Riassunto sondaggi sentiment

Punti chiave della sensitive analysis

UNIVERSITÀ DEGLI STUDI ROMA TRE

8 marzo 2017

Autore: Daniele Mainella

Riassunto sondaggi sentiment

Punti chiave della sensitive analysis

2003 Markov

2 metodi:

- Una statistica globale di espressioni è classificata in una gaussiana usando caratteristiche derivate dalla linea d'intonazione e il profilo dell'energia del segnale parlato;
- Complessità temporale applicando continui modelli di Markov considerando parecchi stati usando caratteristiche istantanee di basso livello anziché statistiche globali.

6 tipi di sentimenti:
rabbia, disgusto, paura, gioia, tristezza e stupore.

In generale l'intonazione mostra più potenza dell'energia.

I suoni non vocali sono eliminati nel profilo dell'intonazione.

Sembra che il primo metodo sia più preciso del primo.

2009 Survey

La stima della dimensione misura quanto un uomo prova sentimenti, positivi o negativi.
La dimensione di attivazione misura se gli uomini sono più o meno propensi a compiere un'azione, attiva o passiva, sotto stati d'animo emotivi.

Alcune emozioni sono indistinguibili (rabbia, paura, ecc.) mentre altre possono mentire (stupore).

Il resto sembra tutto molto teorico e superficiale.

2011 Survey

Il grado di naturalezza del database usato è molto importante per valutare la performance.
Se il database usato è di bassa qualità si può incorrere a conclusioni scorrette.

Tuttavia a certi database è negato l'accesso per motivi di privacy e sicurezza.

Limiti:

- Alcuni database non simulano abbastanza bene le emozioni in modo chiaro e naturale;
- In alcuni db la qualità delle espressioni registrate non è buona;
- Alcuni db non provvedono trascrizioni fonetiche.

Alcuni ricercatori dividono il segnale in piccoli intervalli chiamati frames dai quali si estrae un vettore con caratteristiche locali.
Alcuni preferiscono estrarre statistiche globali dall'intera espressione parlata.

Caratteristiche globali:

- Efficienti solo in emozioni eccitanti (rabia, paura e gioia);

- Falliscono nel classificare emozioni che hanno eccitazioni simili;
- Informazioni temporali presenti nel dominio del tempo vengono perse.

Altri approcci:

- Segmentazione del segnale a fonemi sottostanti e poi calcolare un vettore di caratteristiche per ciascun fonema segmentato (vero per suoni vocali);
- Estrazione di un vettore di caratteristiche per ciascun segmento di voce piuttosto che di ciascun fonema.

Un problema importante nel riconoscimento delle emozioni è l'estrazione di caratteristiche del parlato che caratterizzano il concetto emozionale e allo stesso tempo non dipendono da chi parla o dal contenuto lessicale.

Le caratteristiche del parlato si suddividono in 4 categorie:

- Caratteristiche continue;
- Caratteristiche qualitative;
- Caratteristiche spettrali;
- Caratteristiche TEO (Teager energy operator)-based.

Secondo uno studio, gli stati emotivi di chi parla riguardano l'energia complessiva, la distribuzione dell'energia attraverso lo spettro delle frequenze e le pause e durata delle frequenze del segnale vocale.

Le caratteristiche sonore si possono suddividere nelle seguenti categorie:

- Caratteristiche legate all'intonazione;
- Caratteristiche delle formanti;
- Caratteristiche legate all'energia;
- Caratteristiche del tempo;
- Caratteristiche delle dizioni.

Caratteristiche globali più usate comunemente nell'analisi delle emozioni:

- Frequenza fondamentale;
- Energia;
- Durata;
- Formanti;

Formante = frequenza di risonanza attorno alla quale un suono spettralmente ricco mostra un picco di ampiezza. Una vocale può essere identificata tramite il rapporto tra la prima e la seconda formante.
(cit. Wikipedia)

Le emozioni di rabbia, paura, gioia e stupore hanno caratteristiche simili nella frequenza fondamentale f0:

- Intonazione media: valore medio di f0 per l'espressione;
- Profilo della pendenza: la pendenza del profilo di f0;
- Abbassamento finale: la ripidezza di f0 diminuisce alla fine del profilo cadente o all'inizio della salita;
- Gamma dell'intonazione: la differenza tra il più alto e il più basso valore di f0;
- Linea di riferimento; il valore costante di f0 dopo un'escursione tra un'intonazione alta o bassa.

Il contenuto emozionale di un'espressione è fortemente legato alla qualità della voce.

Caratteristiche acustiche legate alla qualità sonora:

- Livello di voce;
- Intonazione vocale;

- Frasi, fonemi, parole e caratteristiche marginali;
- Strutture temporali.

Il contenuto emozionale di un'espressione ha un impatto sulla distribuzione dell'energia spettrale attraverso la gamma delle frequenze.

Per sfruttare al meglio la distribuzione spettrale circa la gamma delle frequenze udibili, lo spettro stimato è spesso passato in un filtro passa basso.

Il parlato è prodotto da un flusso d'aria non lineare nel sistema vocale.
La tensione dei muscoli di chi parla influenza il flusso d'aria presente nel sistema vocale producendo il suono.
TEO è definita come:
$\Psi\{x(n)\} = x^2(n) - x(n-1)x(n+1)$

È stato osservato che sotto condizioni di stress la frequenza fondamentale cambia.
Caratteristiche TEO-based possono essere usate per la rilevazione dello stress nel parlato, dunque usare TEO solo per quello.

Per equalizzare l'effetto di propagazione del parlato tramite l'aria, un filtro di radiazione preliminare è usato per processare il segnale vocale prima di estrarre le sue caratteristiche.
Per ridurre l'increspatura nello spettro ciascun frame è spesso moltiplicato con Hamming window prima dell'estrazione delle caratteristiche.
Questi intervalli sono solitamente mantenuti intatti nell'analisi delle emozioni.
Qualche post-processo può essere necessario prima che i vettori delle caratteristiche siano usati per testare il classificatore.

Un sistema di ricognizione delle emozioni consiste in due fasi:

- Un'unità di processo del fronte finale che estrae le caratteristiche appropriate dai dati del parlato;
- Un classificatore che decide le emozioni sottostanti delle espressioni parlate.

2012 Emotion

I classificatori usati nel riconoscimento delle emozioni sono KNN (k-nearest neighbors), HMM (Hidden Markov Model), SVM (Support Vector Machine), ANN (Artificial Neural Network) e GMM (Gaussian Mixtures Model).

Il riconoscimento delle emozioni non è nulla senza il riconoscimento dei pattern.
I passi presenti nel riconoscimento dei pattern sono anche presenti nel sistema di riconoscimento delle emozioni.

I passi per il riconoscimento delle emozioni sono:

- Speech input;
- Estrazione delle caratteristiche;
- Selezione delle caratteristiche;
- Classificatori;
- Emozione riconosciuta.

Estrazione e selezione delle caratteristiche

Le caratteristiche del parlato possono essere divise in due principali categorie: lungo termine e breve termine.
Il segnale vocale è diviso in piccoli intervalli che sono attribuiti come un frame.

Ricercare sull'emozione del parlato suggerisce che l'intonazione, l'energia, la durata, i formanti, MFCC (Mel frequency cepstrum coefficient) e LPCC (linear prediction cepstrum coefficient) sono caratteristiche fondamentali.

La **rabbia** ha il valore e varianza del parlato e valore medio di energia più alto.

Nella **felicità** c'è un miglioramento nel valore medio, gamma di variazione del parlato e valore medio di energia.

Nella **tristezza** il valore medio e gamma di variazione dell'intonazione diminuiscono, l'energia è debole, lo speak rate è basso e diminuisce lo spettro dei componenti delle alte frequenze.

La **paura** ha un alto valore medio e gamma di variazione dell'intonazione e un miglioramento dello spettro dei componenti delle alte frequenze.

Una delle caratteristiche principali che indica l'emozione è l'energia, e lo studio dell'energia è dipendente nell'energia a breve termine e nell'ampiezza media a breve termine.

Selezione dei classificatori

Un classificatore riconosce l'emozione nell'espressione di chi parla.

GMM è il più adatto per la rilevazione delle emozioni. Tutte le equazioni sono basate sulla supposizione che tutti i vettori sono indipendenti dal momento che GMM non può formare strutture temporali dei dati acquisiti.

ANN è usato per la sua abilità di trovare margini non lineari separatamente agli altri stati emozionali.

Riguardo gli stati delle emozioni della k espressione, **KNN** alloca un'espressione a una condizione dell'emozione. Il classificatore può classificare tutte le espressioni in un appropriato insieme mirato.

HMM ha ottenuto grande successo per modellare informazioni temporali nello spettro del parlato. Ha l'importante vantaggio che le dinamiche temporali delle caratteristiche del parlato possono essere prese secondo l'accessibilità di procedure ben stabilite per l'ottimizzazione di strutture di riconoscimento.

SVM è generalmente usato nel riconoscimento dei pattern e classificazione dei problemi. Ha una migliore performance di classificazione rispetto agli altri.

2014 Neural

DNN (deep neural network) è una rete neurale feed-forward che è molto più di livelli nascosti tra i suoi input e output (?).
Ha la capacità di apprendimento di rappresentazione ad alti livelli di caratteristiche grezze ed effettivi dati classificati (?):

il riconoscimento delle emozioni mira a identificare effettivi stati ad alto livello di un'espressione a partire dalle caratteristiche a basso livello.

Una promettente caratteristica di DNN è che può apprendere caratteristiche di invarianti ad alto livello da dati grezzi.

2014 Speech

Riguardo le emozioni, le informazioni più rilevanti si trovano nell'intonazione, metrica e qualità della voce.
Il passo successivo a questa strategia è scoprire le caratteristiche che determinano i dati del parlato e scartare quelle inutili.

BerlinEMO DB è un database di emozioni che consiste di 10 speakers (5 maschi e 5 femmine) ai quali è stato chiesto di ripetere 10 testi differenti in tedesco.
I files sono etichettati in 7 categorie emotive: rabbia, noia, disgusto, paura, felicità, tristezza e neutro.

Le caratteristiche estratte sono state performate usando MATLAB e sono:

- **Intonazione**
  Le regioni vocali assomigliano a un segnale periodico nel dominio del tempo.
  La periodicità associata a questi segmenti è detta 'intonazione al periodo t0', che dà la 'frequenza fondamentale f0'
- **Entropia**
  esprime il brusco cambiamento nel segnale
- **Auto Correlazione**
  Correlazione di un segnale con se stesso.
  È la somiglianza tra osservazioni come una funzione di tempo di ritardo tra esse.
  È uno strumento matematico per trovare pattern ripetuti, frequenza fondamentale o rumore nel segnale.
- **Energia**
  L'ampiezza del segnale vocale varia sensibilmente con il tempo.
  Il significato maggiore di ciò è che l'energia fornisce una base per segnali vocali distinguibili da un parlato monovocale.
- **HNR**
  Descrive la qualità del parlato.
  Fornisce un'indicazione della periodicità complessiva del segnale vocale quantificando il rapporto tra componenti periodici (parte armonica) e aperiodici (rumore)
- **ZCR**
  Zero Crossing Rate
  L'indice del cambiamento di segno lungo un segnale, i.e. l'indica al quale il segnale cambia dal positivo al negativo e viceversa.
  È un importante parametro per comprendere la variazione del parlato ed è anche utile per differenziare segnali vocali e non vocali.
- **Statistiche**
  Deviazione standard: mostra quanta varianza o dispersione esiste dalla media;
  Baricentro spettrale: la frequenza media ponderata. Indica dove giace il centro di massa dello spettro;
  Flusso spettrale: misura di quanto velocemente cambia la potenza dello spettro di un segnale, calcolato comparando la potenza dello spettro di ciascun frame contro la potenza dello spettro del frame precedente;
  Spectral roll off: è definito come la n-esima percentuale della distribuzione spettrale (n è di solito 85% o 95%).

SVM è un classificatore statistico che classifica i dati in classi binarie su dati di allenamento.
Costruisce un iperpiano o un insieme di iperpiani in uno spazio multidimensionale o uno spazio a dimensione infinita.
Per classificare usando SVM si converte il problema di multi classificazione in uno di classificazione binaria (ad esempio usando un approccio uno-contro-uno).

2015 Fourier

Nell'analisi di Fourier un segnale è decomposto nelle sue costituenti vibrazioni sinusoidali.
Un segnale periodico può essere descritto in termini di una serie di onde sinusoidali e cosinusoidali armonicamente relazionate.
Un segnale vocale può essere rappresentato come il risultato del passaggio della forma d'onda di un'eccitazione glottale attraverso un filtro lineare di tempo variabile, il quale modella le caratteristiche risonanti del tratto vocale.
La parte armonica del modello è la rappresentazione di una serie di Fourier dei componenti periodici di un segnale vocale: quando una componente aperiodica viene campionata, la sua trasformazione di Fourier diventa una funzione di frequenza periodica e continua.
La DFT (discrete Fourier transform) deriva dal campionamento della trasformata di Fourier di un segnale discreto nel tempo a N frequenze discrete.

I database usati sono: EMODB (corpo emozionale, tedesco), CASIA (database emozionale cinese) e EESDB (il più anziano db di emozioni cinese).

Le armoniche includono frequenza, ampiezza e fase.

I punti medi da H1 a H20 sono differenti guardando le 7 emozioni.
I valori medi di H3 per tristezza, noia e neutrale sono più alti di disgusto, gioia e ansia.
Il picchio di ogni emozione è alle armoniche più basse: H6 per gioia e rabbia e H4 per neutrale, noia, ansia e disgusto.

Normalizzazione:
obiettivo = eliminare la variabilità di chi parla e della registrazione mentre si mantengono le effettive discriminanti delle emozioni.

SVM usa una ottimizzazione convessa quadratica che è vantaggiosa nella creazione di una soluzione ottima globale.
Ci sono due differenti famiglie di soluzioni che mirano a estendere SVM per problemi di multiclasse: la prima segue la strategia "uno-contro-tutti", mentre la seconda segue la strategia "uno-contro-uno".
Il secondo metodo è più conveniente nella pratica.