Clusteranalysis

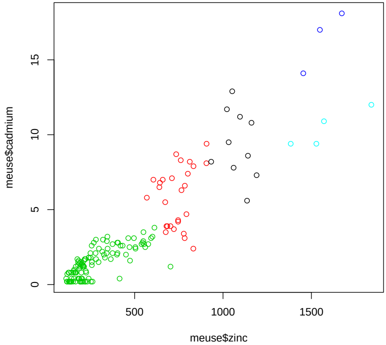
Unsupervised Classification

Kristina Helle

IfGI, WWU Münster

3. December 2008

aim



Split objects in feature space (multivariate) into clusters without prior classification (unsupervised). To achieve

- little variability within clusters.
- big difference between clusters.

unsupervised vs. supervised

	unsupervised	supervised
random variables	$X = (X_1, \dots, X_p)$	$X = (X_1, \dots, X_p)$, Y
data	$x_1, \dots, x_N$, $x_i = (x_{i1}, \dots, x_{iN})$	$(x_1, y_1) = (x_{11}, \dots, x_{1N}, y_1)$, $\dots$, (x_N, y_N)
aim	$P(X)$, where are the values?	$P(Y\|X)$, how is Y for given X?
measure of fit	subjective	reproducing correct Y for given X
examples	PCA, clusteranalysis	regression, nearest neighbour, discriminant analysis

overview

- choice and weighting of variables
- distance measure for points
- clustering algorithms
  - hierarchical
  - iterative
- validation
- spatial clusteranalysis
- software, sources

variables, weighting

- choice of variables (e.g. for vegetation classification blue is of little importance)
- weighting of variables, equal to scaling (stretching one variable puts more weight on it)
  - use of same scale
  - standardisation, replace X_i by $\frac{X_i - E(X_i)}{Var(X_i)}$
    (then equal variance and influence; problem: clusters have probably different means)
  - ...

distance measure between points

L_2^2 $d(x, x') = \sum_{i=1}^{p}(x_i - x_i')^2 = (x - x')^T(x - x')$ (squared euclidean, geometric interpretation, equals Mahalanobis distance if variables uncorrelated and of equal variance)

L_1 $d(x, x') = \sum_{i=1}^{p} |x_i - x_i'|$ (all distances are equally weighted, robust towards outliers)

Mahalanobis $d(x, x') = ((x - x')^T C^{-1}(x - x'))^{1/2}$, C covariance matrix (removes covariance, scale invariant; problem: clusters have probably different covariances)

similarity for categorial variables: frequency of matches ...

- ...

methods

hierarchical (agglomerative, divisive):

- nested clusters of all possible numbers (1 to N) ⇒ clusters depend on clusters on other level
- dendrogram, difference of merged clusters ⇒ aid for choice of good number of clusters

decisions: bottom-up / top-down, measure of cluster distance

pro: aids choice of numbers, hierarchical clustering

iterative:

- optimizes a given number of clusters ⇒ clusters depend on choice of number
- starts with cluster means or clusters ⇒ prior knowledge can be included (closer to supervised), strong dependence on inital clusters

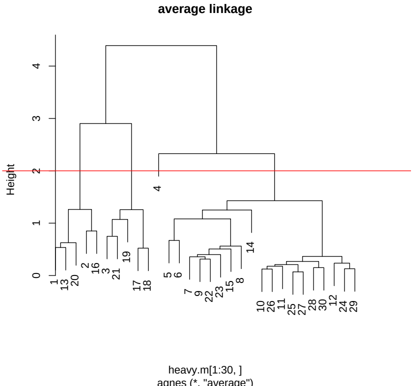
decisions: measure of point-cluster distance

pro: optimization (may end in local optimum)

hierarchic: agglomerative

clusters grow $\Rightarrow$ within cluster heterogenity increases, number of cluster decreases
start: each point is a cluster
repeat: calculate difference of each pair of clusters, merge closest clusters

result: dendrogram



Which clusters were merged?
What was their distance when merging?
Horizontal cut gives clusters for the given minimum cluster difference (e.g. there were four clusters with distance ≥ 2 with 5, 5, 1 and 19 elements respectively).

measures of cluster distance

single linkage $d(G, H) = min_{x \in G, y \in H} d(x, y)$
closest points $\Rightarrow$ chains, detects outliers

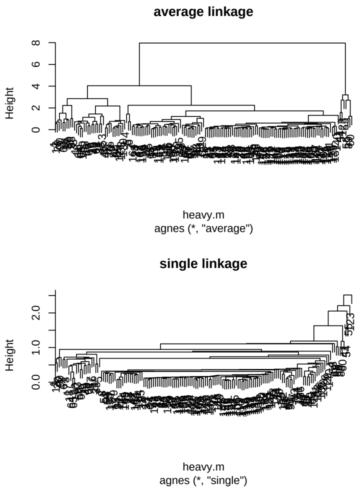
complete linkage $d(G, H) = max_{x \in G, y \in H} d(x, y)$
most distant points $\Rightarrow$ forms compact clusters of equal size

average linkage $d(G, H) = mean_{x \in G, y \in H} d(x, y)$
good compromise between methods above

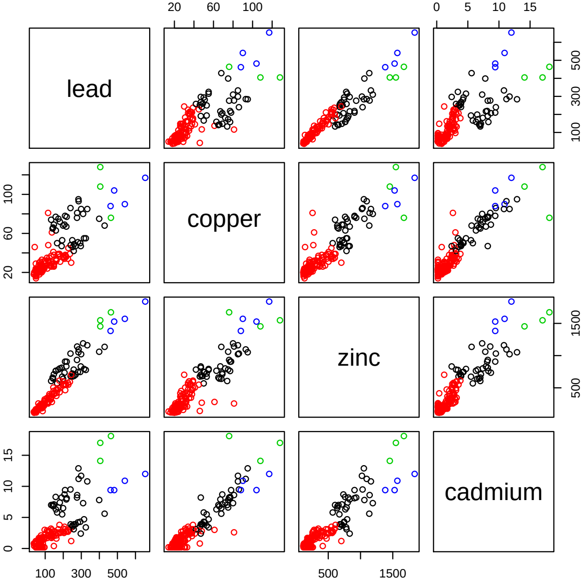
wards method merges the clusters where the measure of heterogenity
$Z(G) = \sum_{l=1}^{k} \sum_{x \in G_l} ||x - \bar{x}_{G_l}||^2$
($G = G_1, \ldots, G_k$ clusters, $\bar{x}_{G_l}$ cluster mean) grows least
$\Rightarrow$ often best, groups of equal size, good aid to find best number of clusters

median, centroid calculate the median / centroid of each cluster merge the closest two
$\Rightarrow$ good geometric interpretation but distance to other clusters may decrease when merging

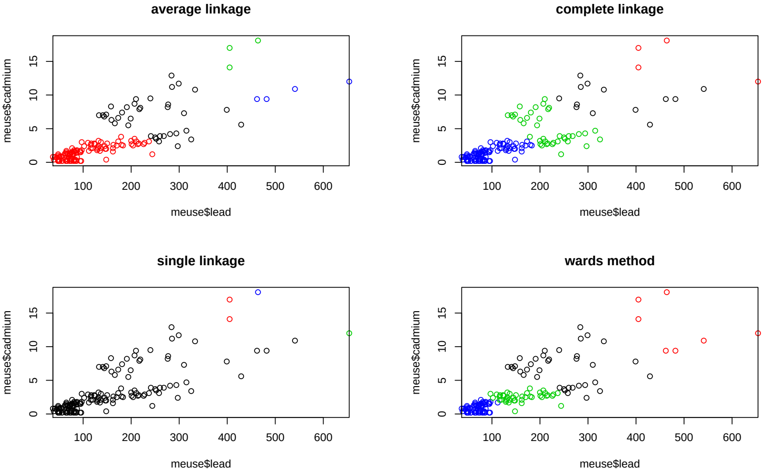
dendrograms



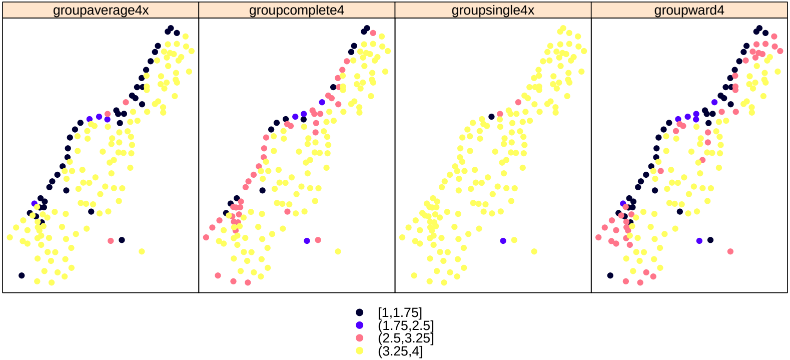
feature space for average linkage



feature space for 4 clusters

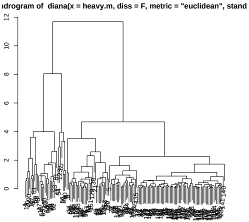


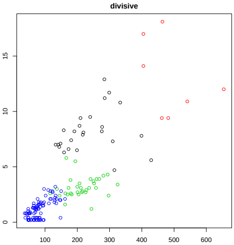
maps for 4 clusters

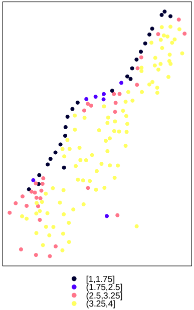


divisive algorithms

start: one cluster containing all points
repeat: split cluster to increase homogenity most
search for points with biggest distance, they are the "germs" of the new clusters
all other points are put togeter with the closer germ
result: dendrogram ⇒ good for few clusters







choice of number of clusters

- interpretability (often 7 or 5)
- "elbows" in the criterion (if next merging is for much more distant clusters / last split increased within cluster homogenity a lot; more accurate: compare with average distances for clusters of uniformly distributed points)

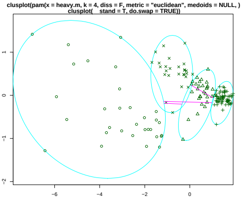
iterative algorithms

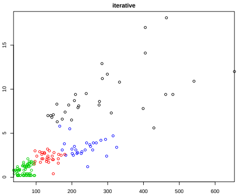
prior choices

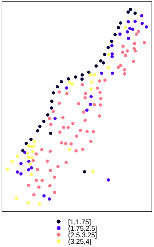
- number of clusters (from hierarchical clustering)
- starting clustercentres / clusters(randomly choosen / from previous hierarchical clustering / from external knowledge)
- measure of point-cluster distance (distance to cluster-mean; distance to cluster-median (central point; robust but high computational effort))

start: n clusters / cluster centres
repeat: for each point find the closest cluster and put it there
update the cluster centres
(alternatively the update can take place after each exchanged point)

results of iterative clustering: k-means







validation

Which clustering is best?

- no objective measure as there is nothing to be reproduced correctly
- minimum within scatter

$$W(C) = \frac{1}{2} \sum_{k=1}^{K} \sum_{C(i)=k} \sum_{C(j)=k} d(x_i, x_j)$$

- or (equivalent) maximal between scatter

$$B(C) = \frac{1}{2} \sum_{k=1}^{K} \sum_{C(i)\neq k} \sum_{C(j)=k} d(x_i, x_j)$$

(C clustering of K clusters, $C(i) = k$ means x_i in cluster k, d distance)

validation

Which clustering is best?

- no objective measure as there is nothing to be reproduced correctly
- minimum within scatter

$$W(C) = \frac{1}{2} \sum_{k=1}^{K} \sum_{C(i)=k} \sum_{C(j)=k} d(x_i, x_j)$$

- or (equivalent) maximal between scatter

$$B(C) = \frac{1}{2} \sum_{k=1}^{K} \sum_{C(i)\neq k} \sum_{C(j)=k} d(x_i, x_j)$$

(C clustering of K clusters, $C(i) = k$ means x_i in cluster k, d distance)

validation

Which clustering is best?

- no objective measure as there is nothing to be reproduced correctly
- minimum within scatter

$$W(C) = \frac{1}{2} \sum_{k=1}^{K} \sum_{C(i)=k} \sum_{C(j)=k} d(x_i, x_j)$$

- or (equivalent) maximal between scatter

$$B(C) = \frac{1}{2} \sum_{k=1}^{K} \sum_{C(i)\neq k} \sum_{C(j)=k} d(x_i, x_j)$$

(C clustering of K clusters, $C(i) = k$ means x_i in cluster k, d distance)

spatial clusteranalysis

analysis of clustering of points in space (epicentres of earthquakes, cases of a disease, ...)
could be applied to the results of a feature-based clusteranalysis to see how the clusters are placed in space

issues

- find cluster centres
- compare clusters of different features (health vs. pollution):
  - similar borders: effects change fast in same regions
  - areas overlapping: correspondance of health cluster and pollution cluster

clusteranalysis with space as a feature: done little (e.g. on fishing, used single linkage to get spatially connected clusters)

other unsupervised classification methods

fuzzy clusteranalysis:
gives for each point membership value for each cluster instead of assigning the point to one cluster only.

self organizing maps:
fit a surface in the feature space and a grid on it so that projection of all points on that surface seperates clusters (points of one cluster are projected to certain cells)

sources

R packages: `(stats)`, `cluster`, `mclust`, `e1071`

literature

- Steinhausen, D.; Langer, K. (1977): Clusteranalyse. de Gruyter, Berlin
- Hastie, T.; Tibshirani, R.; Friedman, J. (2001): The Elements of Statistical Learning, Chapter 14. Springer, New York
- Jacquez, G. (2008): Spatial Cluster Analysis. Chapter 22 In The Handbook of Geographic Information Science, S. Fotheringham and J. Wilson (Eds.). Blackwell Publishing, pages 395-416. (http://www.terraseer.com/pdf/jacquez_ch22_preprint.pdf)
- Mahevas, S.; Ballanger, L.; Trenkel, V. (2008): Cluster analysis of linear model coefficients under contiguity constraints for identifying spatial and temporal fishing effort patterns. Fisheries Research, September 2008, Volume 93 (1-2), Pages 29-38 (http://www.ifremer.fr/docelec/doc/2008/publication-4302.pdf)