Clustering

- Clustering is the process of grouping similar objects into naturally associated subclasses.
- This process results in a set of "clusters" which somehow describe the underlying objects at a more abstract or approximate level.
- The process of clustering is typically based on a "similarity measure" which allows the objects to be classified into separate natural groupings.
- A cluster is then simply a collection of objects that are grouped together because they collectively have a strong internal similarity based on such a measure.
- A similarity measure (or dissimilarity measure) quantifies the conceptual distance between two objects, that is, how alike or disalike a pair of objects are.
  - Determining exactly what type of similarity measure to use is typically a domain dependent problem.

A clustering of a set N is a partition of N, i.e. a set \( C_1, \ldots, C_k \) of subsets of N, such that:
\[
C_1 \cup \ldots \cup C_k = N \quad \text{and} \quad C_i \cap C_j = \emptyset, \text{ for all } i \neq j.
\]

Clustering is used in areas such as:
- medicine, anthropology, economics, data mining
- software engineering (reverse engineering, program comprehension, software maintenance)
- information retrieval
- In general, any field of endeavor that necessitates the analysis and comprehension of large amounts of data may use clustering.

Παράδειγμα ομαδοποίησης αποτελεσμάτων

www.vivisimo.com
Παράδειγμα ομαδοποίησης αποτελεσμάτων

q=Santorini

Τύποι Αλγορίθμων Ομαδοποίησης

• Άναλογα με τη σχέση μεταξύ Ιδιοτήτων και Κλάσεων
  – Monothetic clustering
  – Polythetic clustering

• Άναλογα με τη σχέση μεταξύ Αντικειμένων και Κλάσεων
  – Αποκλειστική (exclusive) ομαδοποίηση
  – Επικαλυπτόμενη (overlapping) ομαδοποίηση
  • Ένα αντικείμενο μπορεί να ανήκει σε παραπάνω από μία κλάση

• Άναλογα με τη σχέση μεταξύ Κλάσεων
  – Χωρίς διάταξη: οι κλάσεις δεν συνδέονται μεταξύ τους
  – Με διάταξη (ιεραρχική): υπάρχουν σχέσεις μεταξύ των κλάσεων

Monothetic vs. Polythetic

• Monothetic
  – Μια κλάση ορίζεται βάσει ενός συνόλου ικανών και αναγκαίων ιδιοτήτων που πρέπει να ικανοποιούν τα μέλη της (Αριστοτελικός ορισμός)

• Polythetic
  – Μια κλάση ορίζεται βάσει ενός συνόλου ιδιοτήτων \( \Phi = \phi_1, \ldots, \phi_n \).
  • Κάθε μέλος της κλάσης πρέπει να έχει ένα μεγάλο αριθμό των ιδιοτήτων \( \Phi \).
  • Κάθε \( \phi \) του \( \Phi \) χαρακτηρίζει πολλά αντικείμενα
  • Δεν είναι αναγκαίο να υπάρχει μια \( \phi \) που να ικανοποιείται από όλα τα μέλη της κλάσης

Στην ΑΠ, έχει δοθεί έμφαση σε αλγόριθμους για αυτόματη παραγωγή polythetic classifications.

Metriκes Συναρτήσεις (Association)

• Μετρικές συναρτήσεις ομοιότητας, συσχέτισης (απόστασης):
  – Pairwise measure
  – Similarity increases as the number or proportion of shared properties increase
  – Typically normalized between 0 and 1
  – \( S(X,Y) = S(Y,X) \)

• Παραδείγματα μετρικών ομοιότητας
  – Οι περισσότερες είναι κανονικοποιημένες εκδόσεις του \( |X \cap Y| \) ή του εωςετερικού γνώμονα \( (\text{εάν έχουμε βεβαρημένους όρους}) \)
  – Dice’s coefficient \( 2|X \cap Y|/(|X| + |Y|) \)
  – Jaccard’s coefficient \( |X \cap Y|/|X \cup Y| \)
  – Cosine correlation
  – Δεν υπάρχει το «καλύτερο» μέτρο (που να δίνει τα καλύτερα αποτελέσματα σε κάθε περίπτωση)
Παραδείγματα Μέτρων για Έγγραφα

- Dice’s coefficient: \( \frac{2 \sum W_{i,j} \cdot W_{m,n}}{\sum W_{i,j} + \sum W_{m,n}} \)
- Jaccard’s coefficient: \( \frac{\sum W_{i,j} \cdot W_{m,n}}{\sum W_{i,j} + \sum W_{m,n} - \sum W_{i,j} \cdot W_{m,n}} \)

Ωμαδοποίηση ως τρόπος Αναπαράστασης (Clustering as Representation)

- Η ομαδοποίηση είναι μια μορφή μη επιτηρούμενης μάθησης (unsupervised learning)
  - Για εμφάνιση της υποκείμενης δομής και κλάσεων

- Η ομαδοποίηση είναι μια μορφή μετασχηματισμού της αναπαράστασης (representation transformation)
  - Τα έγγραφα παριστάνονται όχι μόνο βάσει των όρων αλλά και βάσει των κλάσεων στις οποίες μετέχουν

- Η ομαδοποίηση μπορεί να θεωρηθεί ως μια τεχνική για μείωση των διαστάσεων (dimensionality reduction)
  - Ειδικά το term clustering
  - Latent Semantic Indexing, Factor Analysis είναι παρόμοιες τεχνικές

Ωμαδοποίηση για βελτίωση της απόδοσης (Clustering for Efficiency)

Method:
- 1/ Cluster documents,
- 2/ Represent clusters by mean or average document,
- 3/ compare query to cluster representatives

Ωμαδοποίηση για βελτίωση της Αποτελεσματικότητας (Clustering for Effectiveness)

- By transforming representation, clustering may also result in more effective retrieval
- Retrieval of clusters makes it possible to retrieve documents that may not have many terms in common with the query
  - E.g. LSI

Document Clustering Approaches

- Graph Theoretic
  - Defines clusters based on a graph where documents are nodes and edges exist if similarity greater than some threshold
  - Require at least \( O(n^2) \) computation
  - Naturally hierarchic (agglomerative)
  - Good formal properties
  - Reflect structure of data
- Based on relationships to cluster representatives or means
  - Define criteria for separability of cluster representatives
  - Typically have some measure of goodness of cluster
  - Require only \( O(n \log n) \) or even \( O(n) \) computations
  - Tend to impose structure (e.g. number of clusters)
  - Can have undesirable properties (e.g. order dependence)
  - Usually produce partitions (no overlapping clusters)

Criteria of Adequacy for Clustering Methods

- The method produces a clustering which is unlikely to be altered drastically when further objects are incorporated (stable under growth)
- The method is stable in the sense that small errors in the description of objects lead to small changes in the clustering
- The method is independent of the initial ordering of the objects
Graph Theoretic Clustering Algorithms

Graph Clustering

- Graph clustering deals with the problem of clustering a graph
  - grouping similar nodes of a graph into a set of subgraphs

Quality criteria for graph clustering methods

- Graph clustering methods should produce clusters with **high cohesion** and **low coupling**
  - **high cohesion**: there should be many internal edges
  - **low “cut size”**: The cut size (else called external cost) of a clustering measures how many edges are external to all sub-graphs, that is, how many edges cross cluster boundaries.
  - **Uniformity of cluster size** is also often desirable.
    - A uniform graph clustering is where $|C_i| \approx |C_j|$ for all $i, j \in \{1..k\}$

Quality Measures for Graph Clustering

- There are several. One well known is the CC measure (Coupling-Cohesion measure)

$$CC = \frac{|E^{in}| - |E^{ex}|}{|E|}$$

- $E^{in}$: the “internal” edges: those that connect nodes of the same cluster
- $E^{ex}$: the “external” edges: those that cross cluster boundaries
- maximum value of CC: 1
  - when all edges are internal
- minimum value of CC: -1
  - when all edges are external

Example

- A uniform graph clustering is where $|C_i| \approx |C_j|$ for all $i, j \in \{1..k\}$

Example

- There are several. One well known is the CC measure (Coupling-Cohesion measure)

$$CC = \frac{|E^{in}| - |E^{ex}|}{|E|}$$

- $E^{in}$: the “internal” edges: those that connect nodes of the same cluster
- $E^{ex}$: the “external” edges: those that cross cluster boundaries
- maximum value of CC: 1
  - when all edges are internal
- minimum value of CC: -1
  - when all edges are external
Hierarchical Graph Clustering

- The clusters of the graph can be clustered themselves to form a higher level clustering, and so on.
- A hierarchical clustering is a collection of clusters where any two clusters are either disjoint or nested.

Hierarchical Clustered Graph

A Hierarchical Clustered Graph (HCG) is a pair $(G,T)$ where $G$ is the underlying graph, and $T$ is a rooted tree such that the leaves of $T$ are the nodes of $G$. (the tree $T$ represents an inclusion relationship: the leaves of $T$ are nodes of $G$, the internal nodes of $T$ represent a set of graph nodes, i.e. a cluster)

Graph Theoretic Clustering Approaches

- Given a graph of objects connected by links that represent similarities greater than some threshold, the following cluster definitions are straightforward:
  - Connected Component: subgraph such that each node is connected to at least one other node in the subgraph and the set of nodes is maximal with respect to that property
    - Called single link clusters
  - Maximal complete subgraph: subgraph such that each node is connected to every other node in the subgraph (clique)
    - Complete link clusters
- Others are possible and very common:
  - Average link: each cluster member has a greater average similarity to the remaining members of the cluster than it does to all members of any other cluster

Hierarchical Clustering

- Build a tree-based hierarchical taxonomy (dendrogram) from a set of unlabeled examples.
- Recursive application of a standard clustering algorithm can produce a hierarchical clustering.

Hierarchical Clustering Methods

- Agglomerative (συσσώρευσης) (bottom-up) methods start with each example in its own cluster and iteratively combine them to form larger and larger clusters.
- Divisive (διαίρεσης) (partitional, top-down) separate all examples immediately into clusters.

Implied Edges

Implied edges: edges between the internal nodes. Two clusters are connected iff the nodes that they contain are related. Multiple implied edges (between the same pair of clusters) can be ignored or summed up to form weighted implied edges. Thresholding can be applied in order to filter out some implied edges.

A Hierarchical Compound Graph is a triad $(G,T,I)$ where $(G,T)$ is a hierarchical clustered graph (HCG), and $I$ the set of implied edges set.

An hierarchical (agglomerative) clustering algorithm

1. Βαλε κάθε έγγραφο σε ένα διαφορετικό cluster
2. Υπολόγισε την ομοιότητα μεταξύ όλων των cluster

Hierarchical Clustering Methods

- Agglomerative (συσσώρευσης) (bottom-up) methods start with each example in its own cluster and iteratively combine them to form larger and larger clusters.
- Divisive (διαίρεσης) (partitional, top-down) separate all examples immediately into clusters.
An hierarchical (agglomerative) clustering algorithm

1. Βάλε κάθε έγγραφο σε ένα διαφορετικό cluster
   \[ C := \emptyset ; \text{For } i=1 \text{ to } n \quad C := C \cup \{ d_i \} \]

2. Υπολογίστε την ομοιότητα μεταξύ όλων των εγγράφων cluster
   Compute \( \text{SIM}(c,c') \) for each \( c, c' \in C \)

3. Βρες το ζεύγος \( \{C_u,C_v\} \) με την υψηλότερη (inter-cluster) ομοιότητα

4. Συγχώνεις τα clusters \( C_u, C_v \)

5. Επανάλαβε (από το βήμα 2) όσο ότι να καταλήξουμε να έχουμε 1 μόνο cluster

6. Επέστρεψε την ιεραρχία των clusters (to ιστορικό των συγχωνεύσεων)

---

**Dendogram or Cluster Hierarchy**

**Single Link Example**

**Complete Link Example**
Σύγκριση

- **Single-link**
  - is provably the only method that satisfies criteria of adequacy
  - however it produces "long, straggly (ανάκατα) string" that are not good clusters
    - Only a single-link required to connect
- **Complete link**
  - produces good clusters (more "tight," spherical clusters), but too few of them (many singletons)
- **Average-link**
  - For both searching and browsing applications, average-link clustering has been shown to produce the best overall effectiveness

Ward’s method
(an alternative to single/complete/average link)

- **Cluster merging:**
  - Merge the pair of clusters whose merger minimizes the increase in the total within-group error sum of squares, based on the Euclidean distance between centroids
- **Remarks:**
  - this method tends to create symmetric hierarchies

Computing the Document Similarity Matrix

\[
\begin{array}{cccc}
  d_1 & s_{12} & s_{13} & \ldots \ldots \ldots & s_{1n} \\
  d_2 & s_{22} & & & s_{2n} \\
  \vdots & & & & \vdots & \vdots \\
  d_n & s_{n2} & s_{n3} & \ldots \ldots & s_{nn} \\
  d_1 & d_2 & \ldots & d_{n-1} & d_n \\
\end{array}
\]

- Empty because \( \text{sim}(X,Y)=\text{sim}(Y,X) \)

- Optimization: Compute \( \text{sim}(d_i,d_j) \) only if \( d_i \) and \( d_j \) have at least one term in common (otherwise it is 0)
  - This is done by exploiting the inverted index

Clustering algorithms based on relationships to cluster representatives or means
(Fast Partition Algorithms)

**Fast Partition Methods**

- **Single Pass**
  - Assign the document \( d_1 \) as the representative (centroid, mean) for \( c_1 \)
  - For each \( d_i \), calculate the similarity \( Sim \) with the representative for each existing cluster
    - If \( \text{SimMax} \) is greater than threshold value \( \text{simThres} \), add the document to the corresponding cluster and recalculate the cluster representative; otherwise use \( d_i \) to initiate a new cluster
    - If a document \( d_i \) remains to be clustered, repeat

- **K-means (or reallocation methods)**
  - Select \( K \) cluster representatives
    - For \( i = 1 \) to \( N \), assign \( d_i \) to the most similar centroid
    - For \( j = 1 \) to \( K \), recalculate the cluster centroid \( c_j \)
    - Repeat the above steps until there is little or no change in cluster membership
- **Issues:**
  - How should \( K \) representatives be chosen?
  - Numerous variations on this basic method
    - cluster splitting and merging strategies
    - criteria for cluster coherence
    - seed selection
**K-Means**

- Assumes instances are real-valued vectors.
- Clusters based on centroids, center of gravity, or mean of points in a cluster, $c$:
  
  \[ \overline{\mu(c)} = \frac{1}{|c|} \sum_{x \in c} x \]

- Reassignment of instances to clusters is based on distance to the current cluster centroids.

**Nearest Neighbor Clusters**

- Cluster each document with its $k$ nearest neighbors
- Produces overlapping clusters
- Called “star” clusters by Sparck Jones
- Can be used to produce hierarchic clusters
- cf. “documents like this” in web search

**Complexity Remarks**

- Computing the matrix with document similarities: $O(n^2)$
- Simple reallocation clustering method with $k$ clusters: $O(kn)$
  
  - πολύ γρήγορο από τους αλγορίθμους για ιεραρχική ομαδοποίηση
- Agglomerative or Divisive Hierarchical Clustering:
  
  - απαιτεί $n-1$ συγχωνεύσεις/διαιρέσεις
  - η πολυπλοκότητα του είναι τουλάχιστον $O(n^2)$

**Cluster Searching**

**Top-down searching:**
- start at top of cluster hierarchy, choose one of more of the best matching clusters to expand at the next level
  - tends to get lost

**Bottom-up searching:**
- create inverted file of “lowest level” clusters and rank them
  
  - more effective
  - indicates that highest similarity clusters (such as nearest neighbor) are the most useful for searching

- After clusters are retrieved in order, documents in those clusters are ranked
- Cluster search produces similar level of effectiveness to document search, finds different relevant documents
Human Clustering

• Ερωτήματα
  – Is there a clustering that people will agree on?
  – Is clustering something that people do consistently?
  – Yahoo suggests there’s value in creating categories
    • Fixed hierarchy that people like
• “Human performance on clustering Web pages”
  – Macskassy, Banerjee, Davison, and Hirsh (Rutgers)
  – KDD 1998, and extended technical report
• Αποτελέσματα: Μάλλον δεν υπάρχει μεγάλη συμφωνία
  • γενικά προτίμηση σε μικρά clusters
  • άλλοι χρήστες προτίμονται/δημιουργούν επικαλυπτόμενα, άλλοι αποκλειστικά clusters
  • τα περιεχόμενα των clusters διέφεραν αρκετά
  • γενική αμβλύπηση (ανεξαρτήτου επερώτησης) δεν φαίνεται να είναι πολύ χρήσιμη

Text Clustering

• HAC and K-Means have been applied to text in a straightforward way.
• Typically use normalized, TF/IDF-weighted vectors and cosine similarity.
• Optimize computations for sparse vectors.
• Applications:
  – During retrieval, add other documents in the same cluster as the initial retrieved documents to improve recall.
  – Clustering of results of retrieval to present more organized results to the user (e.g., vivisimo search engine)
  – Automated production of hierarchical taxonomies of documents for browsing purposes (like Yahoo & DMOZ).

Clustering vs Classification

• Clustering
  – Unsupervised
  – Input
    • Clustering algorithm
    • Similarity measure
    • Number of clusters (e.g. in K Means)
  – No specific information for each document
• Classification (or categorization)
  – Supervised
    – Each document is labeled with a class
    – Build a classifier that assigns documents to one of the classes

Supervised vs Unsupervised Learning

• This setup is called supervised learning in the terminology of Machine Learning
• In the domain of text, various names
  – Text classification, text categorization
  – Document classification/categorization
  – “Automatic” categorization
  – Routing, filtering ...
• In contrast, the earlier setting of clustering is called unsupervised learning
  – Presumes no availability of training samples
  – Clusters output may not be thematically unified.
Text Categorization Examples

Assign labels to each document or web-page:

- Labels are most often topics such as Yahoo-categories
  e.g., "finance," "sports," "news>world>asia>business"
- Labels may be genres
  e.g., "editorials" "movie-reviews" "news"
- Labels may be opinion
  e.g., "like", "hate", "neutral"
- Labels may be domain-specific binary
  e.g., "interesting-to-me", "not-interesting-to-me"
  e.g., "spam": "not-spam"
  e.g., "contains adult language": "doesn’t"

Classification Methods

- Manual classification
  - Used by Yahoo!, Looksmart, about.com, ODP, Medline
  - very accurate when job is done by experts
  - consistent when the problem size and team is small
  - difficult and expensive to scale
- Automatic document classification
  - Hand-coded rule-based systems
    - Used by spam filters, Reuters, CIA, Verity, ...
    - E.g., assign category if document contains a given boolean combination of words
    - Commercial systems have complex query languages (everything in IR query languages + accumulators)
    - Accuracy is often very high if a query has been carefully refined over time by a subject expert
    - Building and maintaining these queries is expensive

Classification Methods (II)

- Supervised learning of document-label assignment function
  Many new systems rely on machine learning (Autonomy, Kana, MSN, Verity, Enkata, ...)
  - k-Nearest Neighbors (simple, powerful)
  - Naive Bayes (simple, common method)
  - Support-vector machines (new, more powerful)
  - ... plus many other methods
  - No free lunch: requires hand-classified training data
  - But can be built (and refined) by amateurs