Web Mining

Mining di documenti (testi)

- rappresentazioni di testi
- classificazione di testi
- regole di associazione per collezioni di testi
- clustering di collezioni di testi

Mining di documenti web e applicazioni multimediali

Rappresentazioni di Testi

Test rappresentato da un vettore di attributi (termini, features)

- ogni termine rappresenta una parola che accade nella collezione di testi

termine è unità linguistica

- parola come si trova nel testo
- stem (radice: suffisso eliminato)
- parola o stem + tag grammaticale

Misure di frequenza di termini

Selezionare features (cancellare certe parole dei testi)

Costruire features (identificare termini composti)

Unità linguistiche: Stems e Tags

Lemmatizzazione

- le parole sono proiettate ad un'entrata standardizzata di un lessico
- p.e. riduzione delle parole alle loro radice (stemming)
  - computes, computing, computer -> comput
- lemmatizzatori esistono per le lingue differenti
- alcune lingue con l'inventario limitato delle desinenze grammaticali inglese
- Tedesco: riduzione di una dimensione di 70%

Tags

- termini con tags (parola, lemma, stem)
- tag indica la funzione grammaticale della parola
- tags: p.e. sostantivo.dativo.singulare.mascolino verbo.3.singulare
- Tedesco: tagging aumenta dimensione di 200% (risolvendo le ambiguità)

Frequenze

N testi in una collezione di testi

K termini in collezione (dimensione dei vettori)

- numero di casi del termine w_k in testo t_i: $TF(w_k, t_i)$
- numero di casi del termine w_k in collezione: $TF(w_k) = \sum TF(w_k, t_i)$ $i=1,...,K$
- term-frequency vettore di documento t_i: $TF_i=(TF(w_1,t_i),...,TF(w_K,t_i))$

Ordine di parole in testo

- tipicamente non considerato, spesso non così importante

Misure di Frequenze

Le frequenze grezze sono elaborate ad una misura di frequenza applicando 3 passi:

Trasformazioni:

bijective mappings delle frequenze grezze

Pesi d´importanza:

il vettore trasformato di frequenze è moltiplicato per un vettore di pesi di importanza (termini hanno pesi differenti)

Normalizzazione:

il vettore risultante è normalizzato alla lunghezza di unità

Distribuzioni di Frequenze

La distribuzione delle frequenze di parole in un testo è estremamente irregolare

Alcuni termini accadono molto spesso

Circa la metà dei termini accade solo una volta

Ma le parole rare contengono le informazioni altamente specifiche sul contenuto del testo

Leggi empiriche (e.g. Zipf Mandelbrot) sul percentuale delle parole con frequenza f in un testo

(tipicamente una certa funzione di $1/f^2$)

Trasformazioni di Frequenze

- frequenze gregge: nessun trasformazione, usa $TF(w_k, t_i)$
  = frequenza del termine k in testo i

- frequenze logaritmiche: $\log(1+TF(w_k, t_i))$

- frequenze inverse: $1 - 1 / (TF(w_k, t_i) +1)$ fra 0 e 1

Pesi d´Importanza

Origine in information retrieval (indexing): Salton

Misura: Come specifico ai testi della collezione è un termine

un termine che è uniformemente distribuito nella collezione è meno specifico e dovrebbe ottenere un peso basso di importanza

un termine usato soltanto in alcuni testi dovrebbe essere dato un alto peso di importanza

la frequenza trasformata di TF in un testo è moltiplicato per il peso di importanza del termine

- nessun pesi di importanza
- inverse document frequency idf, di termine w_k : $idf_k = \log(N/df_k)$
  N testi, df_k testi contengono termine w_k . Ma ignora le frequenze dei termini all'interno del testo
- ridondanza di w_k : $r_k = \log N + \sum_{i=1,...,N} (TF(w_k,t_i)/ TF(w_k) \log(TFw_k,t_i)/ TF(w_k))$

Normalizzazione

I testi differiscono nella loro lunghezza (p. e. da una dozzina a parecchie migliaia delle parole)

Le frequenze dei termini devono essere normalizzate per rendere paragonabili i testi con lunghezze differenti

term-frequency vector (frequenze grezze o trasformate e pesate) del testo t_i
$f_i=(TF(w_1,t_i),...,TF(w_K,t_i))$

f -> f / ||f|| L_1 norma divide le frequenze di termini per il numero totale di termini in testo

f -> f / ||f|| L_2 norma euklidiana

Feature Selection: Selezione di termini

Le parole/termini sono eliminate se si presentano nella collezione meno di 3, … volte

Le parole di stop sono eliminate (and, or, the, etc.)

Ranking features (ordinare secondo information gain)

seleziona il sottoinsieme delle termini *top n*

ma: poche termini irrilevanti nei documenti

Proprietà dei Testi

Alte dimensioni di termini: 10.000 -100.000

Pochi termini irrilevanti
l'ordinamento dei termini e la classificazione con i segmenti più bassi ancora dà risultati molto più meglio dei casuali

i vettori sono sparsi
ogni documento contiene soltanto alcune entrate che non sono zero

sparsi anche per tuples: qualche termine solo in pochi documenti

Compiti di Mining di Testi

Classificazione

Regole di associazione

Clustering

Applicazioni di Classificazione di Testi

Classificazione dei documenti in un numero fisso di categorie predefinite
- ogni documento può essere in una o nessun o multiple categorie
- imparare i classificatori dagli esempi
- i classificatori fanno automaticamente le assegnazioni delle categorie

ogni categoria trattata come un problema separato di classificazione binaria

rispondendo se un documento dovrebbe essere assegnato ad una categoria particolare oppure no

Classificare news stories

Trova informazioni interessanti sul web
- pagine di web su text mining (che trovo interessante: i miei esempi di addestramento)

Preprocessing dei testi: riduzione di dimensionalità (rappresenta un testo con un piccolo numero di categorie)

Classificazione di Testi

Topic identification: identifica i documenti di un soggetto dato

Collezione di documenti Reuters:
- 8762 training e 3009 test documenti, raccolti da Carnegie group (1987)
- uso standardizzato della lingua
- lunghezza del testo solitamente piccola (< 100 parole)
- parole che avvengono soltanto una volta in testo (25%)
  - molto più basso di usuale (50%)

p.e. Testi tedeschi di giornali

classe (topics): economia, sport, affari locali ecc.

O Giugno1988 vs. Giugno1991

Metodi per Classificazione di Testi

Macchine a supporto vettoriale

Classificatori *Naive Bayes*

Algoritmo di Rocchio (information retrieval)

usa combinazione lineare lc di somme degli esempi positivi (negativi) normalizzati e coseno fra il nuovo documento e lc

k esempi più vicini

misura di similarità: coseno dei vettori normalizzati di 2 documenti

Alberi di decisione

Reti neuronali

Macchine a Supporto Vettoriale e Classificazione di Testi

SVM possono efficacemente trattare vettori di 100,000 dimensioni, dato che questi sono sparsi

opzioni:

funzioni di kernel

preprocessing linguistico

Esperimenti con i testi inglesi di Reuters:

per applicare SVM, non si deve pretrattare (stemming ed eliminazione delle parole di stop) per l'inglese

anche per le lingue che sono morfologicamente più ricche, p.e. forme grammaticale?

Esperimenti con giornali tedeschi

Kernel functions di Macchine a Supporto Vettoriale

Linear kernel: $K(x,y)=\langle x,y\rangle$ prodotto scalare dei vettori x, y

second order polynomial kernel: $K(x,y)=\langle x,y\rangle^2$

Gaussian rbf kernel: $K(x,y)=e^{-||x-y||/\ 2\sigma^2}$

Confrontare Risultati

precisione: probabilità che il documento previsto per essere in categoria + appartiene a questa categoria |p & +| / |p|

recall: probabilità che documento che appartiene alla categoria + è classificato in questa categoria |p & +| / |+|

trade-off fra precision e recall (precisione più alta, recall più bassa)

precision/recall break even point

variare un certo parametro del metodo di classificazione: realizza i punti differenti di precision/recall, seleziona punto interpolato con precision=recall

definito soltanto per la classificazione binaria

si deve prendere la media dei risultati di un problema di multipli classi : microaveraging

media di tavole binarie di contingenza, non di precision/recall

Esperimenti di Macchine a Supporto Vettoriale per Classificare Testi

3 kernel functions,
3 trasformazioni di frequenza, 3 pesi di importanza, 2 normalizzazioni:

quali combinazioni sono efficaci:
- nessuna differenza fra i noccioli
- *ridondanza* migliore di *inverse document frequency*
- L2 migliore di L1
- combinazione migliore:
  - frequenza logaritmica con ridondanza e L2

Esperimenti di Macchine a Supporto Vettoriale per Classificare Testi

il preprocessing linguistico è costoso (p.e. lemmatizzazione, tagging, pesi di importanza)

SVM può trattare gli spazi altodimensionali (termini con tags di p.e. 200.000), cancellare parole rare o di stop non necessario

la lemmatizzazione è molto lento, parole al posto di lemma: nessuna perdita notevole di precision/recall

tagging richiede più sforzo che la lemmatizzazione ed i risultati non sono migliori

nessun preprocessing necessario per SVM

Altri metodi per classificare testi

Reti neuronali (testi francesi: la lemmatizzazione può essere saltata)
ma la selezione di termini necessaria (evita grandi dimensioni)
dimensione < 1000 anche per:

- alberi di decisione, reti bayesiane
- k esempi più vicini (O(l•m) tempo di classificazione)
- l training esempi, m test esempi: Troppo costoso

Teorema di Bayes e ipotesi MAP

Teorema di Bayes: $P(h|D) = P(D|h)\,(P(h) / P(D))$

Trova ipotesi migliore $h \in H$ (spazio d´ipotesi) per dati osservati D

P(h) *prior probability* di h (p.e. ogni h ha la stessa probabilità)
P(D) *prior probability* di dati osservati (nessun conoscenza su h)
P(h|D) *posterior probability* di h, per dati osservati D
MAP ipotesi *maximum a posteriori hypothesis*
P(D|h) *likelihood* di D dato h *maximum likelihood hypothesis*

Naive Bayes classifier

MAP hypothesis *maximum a posteriori hypothesis*

{argmax $P(v_j|a_1,...,a_n)$ $v_j \in V$} V attributo di target (categoriale), $a_i \in A_i$ attributi

{argmax $P(a_1,...,a_n|v_j)\ P(v_j)$ $v_j \in V$}

naive Bayes (presupposto semplificato):
valori dei attributi sono condizionalmente indipendenti, dato il valore di target
{argmax $P(v_j)\ \prod_j P(a_j|v_j)$ $v_j \in V$}

includere frequenze delle parole/termini

Risultati di Classificazione

Joachims: collezione di Reuters , topic prediction (10 topics più frequenti)
Ohsumed corpus 50.216 documenti (categorie di malattie)

microaveraged performance of precision / recall breakeven point

Bayes	72%	57	
Rocchio	80	57	
C4.5	80	50	
k-NN	82	59	k=30 (1,15,**30**,45, 60)
SVM polyn.	86	66	d=1,...,5 (risultati più o meno uguali)
SVM rbf	86	66	γ= 0.6,0.8,1,1.2 risultati uguali

Classificazione da preprocessing e data mining su vettori di topic

Data mining metodo: regole di associazione

Regole di associazione su stem/parole sono spesso inutili
associazioni fra parole composte (p.e. wall street) o non interpretabili

I termini che rappresentano un documento sono topics (soggetti, concetti)
p.e. 1000 topics vs. 100.000 parole

Regole di associazione offrono
- conoscenze sul dominio rappresentato dai documenti
  associazioni fra concetti
- accesso ai dati: documenti che osservano le regole

Document Explorer

Text mining con regole di associazione (Ronen Feldman)

Document Explorer con nuove possibilità di navigare in collezioni
- insiemi frequenti
- regole di associazione
- keyword graphs (nodo: concetto, arco: associazione)
- pattern di trend e cambiamento (collezioni dinamiche)

strumenti per l´estrazione di concetti e amministrazione delle gerarchie di concetti

Documenti e Regole di Associazione

Utente sia interessato alle alleanze di affari fra aziende

Collezione: articoli finanziari

Aziende: lati sinistri di regole
Concetti di alleanze di affari: lati destri

Regole:

America online inc, bertelsmann ag -> joint venture 13/0.72
Apple computer inc, sun microsystems inc. -> merger talk 22/0.27

Navigare: Documenti che (non) osservano regola, lato sinistra, …

Estrazione di Concetti

Metodi di classificazione di testi: spesso troppi concetti, concetti non dati

Metodi per estrarre concetti dai documenti, p.e.

profits, Canada, big banks, Canadian Imperial Bank of Commerce,
earnings season, net income, bank shares, mutual funds, pension plans

riconoscere nomi di posizioni geografiche, di persone, di aziende …

Estrazione di Concetti: Fasi

Preprocessing linguistico
- tokenization (identificare parole, nomi in testo, per differenti forme di testi)
- part of speech tagging
  - basato su regole (Brill tagger deriva regole da esempi= testi tagged a mano)
- lemmatization

Generare termini
- generare candidati semplici
- selezionare sequenze di termini con pattern: Sost-Sost o Sost-prepos-Sost
- coefficiento di associazione per paio di termini

Filtrazione dei termini → concetti
- annulli i termini ugualmente distribuiti (p.e. right direction, other issue, same time)
  - test statistici di frequenze (deviazione standard, chi-2)
  - information retrieval scores: massimo TF-IDF score (su tutti i documenti)

Tassonomie di Concetti

Diverse tassonomie
- Persone
- Aziende
- Locazioni geografiche
- Concetti economici

Entrate delle tassonomie sono usate come:
- oggetti (items) nelle regole regole gerarchiche
- vincoli per limitare applicazioni
  - p.e. solo conclusioni per un seletto concetto

Strumenti semiautomatici per costruire tassonomie
- editori
- usare insiemi frequenti, clustering di concetti

Clustering di Documenti

Usato gia nel information retrieval

Metodi numerici

- misura di similarità di documenti (coseno delle frequenze trasformate)
- agglomerazioni (complete, single linkage), k-means, etc.
- anche per documenti strutturati
  - multirelazionali: intestazione, estratto, sezioni, …

Self organizing maps per analizzare un achivio di documenti
(mappa auto-organizzante) Kohonen

Mappa Auto-Organizzante

Visualizzazione intuitiva delle similarità dei documenti
ordinamento spaziale dei documenti sulla mappa

Approcci: rete neuronale, tecnica di descrizione di dati

- rete neuronale non supervisionato
- mappa bidimensionale
- similarità da distanze in spazio di due dimensioni
- documenti simili nelle regioni vicine
- cluster (regione sulla mappa) descritto con termini rappresentativi

Mappa Auto-Organizzante: Esempio

Merkl

Time Magazine Collection: 420 documenti

stemming, poi escludere termini che compaiono in meno di 10 % dei documenti e in più di 90 % dei documenti

pesi: TFxIDF

primi 6000 termini

griglia bidimensionale 10X15 cellule

Ogni cellula rappresenta un cluster di documenti

Cellula descritta da un insieme di concetti tipici

Mappa Auto-Organizzante: Metodo

Ogni cellula i ha un vettore di pesi m_i dimensione di vettori =n= numero di termini

Metodo iterativo (t)

seleziona vettore d´un documento x (frequenze trasformate di termini)

trova cellula con distanza minima fra x e m_i

associa x alla cellula i

adotta il vettore del peso di i e di cellule vicine

$$m_i(t+1)=m_i(t)+f_1(t)\, f_{ci}(t)\, [x(t)-m_i(t)]$$

f_1 diminuisce con t, f_{ci} funzione di vicinanza

Mappa Auto-Organizzante: Descrizioni

mappa dallo spazio altodimensionale dell'input allo spazio bidimensionale della proiezione

con la topologia preservata

sommario dei documenti che sono proiettati ad una cellula

metodo euristico

- termini con deviazioni sommate basse dei vettori di documenti e del vettore di peso della cellula
- vettori sparsi: escludere termini con deviazioni sommate troppo basse

Progetto simile: Websom, ma produce mappe di categorie di parole

- parole simili sono proiettate ad una cellula

Web mining

Documenti web come testi non strutturati

Documenti strutturati

log files: internet access logs (IP address, URL, time stamp)

Esempi di due applicazioni

comportamento di navigazione di utenti (Anand)

predizione di indici di borsa quotidiani (Wüthrich)

Comportamento di navigazione di utenti

Online bookshop

identificare le sequenze delle navigazioni di clienti

deriva regole di sequenze (sequential Apriori) dalle logfiles

clienti che avevano comprato un articolo

identifica le loro sequenze tipiche quando comprano un altro articolo

applicare queste regole quando clienti navigano

offrire dinamicamente articoli specialmente adattati

Usa gerarchie per gli attributi di logfiles

Predizione di Indici di Borsa

notizie finanziarie

Wall Street Journal, Financial Times, Reuters, CNN, etc offrono versioni elettroniche

download di pagine di web e di ultimi valori di chiusura delle borse da agenti (ogni mattino)

genera vettori di frequenze (trasformate) di 400 termini (predefiniti)

applica regole che concludono i valori di chiusura odierna

Le regole sono attualizzate periodicamente

Dati e Applicazioni Multimediali

Dati audivisionali (audio-video streams vs. images)

Mining di conoscenze nascoste in questi documenti multimediali

- call center: categorizzare telefonate dei clienti
- rete della TV: categorizzare servizi
- immagini mediche
- immagini del sattelite
- video di sorveglianza
  - identificare faccia
  - identificare comportamento anormale (shoplifter, pattern di traffico e emergenza)
- news on demand agenti intelligenti controllano news channels e presentano servizi interessanti
- generare astratti di videos (video frames rappresentativi)
- classificare videos con categorie di violenza, adulto, soap
- query by image content, indexing multimedia databases

Applicazioni Multimediali: Compiti

Fusione di dati dai medi differenti

Estrazione delle informazioni semantiche (ancora con successo limitato)

- p.e. scoperta del comportamento anormale nella sorveglianza
- può essere contenuto in più di un flusso dei dati
  - visione (timore) ed audio (gridare)
- da coordinare

Tecniche di base

- analisi e riconoscimento di lingua e discorso
- traduzione dall´audio allo testo
- analisi di immagine
- scoperta degli oggetti e del movimento
- scoperta dei tagli di scena