Text Classification, K-NN, Evaluation

Wm&R a.a. 2009/10

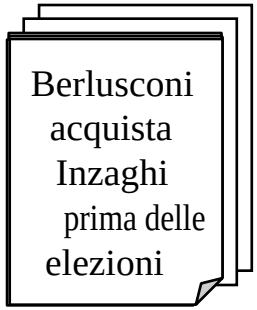
R. Basili

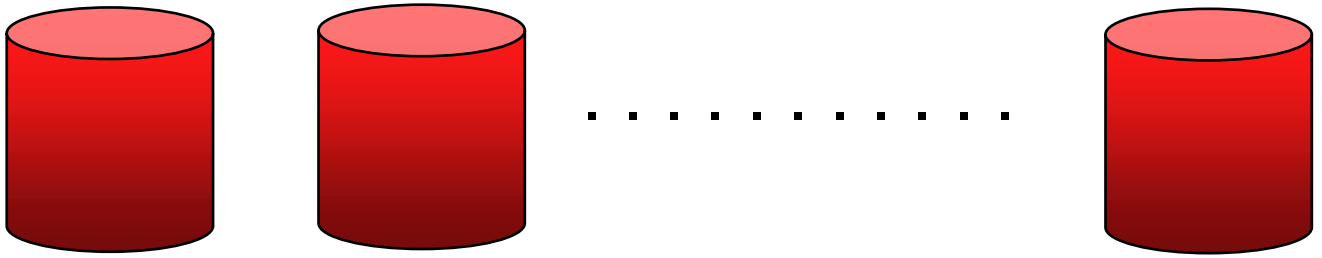
Dipartimento di Informatica Sistemi e produzione
Università di Roma "Tor Vergata"
Email: basili@info.uniroma2.it

Sommario

- Categorizzazione e Ottimizzazione del Testo
  - Introduzione al TC
  - TC: Valutazione Prestazioni
  - I passi nello sviluppo di un sistema TC
  - Il classificatore Rocchio
  - Apprendimento Pigro: K-NN
  - Il Classificatore Parametrizzato Rocchio (PRC)
  - Valutazione Comparativa: Rocchio, PRC e SVM

Introduzione alla Categorizzazione del Testo





Politica
C_1

Economia
C_2

Sport
C_n

Problema di Classificazione del Testo

- Dato:
  - un insieme di categorie obiettivo: $C = \left\{ C^1,..,C^n \right\}$
  - L'insieme *T* dei documenti,

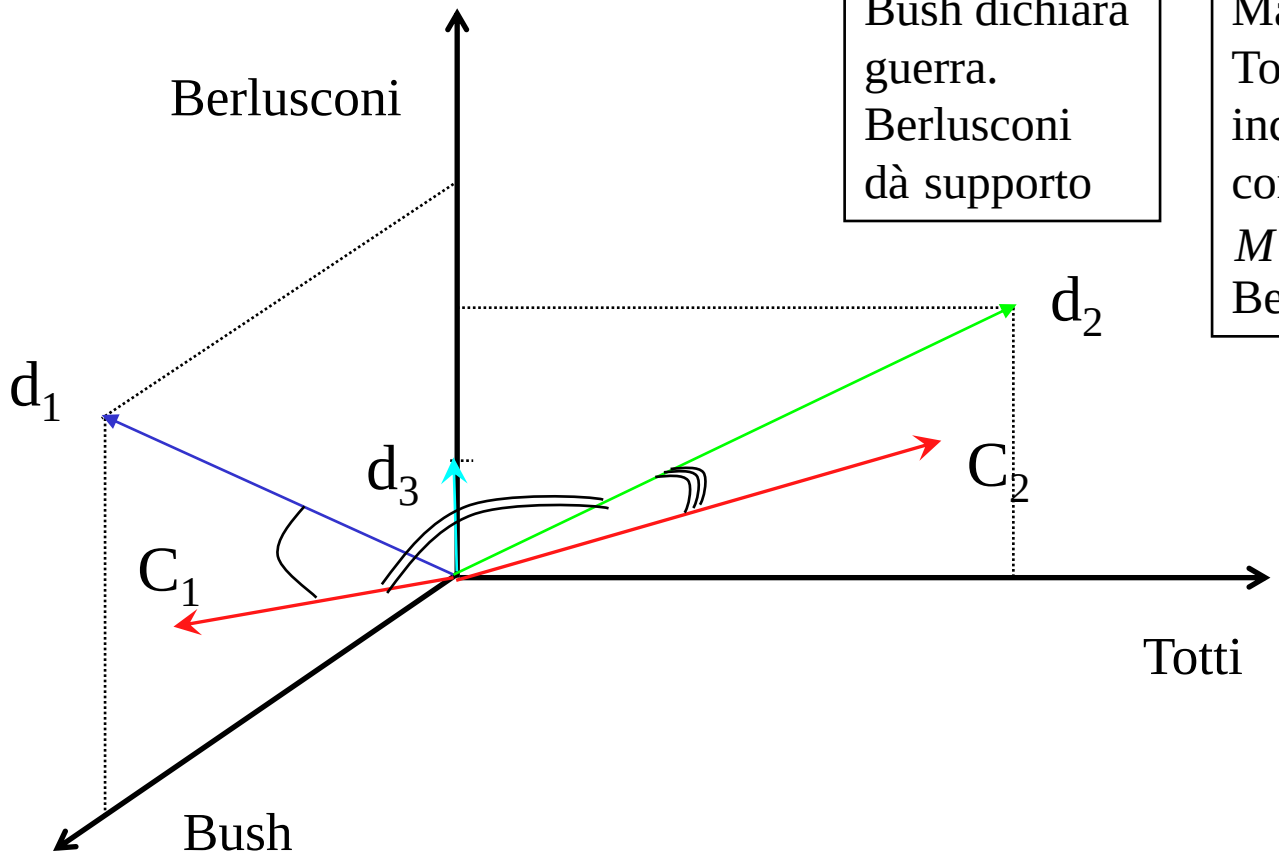
  definisce

$$f : T \rightarrow 2^C$$

- VSM (Salton89')
  - Le caratteristiche sono dimensioni di uno Spazio Vettoriale.
  - Documenti e Categorie sono vettori di pesi caratteristici.
  - *d* è assegnato a C^i se

$$\vec{d} \times \vec{C}^i > th$$

Il Modello a Spazio Vettoriale



d_1: Politica

Bush dichiara guerra. Berlusconi dà supporto

d_2: Sport

Magico Totti nell' incontro di ieri contro il *Milan di* Berlusconi

d_3:Economia

Berlusconi acquista Inzaghi prima delle elezioni

C_1 : Categoria Politica

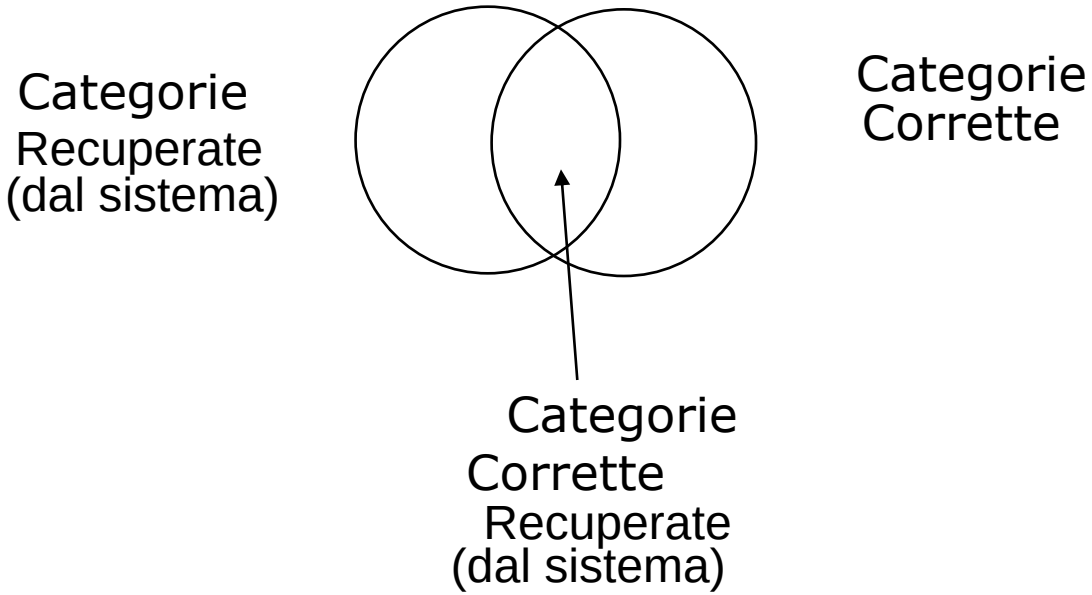
C_2 : Categoria Sport

Categorizzazione del Testo Automatizzata

- Un corpo di documenti pre-categoricizzati
- Divide i documenti in due parti:
  - Insieme di Training
  - Insieme di Test
- Applica un modello di apprendimento a macchina supervisionata all' insieme di apprendimento
  - Esempi Positivi
  - Esempi Negativi
- Misura le prestazioni sull' insieme di test
  - es., Precisione e Richiamo

Misurazione delle Prestazioni

- Dato un insieme di documenti *T*
- Precisione = # Categorie Corrette Recuperate / # Categorie Recuperate
- Richiamo = # Categorie Corrette Recuperate / # Categorie Corrette



Precisione e Richiamo di C_i

- a, numero di etichettature/classificazioni corrette
- b, numero di sbagli, etichettature sbagliate
- c, numero di etichette non pervenute

The *Precision* and *Recall* are defined by the above counts:

$$Precision_i = \frac{a_i}{a_i + b_i}$$

$$Recall_i = \frac{a_i}{a_i + c_i}$$

Misurazione delle Prestazioni

- Punto di Pareggio
  - Trova la soglia per cui

    Richiamo = Precisione
  - Interpolazione
- f-misura
  - Significato armonico tra precisione e richiamo
- Prestazione globale su più di due categorie
  - Micro-media
    - Il conteggio si riferisce al classificatore
  - Macro-media (misure medie su tutte le categorie)

F-misura e Micro-Media

$$F_1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

$$\mu Precision = \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} a_i + b_i}$$

$$\mu Recall = \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} a_i + c_i}$$

$$\mu BEP = \frac{\mu Precision + \mu Recall}{2}$$

$$\mu f_1 = \frac{2 \times \mu Precision \times \mu Recall}{\mu Precision + \mu Recall}$$

N-fold - Validazione Incrociata

- In modo da ottenere una misura più stabile per F1, è naturale ripetere la valutazione su divisioni multiple
- Per ogni divisione, l'intero oracolo (es. lo standard d'oro degli esempi etichettati) è diviso in *n* parti
  - Una parte e' usata per il testing
  - Le rimanenti *n -1* parti sono usate per far apprendere il classificatore
- Il risultato finale è il valore (aspettato) significato di F1 più/meno la deviazione standard ottenuto sulle *n* ripetute misurazioni

Fasi di Categorizzazione del Testo

- Pre-Processamenteo del Corpo (tokenizzazione, stemming)
- Selezione di Caratteristica (opzionalmente)
  - Frequenza di Documento, Recupero Informazioni, χ_2 , informazione mutuale ,...
- Peso delle caratteristiche
  - per documenti e profili
- Misura di similarità
  - tra documenti e profili (es. prodotto scalare)
- Inferenza Statistica
  - applicazione della soglia
- Valutazione della Prestazione
  - Accuratezza, Precisione/Richiamo, BEP,f-misura,...

Selezione Caratteristica

- Un certo numero di feature potrebbero non essere rilevanti
- Per esempio, in TC le “function words” come “the”, “on”,”those”…
- L’algoritmo di learning ha due benefici:
  - Migliora l’efficienza
  - Migliora l’accuratezza
- Ordinare le feature per rilevanza e prendere le *m* più rilevanti

Quantità statistica per ordinare le caratteristiche

- Basato sul conti corpi della coppia "caratteristica-categoria"

- A is the number of documents in which both f and c occur, i.e. (f, c);
- B is the number of documents in which only f occurs, i.e. $(f, \bar{c})$;
- C is the number of documents in which only c occurs, i.e. $(\bar{f}, c)$;
- D is the number of documents in which neither f nor c occur, i.e. $(\bar{f}, \bar{c})$;
- N is the total number of documents, i.e. $A + B + C + D$.

Selettore Statistico

- Chi-square, MI Puntuale e MI

$$\chi^2(f,c) = \frac{N \times (AD - CB)^2}{(A+C)(B+D)(A+B)(C+D)}$$

$$PMI(f,c) = log\frac{P(f,c)}{P(f) \times P(c)}$$

$$MI(f) = -\sum_{c \in \mathcal{C}} P(c)log(P(c)) + P(f)\sum_{c \in \mathcal{C}} P(c|f)log(P(c|f)) + P(\bar{f})\sum_{c \in \mathcal{C}} P(c|\bar{f})log(P(c|\bar{f}))$$

Stima delle Probabilità

- $P(f, c)$ is the probability that f and c co-occurs and can be estimated by A/N;
- $P(f)$ is the probability of f, estimated by $(A + B)/N$;
- $P(c)$ is the probability of f, estimated by $(A + C)/N$;
- $P(c|f)$ is the probability of c by considering only the documents that contain f. It can be estimated by $\frac{P(f,c)}{P(f)}$.
- $P(\bar{f})$ is the probability that f does not occur, estimated by $(C + D)/N$;

Stima delle Probabilità (cont)

- $P(c|\bar{f})$ is the probability of c by considering only the documents that do not contain f. It can be estimated by $\frac{P(\bar{f},c)}{P(\bar{f})}$. In turn, $P(\bar{f},c)$ is estimated by C/N.

- $\mathcal{C}$ is the collection of categories, i.e. $\{c_1, c_2, ..., c_n\}$. Note that PMI and χ^2 are defined on only two categories, i.e. c and *not* c whereas MI can be evaluated on $n > 2$ categories[7].

For example, we can apply the above formulas to evaluate the PMI as follows:

$$PMI(f,c) = \frac{A \times N}{(A+C)(A+B)}$$

Selettore Globale

$$PMI_{max}(f) = \max_{c \in \mathcal{C}} \quad MI(f,c)$$

$$PMI_{avg}(f) = \sum_{c \in \mathcal{C}} P(c) \times MI(f,c)$$

$$\chi^2_{max}(f) = \max_{c \in \mathcal{C}} \quad \chi^2(f,c)$$

$$\chi^2_{avg}(f) = \sum_{c \in \mathcal{C}} P(c) \times \chi^2(f,c)$$

Pesaggio del documento: un esempio

- N, il numero generale di documenti,
- N_f, il numero di documenti che contiene la caratteristica *f*
- o_f^d le occorrenze della caratteristica *f* nel documento *d*

- Il peso *f* in un documento è:

$$w_f^d = (\log \frac{N}{N_f}) o_f^d = IDF(f) \cdot o_f^d$$

- Il peso puo' essere normalizzato:

$$w'^d_f = \frac{w_f^d}{\sqrt{\sum_{t \in d} (w_t^d)^2}}$$

Stima della Similarità

- Data la rappresentazione del documento

$$\vec{d} = < w_{f_1}^d, \ldots, w_{f_n}^d >$$

- Data la rappresentazione della categoria

$$\vec{C}_i = < W_{f_1}^i, \ldots, W_{f_n}^i >$$

- Può essere definità la seguente funzione di similarità (misura del coseno)

$$s_{di} = \cos(\angle \vec{d}, \vec{C}_i) = \frac{\vec{d} \times \vec{C}^i}{\|\vec{d}\| \times \|\vec{C}_i\|} = \frac{\sum_f w_f^d \times W_f^i}{\|\vec{d}\| \times \|\vec{C}_i\|}$$

Il Classificatore Rocchio

- $\vec{d}_f$, il peso di *f* in *d*
  - Icuni schemi di pesaggio (es. TF * IDF, Salton 91')

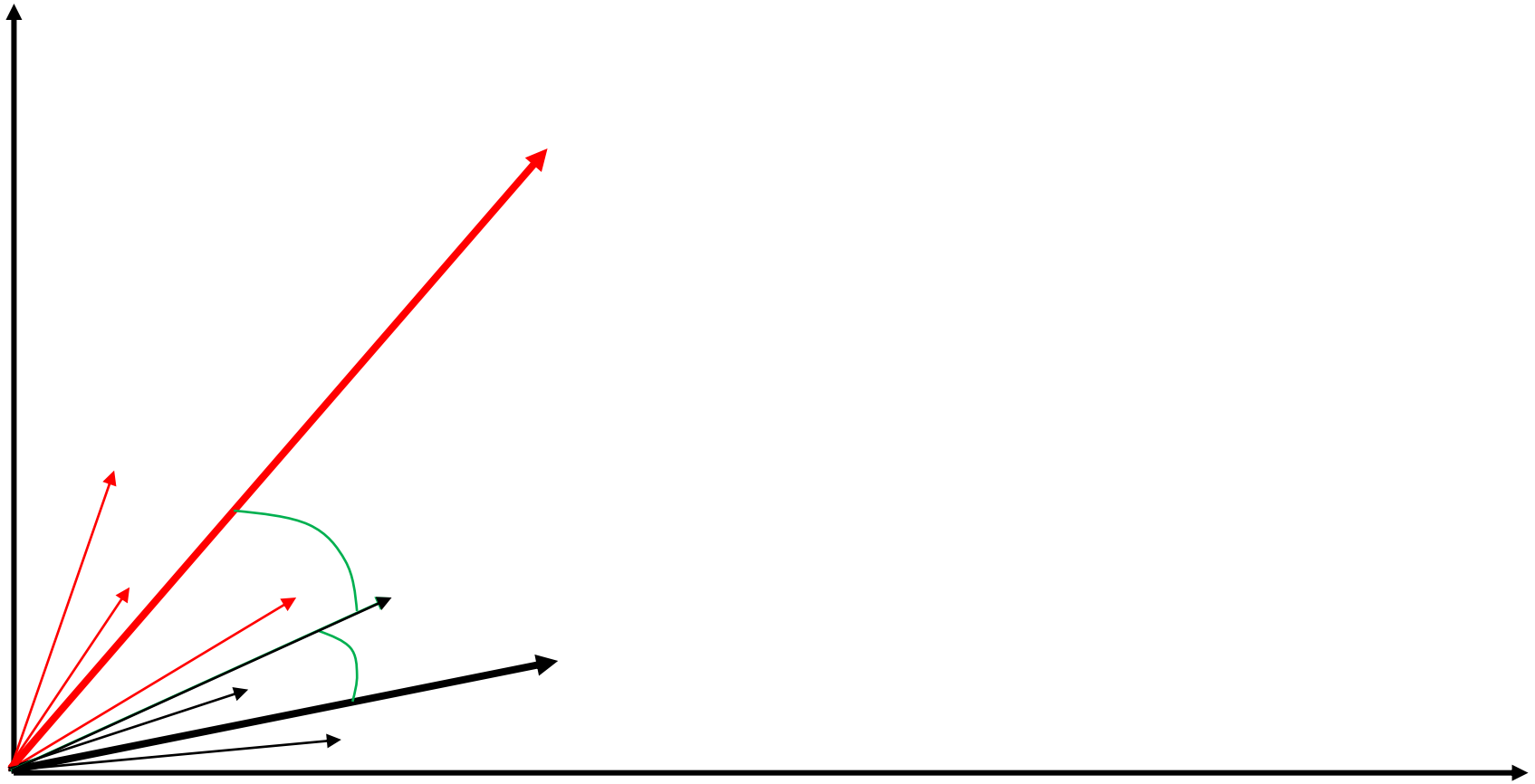
- $\vec{C}_f^i$, i pesi del profilo di *f* in C_i:

$$\vec{C}_f^i = \max\left\{ 0, \quad \frac{\beta}{|T_i|}\sum_{d\in T_i}\vec{d}_f - \frac{\gamma}{|\overline{T}_i|}\sum_{d\in \overline{T}_i}\vec{d}_f \right\}$$

- T_i, i documenti di training in C^i

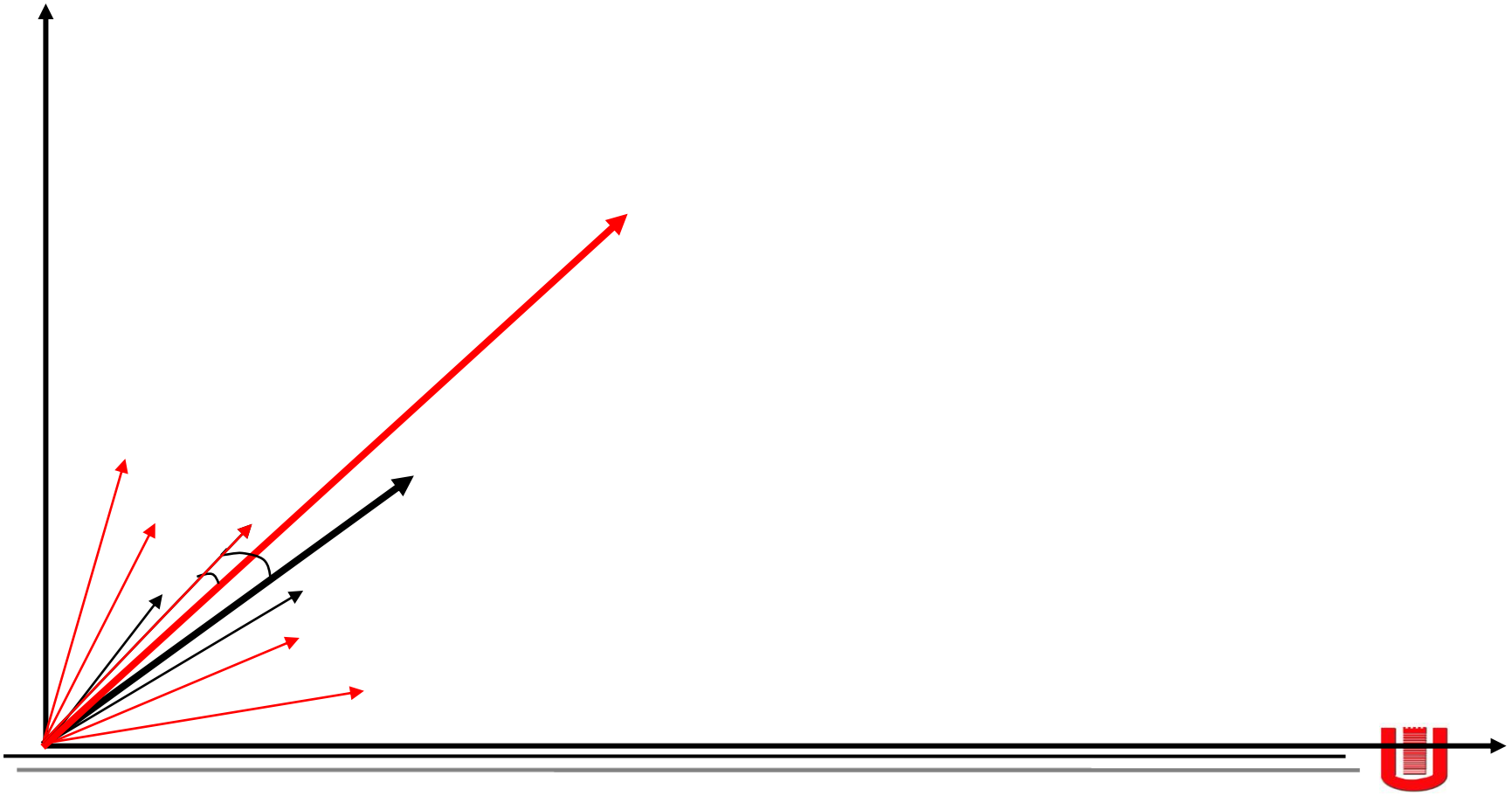
- *d* è assegnato in C^i se $\vec{d}\times\vec{C}^i > th$

Visione bidimensionale della categorizzazione di Rocchio



Limitazioni del TC di Rocchio

- Modelli basati sul prototipo hanno problemi con le categorie polimorfiche (es. disgiuntive)



Interpretazione dei parametri di Rocchio con la Selezione delle Caratteristiche

- Il lavoro di letteratura usa un gruppo di valori per β e γ
- Interpretazioni delle informazioni di positivo (β) vs. negativi (γ)
- $\Rightarrow$ valore di $\beta > \gamma$ (e.g. 16, 4)

- La nostra interpretazione [IJAIT 2002, ECIR 2003]:
- Rimuovi un parametro

$$\vec{C}_f^i = \max\left\{ 0, \quad \frac{1}{|T_i|}\sum_{d\in T_i}\vec{d}_f - \frac{\rho}{|\overline{T}_i|}\sum_{d\in \overline{T}_i}\vec{d}_f \right\}$$

- 0 caratteristiche pesate non affliggono la stima di similarità
- aumentare ρ causa che molte caratteristiche siano settate a 0 $\Rightarrow$ sono rimosse

Interpretazione dei parametri di Rocchio con la Selezione delle Caratteristiche

- Aumentando ρ :
  - Caratteristiche che hanno grandi pesi negativi ottengono prima un valore zero
  - Grande peso negativo occorre molto di frequente per altre categorie
  - $\Rightarrow$ zero peso per caratteristiche irrilevanti
- Se ρ e'un selettore di caratteristica, settalo in accordo con la strategia di selezione caratteristica standard [Yang, 97]
- In più, possiamo trovare un valore massimo ρ_{max} (associato con tutte le caratteristiche rimosse)
- Questa interpretazione abilita $\gamma >> \beta$

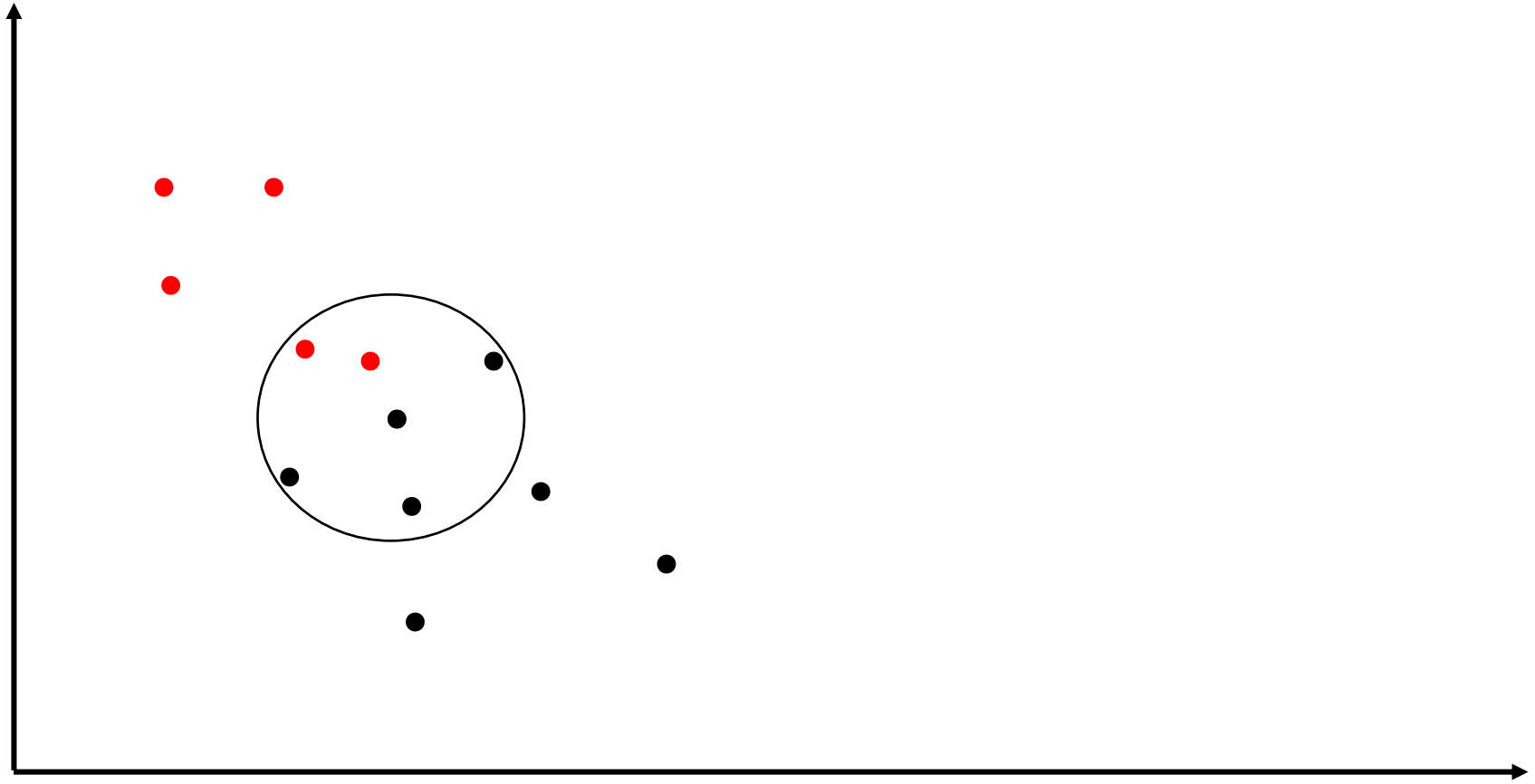
Apprendimento Nearest-Neighbor

- Apprendimento basato sulla memoria: l'apprendimento è solo immagazzinare le rappresentazioni degli esempi di training in *D*.
- Testando l'istanza *x*:
  - Computa la similarità tra *x* e tutti gli esempi in *D*.
  - Assegna *x* alla categoria dell'esempio più simile in *D*.
- Non computa esplicitamente una generalizzazione o un prototipo di categoria.
- Anche chiamato:
  - Case-based (basato sul caso)
  - Memory-based (basato sulla memoria)
  - Lazy learning (apprendimento pigro)

K Nearest-Neighbor

- Usando solo l'esempio più vicino per determinare la categorizzazione siamo soggetti ad errori a causa di:
  - Un singolo esempio atipico.
  - Rumore (es. errore) nell'etichetta di categoria di un singolo esempio di training.
- Un'alternativa più robusta è di trovare il *k* più simile esempio e restituire la maggiore categoria di questi *k* esempi.
- Il valore di *k* è tipicamente dispari, 3 e 5 sono i più comuni.

5 Illustrazioni di Nearest Neighbor (Distanza Euclidea)



K Nearest Neighbor per il Testo

Apprendimento:

Per ogni esempio di training $\langle x, c(x) \rangle \in D$

Computa il corrispondente vettore TF - IDF, $\mathbf{d}_x$, per il documento x

Istanza di test y:

Computa il vettore TF-IDF $\mathbf{d}$ per il documento y

Per ogni $\langle x, c(x) \rangle \in D$

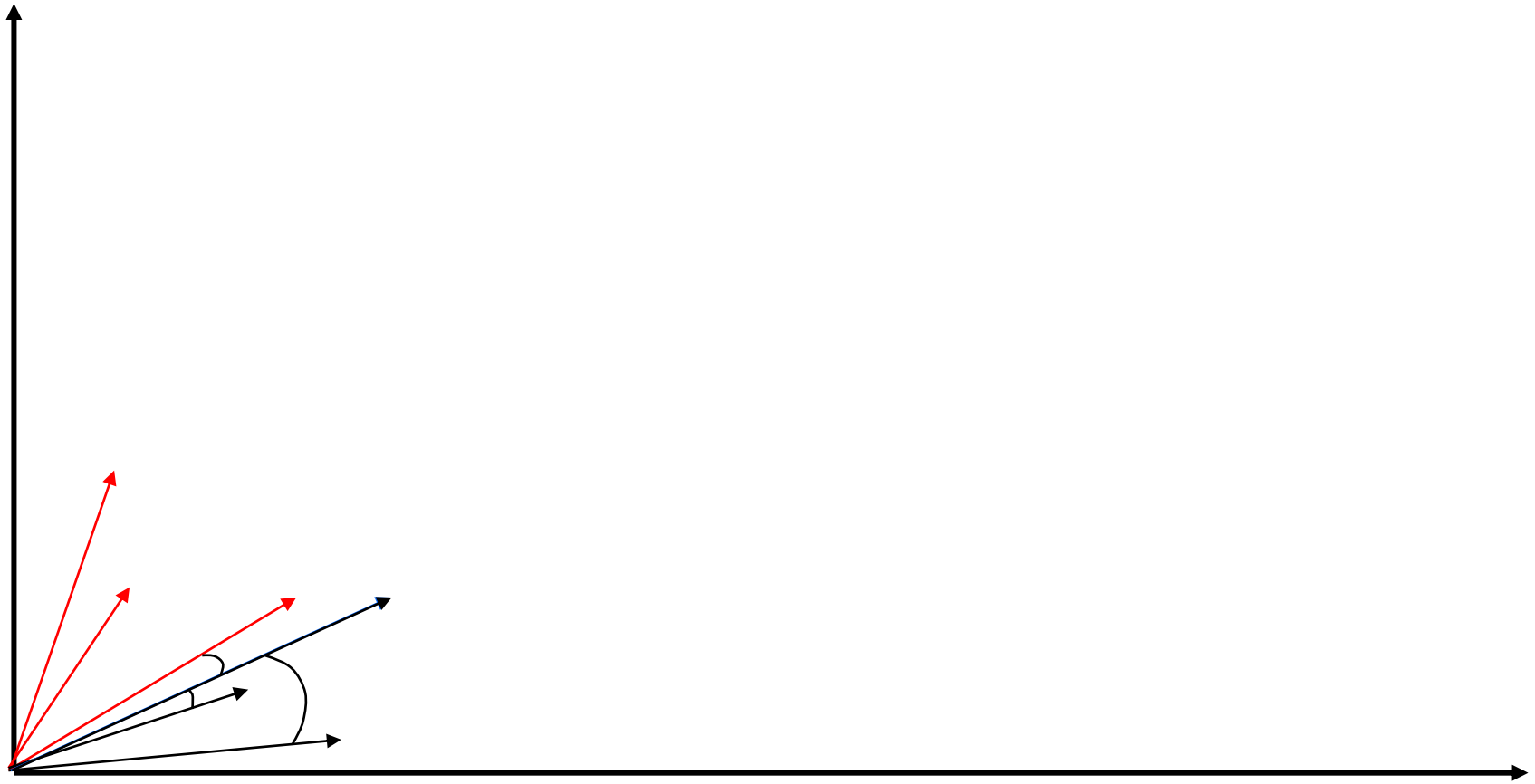
Sia $s_x = \text{cosSim}(\mathbf{d}, \mathbf{d}_x)$

Ordina esempi, x, in D decrescendo valori di s_x

Sia N il più vicino esempio k (il primo) in D. *(ottieni k più simili vicini)*

Restituisci la classe maggioritaria degli esempi in N

Illustrazione di 3 Nearest Neighbor per il Testo



Altri classificatori di testo

- *RIPPER* [Cohen and Singer, 1999] usa una nozione estesa di un profilo. Impara i contesti che sono positivamente correlati con le classi obiettivo, es. co-occorrenze di parole.
- EXPERT usa come contesto parole vicine (sequenze di parole).
- *CLASSI* e'un sistema che usa un approccio basato sulle reti neurali per la categorizzazione di testo [Ng *et al.*, 1997]. Le unità base della rete sono solo percettroni.
- *Dtree* [Quinlan, 1986] e' un sistema basato su un modello ben conosciuto di apprendimento macchina.
- *CHARADE* [I. Moulinier and Ganascia, 1996] e *SWAP*1 [Apt´e *et al.*, 1994] usa algoritmi di apprendimento macchina per estrarre induttivamente le regole Disgiuntive di Form Normale dai documenti di training.

Esperimenti

- Reuters Collection 21578 Apté split (Apté94)
  - 90 classi (12,902 documenti)
  - Un taglio fisso tra l'insieme di training e di test
  - 9603 contro 3299 documenti
- Token
  - circa 30,000 differenti
- altre differenti versioni sono state usate ma ...

  la maggioranza dei risultati TC sono relativi al 21578 Apté
  - [Joachims 1998], [Lam and Ho 1998], [Dumais et al. 1998], [Li Yamanishi 1999], [Weiss et al. 1999],

    [Cohen and Singer 1999]…

Un documento Reuters: Categoria Acquisizione

CRA SOLD FORREST GOLD FOR 76 MLN DLRS - WHIM CREEK

SYDNEY, April 8 - <Whim Creek Consolidated NL> said the consortium it is leading will pay 76.55 mln dlrs for the acquisition of CRA Ltd's <CRAA.S> <Forrest Gold Pty Ltd> unit, reported yesterday.

CRA and Whim Creek did not disclose the price yesterday.

Whim Creek will hold 44 pct of the consortium, while <Austwhim Resources NL> will hold 27 pct and <Croesus Mining NL> 29 pct, it said in a statement.

As reported, Forrest Gold owns two mines in Western Australia producing a combined 37,000 ounces of gold a year. It also owns an undeveloped gold project.

Un documento Reuters: Categoria Olio Crudo

FTC URGES VETO OF GEORGIA GASOLINE STATION BILL

WASHINGTON, March 20 - The Federal Trade Commission said its staff has urged the governor of Georgia to veto a bill that would prohibit petroleum refiners from owning and operating retail gasoline stations.

The proposed legislation is aimed at preventing large oil refiners and marketers from using predatory or monopolistic practices against franchised dealers.

But the FTC said fears of refiner-owned stations as part of a scheme of predatory or monopolistic practices are unfounded. It called the bill anticompetitive and warned that it would force higher gasoline prices for Georgia motorists.

Interpretazione dei parametri di Rocchio con la Selezione delle Caratteristiche

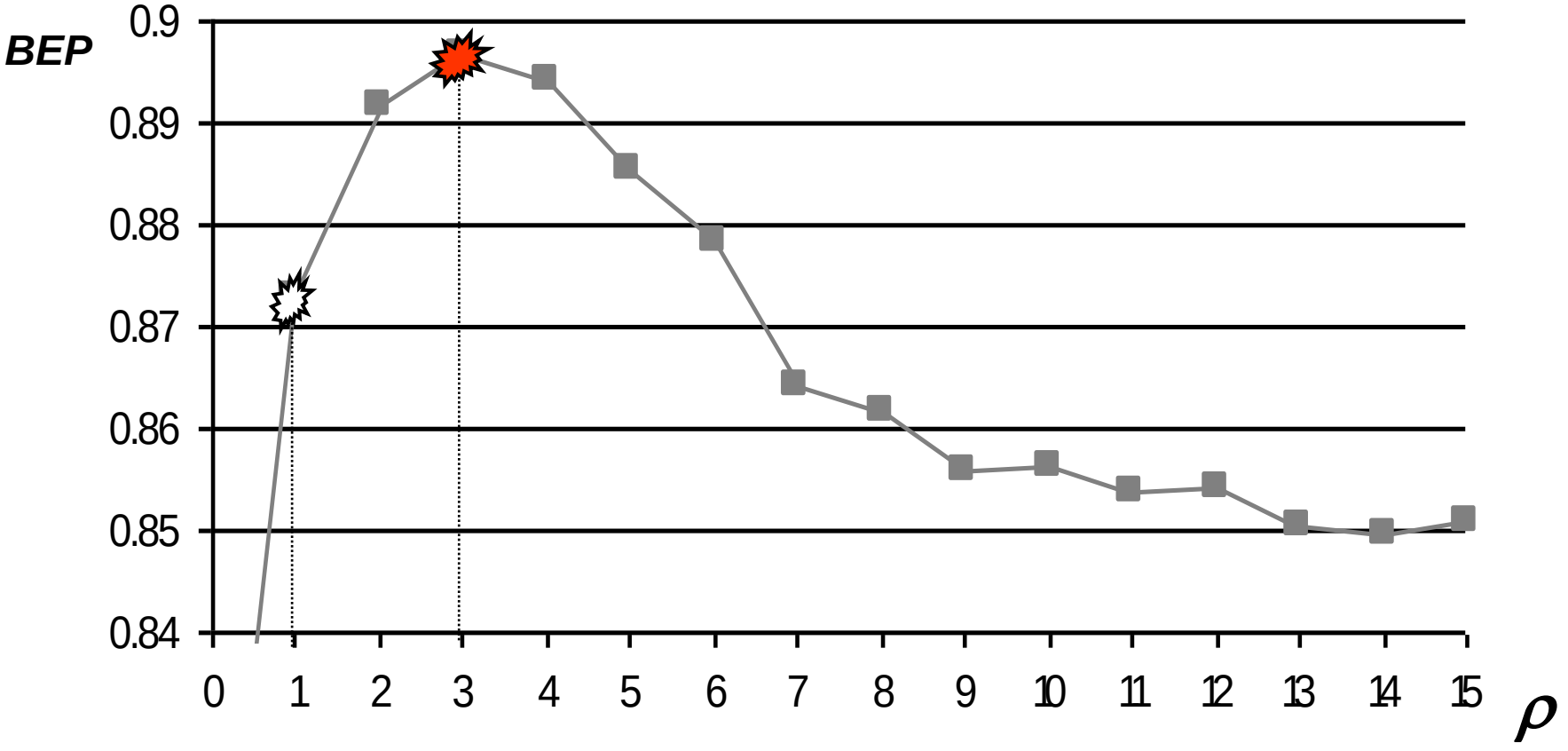
- Il lavoro di letteratura usa un gruppo di valori per β e γ
- Interpretazioni delle informazioni di positivo (β) vs. negativi (γ)
- $\Rightarrow$ valore di $\beta > \gamma$ (e.g. 16, 4)

- La nostra interpretazione [IJAIT 2002, ECIR 2003]:
- Rimuovi un parametro

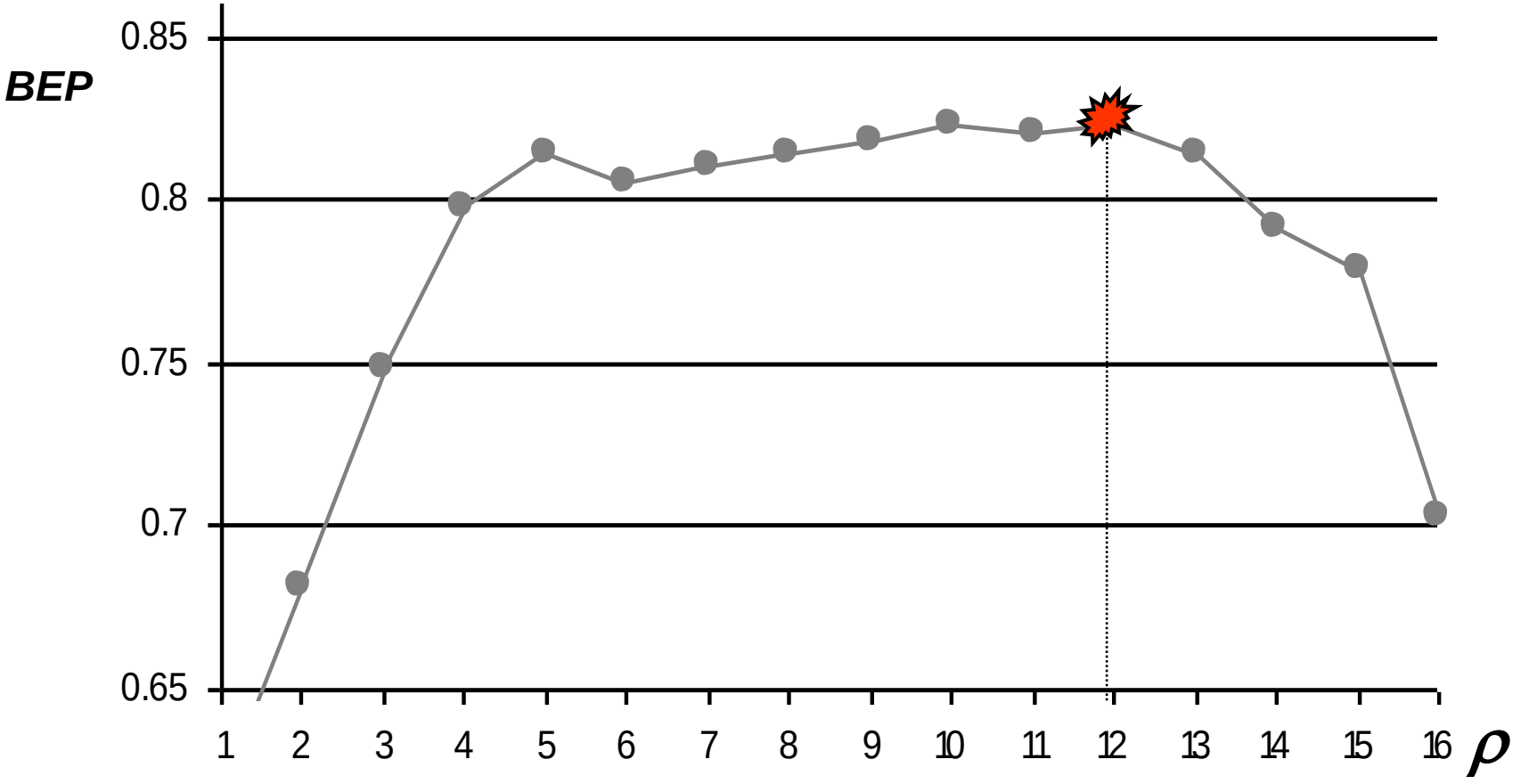
$$\vec{C}_f^i = \max\left\{ 0, \quad \frac{1}{|T_i|}\sum_{d\in T_i}\vec{d}_f - \frac{\rho}{|\overline{T}_i|}\sum_{d\in \overline{T}_i}\vec{d}_f \right\}$$

- 0 caratteristiche pesate non affliggono la stima di similarità
- aumentare ρ causa che molte caratteristiche siano settate a 0 $\Rightarrow$ sono rimosse

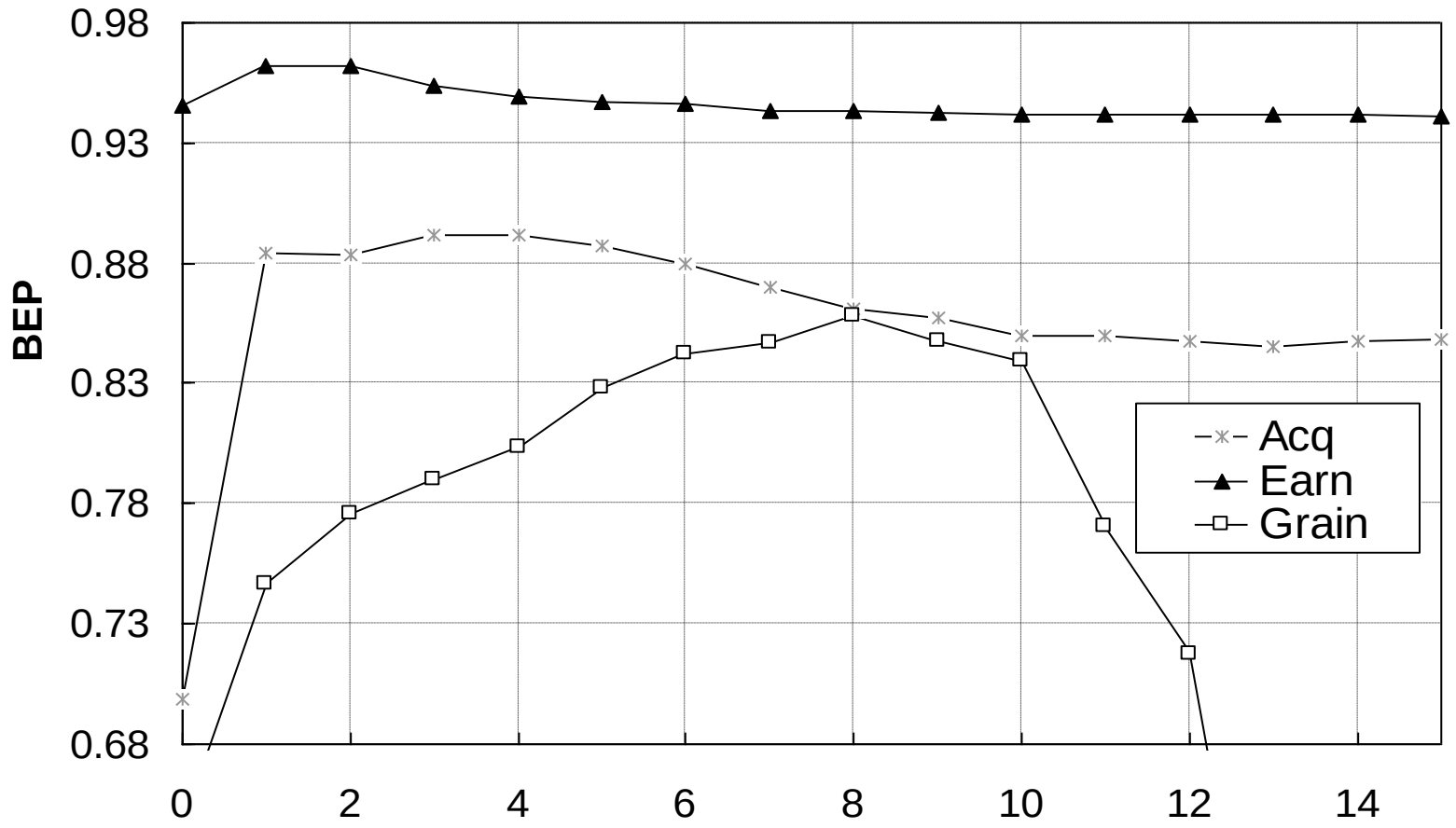
L'impatto dei ρ parametri sulla Categoria Acquisizione



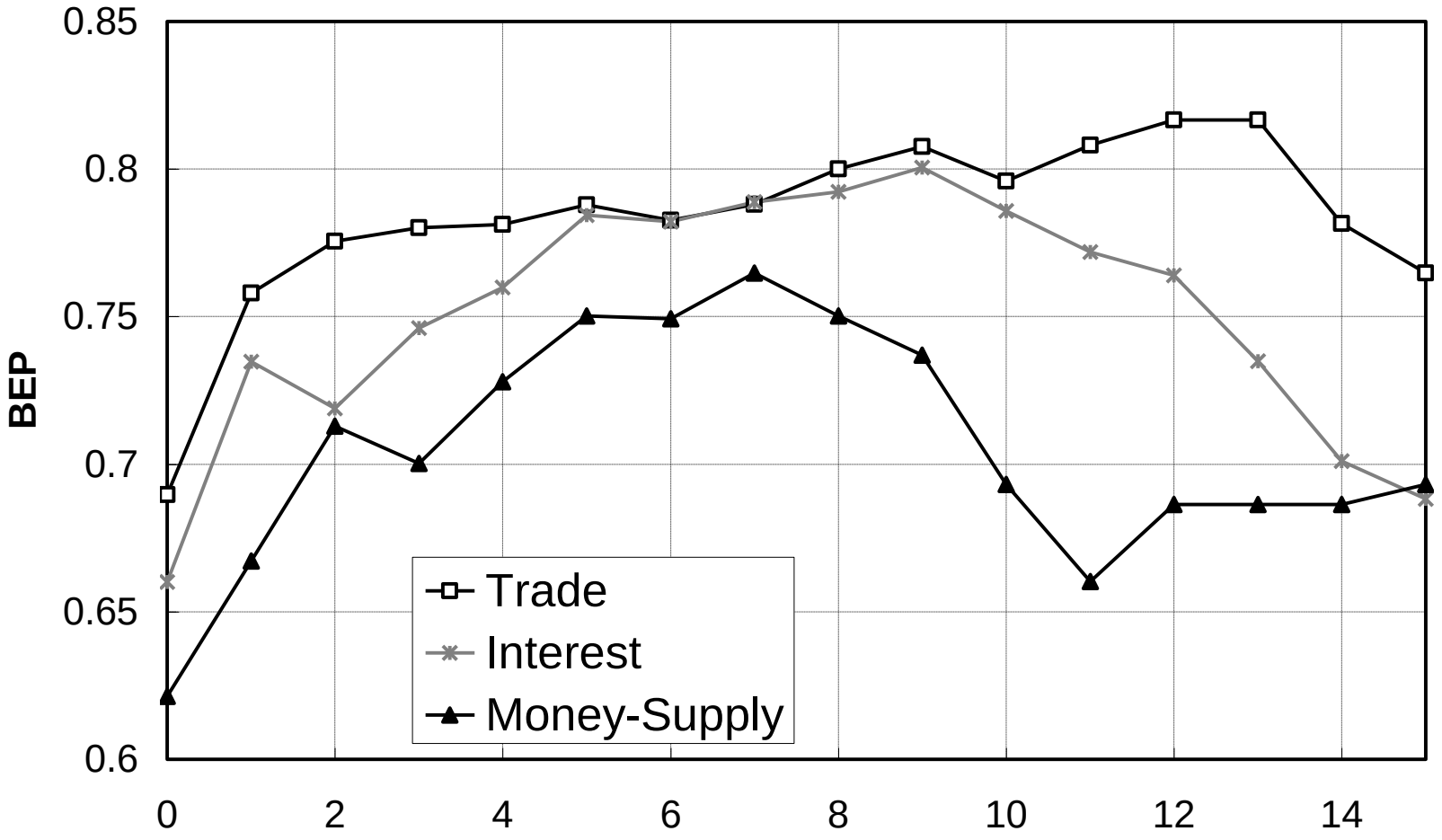
L'impatto dei ρ parametri sulla Categoria Commercio



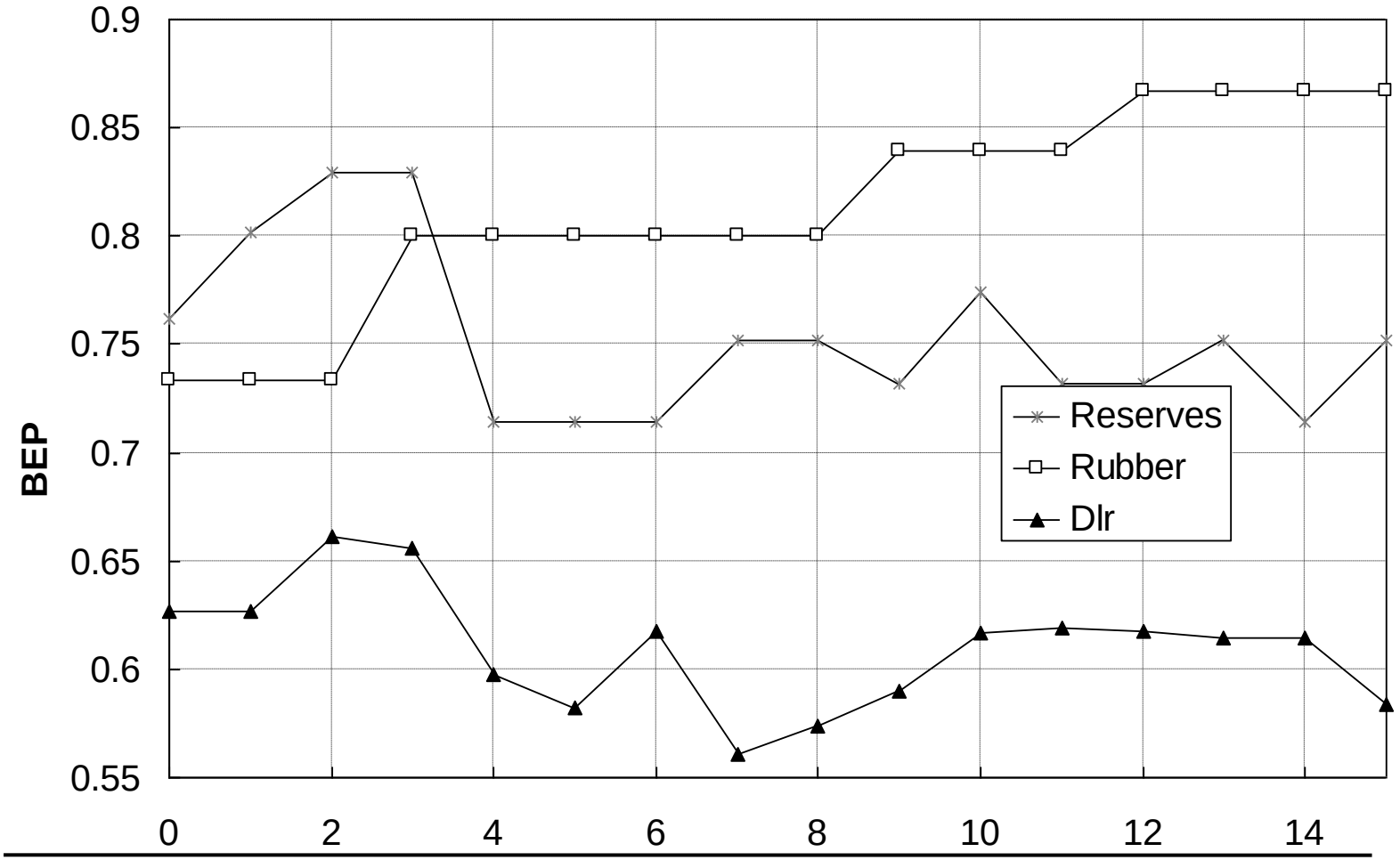
Categorie più popolate



Categorie di Media Grandezza



Categorie di Piccola Grandezza



Procedura di Stima Parametri

- Insieme di Validazione di circa 30% del corpo di training
- per tutti i $\rho \in [0,30]$
  - ADDESTRA il sistema sul materiale rimanente
  - Misura il BEP nell'insieme di validazione
- Prendi la ρ associata al più grande *BEP*
- ri -ADDESTRA il sistema sull'intero insieme di training
- TESTA il sistema basato sul modello ottenuto parametrizzato
- Per risultati più affidabili:
  - 20 insiemi di validazione e rendi ρ medio
- Il Classificazione Parametrizzato di Rocchio si riferirà come PRC

Analisi Comparativa

- Parametrizzazione della letteratura di Rocchio
  - $\rho = 1$ ($\gamma = \beta = 1$) e $\rho = 1/4$ ($\gamma = 4$, $\beta = 16$)
- Insiemi di test fissati di Reuters
  - Altri risultati di letteratura
- SVM
  - Per meglio collocare i nostri risultati
- Validazione Incrociata (20 esempi)
  - Risultati più affidabili
- Validazione incrociata corpo/linguaggio
  - Reuters, Ohsumed (Inglese) e ANSA (Italiano)

Risultati sul taglio fisso di Reuters

Insieme Caratteristiche (~30.000)	PRC	Std Rocchio ($\gamma = ¼\ \beta$ or $\gamma = \beta$)	SVM
Token	82.83 %	72.71% - 78.79 %	85.34 %
Letteratura (stemi)	-	75 % - 79.9%	84.2 %

- Risultati della letteratura di Rocchio (Yang 99', Choen 98', Joachims98')
- Risultati della letteratura SVM (Joachims 98')

Punti di Parità su ben conosciuti classificatori su Reuters

SVM	**PRC**	KNN	RIPPER	CLASSI*	Dtree
85.34%	82.83%	82.3%	82%	80.2%	79.4%

SWAP1*	CHARADE*	EXPERT	Rocchio	Naive Bayes
80.5%	78.3%	82.7%	72%-79.5%	75 % - 79.9%

* Valutazione su differenti versioni di Reuters

Validazione Incrociata

1. Generate n random splits of the corpus. For each split j, 70% of data can be used for training (LS^j) and 30% for testing (TS^j).

2. For each split j
   (a) Generate m validation sets, ES^j_k of about 10/30% of LS^j.
   (b) Learn the classifiers on $LS^j - {ES^j}_k$ and for each ${ES^j}_k$ evaluate: (i) the threshold associated to the BEP and (ii) the optimal parameter ρ.
   (c) Learn the classifiers Rocchio, $SVMs$ and PRC on LS^j: in case of PRC use the estimated $\bar{\rho}$.
   (d) Evaluate f_1 on TS_j (use the estimated thresholds for Rocchio and PRC) for each category and account data for the final processing of the global μf_1.

3. For each classifier evaluate the mean and the Standard Deviation for f_1 and μf_1 over the TS_j sets.

Validazione Incrociata su Reuters (20 esempi)

	Rocchio				PRC		SVM	
	RTS		TS^{σ}		RTS	TS^{σ}	RTS	TS^{σ}
	ρ=.25	ρ=1	ρ=.25	ρ=1				
earn	95.69	95.61	92.57±0.51	93.71±0.42	95.31	94.01±0.33	98.29	97.70±0.31
acq	59.85	82.71	60.02±1.22	77.69±1.15	85.95	83.92±1.01	95.10	94.14±0.57
money-fx	53.74	57.76	67.38±2.84	71.60±2.78	62.31	77.65±2.72	75.96	84.68±2.42
grain	73.64	80.69	70.76±2.05	77.54±1.61	89.12	91.46±1.26	92.47	93.43±1.38
crude	73.58	80.45	75.91±2.54	81.56±1.97	81.54	81.18±2.20	87.09	86.77±1.65
trade	53.00	69.26	61.41±3.21	71.76±2.73	80.33	79.61±2.28	80.18	80.57±1.90
interest	51.02	58.25	59.12±3.44	64.05±3.81	70.22	69.02±3.40	71.82	75.74±2.27
ship	69.86	84.04	65.93±4.69	75.33±4.41	86.77	81.86±2.95	84.15	85.97±2.83
wheat	70.23	74.48	76.13±3.53	78.93±3.00	84.29	89.19±1.98	84.44	87.61±2.39
corn	64.81	66.12	66.04±4.80	68.21±4.82	89.91	88.32±2.39	89.53	85.73±3.79
MicroAvg. 90 cat.	72.61	78.79	73.87±0.51	78.92±0.47	82.83	83.51±0.44	85.42	87.64±0.55

Corpora: Ohsumed e ANSA

- Ohsumed:
  - Include 50,216 abstract medici
  - I primi 20.000 documenti dell'anno 91
  - 23 *categorie MeSH* di malattie [Joachims, 1998]

- ANSA:
  - 16,000 oggetti notizia in Italiano dall'agenzia notizie dell'ANSA,
  - 8 categorie obiettivo,
  - 2,000 documenti ognuno,
  - es. Politica, Sport o Economia.

- Testaggio al 30 %

Un documento Ohsumed:
Infezioni Batteriche e Micosi

Replacement of an aortic valve cusp after neonatal endocarditis.
Septic arthritis developed in a neonate after an infection of her hand.
Despite medical and surgical treatment endocarditis of her aortic valve developed and the resultant regurgitation required emergency surgery.
At operation a new valve cusp was fashioned from preserved calf pericardium.
Nine years later she was well and had full exercise tolerance with minimal aortic regurgitation.

Validazione Incrociata su Ohsumed/ANSA (20 esempi)

	Rocchio		PRC	SVM
Ohsumed	BEP		f1	f1
MicroAvg.	ρ=.25	ρ=1		
(23 cat.)	54.4 $\pm$ .5	61.8 $\pm$.5	65.8$\pm$.4	68.37$\pm$.5

	Rocchio		PRC
ANSA	BEP		f1
MicroAvg.	ρ=.25	ρ=1	
(8 cat.)	61.76 $\pm$.5	67.23 $\pm$.5	71.00 $\pm$.4

Complessità Computazionale

- PRC
  - Facile da Implementare
  - Bassa complessità di addestramento O(*n*m log n*m*)
    - (*n* = numero di doc. e *m* = massimo numero di caratteristiche in un documento)
  - Bassa classificazione di complessità

    *min{O(M), O(m*log(M))}* (*M* e' il max numero di caratteristiche in un profilo)
  - *Buone prestazioni: il secondo classificatore più accurato su Reuters*
- SVM
  - Implementazione più complessa
  - Tempo di Apprendimento alto > $O(n^2)$ (per risolvere il problema di ottimizzazione quadratica)
  - Bassa complessità di classificazione fase (per SVM lineari)

    *min{O(M), O(m*log(M))}*

Sommario

- La valutazione di prestazioni dei classificatori di testo è eseguita contro una porzione dello standard oro chiamato insieme di test.
- Indici di prestazione sono in genere prodotti in una base per classe (p_i, r_i, F_i) e poi possono essere computati globalmente attraverso micro-media lungo le classi
- In questa lezione abbiamo discusso 2 approcci geometrici alla classificazione automatica del testo
- Classificatore Rocchio
  - *Classificatore non-parametrico basato su criteri empirici*
  - *Versione parametrizzata del PRC ottimizza il ruolo di esempi negativi lungo le classi*
  - *Buone Prestazioni: il secondo miglior accurato classificatore su Reuters*
- Lazy learning: K-NN
  - Nessuna generalizzazione e'tentata
  - *Bassa complessita'se la ricerca per il migliore esempio k è ottimizzata*