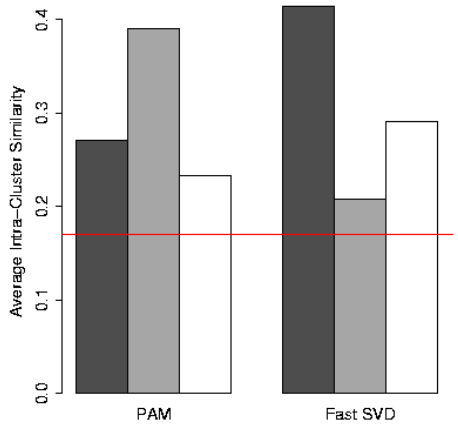


- Plot of the first two singular vectors obtained from a 2 class clustering
- The class structure is not very clear.

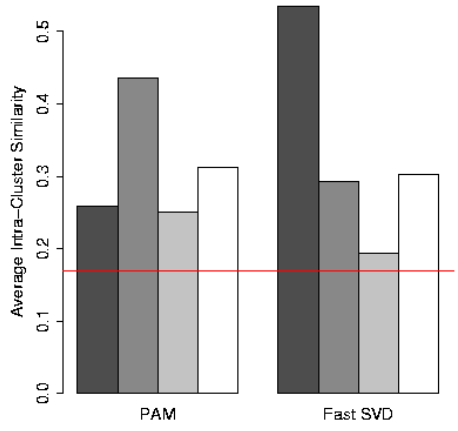
Ames Mutagenecity - Class Similarity



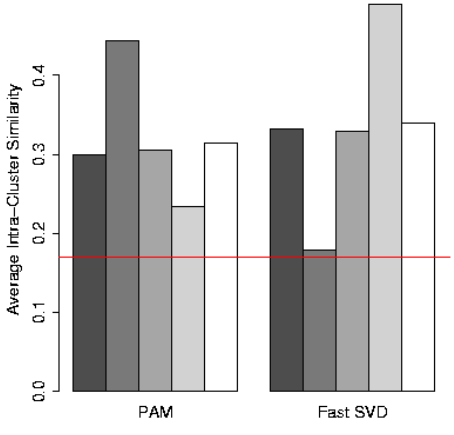
2 Cluster



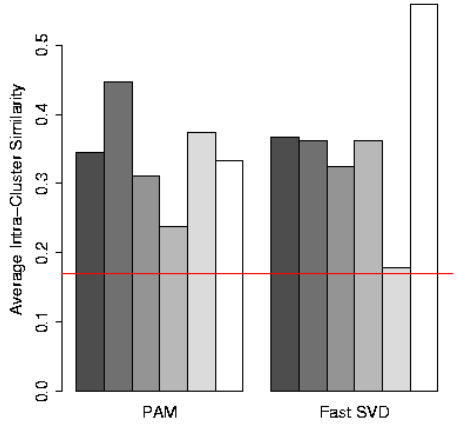
3 Cluster



4 Cluster



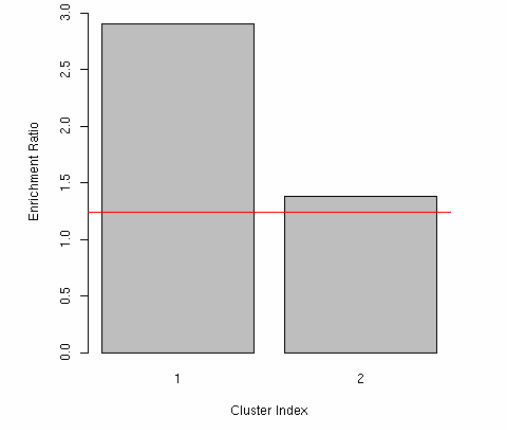
5 Cluster



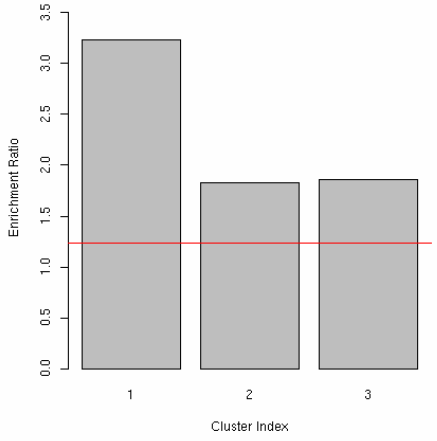
6 Cluster

- Spectral clustering generally improves over PAM
- Certain clusters are of poor quality

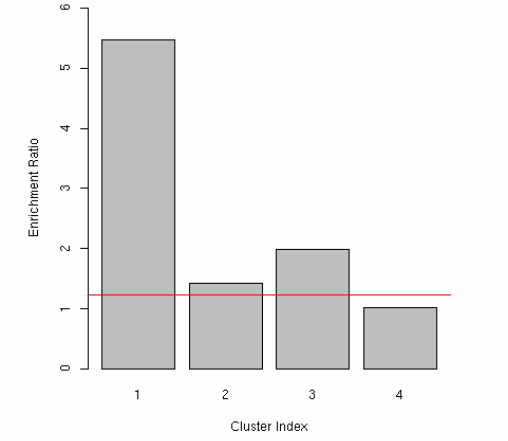
Ames Mutagenecity - Class Enrichment



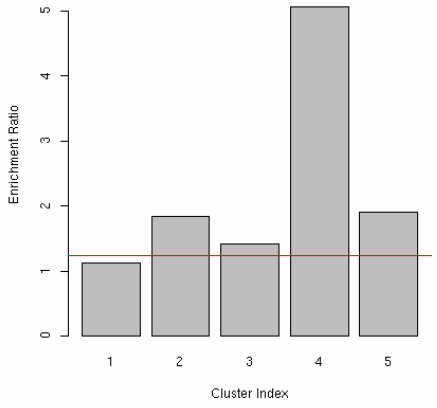
2 Cluster



3 Cluster



4 Cluster

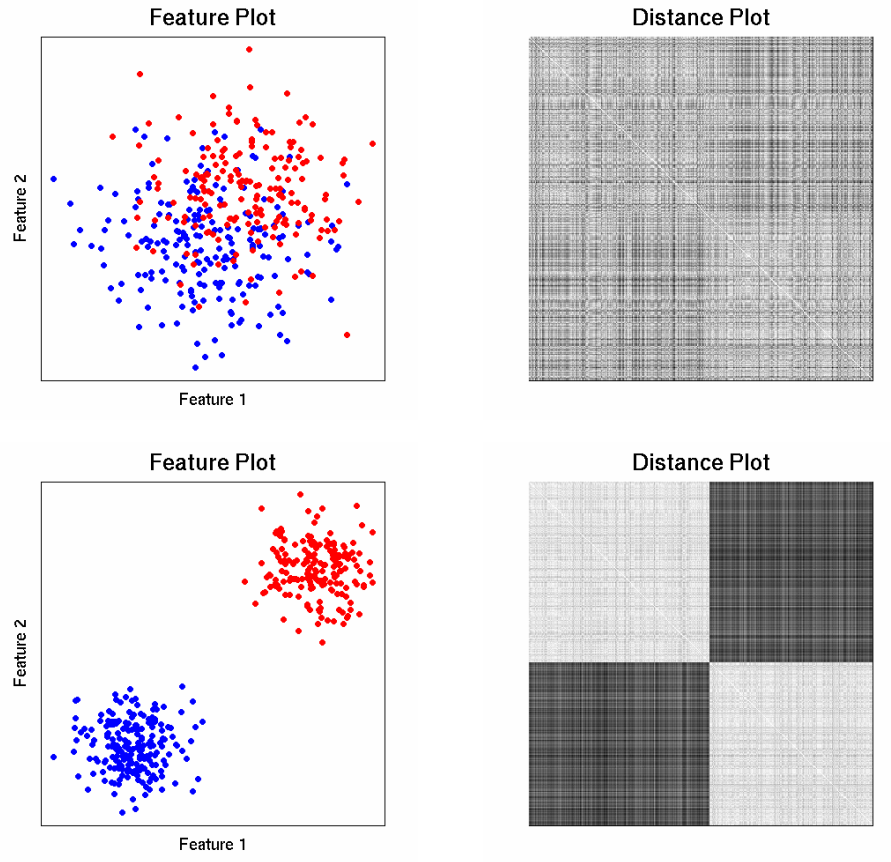


5 Cluster

- Class enrichment is defined by the ratio of the sizes of the larger class to the smaller class
- For the overall dataset *mutagen : non-mutagen = 1.24*
- The spectral clustering algorithm enriches specific clusters
- The class enrichment does not always correspond to average cluster similarity

Block Structure & Spectral Clustering

- It has been shown that good clusterings correspond to block diagonal distance (affinity) matrices
- The aim is to enhance the block diagonal character of a distance matrix
- Analysis of the block structure can be also be used for hierarchical clustering



Scott, G.L.; Longuet-Higgins, H.C., *British Machine Vision Conf.*, **1990**, 103-108

Ng, A., et al., in *Adv. in Neural Inf. Proc.*, **2002**, *14*, MIT Press

Asymmetric Spectral Clustering

- A Gaussian kernal leads to an asymmetric affinity matrix

$$A_{i,j} = \exp\left(\frac{\|x_i - x_j\|^2}{2\sigma_i^2}\right)$$

- Bandwidth is different for each observation
- Individual bandwidths determined by a neighborhood size, τ

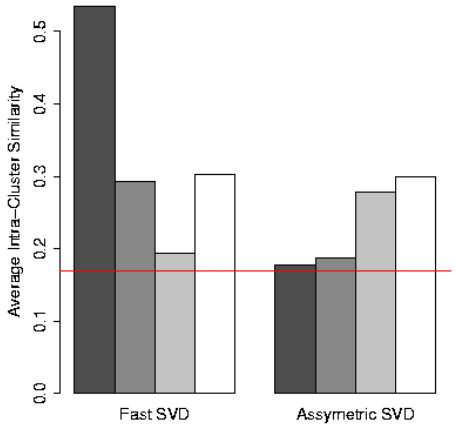
$$\sum_{j=1}^{n} \exp\left(\frac{\|x_i - x_j\|^2}{2\sigma_i^2}\right) = \tau$$

- Optionally evaluate the conductance matrix, C
- Cluster A or C

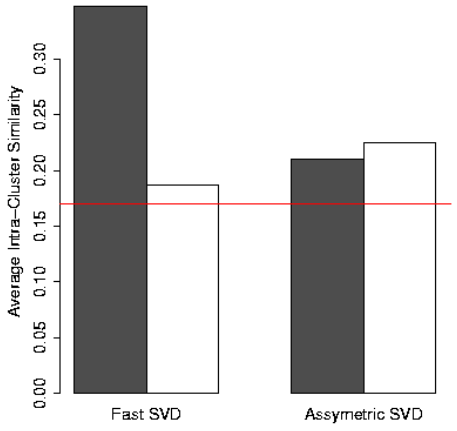
Fischer, I., Poland, J., *Proc. 14th Annual Machine Conf. Of Belgium and the Netherlands*, **2005**, 21-28

Ben-Hur, A. et al., *J. Mach. Learn.*, **2001**, *2*, 125-137

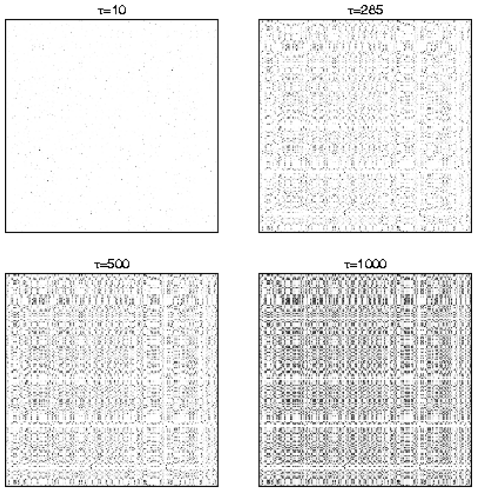
Asymmetric Spectral Clustering



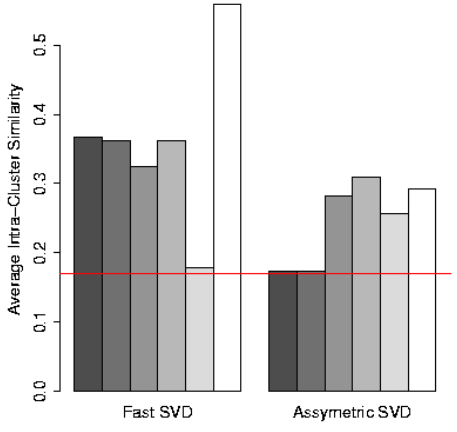
4 Cluster



2 Cluster



- For the AMES dataset the asymmetric approach does not work very well
- Possibly due to poor band amplification



6 Cluster

- Asymmetric distance matrices for different τ
- White indicates minimum affinity and black is the maximum

Summary

- Fast SVD based clustering gives nearly identical results compared to slow SVD based clustering at significantly higher speed
- The average intra-cluster similarities are comparable to PAM and k-means
- SVD based clustering appears to emphasize specific clusters over others
- The algorithm appears to handle correlated and information-poor descriptors well

Acknowledgements

- Dr. Petros Drineas for clarifications about the fast SVD algorithm
- Dr. Igor Fischer for providing code to test the asymmetric affinity approach
- Chemical Computing Group for providing MOE