Data Mining
Cluster Analysis: Advanced Concepts and Algorithms

Lecture Notes for Chapter 9

Introduction to Data Mining

by

Tan, Steinbach, Kumar

Overall Issues

- Subjective Interpretations
  - Data
  - Cluster
  - Application
- Objective Measures
  - Infeasible Computation

Characteristics

- Data
- Clusters
- Clustering Algorithms

K-means vs DBSCAN

- Perform poorly when clusters have widely differing density
- Designed for Euclidean data, but have been extended to handle other type
- Use all attributes

K-means vs DBSCAN

	K-means	DBSCAN
Inclusion of data	Keep all the objects	Discard noisy objects
Algorithmic type	Prototype-based	Density-based
Different sizes and shapes	No	Yes
Data type	Well-defined centroid (mean or median)	Definition of density (Euclidean)
High-dimensional data	Yes	No
Distribution of data	Spherical Gaussian	No assumption
Overlapping clusters	Separate	Merge
Time complexity	O(m)	$O(m^2)$
Stable results	No	Yes
Parameters	K	Eps and MinPts
Problem formation	Optimization	None

Data Characteristics

- High Dimensionality
  - Density: Volume ≈ r^d, r is radius and d is dim
  - Proximity: Attributes
  - Dimensionality Reduction
- Size
  - Well is small and medium
  - Scalable Algorithms
- Sparseness
  - Asymmetric attributes
  - Similarity Measures

Data Characteristics (Cont.)

- Noise and Outliers
  - Atypical points
  - Detection/Deletion
    - Preprocess - DBSCAN
    - During process - Chameleon, SNN, CURE
- Type of Attributes and Data Set
  - Attributes: Categorical, quantitative, binary, discrete, continuous
  - Data Set: Structured, graph, ordered
  - Proximity and Density Measures
  - Data Structure and Algorithms

Data Characteristics (Cont.)

- Scale
  - Different attributes in different scales.
  - Standardization/Normalization
    - Mean and standard deviation: N(0,1)
- Mathematical Properties of the Data Space
  - Mean
  - Density
  - Meaningful Mathematical Operations

Cluster Characteristics

- Prototype-, Graph-, Density-Based

- Data Distribution: Mixture of distributions
  - Mixture Models
- Shape: Arbitrary
  - Chameleon, CURE
- Sizes: Different
- Densities: Various
  - SNN

Cluster Characteristics (Cont.)

- Clusters: Poorly Separated – Touched/Overlapped
  – Fuzzy clustering
- Relationships among Clusters: Points to Clusters
  – Self-organizing maps (SOM)
- Subspace Clusters: Subset of Attributes
  – Feature Selection

General Characteristics

- Order (of Data) Dependence
  - SOM
- Non-determinism: Random Initialization
  - K-means
- Scalability
  - Non-Linear Time
  - Random Access
- Parameter Selection
  - The fewer, the better, but more drastic
  - Trial and Error
  - “Choosing the right number of clusters”

General Characteristics (Cont.)

- Transformation
  - Graph-Based
    - Proximity graph -> connected components
- Optimization
  - Exhaustive approach – Computationally Infeasible
  - Heuristic – Not Optimal
  - Greedy – Local / Not Global

Algorithms

- Prototype-Based
  - Fuzzy
  - Mixture Models
  - Self-Organizing Maps (SOM)
- Density-Based
  - Grid-Based
  - Subspace
  - DENCLUE

Algorithms (Cont.)

- Graph-Based
  - MST
  - OPOSSUM
  - Chameleon
  - Shared Nearest Neighbor
  - Jarvis-Patrick
- Scalable
  - BIRCH
  - CURE

Prototype-Based Clustering

- A cluster is defined by a prototype
  - Prototype of K-means: centroid

- Objects belong to more than one cluster
- A Cluster modeled as a statistical distribution
  - Mean and Variance
- Clusters constrained to fixed relationships
  - Neighborhood

Fuzzy Sets

- An object belongs to a set with a degree of membership between 0 and 1
- Example:
  - 25% “cloudy days”, 75% “non-cloudy days”

- Fuzzy pseudo-partition:
  - Object $\mathbf{x}_i$, Cluster C_j, membership weight w_{ij}
  - $\Sigma_{j=1,k}\, w_{ij} = 1$
  - $0 < \Sigma_{i=1,m}\, w_{ij} < m$

Basic fuzzy c-means algorithm

1. Select an initial fuzzy pseudo-partition (assign values to all the w_{ij})
2. **Repeat**
3. Compute the centroid of each cluster (using the fuzzy pseudo-partition)
4. Re-compute the fuzzy pseudo-partition, w_{ij}
5. **Until** The centroids don't change much

- Computing the Sum of the Squared Error:

$$SSE(C_1, C_2, \ldots, C_k) = \Sigma_{j=1,k}\Sigma_{i=1,m}\, w^p{}_{ij}\, dist(\mathbf{x}_i, \mathbf{c}_j)^2$$

p is the influence of the weights between 1 and ∞

- Initialization
  - Random (K-means)
  - Weights

- Computing Centroids

$$\mathbf{c}_j = \Sigma_{i=1,m}\, w^p_{ij}\, \mathbf{x}_i \,/\, \Sigma_{i=1,m}\, w^p_{ij}$$

- Updating the Fuzzy Pseudo-Partition

For $p = 2$,

$$W_{ij} = 1/dist(\mathbf{x}_i, \mathbf{c}_j)^2 \Big/ \Sigma_{q=1,k} 1/\mathrm{dis}(\mathbf{x}_i, \mathbf{c}_q)^2$$

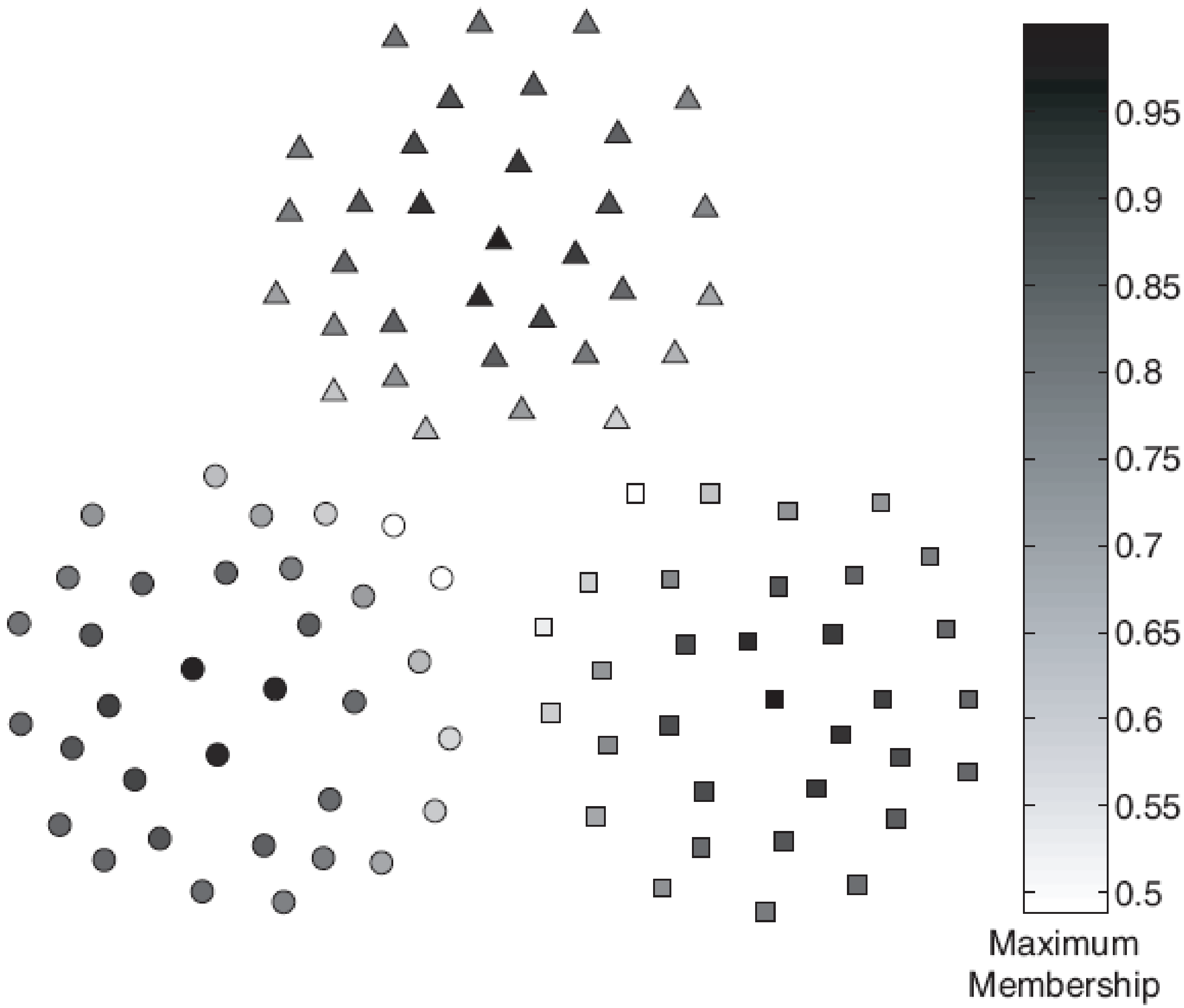
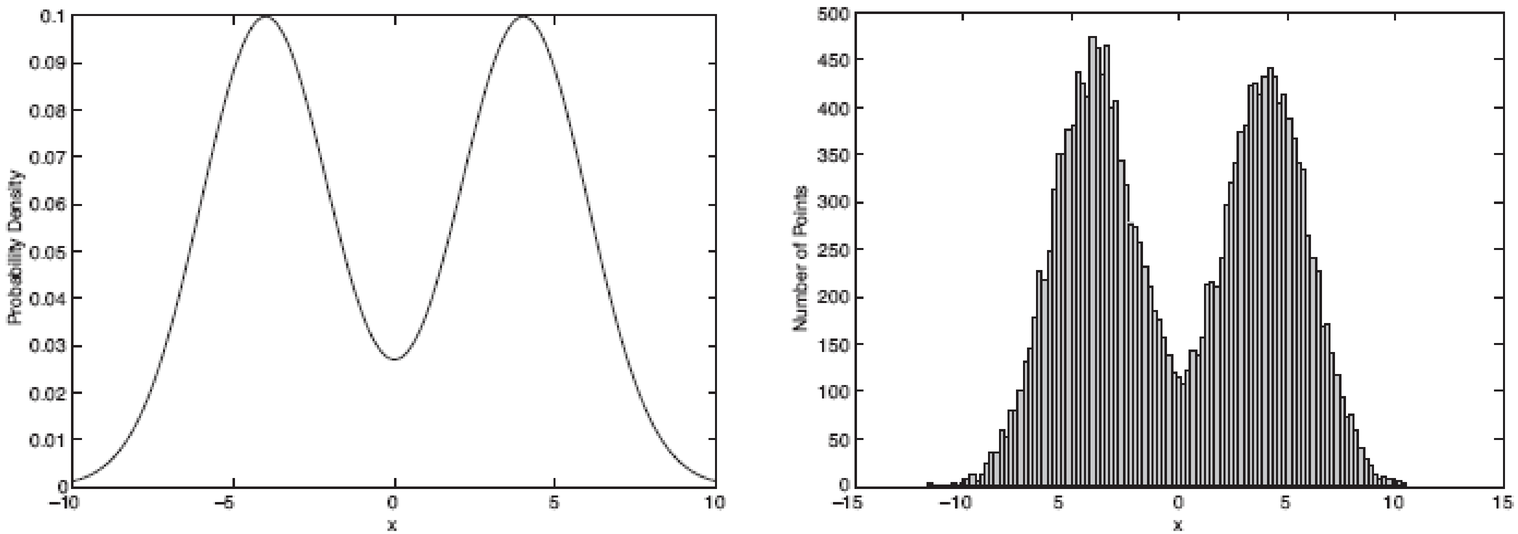


Figure 9.1. Fuzzy c-means clustering of a two-dimensional point set.

Strengths and Limitations

- Similar to K-means
- More computationally intensive

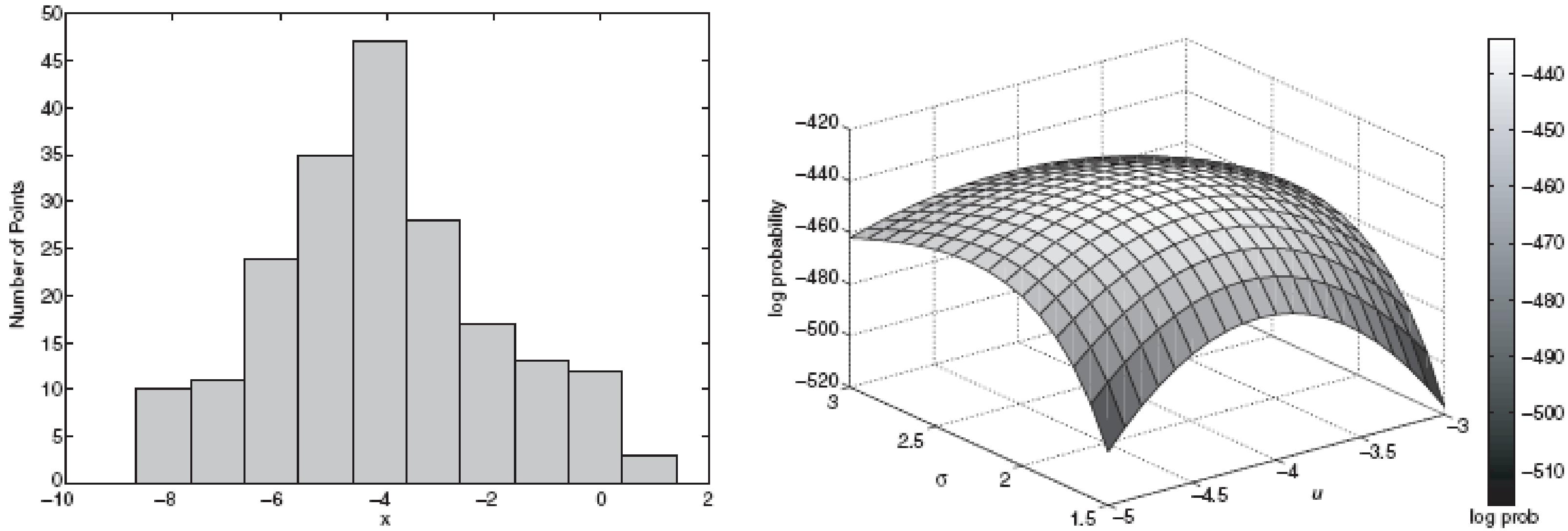
Parameters: Means and Standard Deviation



(a) Probability density function for the mixture model.

(b) 20,000 points generated from the mixture model.

Figure 9.2. Mixture model consisting of two normal distributions with means of -4 and 4, respectively. Both distributions have a standard deviation of 2.



(a) Histogram of 200 points from a Gaussian distribution.

(b) Log likelihood plot of the 200 points for different values of the mean and standard deviation.

Figure 9.3. 200 points from a Gaussian distribution and their log probability for different parameter values.

Expectation-Maximization Algorithm

1. Select an initial set of model parameters
2. **Repeat**
3. Expectation Step:

For each object, calculate the probability

$prob(distribution\ j \mid \mathbf{x}_i, \mathbf{\Theta})$

4. Maximization Step:

Given the probabilities from the expectation step, find the new estimates of the parameters that maximize the expected likelihood

5. **Until** the parameter do not change much

Weights vs Probabilities

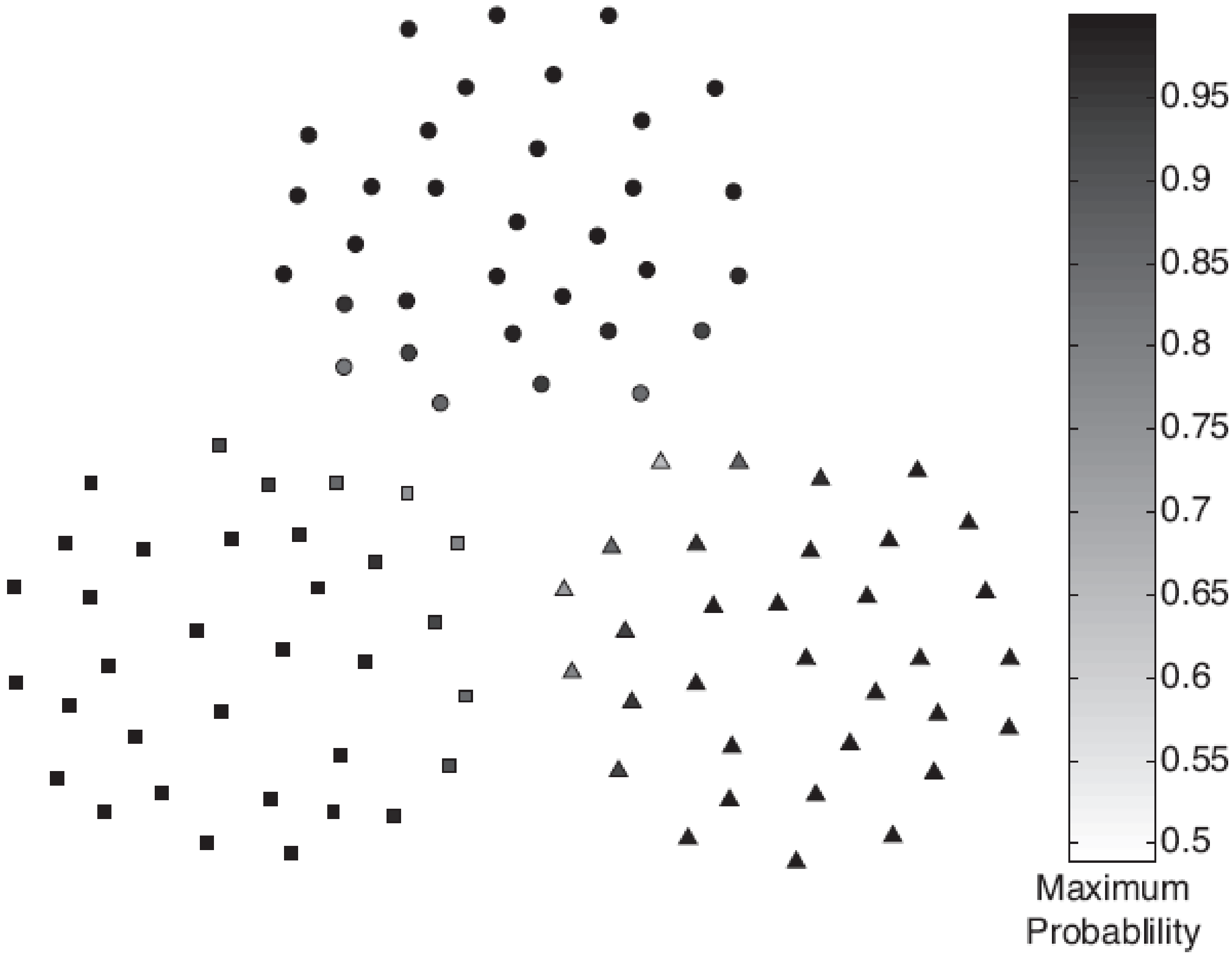


Figure 9.4. EM clustering of a two-dimensional point set with three clusters.

((-4,1), 2) vs ((0,0), 0.5)

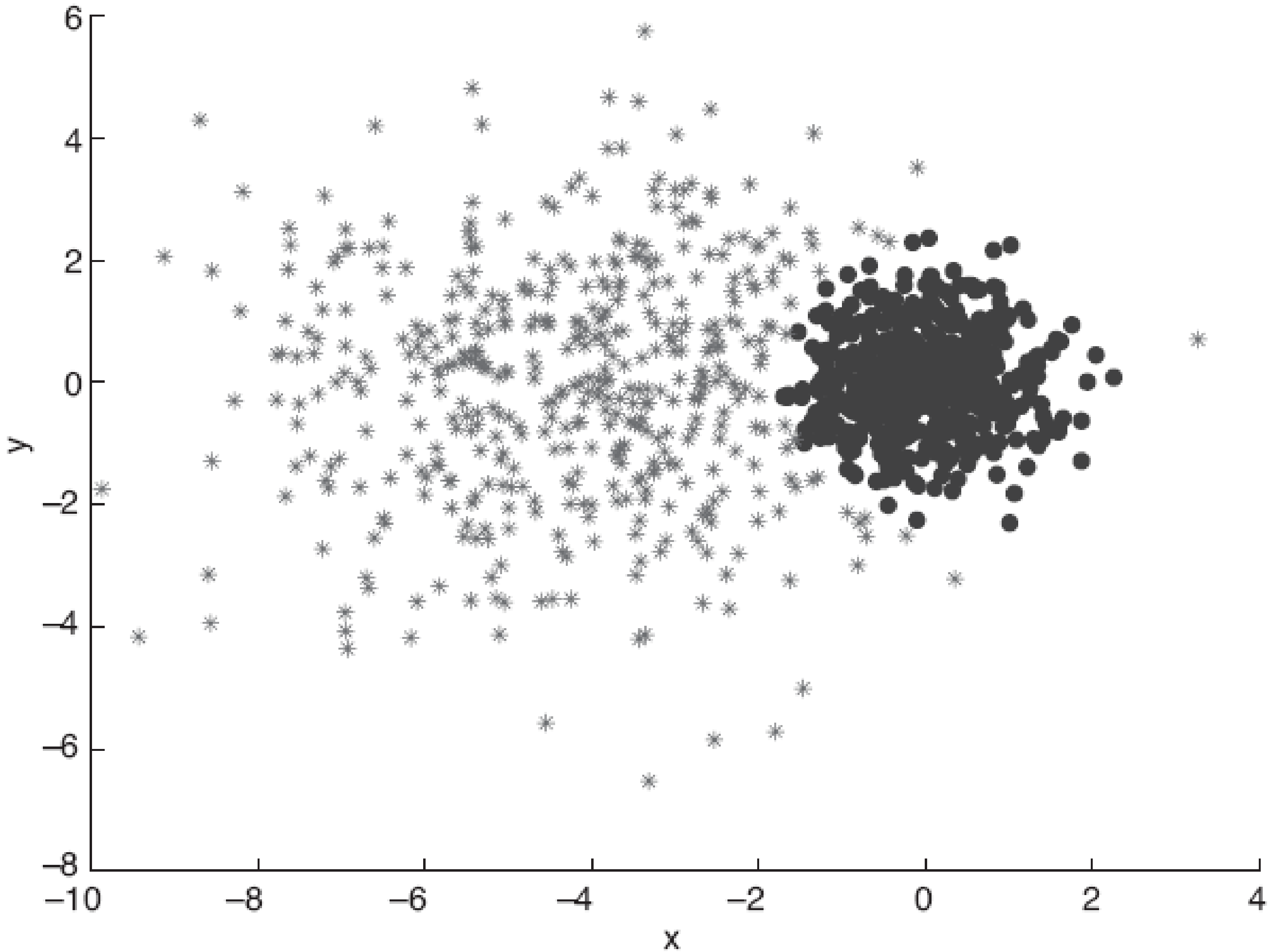
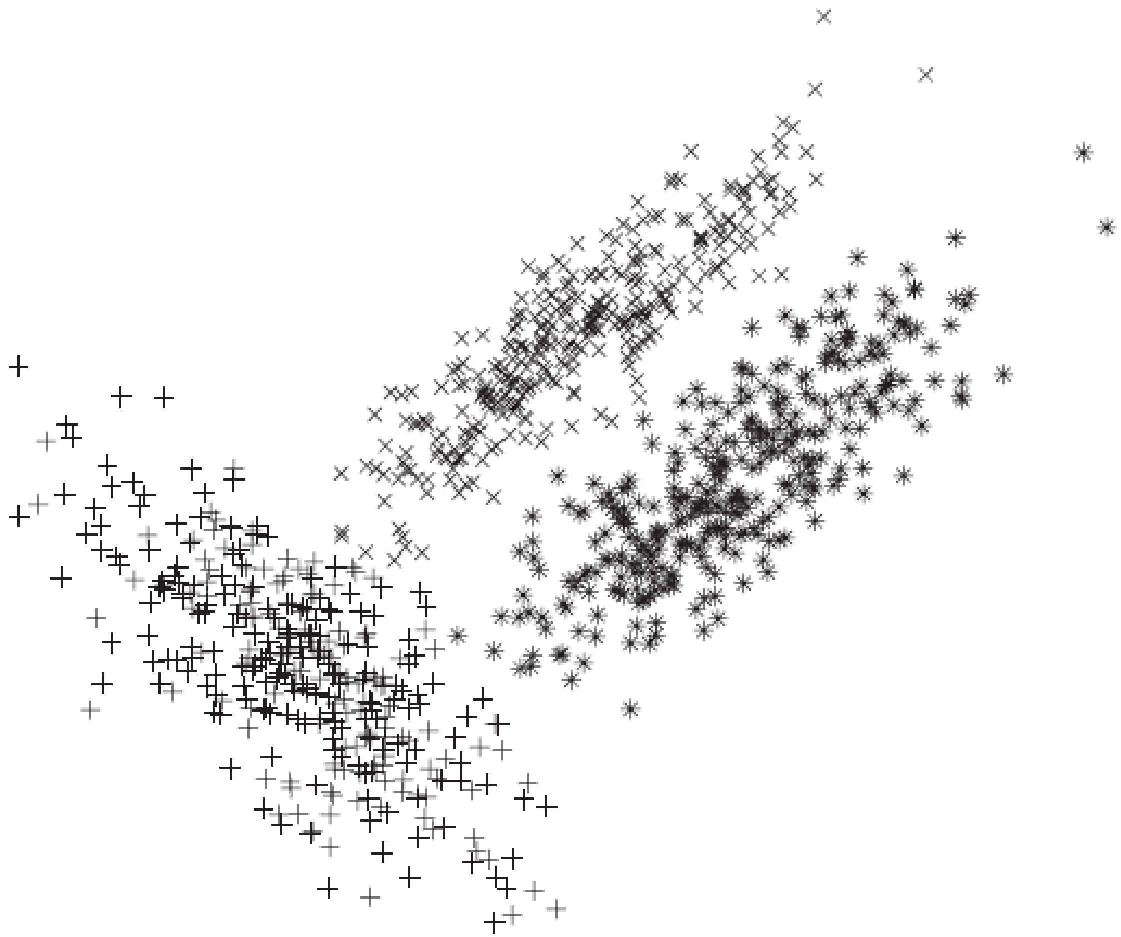


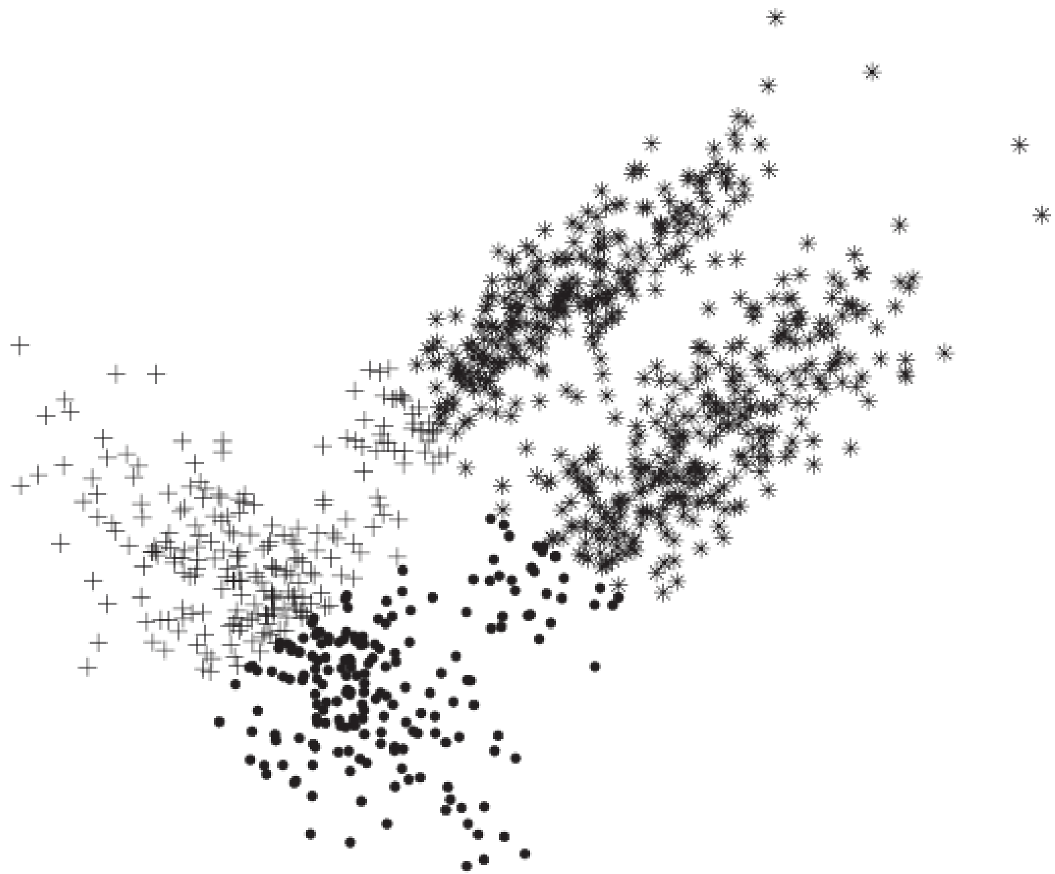
Figure 9.5. EM clustering of a two-dimensional point set with two clusters of differing density.

Highest Probability



(a) Clusters produced by mixture model clustering.

Improperly Handling



(b) Clusters produced by K-means clustering.

Figure 9.6. Mixture model and K-means clustering of a set of two-dimensional points.

Strengths and Limitations

- Positive
  - More general than fuzzy c-means
    - Different Sizes
    - Elliptical shapes
  - Eliminating certain complexity of data by simplification
  - Clusters described by a small number of parameters
- Negative
  - Time Complexity: Large Numbers of Components
  - Data Points: Small or Nearly Co-Linear
  - Model Selection: Bayesian
  - Noise and Outliers

An Example Organized in a Rectangular Lattice

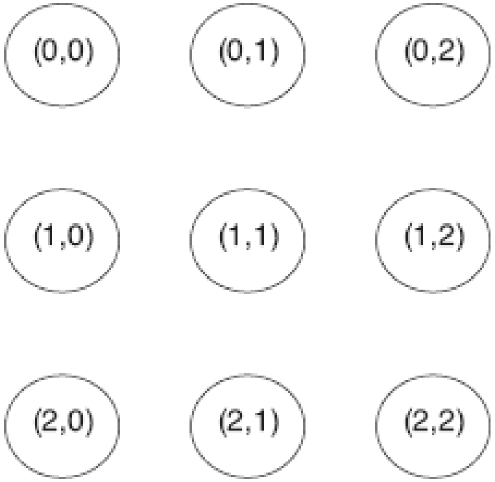


Figure 9.7. Two-dimensional 3-by-3 rectangular SOM neural network.

Kohonen Self-Organizing Feature Maps

1. Initialize the centroids
2. **Repeat**
3. Select the next object (point)
4. Determine the closest centroid to the object
5. Update this centroid and the centroids that are close, i.e., in a specified neighborhood
6. **Until** the centroids do not change much
7. Assign each object to its closest centroid and return the centroids and clusters

Update Step (Line 5)

- $\mathbf{m}_1$, $\mathbf{m}_2$, …, $\mathbf{m}_k$ be the centroids, k = rows * cols
- time step t, current object (point) $\mathbf{p}(t)$
- For time $t + 1$,
  - $\mathbf{m}_j(t + 1) = \mathbf{m}_j(t) + h_j(t)(\mathbf{p}(t) - \mathbf{m}_j(t))$

- $h_j(t)$ diminishes with time, and enforces a neighborhood effect
  - Step function: $\alpha(t)$ if $dist(\mathbf{r}_j, \mathbf{r}_k) <= \tau$, 0 otherwise
  - Gaussian function

Document Data: 3204 / 6 Sections into 4 by 4 Hexagonal Grid

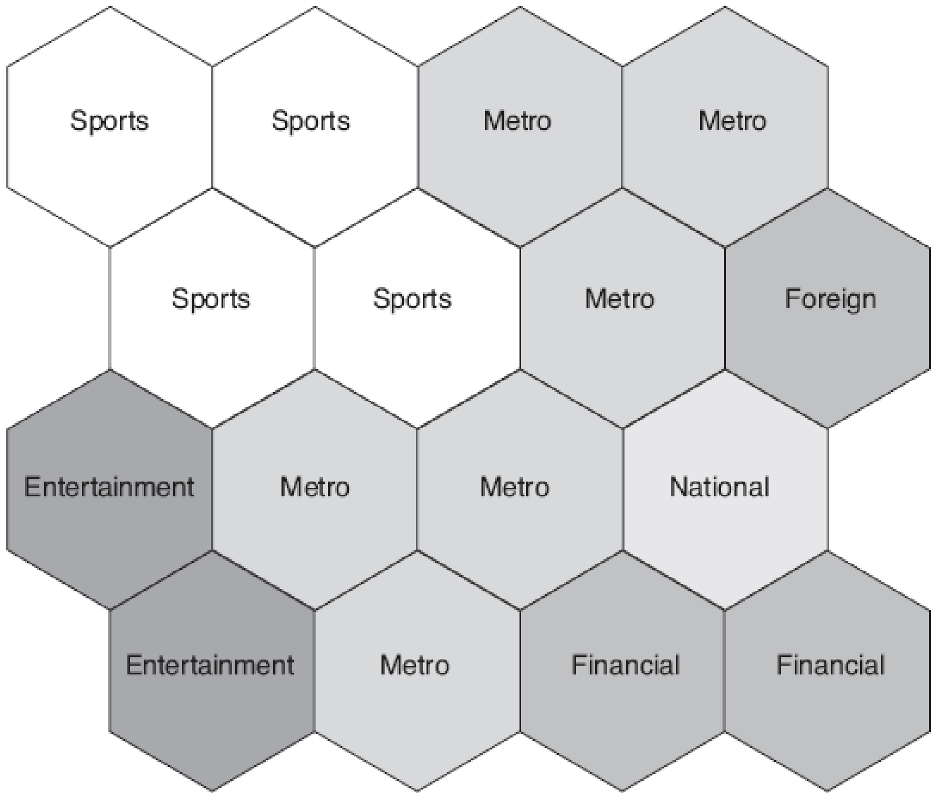
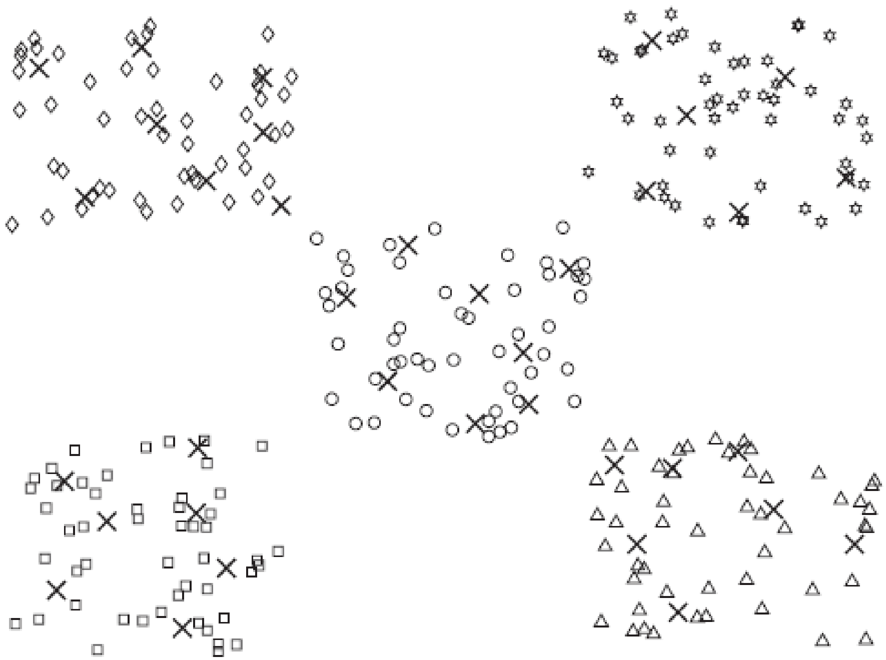


Figure 9.8. Visualization of the relationships between SOM cluster for *Los Angeles Times* document data set.



(a) Distribution of SOM reference vectors (**X**'s) for a two-dimensional point set.

diamond	diamond	diamond	hexagon	hexagon	hexagon
diamond	diamond	diamond	circle	hexagon	hexagon
diamond	diamond	circle	circle	circle	hexagon
square	square	circle	circle	triangle	triangle
square	square	circle	circle	triangle	triangle
square	square	square	triangle	triangle	triangle

(b) Classes of the SOM centroids.

Figure 9.9. SOM applied to two-dimensional data points.

Strengths and Limitations

- Positive
  - Facilitate the interpretation and visualization of the clustering results
- Negative
  - Decisions on parameters, the neighborhood function, the grid type, the number of centroids
  - SOM cluster ≠ natural cluster
  - Lacking of specific objective function
  - No guarantee of convergence