

Riconoscimento e Recupero dell'informazione per Bioinformatica

Manuele Bicego

Corso di Laurea in Bioinformatica
Dipartimento di Informatica - Università di Verona

Il docente

Manuele Bicego

Dipartimento di informatica

Ca Vignal 1 – Primo Piano – Stanza 1.55

Telefono: 045 8027072 e-mail: manuele.bicego@univr.it

Ricevimento:

⇒ giovedì ore 14.30-17.00

⇒ su appuntamento concordato via e-mail

Il corso

- ⇒ Corso nuovo che unisce due corsi del vecchio ordinamento:
 - ⇒ Riconoscimento e Classificazione per la Bioinformatica (7 CFU)
 - ⇒ Recupero dell'informazione (5 CFU)

- ⇒ Corso da 12 CFU: 9 teoria + 3 laboratorio
 - ⇒ Lezioni di teoria in aula (da tre ore!)
 - ⇒ Esercitazioni pratiche in laboratorio Delta (da tre ore)

Il corso

Orari:

Mercoledì 14.30 - 17.30	Aula F
Giovedì 10.30 - 13.30	Aula D
Venerdì 10.30 - 13.30	Laboratorio Delta

NOTA: per questa settimana, la lezione di venerdì si svolgerà in aula F

Pre-requisiti

Pre-requisiti per le attività in aula:

- ⇒ Conoscenze base di Probabilità, Statistica, Analisi
- ⇒ Aver seguito il corso di Elaborazione delle Immagini può essere utile (ma non fondamentale)

Pre-requisiti per le attività in laboratorio:

- ⇒ Capacità di programmare in Matlab (verranno forniti i fondamenti, in ogni caso)

Obiettivi formativi

- ⇒ Fornire le basi delle metodologie di Pattern Recognition
 - ⇒ Capire cos'è la pattern recognition
 - ⇒ Capire la differenza tra pattern recognition supervisionata (classificazione) e non supervisionata (clustering)
 - ⇒ Capire come fare classificazione e clustering
 - ⇒ Capire come validare i risultati di classificazione o di clustering
 - ⇒ Vedere esempi di applicazione di tecniche di Pattern Recognition a problemi di bioinformatica
- ⇒ L'attenzione è rivolta principalmente alla descrizione delle metodologie piuttosto che ai dettagli dei programmi applicativi (già visti in altri corsi – ex. Lab. bioinformatica).

Programma

- ⇒ Il corso si svilupperà secondo diverse linee (che si intersecano)
- ⇒ Metodologie:
 - ⇒ Verranno presentate diverse metodologie di classificazione e clustering, i problemi sottostanti e le motivazioni che portano allo studio di queste tecniche
- ⇒ Applicazioni:
 - ⇒ Verranno analizzati alcuni problemi bioinformatici che sono classicamente risolti con metodologie di classificazione e clustering.
 - ⇒ Per ogni applicazione, verranno fornite una descrizione del problema e una soluzione classica.
- ⇒ Laboratorio:
 - ⇒ verranno implementati semplici algoritmi di classificazione e clustering

Programma dettagliato

INTRODUZIONE:

- ⇒ Introduzione generale alla Pattern Recognition: cos'è, cosa serve, tipico sistema di PR (Ric/Rec)
- ⇒ Rappresentazione e visualizzazione dei dati (Ric/Rec)

CLASSIFICAZIONE:

- ⇒ Teoria di Bayes per la classificazione (Ric)
- ⇒ Classificatori generativi (Ric)
- ⇒ Classificatori discriminativi (Ric)

Programma dettagliato

CLUSTERING:

- ⇒ Misure di similarità tra dati (Rec)
- ⇒ Tecniche di clustering: tassonomia e dettagli delle tecniche più utilizzate (Rec)
- ⇒ Validazione (Rec)

APPLICAZIONI:

- ⇒ classificazione e clustering di dati microarray (con cenni alle tecniche di biclustering) (Ric/Rec)
- ⇒ analisi evolutiva (filogenesi-filogenomica) (Rec)
- ⇒ analisi di immagini biomedicali (Rec)
- ⇒ ...

Materiale didattico

- ⇒ Materiale didattico: lucidi del corso, appunti presi a lezione (per lezioni alla lavagna), libri suggeriti, articoli, internet in generale.
 - ⇒ I lucidi del corso saranno messi (*possibilmente*) in linea il giorno prima della lezione
- ⇒ Gli eventuali seminari di docenti esterni sono inclusi nel materiale didattico del corso.
- ⇒ Laboratorio di riferimento: VIPS (Vision, Image Processing & Sound), CV2, piano -2
Info: <http://vips.sci.univr.it>

Testi consigliati

- ⇒ R. Duda, P. Hart, D. Stork *Pattern Classification*. Wiley, 2001 (2nd edition).
- ⇒ P. Baldi, S. Brunak, *Bioinformatics, The Machine Learning Approach*. MIT Press, 2001
- ⇒ G. Gan, C. Ma, J. Wu: *Data Clustering: Theory, Algorithms and Applications*, ASA-SIAM Series on Statistics and Applied Probability, 2007
- ⇒ A.K. Jain and R.C. Dubes, *Algorithms for Clustering Data*, Prentice-Hall, 1988.
 - ⇒ Disponibile on line
http://homepages.inf.ed.ac.uk/rbf/BOOKS/JAIN/Clustering_Jain_Dubes.pdf
- ⇒ S. Theodoridis, K. Koutroubas: *Pattern Recognition*, Second edition, Academic press, 2003

Testi consigliati

- ⇒ N. Cristianini, M.W. Hahn: *Introduction to Computational Genomics*, Cambridge University Press, 2007
- ⇒ W.J. Ewens, G.R. Grant: *Statistical Methods in Bioinformatics*, Springer 2001
- ⇒ C.M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.
- ⇒ W.J. Ewens, G.R. Grant, *Statistical Methods in Bioinformatics*. Springer, 2001
- ⇒ E. Keedwell, A. Narayanan, *Intelligent Bioinformatics*. Wiley, 2005
- ⇒ M. Berthold, D.J. Hand, *Intelligent Data Analysis*. Springer, 2003 (2nd edition).

Modalità d'esame

- ⇒ Seminario di approfondimento + breve scritto
oppure (per quelli particolarmente *motivati*)
- ⇒ Progetto + breve scritto

(esame a parte per chi deve fare solo uno o solo l'altro)

NOTA:

- ⇒ Seminario da fare singolarmente
- ⇒ Il progetto si può fare in due (è più impegnativo)
 - ⇒ Possibilità di tesi di laurea

⇒ NOTA: Tutti le informazioni, il materiale didattico, gli aggiornamenti e gli avvisi inerenti al corso sono pubblicati alla pagina web del corso

⇒ Raggiungibile dal sito di scienze
www.scienze.univr.it -> corsi di laurea -> Laurea in Bioinformatica -> insegnamenti -> pagina web del corso.

Introduzione

Sommario

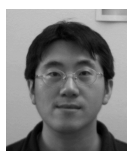
- ⇒ Introduzione alla Pattern Recognition
- ⇒ Pattern Recognition e Bioinformatica: perché?
- ⇒ Difficoltà della Pattern Recognition non supervisionata (clustering)

Pattern Recognition

- ⇒ Definizione storica: "il processo che prende in input dati grezzi (raw) ed effettua un'azione sulla base della *categoria* dei dati" [Duda et al., 2001]
 - ⇒ riconoscimento di pattern
 - ⇒ rilevazione di pattern
 - ⇒ clustering di pattern
- ⇒ Prospettiva informatica: fare Pattern Recognition con i calcolatori

Esempi classici

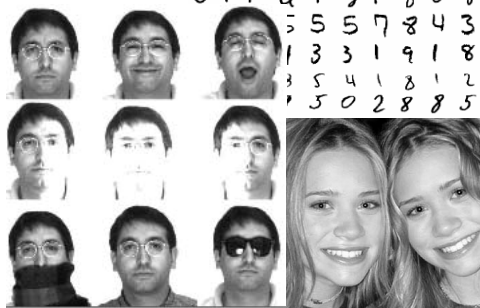
0 1 2 3 4 5 6 7 8 9
riconoscere caratteri scritti a mano



distinguere diverse persone sulla base del volto

Problemi facilmente risolvibili dagli esseri umani

0 1 2 3 4 5 6 7 8 9
0 1 2 3 4 5 6 7 8 9
0 1 2 3 4 5 6 7 8 9
0 1 2 3 4 5 6 7 8 9
0 1 1 4 1 3 1 8 0 8
5 5 5 7 8 4 3
1 3 3 1 9 1 8
3 5 4 1 8 1 2
5 0 2 8 8 5



Veramente difficili per i calcolatori!

Altre applicazioni classiche

- ⇒ Riconoscimento del parlato
- ⇒ Riconoscimento di oggetti da immagini
- ⇒ Classificazione di video
- ⇒ Riconoscimento di impronte digitali
- ⇒ Riconoscimento di gesti
- ⇒ Segmentazione di immagini

- ⇒ Nel laboratorio VIPS:
 - ⇒ classificazione e clustering di oggetti/immagini/video/audio
 - ⇒ classificazione di situazioni (videosorveglianza)
 - ⇒ Bioinformatica: filogenesi, genetica computazionale, analisi di spettri NMR, classificazione di immagini MRI,...

Il problema principale

Capire e modellare i diversi pattern di un problema
(tipicamente in termini di classi / gruppi / categorie)

Il paradigma principale

Il problema è risolto usando il cosiddetto paradigma
"apprendimento da esempi"

I modelli sono stimati attraverso un insieme di esempi
campionati dal problema (il training set – insieme di
addestramento)

L'obiettivo principale

GENERALIZZAZIONE: capacità di generalizzare anche a oggetti sconosciuti (non presenti nel training set)

Il vero problema

- ⇒ Derivare un modello per il problema a partire da esempi
- ⇒ Tipicamente il problema è risolto con una procedura di ottimizzazione

$$\text{Model} \leftarrow \max E(T, P, \Theta)$$

T = training set
P = informazioni a priori
 Θ = parametri

Il vero problema

Sotto problemi tosti

- ⇒ definire la funzione E
 - ⇒ compromesso tra la capacità di spiegare il training set e la complessità

- ⇒ ottimizzare E (tipicamente una funzione difficile da ottimizzare)
 - ⇒ discesa lungo il gradiente
 - ⇒ Expectation – Maximization
 - ⇒ Simulated annealing
 - ⇒ Tabu Search (Reactive Tabu search)
 - ⇒ Algoritmi genetici

Altri problemi

- ⇒ Aspetti teorici
 - ⇒ convergenze del learning
 - ⇒ comportamenti asintotici
 - ⇒ ottimalità delle soluzioni
 - ⇒ ...
- ⇒ Aspetti pratici
 - ⇒ accuratezza
 - ⇒ requisiti computazionali (tempo e spazio)
 - ⇒ flessibilità
 - ⇒ usabilità
 - ⇒ ...

Tipologie di problemi in PR

⇒ Classificazione

⇒ Detection

⇒ Clustering

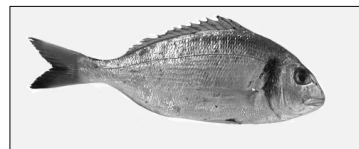
Un esempio

[Duda Hart Stork, Pattern Classification, Second Edition, Wiley 2001]

Problema: modellare pesci



bass

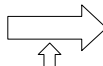


gilthead bream

Tipologie di problemi in PR

Classificazione

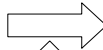
Trovare due modelli M1 e M2, uno per l'orata e uno per la spigola



spigola o orata?

M1, M2

Detection

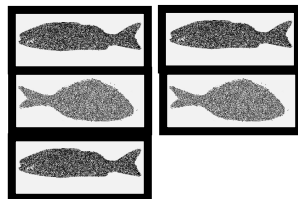


Questo pesce è una spigola?

M

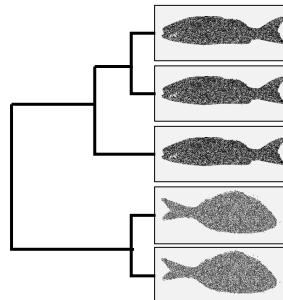
Tipologie di problemi in PR

Clustering



1. identificare pesci simili (identificare tutti i gruppi "naturali" e creare i modelli)

2. identificare le relazioni tra i pesci (clustering gerarchico)



Sistema di Pattern Recognition Statistica

⇒ Componenti:

⇒ Rappresentazione dei dati

⇒ rappresentare ogni oggetto del problema in esame (features)

⇒ Addestramento/Costruzione del modello

⇒ costruire il modello / i modelli dall'insieme di addestramento

⇒ differente a seconda che si tratti di classificazione, clustering o detection

⇒ Task: classificazione, detection, clustering

Tipico sistema di PR

Rappresentazione

Problema

↓ campionamento

Dati grezzi

↓ estrazione di feature

Dati rappresentati

Costruzione del modello

Training Set

Informazioni a priori

↓ addestramento

Modelli

Alcuni commenti

- ⇒ "Il modello deve spiegare il problema": capacità di generalizzare anche a pattern mai visti (capacità di generalizzazione)

- ⇒ Il training set deve essere adeguatamente:
 - ⇒ largo (molti pattern)
 - ⇒ completo (tutte le categorie devono essere ragionevolmente rappresentate)
 - ⇒ variabile (deve tenere in considerazione la variabilità dei pattern nelle categorie)

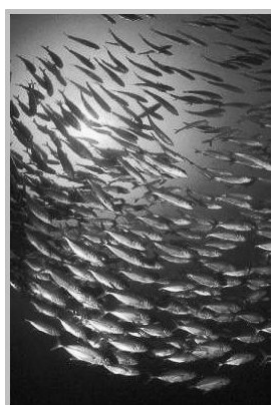
Rappresentazione

- ⇒ Pattern: misure dell'oggetto
 - ⇒ spesso è un vettore d-dimensionale $x = [x_1, x_2, \dots, x_d]$ (dove ogni x_i rappresenta una feature)
 - ⇒ ogni pattern è un punto in uno spazio d-dimensionale
 - ⇒ Esempi: misure fisiche dell'oggetto, immagini, spettri
 - ⇒ feature: semplici da calcolare, invarianti a trasformazioni irrilevanti, affidabili, indipendenti, discriminanti, poche (problema della *curse of dimensionality*);
- ⇒ Spesso definiti come "dati grezzi"
- ⇒ Altre possibili rappresentazioni:
 - ⇒ stringhe (exe DNA)
 - ⇒ dati tridimensionali (exe: proteine foldate (ripiegate))

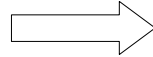
Rappresentazione

- ⇒ Pre-processing dei dati: "migliorare" la rappresentazione:
 - ⇒ ridurre il numero dello spazio delle features (per visualizzare, per ridurre il carico computazionale, ...)
 - ⇒ mettere in evidenza particolari strutture o migliorare le capacità discriminative dello spazio
- ⇒ Estrazione di feature: trasformazione dello spazio originale (PCA, ICA, LDA, MDS,...)
- ⇒ Selezione di features: selezionare le feature migliori (SFFS, SBFS, Matching Pursuit, ...)

Esempio: rappresentazione dei dati



Il problema



campionamento

dati grezzi (un'immagine)



estrazione delle features

(altezza, lunghezza)

$$\mathbf{x}_1 = [5, 10]$$

dati pre-processati

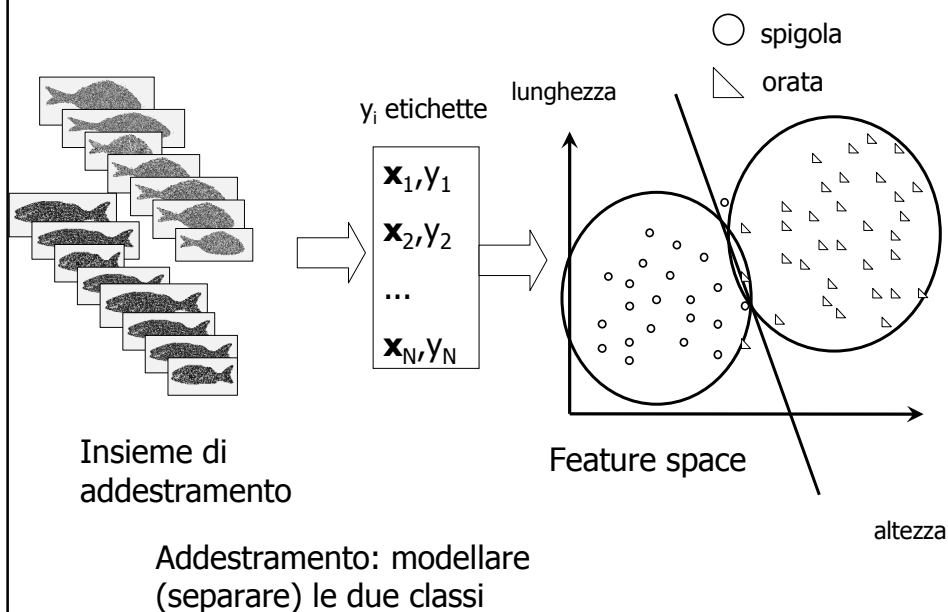
Costruzione del modello

- ⇒ Costruzione di un modello in grado di spiegare i dati del training set
 - ⇒ training/learning/addestramento

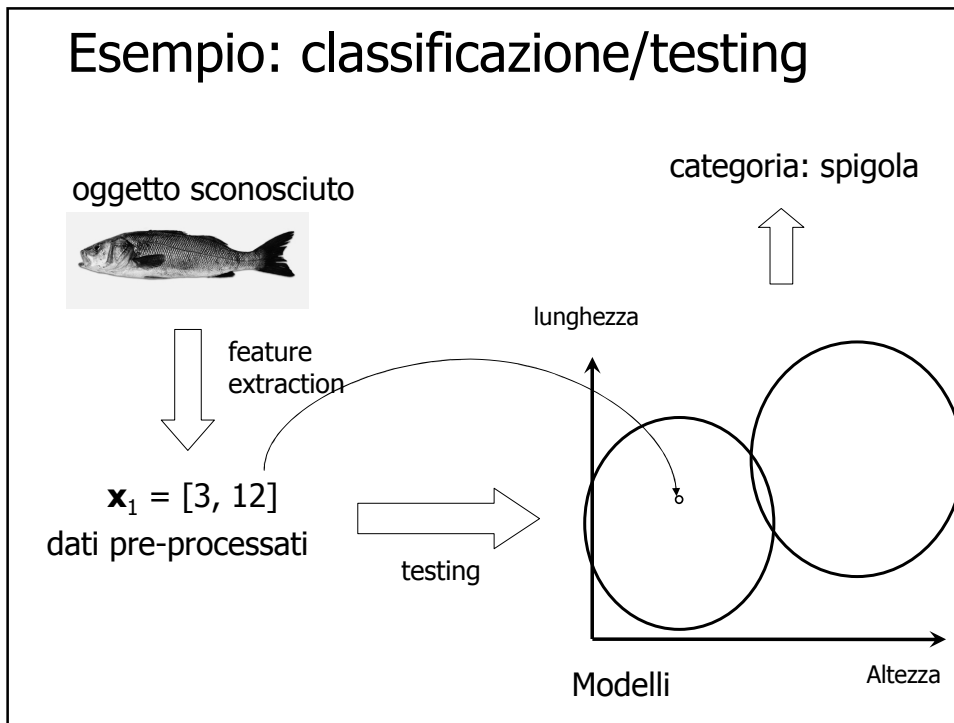
- ⇒ Possibili scelte:
 - ⇒ tipo di modello
 - ⇒ parametri del modello
 - ⇒ dimensione del modello
 - ⇒ metodo di addestramento (funzione da ottimizzare, metodo di ottimizzazione)
 - ⇒ metodo di validazione (come capire se il modello scelto effettivamente rappresenta il fenomeno in questione)

- ⇒ procedura diversa a seconda che si parli di classificazione, clustering o detection

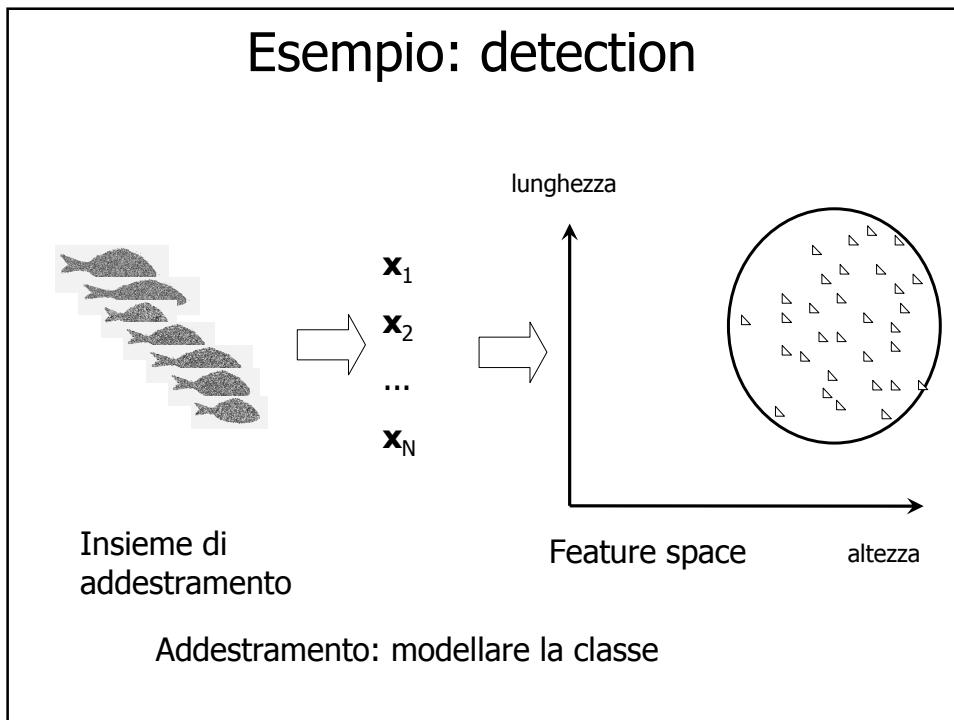
Esempio: classificazione



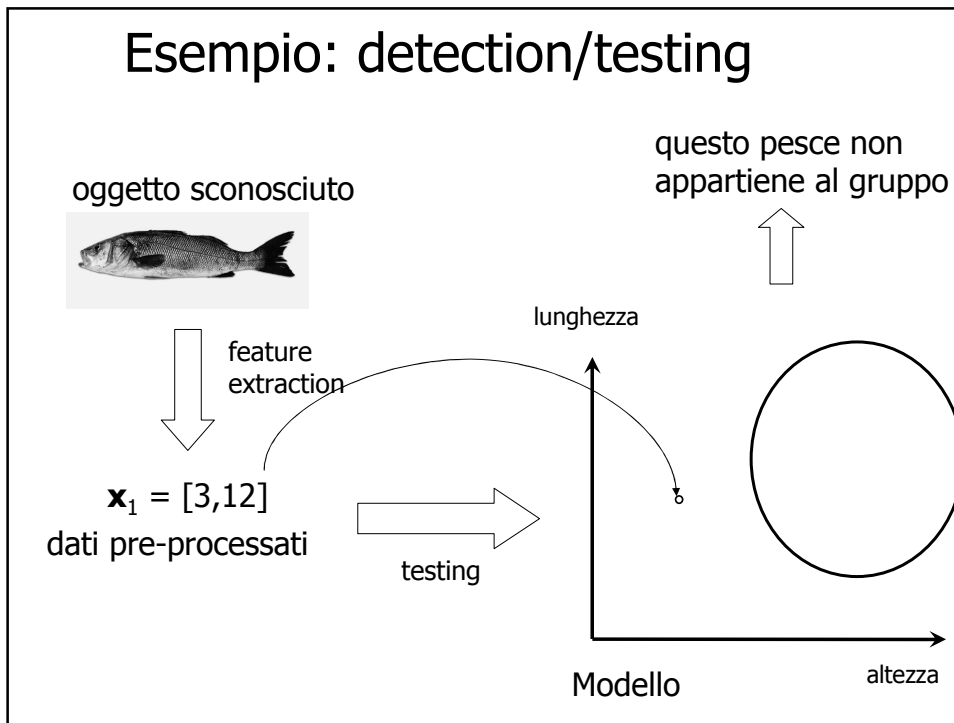
Esempio: classificazione/testing



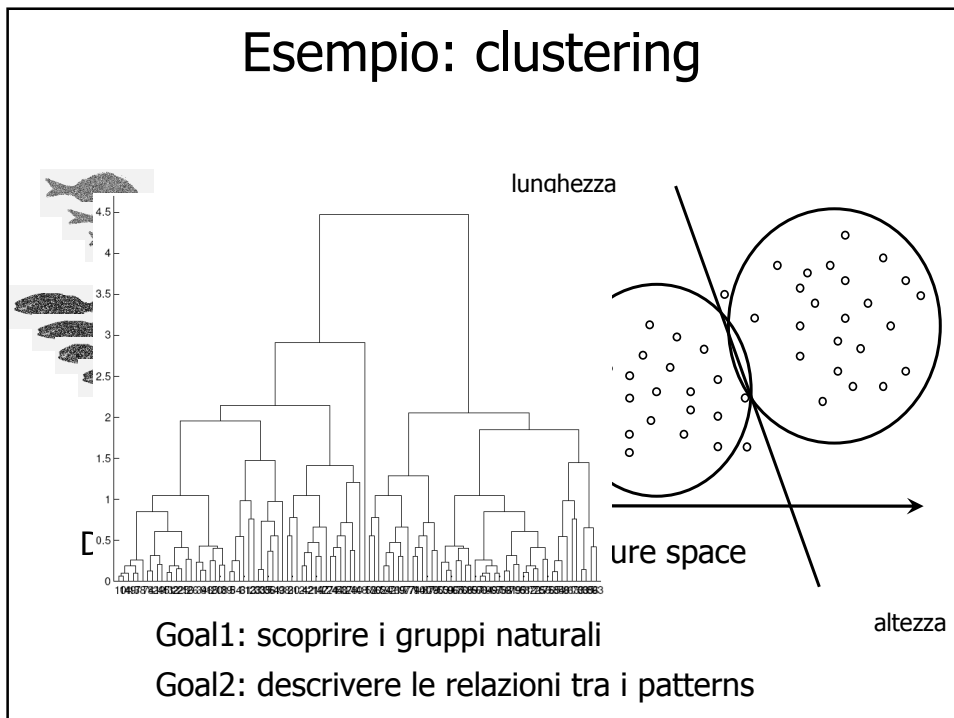
Esempio: detection



Esempio: detection/testing



Esempio: clustering



Sommario

- ⇒ La costruzione del modello può avvenire in modo supervisionato (classificazione) o non supervisionato (clustering)
 - ⇒ Supervisionato: per ogni oggetto del training set si conosce l'esatta categoria
 - ⇒ Non supervisionato: non si conosce nulla

- ⇒ *Reinforcement learning* (per classificazione)
 - ⇒ a metà strada tra le due: non viene fornita alcuna informazione sulla categoria esatta, viene dato un giudizio sulla correttezza della classificazione

Interpretazione dei risultati

- ⇒ L'obiettivo finale è quella di estrarre / recuperare conoscenza
 - ⇒ ottenere intuizioni dal data set

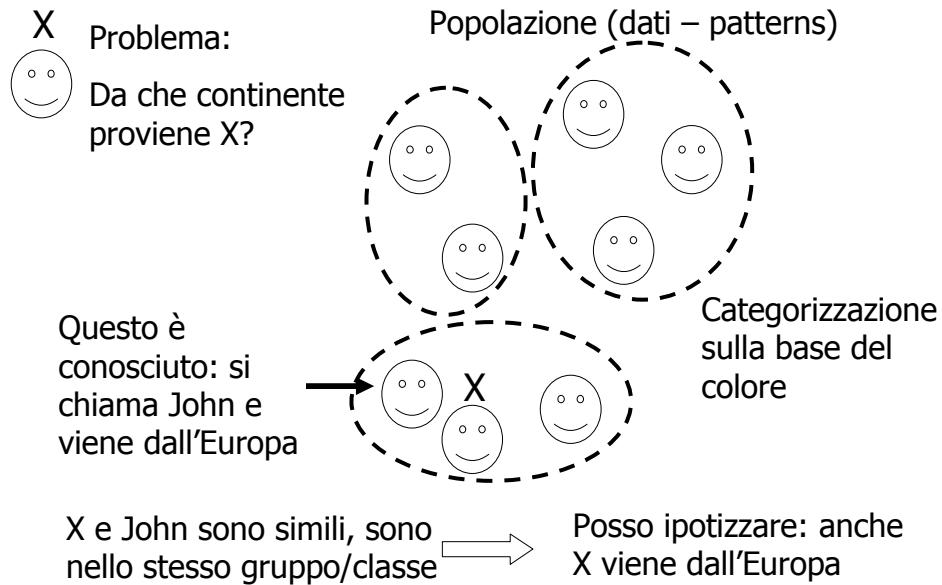
- ⇒ Il fuoco deve essere sulla "interpretabilità" dei prodotti
 - ⇒ interpretabilità dei metodi
 - ⇒ mette a proprio agio l'utente
 - ⇒ interpretabilità delle soluzioni
 - ⇒ permette di capire gli errori

Pattern Recognition e bioinformatica: perché?

PR e bioinformatica: perché?

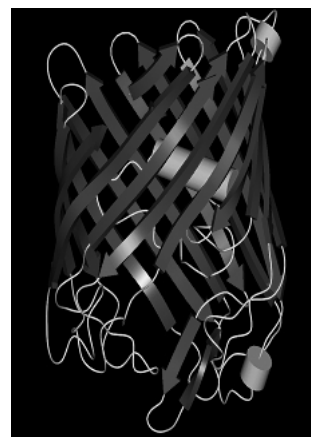
- ⇒ IDEA PRINCIPALE: la caratterizzazione di una popolazione in termini di gruppi/classi/categorie può essere utilizzata per inferire alcune proprietà di oggetti sconosciuti guardando ad oggetti conosciuti nello stesso gruppo
- ⇒ In altre parole: vengono inferite informazioni su entità sconosciute a partire da informazioni note su entità conosciute che siano "simili"

Esempio 1



Esempio 2

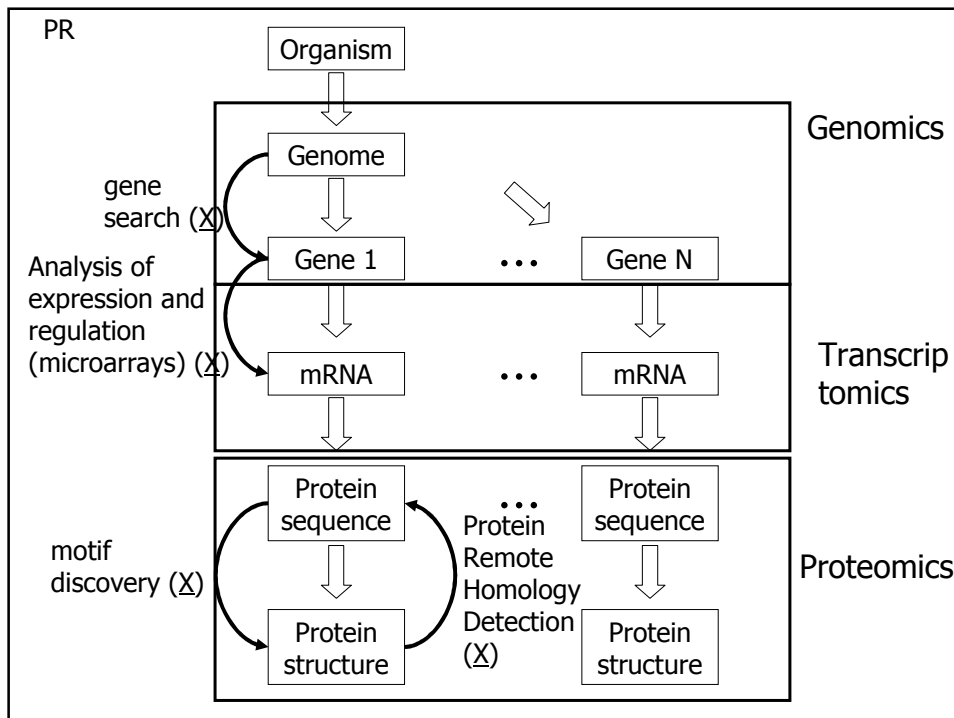
- ⇒ Esempio:
- ⇒ ho una proteina B che non so cosa faccia
 - ⇒ trovo una proteina A che ha una struttura/sequenza molto simile (misura di similarità, clustering)
 - ⇒ Posso ipotizzare che la proteina B abbia una funzione simile.

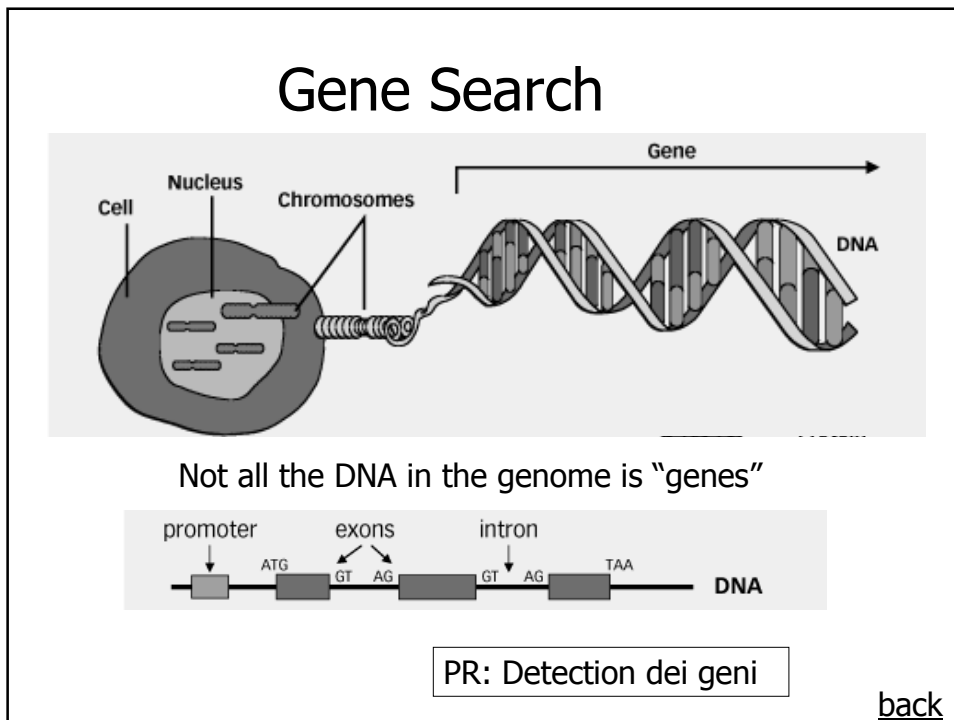
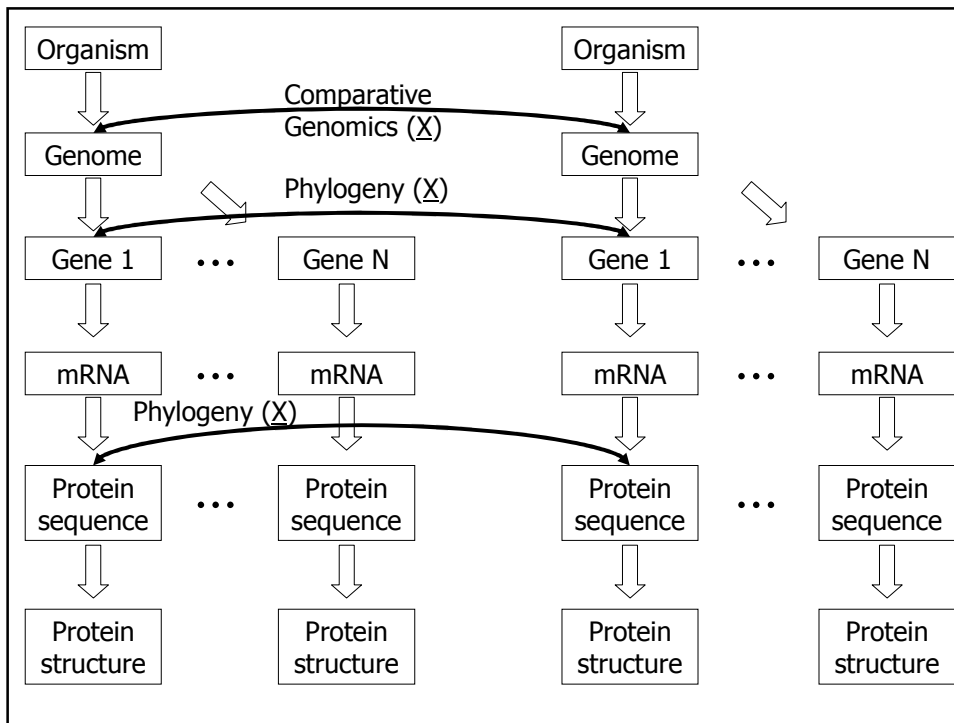


PR e bioinformatica: perché?

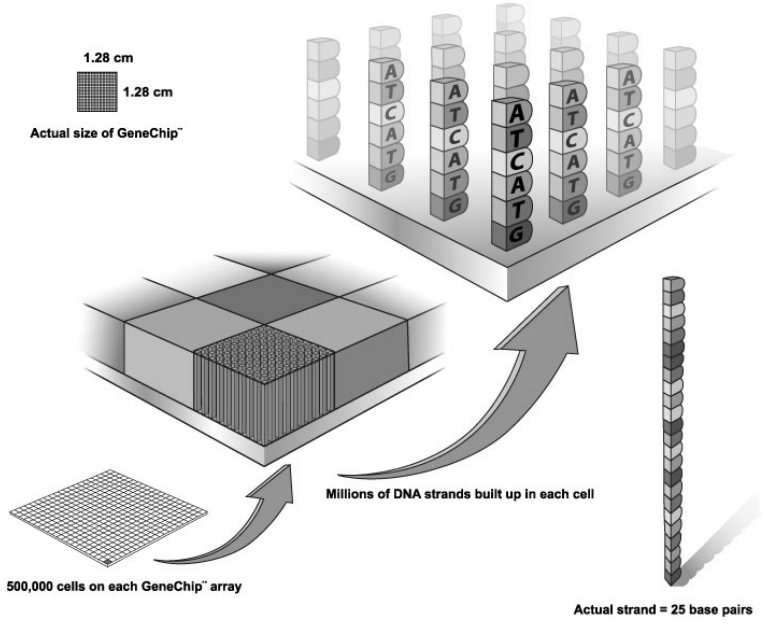
Altre motivazioni:

1. In bioinformatica ci sono molti problemi di classificazione, clustering e detection
2. Possibilità di derivare modelli per i dati tramite esempi (paradigma di apprendimento da esempi)
3. Ci sono problemi di classificazione (onerosi in termini di tempo) che possono essere automatizzati
 - ⇒ apprendimento da esempi che possono essere giudicati da esperti / validati sperimentalmente

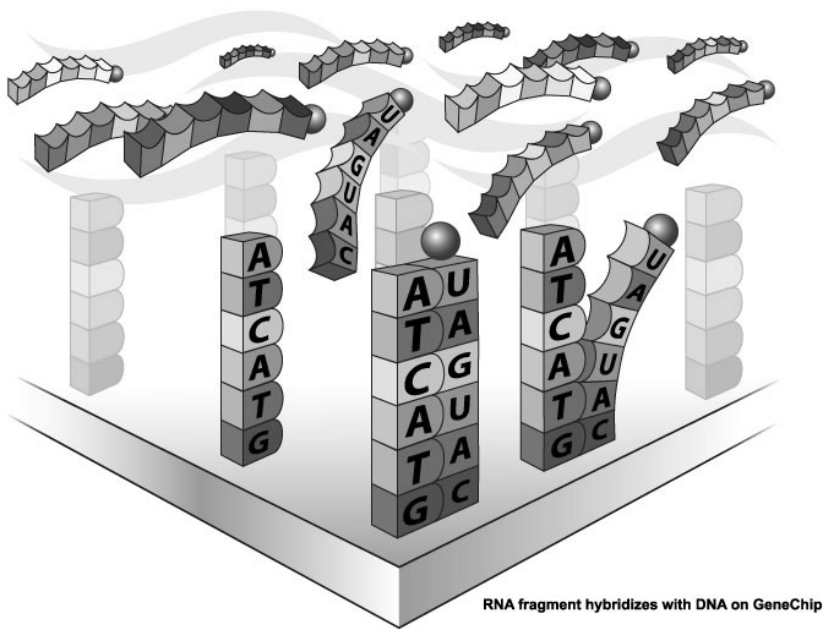




Microarray



RNA fragments with fluorescent tags from sample to be tested



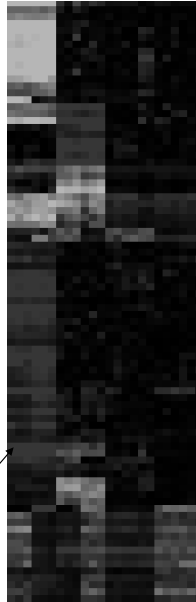
⇒ Microarray: tecnologia in grado di analizzare molte migliaia di geni simultaneamente

⇒ Expression microarray: l'obiettivo è quello di trovare il livello di espressione di diversi geni

⇒ Di solito vengono effettuati diversi esperimenti:

- ⇒ differenti condizioni di crescita
- ⇒ soggetti diversi
- ⇒ malattie diverse

lo spot (k,i) rappresenta quanto il gene k è espresso nell'esperimento i

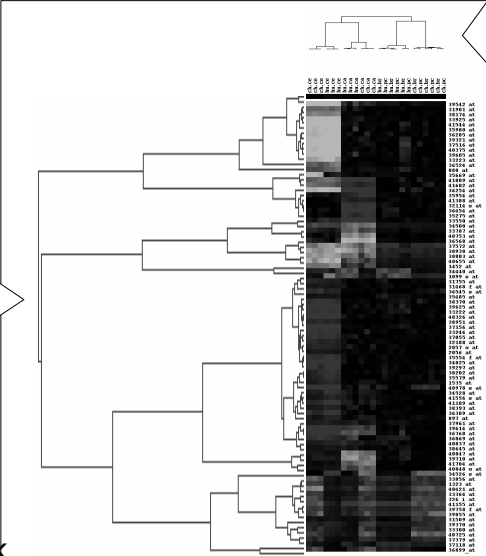


Clustering

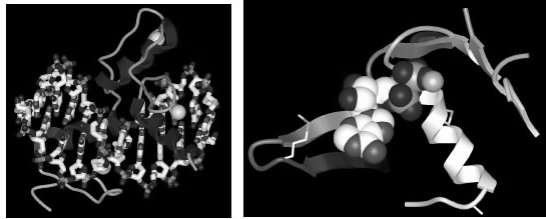
clustering di esperimenti
trovare esperimenti con geni espressi in modo simile

clustering di geni
Trovare geni con pattern di espressione simili (quindi con funzioni ipoteticamente correlate)

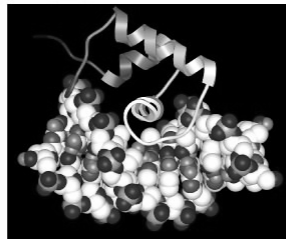
back



Motif Discovery



-PHE-ALA-ARG-SER-ASP-GLU-ARG-LYS-ASP-HIS-



[back](#)

parti simili in
strutture di
diverse proteine



pattern simili nelle
sequenze



PR: rilevamento di
questi pattern
(motif)

Protein Remote Homology Detection

Stessa
Funzione "A"



Molte proteine
con strutture
simili



Le sequenze
corrispondenti
condividono una
similarità remota

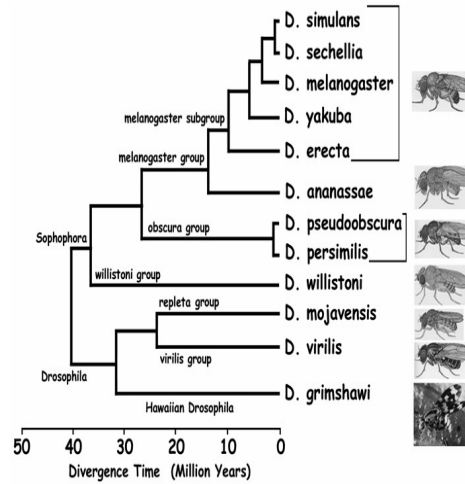
PR: caratterizzare ogni classe di sequenze
omologhe (in senso remoto)

[back](#)

Filogenesi

⇒ Filogenesi: inferire le relazioni genealogiche tra gli organismi

⇒ clustering gerarchico di sequenze geniche o proteiche



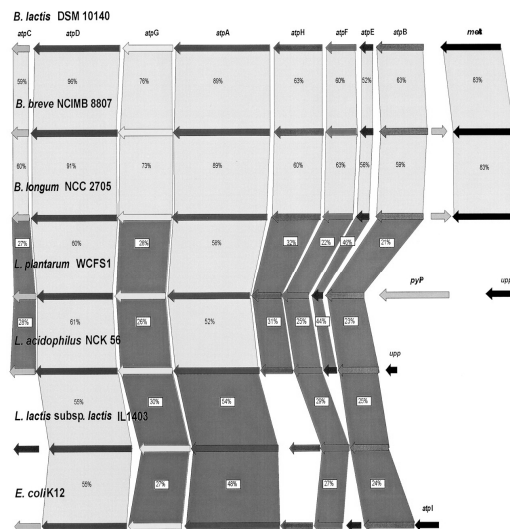
⇒ Filogenomica: ha lo stesso obiettivo della filogenesi ma viene effettuata a livello di genoma

Problemi:

⇒ Presenza/assenza di geni

⇒ traslazioni spaziali di geni

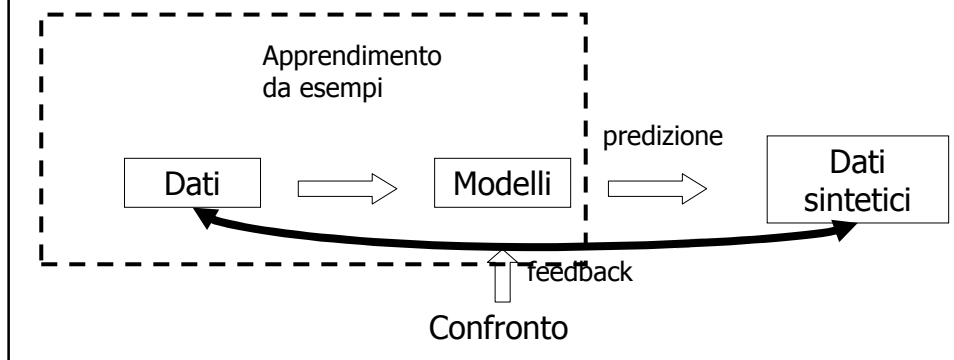
⇒ duplicazione di geni



PR e bioinformatica: perché?

⇒ Possibilità di derivare modelli per i dati attraverso esempi

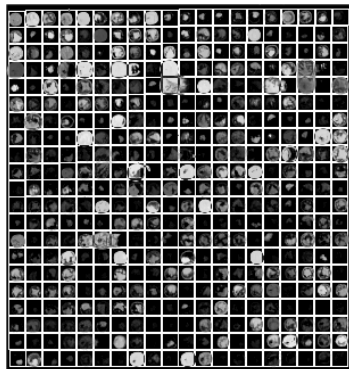
⇒ Il paradigma della Pattern Recognition "apprendimento da esempi"



PR e Bioinformatica: perché?

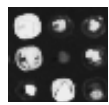
⇒ Possibilità di automatizzare procedure di classificazione onerose dal punto di vista del tempo richiesto

⇒ Esempio: qualità degli spot dei microarray

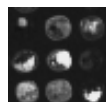


Problema:

⇒ Trovare gli spot dei microarray con bassa qualità



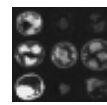
size



roundness



intensity



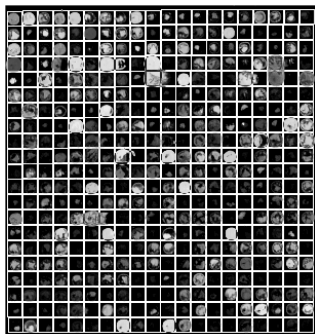
pixel
distribution

Approccio tipico:

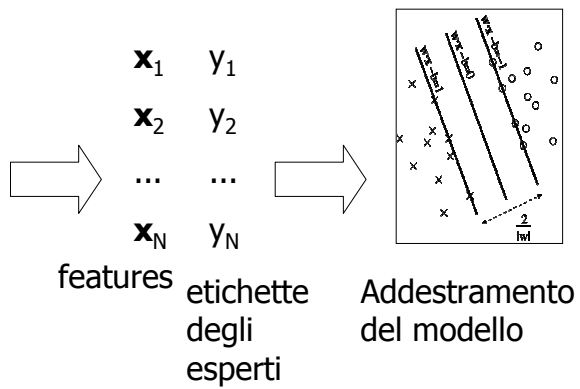
⇒ Annotazione manuale da parte di esperti

L'approccio PR

Imparare un modello, usando i giudizi dell'esperto, in un esperimento

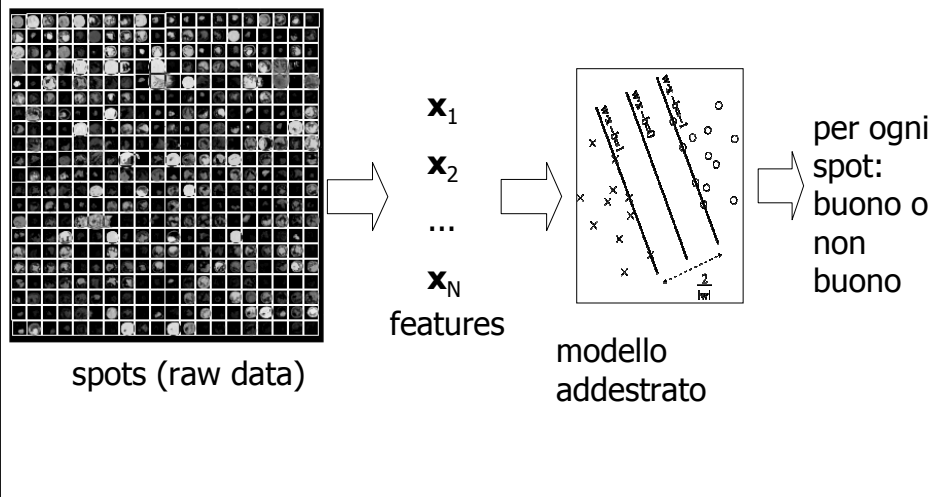


spots (raw data)



L'approccio PR

⇒ Testing: per un esperimento qualsiasi



PR e Bioinformatica (more)

⇒ Applicazioni legate alle immagini:

- ⇒ rilevamento di parti interessanti in immagini
 - ⇒ spots in microarray
 - ⇒ gel
 - ⇒ immagini mediche
- ⇒ misure
 - ⇒ calcolo di feature (e.g. rotondità degli spot nei microarray)
 - ⇒ presenza / assenza di elementi

PR e Bioinformatica (more)

Ricerche nei database (GenBank, PDB)

⇒ sequenze:

- ⇒ trovare similarità tra sequenze (e.g. BLAST, FASTA)
- ⇒ PR: similarità trovate utilizzando modelli addestrati

⇒ documenti:

- ⇒ estrazione di informazioni rilevanti
- ⇒ PR: retrieval by content (invece del classico ritrovamento per keywords)
- ⇒ PR: classificazione di documenti

Sfide

⇒ Enorme quantità di dati:

- ⇒ 2007: 2596 progetti genoma (il primo nel 1995)
(www.genomesonline.org)
- ⇒ Procarioti (e.g. batteri) 0.5 – 8 Mega bp
- ⇒ Eucarioti (e.g. uomo) 12 Mbp – 16 Gbp
- ⇒ Uomo: 3 Gbp – circa 30.000 geni

⇒ Potenziale crescita esplosiva (e.g. il 95% della biodiversità microbica è sconosciuta)

Sfide

- ⇒ Notevole complessità e diversità dei sistemi biologici
- ⇒ Necessità di utilizzare il più possibile le informazioni biologiche note a priori
- ⇒ Ricerca della "interpretabilità biologica" di:
 - ⇒ metodologie sviluppate
 - ⇒ soluzioni ottenute

Un commento sul clustering

- ⇒ Il clustering è un problema più difficile della classificazione
 - ⇒ Il processo è non supervisionato: non sappiamo se facciamo giusto! (differentemente dalla classificazione)
- ⇒ Il clustering rappresenta l'organizzazione di un insieme di patterns (entità) in gruppi (clusters) sulla base della similarità

- ⇒ Qual'è la similarità più appropriata?
 - ⇒ Cambiare la similarità cambia il risultato
- ⇒ Cosa deve rappresentare un "buon gruppo"?
 - ⇒ Il concetto di gruppo è definito in modo vago e assolutamente soggettivo

Esempio: Oggetti da clusterizzare



Primo gruppo: bicchieri



Secondo gruppo: padelle



Primo gruppo: contenitori vuoti



Secondo gruppo: contenitori pieni



Quindi

- ⇒ Il concetto di cluster è vago:
- ⇒ Dipendentemente dalle misure di similarità utilizzate cambia il risultato
- ⇒ La scelta della misura di similarità è cruciale.
 - ⇒ Dovrebbe essere fatta in modo da inglobare la maggior quantità possibile di informazione a priori.
- ⇒ Il risultato può cambiare anche a seconda della metodologia utilizzata per fare clustering (il concetto sarà più chiaro in seguito)

Riviste e convegni principali (sul tema generale)

- ⇒ Convegni
 - ⇒ NIPS, ICML, ECML, CVPR, AI, ICPR, ICCV, ECCV, ICIP, etc.
- ⇒ Journals
 - ⇒ PAMI, IEEE Trans. on Pattern Analysis & Machine Intelligence
 - ⇒ Artificial Intelligence
 - ⇒ Machine Learning
 - ⇒ Journal of Machine Learning Research
 - ⇒ CVIU, Computer Vision and Image Understanding
 - ⇒ GMIP, Graphical Models & Image Processing
 - ⇒ IVC, Image and Vision Computing
 - ⇒ PR, Pattern Recognition
 - ⇒ PRL, Pattern Recognition Letters
 - ⇒ IEEE Trans. on Image Processing
 - ⇒ IEEE Trans. on Systems, Man, & Cybernetics
 - ⇒ Int. J. on Pattern Recognition & Artificial Intelligence
 - ⇒ IEEE Trans. on Neural Networks
 - ⇒ Neural Computation
 - ⇒ Proceedings of the IEEE

Riviste e convegni principali (specifici su Bioinfo)

- ⇒ Convegni
 - ⇒ molti, parole chiave bioinformatics, computational biology, medical informatics, pattern matching, systems biology, AI in Medicine, etc.
- ⇒ Journals
 - ⇒ Bioinformatics
 - ⇒ BMC Bioinformatics
 - ⇒ Journal of Bioinformatics & Computational Biology
 - ⇒ IEEE/ACM Trans. on Computational Biology & Bioinformatics
 - ⇒ Int'l Journal of Data Mining & Bioinformatics
 - ⇒ Eurasip Journal of Bioinformatics & Systems Biology
 - ⇒ Int'l Journal of Bioinformatics Research and Applications
 - ⇒ Journal of Biomedical Informatics
 - ⇒ Journal of Computational Biology
 - ⇒ Journal of Proteomics & Bioinformatics
 - ⇒ Journal of Integrative Bioinformatics
 - ⇒ The Open Bioinformatics journal