

# Data Mining Techniques

## Chapter 11: Automatic Cluster Detection

Clustering . . . . .	2
k-Means Clustering . . . . .	3
Similarity and distance . . . . .	4
Data preparation . . . . .	5
Distance calculations . . . . .	6
Graphical view . . . . .	7
Hierarchical (agglomerative) clustering . . . . .	8
Family gathering example . . . . .	9
Hierarchical clustering considerations . . . . .	10
Other clustering approaches . . . . .	11
Evaluating clusters . . . . .	12

## Clustering

- Clustering provides a way to learn about the structure of complex data by splitting it up into smaller pieces, each of which can be explained more simply (than the whole).
- Clustering algorithms are undirected DM techniques used to search for groups of records (clusters) composed of records similar to each other.
- Clustering results sometimes feed into follow-up analyses with a more specific focus, e.g., segmentation in marketing.
- Examples: stellar evolution (p. 351), military clothing design (p. 352).

© Iain Pardoe, 2006

2 / 12

## k-Means Clustering

- Algorithm looks for fixed number of clusters ( $k$ ) defined in terms of proximity of data points to each other:
  1. (randomly) select  $k$  data points to be *seeds* for the cluster centers;
  2. assign each record to the “closest” seed (requires some measure of “distance”);
  3. (re)calculate centroids of each cluster;
  4. centroids become seeds for the next iteration (continue until no more changes).
- What does  $k$  mean? Try solutions for different values of  $k$  to see which makes the most sense:
  - smallest within-cluster distances relative to between-cluster distances;
  - most useful from a business perspective.

© Iain Pardoe, 2006

3 / 12

## Similarity and distance

- Clustering algorithms require a numeric measure of similarity or distance (between two records):
  - quantitative variables (interval or ratio/true measures) can use geometric distance metrics, e.g., Euclidean distance;
  - categorical (nominal) and rank (ordinal) variables require other metrics.
- Some formal similarity measures:
  - geometric distance between two points (Euclidean);
  - angle between two vectors (e.g., figure 11.7 on p. 362)—related to correlation;
  - Manhattan distance;
  - for categorical variables,  $\# \text{ features in common} = \# \text{ matches} / \# \text{ variables}$ .

© Iain Pardoe, 2006

4 / 12

## Data preparation

- Scaling for consistency, e.g., standardize/rescale to (0,1) or (-1,+1) or Z-scale.
- For example, consider clustering three employees on experience (in years) and salary (in \$k):
  - A has 2 years experience, salary \$20k;
  - is she closer to B (3 years, salary \$40k) or C (6 years, salary \$30k)?
  - without standardizing, C would be closer (seems incorrect if variables should weigh equally);
  - with standardizing, B would be closer (seems more correct since B has 50% more experience and 100% more salary, but C has 200% more experience and 50% more salary).
- Alternatively, use weights to encode business context knowledge (that one variable is more/less important than others, say).

© Iain Pardoe, 2006

5 / 12

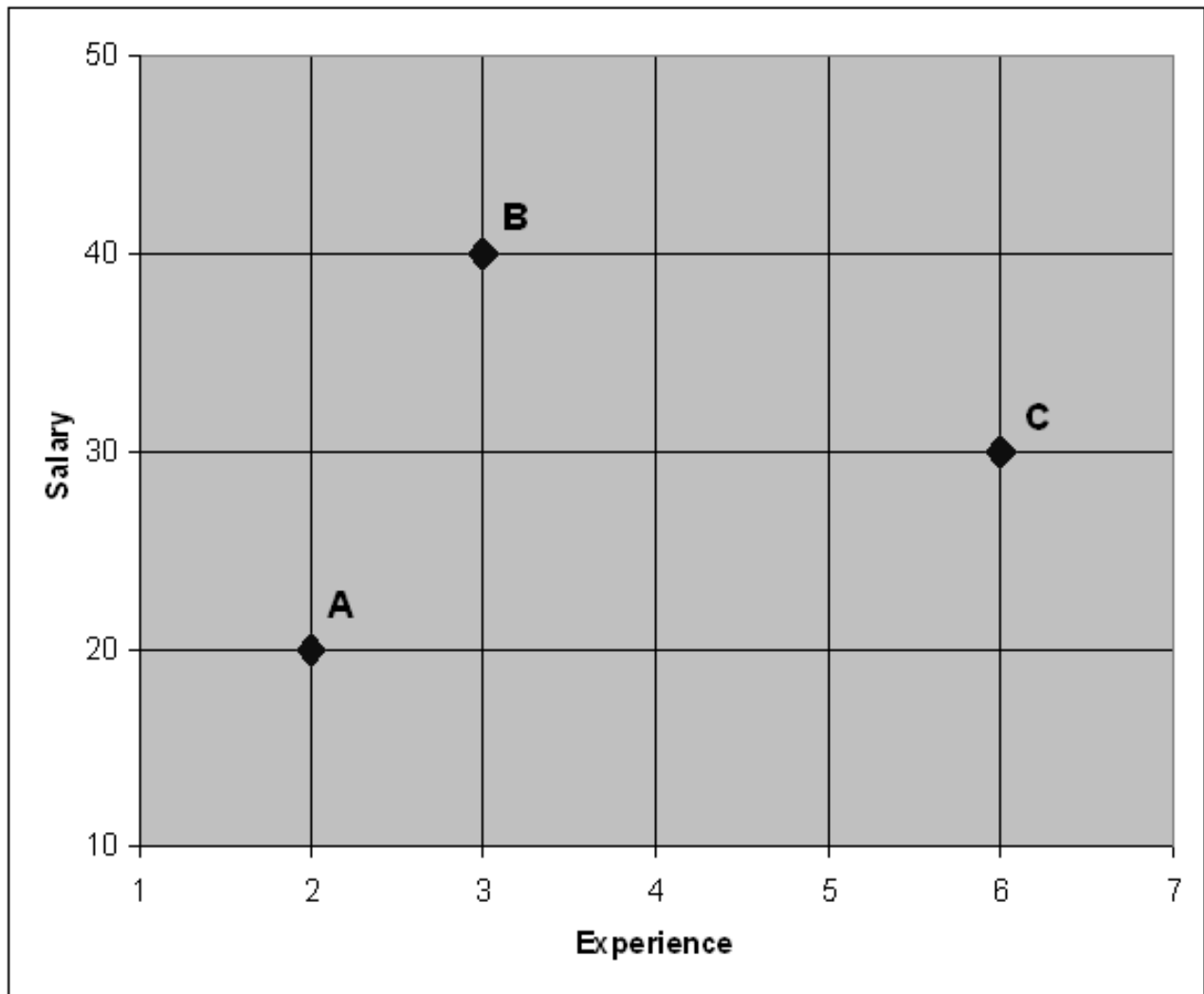
## Distance calculations

	Employee	Experience	Salary	$D_{\text{exp}}^2$	$D_{\text{sal}}^2$	$\sqrt{D_{\text{exp}}^2 + D_{\text{sal}}^2}$
• Unstandardized:	A	2	20			
	B	3	40	1	400	20.0
	C	6	30	16	100	<b>10.8</b>
	Employee	$Z_{\text{exp}}$	$Z_{\text{sal}}$	$D_{\text{exp}}^2$	$D_{\text{sal}}^2$	$\sqrt{D_{\text{exp}}^2 + D_{\text{sal}}^2}$
• Standardized:	A	-0.80	-1			
	B	-0.32	1	0.23	4	<b>2.06</b>
	C	1.12	0	3.69	1	2.17

© Iain Pardoe, 2006

6 / 12

## Graphical view



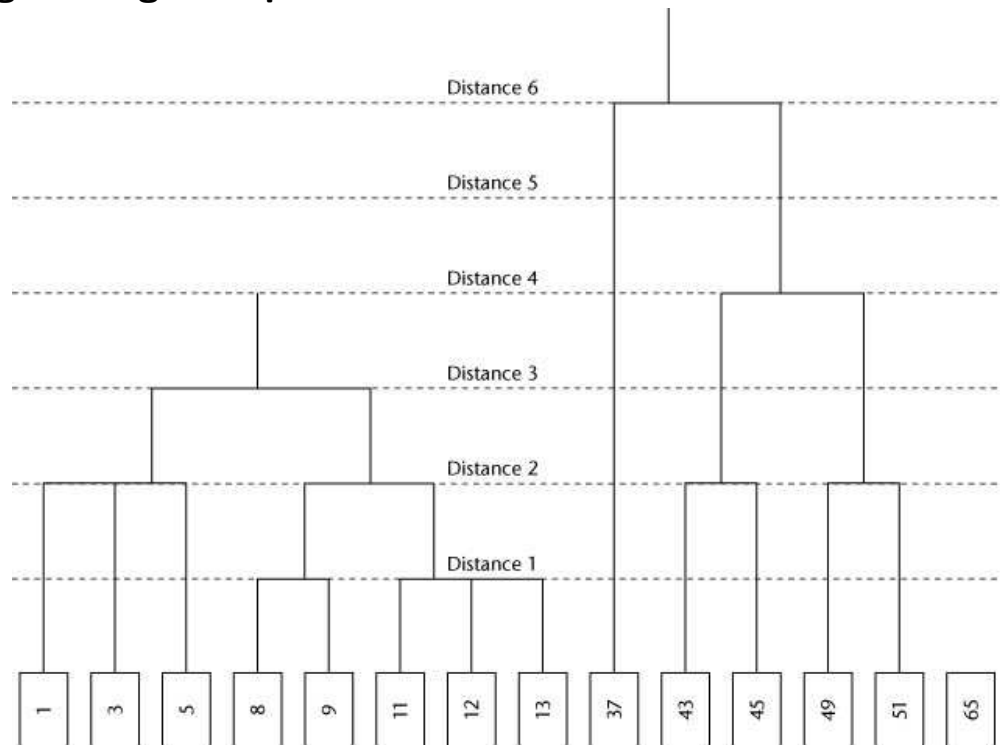
## Hierarchical (agglomerative) clustering

- Start with each point in its own cluster.
- Gradually merge points that are close together into fewer, larger clusters.
- Stop when remaining clusters are all well separated and joining any more would be counter-productive.
- Steps:
  1. calculate all pairwise similarities/distances;
  2. find smallest and join these two points/clusters;
  3. update pairwise similarities and repeat.
- Distance between clusters:
  - single linkage (nearest neighbors);
  - complete linkage (farthest neighbors);
  - centroid distance.
- Dendrogram: tree showing results (e.g., next slide).

© Iain Pardoe, 2006

8 / 12

## Family gathering example



© Iain Pardoe, 2006

9 / 12

## Hierarchical clustering considerations

- Trade-off between too many, pure clusters (low within-cluster distances) and too few, poorly defined clusters (high within-cluster distances):
  - stop joining more clusters when average within-cluster distance would make a large jump after the join.
- Would like final clusters to be distinct and “far apart” (high between-cluster distances).
- Alternative: divisive clustering (agglomerative clustering in reverse, more like decision trees).
- Self-organizing maps: application of clustering using neural networks.

© Iain Pardoe, 2006

10 / 12

## Other clustering approaches

- Disadvantages of k-means and hierarchical:
  - perform poorly with overlapping clusters;
  - sensitive to outliers;
  - each record is either in or out of a given cluster.
- Combined approach:
  - use hierarchical to determine k and cluster centers;
  - follow up with k-means to fine-tune results.
- Gaussian (normal) mixture models:
  - a probabilistic variant of k-means;
  - points not uniquely identified with a given cluster, but rather given probabilities of belonging to various clusters.

© Iain Pardoe, 2006

11 / 12

## Evaluating clusters

- Goal is to find clusters:
  - whose records *within* clusters are very similar;
  - and there are large, actionable differences *between* clusters.
- Inside each cluster: profile standardized variable means across the clusters.
- Outside the clusters: find anomalies that lie outside the “standard” clusters.
- Case study: clustering towns (p. 374–380).

© Iain Pardoe, 2006

12 / 12