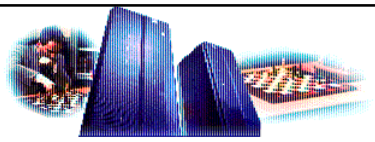
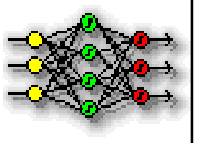




Sistemi Intelligenti
Learning and Clustering

Alberto Borghese

Università degli Studi di Milano
Laboratorio di Sistemi Intelligenti Applicati (AIS-Lab)
Dipartimento di Scienze dell'Informazione
alberto.borghese@unimi.it

Riassunto

- I tipi di apprendimento
- Il clustering
- Le feature
- Clustering gerarchico: algoritmi agglomerativi
- Clustering gerarchico: algoritmi divisivi

I vari tipi di apprendimento

$x(t+1) = f[x(t), a(t)]$ Ambiente
$a(t) = \mathbf{g}[x(t)]$ Agente

Supervisionato (learning with a teacher). Viene specificato per ogni pattern di input, il pattern desiderato in output.

Semi-Supervisionato. Viene specificato solamente per **alcuni** pattern di input, il pattern desiderato in output.

Non-supervisionato (learning without a teacher). Estrazione di similitudine statistiche tra pattern di input. Clustering. Mappe neurali.

Apprendimento con rinforzo (reinforcement learning, learning with a critic). L'ambiente fornisce un'informazione puntuale, di tipo qualitativo, ad esempio success or fail.

I gruppi di algoritmi

Clustering (data mining)

Classification

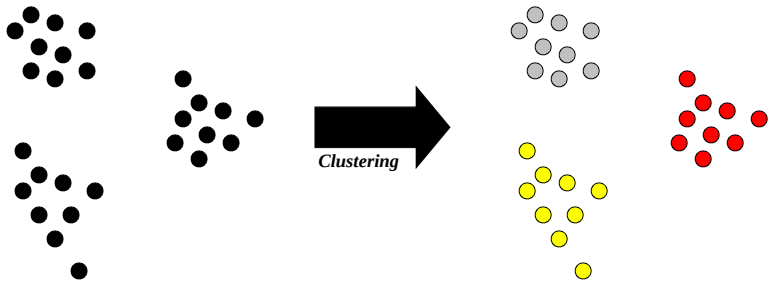
Predictive regression

Riassunto

- I tipi di apprendimento
- Il clustering
- Le feature
- Clustering gerarchico: algoritmi agglomerativi
- Clustering gerarchico: algoritmi divisivi

Clustering

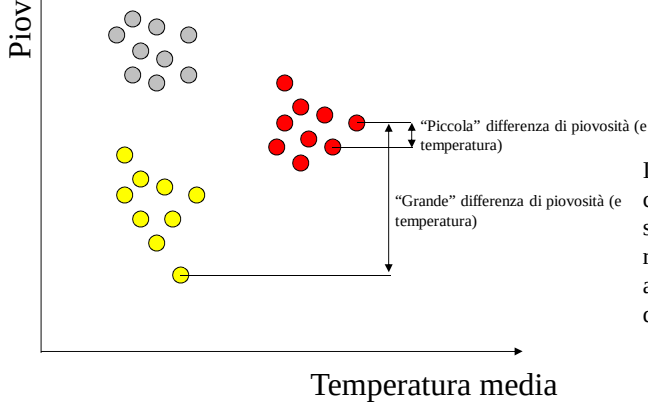
- Clustering: raggruppamento degli “oggetti” in classi omogenee tra loro.
  - Raggruppamento per colore
  - Raggruppamento per forme
  - Raggruppamento per tipi
  - …..



Novel name: **data mining**

Clustering

Processo attraverso il quale i dati (pattern, vettori) vengono organizzati in cluster, basata sulla similarità. L'elaborazione verrà poi effettuata sui prototipi che rappresentano ciascun cluster.



I pattern appartenenti ad un cluster valido sono più simili l'uno con l'altro rispetto ai pattern appartenenti ad un cluster differente.

Esempio di clustering

Clustering -> Indicizzazione

Ricerca immagini su WEB.

Il clustering per…

… Confermare ipotesi sui dati (es. “E’ possibile identificare tre diversi tipi di clima in Italia: mediterraneo, continentale, alpino…”);

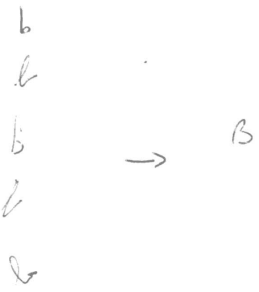
… Esplorare lo spazio dei dati (es. “Quanti tipi diversi di clima sono presenti in Italia? Quante sfere sono presenti in un’immagine?”);

… Semplificare l’interpretazione dei dati (“Il clima di ogni città d’Italia è approssimativamente mediterraneo, continentale o alpino.”).

…“Ragionare” sui dati o elaborare i dati in modo stereotipato.

Clustering: definizioni

- **Pattern**: un singolo dato $\mathbf{X} = [x_1, x_2, \ldots x_D]$. Il dato appartiene quindi ad uno spazio multi-dimensionale, solitamente eterogeneo.
- **Feature**: le caratteristiche dei dati significative per il clustering, possono costituire anch’esso un vettore, il vettore delle feature: $f_1, f_2, \ldots f_M$. Questo vettore costituisce l’input agli algoritmi di clustering.



Inclinazione, occhielli, lunghezza, linee orizzontali, archi di cerchio ...

Clustering: definizioni

- **D**: dimensione dello spazio dei pattern;
- **M**: dimensione dello spazio delle feature;
- **Cluster**: in generale, insieme che raggruppa dati simili tra loro, valutati in base alle feature;
- **Funzione di similarità o distanza**: una metrica (o quasi metrica) nello spazio delle feature, usata per quantificare la similarità tra due pattern.
- **Algoritmo**: scelta di come effettuare il clustering (motore di clustering).

Clustering

- Data, $\{X_1 ..... X_N\} \in R^D$
- Cluster $\{C_1 ..... C_M\}$ ➔ $\{P_1 ..... P_M\} \in R^D$

$\mathbf{P_j}$ represents the set of data inside its cluster.

The set of data inside its cluster has to be determined

The cluster are determined considering features associated to the data.

Tassonomia (sintetica) degli algoritmi di clustering

- Algoritmi gerarchici (agglomerativi, divisivi), e.g. **Hirarchical clustering**.
- Algoritmi partizionali, hard: **K-means, quad-tree decomposition**.
- Algoritmi partizionali, soft: fuzzy c-mean, neural-gas, enhanced vector quantization, **mappe di Kohonen**.
- Algoritmi statistici: **mixture models**.

Il clustering

Per una buona review: Xu and Wunsch, IEEE Transactions on Neural Networks, vol. 16, no. 3, 2005.

Il clustering non è di per sé un problema ben posto. Ci sono diversi gradi di libertà da fissare su come effettuare un clustering.

Rappresentazione dei pattern;

Calcolo delle feature;

Definizione di una misura di prossimità dei pattern attraverso le feature;

Tipo di algoritmo di clustering (gerachico o partizionale)

Validazione dell'output (se necessario) -> Testing.

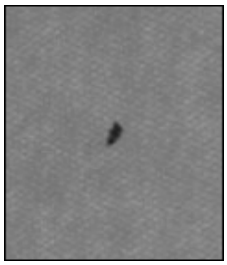
Problema a cui non risponderemo: **quanti cluster**? Soluzione teorica (criterio di Akaike), soluzione empirica (growing networks di Fritzke).

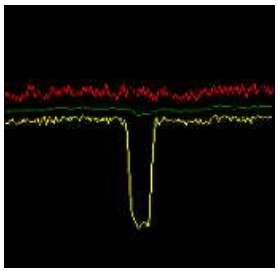
Riassunto

- I tipi di apprendimento
- Il clustering
- Le feature
- Clustering gerarchico: algoritmi agglomerativi
- Clustering gerarchico: algoritmi divisivi

Features

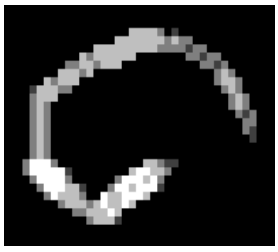
Macchie dense





- *Località.*
- *Significatività.*
- *Rinoscibilità.*

Fili



Rappresentazione dei pattern

- Feature selection: identificazione delle feature più significative per la descrizione dei pattern.

  Esempio: descrizione del clima e della città di Roma.
  Roma: [17°; 500mm; 1.500.000 ab.]

- Come scegliere i pattern?
  - Vicinanza ai bordi di ciascun cluster (Support Vector Machines)
  - Tutti i pattern

- Come valutare le feature?
  - Analisi statistica del potere discriminante: correlazione tra feature e loro significatività.

Feature & feature

- Feature extraction: trasformazione delle feature per creare nuove, significative feature;

- Elaborazione di primo livello, per ottenere informazioni caratteristiche del fenomeno che, ad esempio, siano inverianti.

  Esempio: descrizione di oggetti circolari.
  Posso misurare l'area e il perimetro, ma il loro rapporto è più significativo.

  Esempio: descrizione del clima.
  Milano: [13°; 900mm; 265 giorni sole; 100 giorni pioggia]
  oppure
  Milano: [13°; 900mm / 100 giorni pioggia; 265 giorni sole]

Similarità tra feature

- Definizione di una **misura di distanza tra due features**;

  Esempio:
  Distanza euclidea…
  dist (Roma, Milano) = …
  dist ([17°; 500mm], [13°; 900mm]) = …
  = … Distanza euclidea? = …
  = $((17-13)^2+(500-900)^2)^{½}$ = 400.02 ~ 400

Ha senso?

Normalizzazione feature

Att.ne!
dist (Roma, Milano) = …
dist ([17°; 500mm], [13°; 900mm]) = …
= … Distanza euclidea? = …
= $((17-13)^2+(500-900)^2)^{½}$ = 400.02 ~ 400

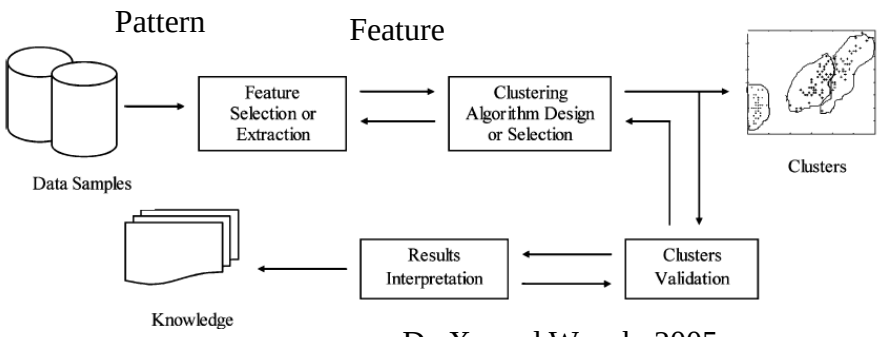
La distanza tra le due città in termini di gradi è insignificante nel nostro calcolo… **E' necessario trovare una metrica corretta per la rappresentazione dei dati. Ad esempio, normalizzare le feature**!

T_{Max}= 20° T_{Min}= 5° → $T_{Norm} = (T - T_{Min})/(T_{Max} - T_{Min})$
P_{Max}= 1000mm P_{Min}= 0mm → $P_{Norm} = (P - P_{Min})/(P_{Max} - P_{Min})$
$Roma_{Norm}$ = [0.8 0.5]
$Milano_{Norm}$ = [0.53 0.9]
$dist(Roma_{Norm}, Milano_{Norm}) = ((0.8-0.53)^2+(0.5-0.9)^2)^{½}$ = 0.4826

Altre funzioni di distanza

- Distanza euclidea:
  $dist(x,y)=[\Sigma_{k=1..d}(x_k-y_k)^2]^{1/2}$
- Minkowski:
  $dist(x,y)=[\Sigma_{k=1..d}(x_k-y_k)^p]^{1/p}$
- Mahalanobis:
  $dist(x,y)= (x_k-y_k)S^{-1}(x_k-y_k)$, con S matrice di covarianza.
- Context dependent:
  $dist(x,y)= f(x, y, context)$

Analisi mediante clustering



Da Xu and Wunsh, 2005

I cluster ottenuti sono significativi?
Il clustering ha operato con successo?

NB i cammini all'indietro consentono di fare la sintonizzazione dei diversi passi.

Riassunto

- I tipi di apprendimento
- Il clustering
- Le feature
- Clustering gerarchico: algoritmi agglomerativi
- Clustering gerarchico: algoritmi divisivi

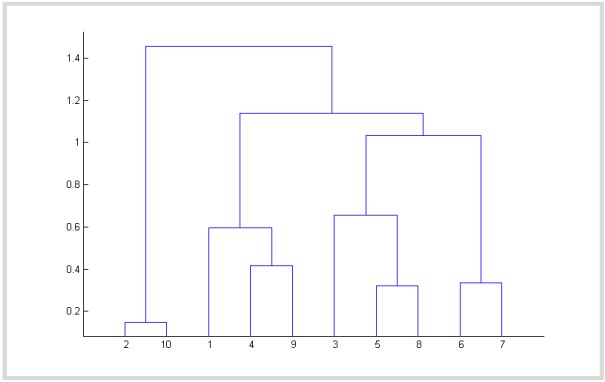
Hierachical Clustering

- In brief, HC algorithms build a whole hierarchy of clustering solutions
  - Solution at level k is a *refinement* of solution at level k-1
- Two main classes of HC approaches:
  - Agglomerative: solution at level k is obtained from solution at level k-1 by merging two clusters
  - Divisive: solution at level k is obtained from solution at level k-1 by splitting a cluster into two parts
    - Less used because of computational load

The 3 steps of agglomerative clustering

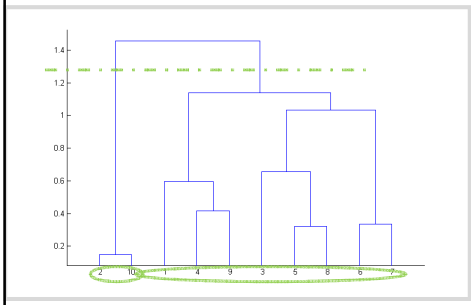
1. At start, each input pattern is assigned to a singleton cluster
2. At each step, the two *closest* clusters are merged into one
   - So the number of clusters is decreased by one at each step
3. At the last step, only one cluster is obtained

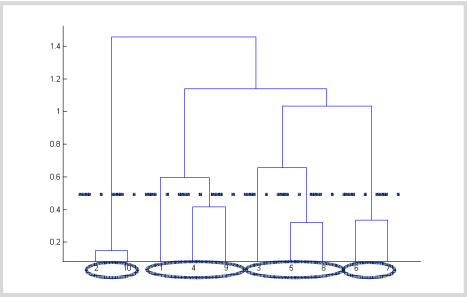
The clustering process is represented by a *dendrogram*



How to obtain the final solution

- The resulting dendrogram has to be cut at some level to get the final clustering:
  - Cut criterion: number of desired clusters, or threshold on some features of resulting clusters





Dissimilarity criteria

- Different distances/indices of dissimilarity (***point wise***) …
  - E.g. euclidean, city-block, correlation…
- … and agglomeration criteria: Merge clusters C_i and C_j such that *diss(i, j)* is minimum (***cluster wise***)
  - Single linkage:
    - diss(i,j) = min d(x, y), where x is in C_i , y in cluster C_j
  - Complete linkage:
    - diss(i,j) = max d(x, y), where x is in cluster i, y in cluster j
  - Group Average (GA) and Weighted Average (WA) Linkage:
    - diss(i j) = $\sum_{x\in C_i}\sum_{y\in C_j} w_i w_j d(x,y) \Big/ \sum_{x\in C_i}\sum_{y\in C_j} w_i w_j$

GA: $w_i = w_j = 1$
WA: $w_i = n_i$, $w_j = n_j$

Cluster wise dissimilarity

- Other agglomeration criteria: Merge clusters C_i and C_j such that *diss(i, j)* is minimum
  - Centroid Linkage:
    - diss(i, j) = $d(\mu_i, \mu_j)$
  - Median Linkage:
    - diss(i,j) = $d(center_i, center_j)$, where each $center_i$ is the average of the centers of the clusters composing C_i
  - Ward's: Method:
    - diss(i, j) = increase in the total error sum of squares (ESS) due to the merging of C_i and C_j
- Single, complete, and average linkage: *graph methods*
  - *All points in clusters are considered*
- Centroid, median, and Ward's linkage: *geometric methods*
  - *Clusters are summed up by their centers*

Ward's method

It is also known as minimum variance method.

Each merging step minimizes the increase in the total ESS:

$$ESS_i = \sum_{x \in C_i} (x - \mu_i)^2 \qquad ESS = \sum_i ESS_i$$

When merging clusters C_i and C_j, the increase in the total ESS is:

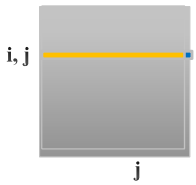
$$\Delta ESS = ESS_{i,j} - ESS_i - ESS_j$$

Spherical, compact clusters are obtained.

The solution at each level k is an approximation to the optimal solution for that level (the one minimizing ESS)

How HC operates

- HC algorithms operate on a dissimilarity matrix:
  - For each pair of existant clusters, their dissimilarity value is stored
- When clusters C_i and C_j are merged, only dissimilarities for the new resulting cluster have to be computed
  - The rest of the matrix is left untouched



The Lance-William recursive formulation

Used for iterative implementation. The dissimilarity value between newly formed cluster $\{C_i, C_j\}$ and every other cluster C_k is computed as:

$$diss(k,(i,j)) = \alpha_i diss(k,i) + \alpha_j diss(k,j) + \beta diss(i,j) + \\ + \gamma |diss(k,i) - diss(k,j)|$$

Only values already stored in the dissimilarity matrix are used. Different sets of coefficients correspond to different criteria.

Criterion	α_i	α_j	β	γ
Single Link.	½	½	0	-½
Complete Link.	½	½	0	½
Group Avg.	$n_i/(n_i+n_j)$	$n_j/(n_i+n_j)$	0	0
Weighted Avg.	½	½	0	0
Centroid	$n_i/(n_i+n_j)$	$n_j/(n_i+n_j)$	$-n_in_j/(n_i+n_j)^2$	0
Median	½	½	- ¼	0
Ward	$(n_i+n_k)/(n_i+n_j+n_k)$	$(n_j+n_k)/(n_i+n_j+n_k)$	$-n_k/(n_i+n_j+n_k)$	0

Characteristics of HC

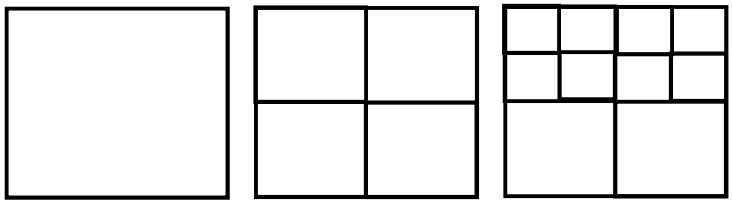
- Pros:
  - Indipendence from initialization
  - No need to specify a desired number of clusters from the beginning
- Cons:
  - Computational complexity at least $O(N^2)$
  - Sensitivity to outliers
  - No reconsideration of possibly misclassified points
  - Possibility of inversion phenomena and multiple solutions

Riassunto

- I tipi di apprendimento
- Il clustering
- Le feature
- Clustering gerarchico: algoritmi agglomerativi
- Clustering gerarchico: algoritmi divisivi

Algoritmi gerarchici: QTD

- Quad Tree Decomposition;
- Suddivisione gerarchica dello spazio delle feature, mediante splitting dei cluster;
- Criterio di splitting (~distanza tra cluster).



Algoritmi gerarchici: QTD

- Clusterizzazione immagini RGB, 512x512;
- Pattern: pixel (x,y);
- Feature: canali R, G, B.
- Distanza tra due pattern (non euclidea):
  dist (p1, p2) =
  dist ([R1 G1 B1], [R2 G2 B2]) =
  max (|R1-R2|, |G1-G2|, |B1-B2|).

Algoritmi gerarchici: QTD

p1 = [0 100 250]
p2 = [50 100 200]
p3 = [255 150 50]

dist (p1, p2) = dist ([R1 G1 B1], [R2 G2 B2]) =
max (|R1-R2|, |G1-G2|, |B1-B2|) = max([50 0 50]) = 50.

dist (p2, p3) = 205.

dist (p3, p1) = 255.

Algoritmi gerarchici: QTD

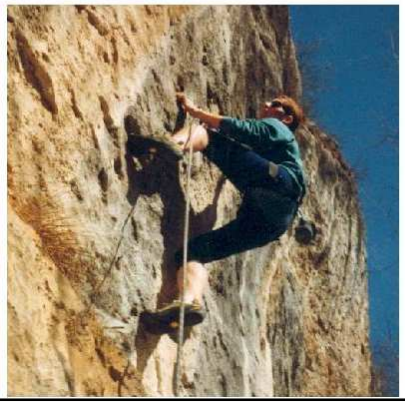
Criterio di splitting: se due pixel all'interno dello stesso cluster distano più di una determinata soglia, il cluster viene diviso in 4 cluster.

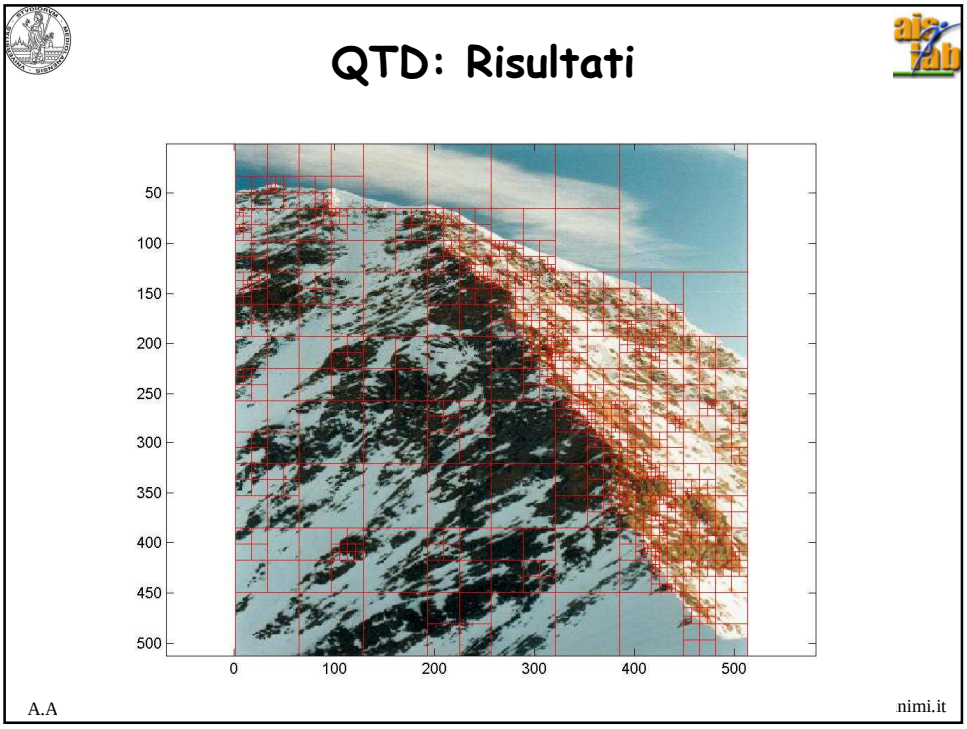
Esempio applicazione: segmentazione immagini, compressione immagini, analisi locale frequenze immagini…

QTD: Risultati

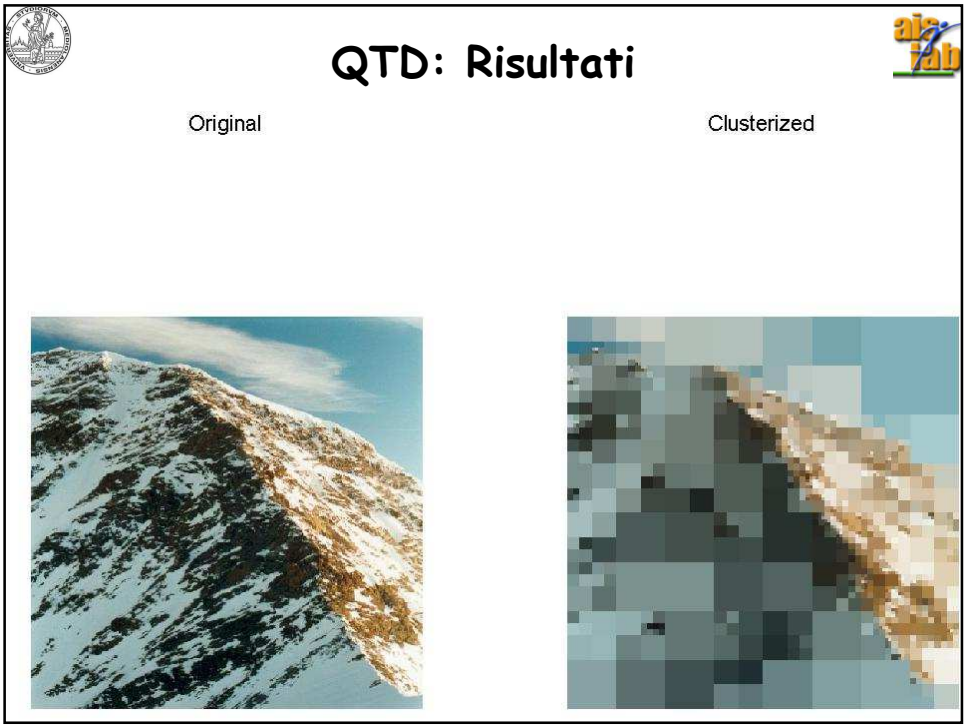
Original

Clusterized





QTD: Risultati



QTD: Risultati

Original

Clusterized

Riassunto

- I tipi di apprendimento
- Il clustering
- Le feature
- Clustering gerarchico: algoritmi agglomerativi
- Clustering gerarchico: algoritmi divisivi