The PMEmo dataset for MER

PMEmo dataset contiene annotazioni musicali e segnali EDA (ElectroDermal Activity) per 794 brani collezionando i dati di 457 soggetti.

I brani sono stati pre-computati con audio features e selezionati solo le parti di chorus.

Le ricerche su MER si basano principalmente su lavori di machine learning prediction come nel lavoro [23] che però richiede un grande numero di brani con emotional labels per fare training del modello. Questo però ovviamente è costoso a livello di tempo e laborioso per il fatto di acquisizione dei dati.

Per eliminare il problema delle differenze culturali nella percezione della musica come descritto in [8] hanno fatto partecipare persone di diverse nazioni.

Il dataset fornisce:

* Metadata del brano:
  + Titolo
  + Artista
  + Inizio e fine della parte di chorus
* Selezioni manuali delle clip in MP3
* Pre-computed features
* Emotion labels:
  + Per l’intera clip
  + Dinamiche, con segmenti di 0.5 s

Partendo da altri dataset come Emotify, MoodSwing etc. e dallo studio di psicologi le emozioni vengono divise in tre categorie come descritto in [22]:

* Expressed emotion
* Perceived emotion
* Felt (o evoked) emotion

Di solito ci si concentra su perceived emotion e ci sono vari approcci che descrivono come riportare le risposte degli ascoltatori, di solito usando:

* cathegorical method: considera le emozioni come categorie distinte, ad esempio Emotify [1] applica GEMS classifica le emozioni nei nove fattori per music-induced emotions.

Il problema del cathegorical approach è che le emozioni sono molto più varie e hanno molte più sfaccettature come spiegato in [22].

* dimensional method: Valence-Arousal method bi-dimensionale propost da Russel [12] è il più usato.

Oltre alle annotazioni soggettive lavori [3] e [7] usano dati di segnali EEG (electroencephalogram signals). DEAP dataset [10] è un dataset che ha relazionato segnali EEG e emotion rating da 32 volontari con 22 persone con anche il video della faccia.

Il dataset ha metadata in CSV delimitati da virgole.

I brani sono presi da classifiche popolari come Billiboard Hot 100, iTunes Top 100 songs etc. hanno poi collezionato 1000 brani ma c’erano alcuni duplicati e quindi sono arrivati a 794 brani pop.

Le clip sono state poi tagliate nella parte di chorus e sono di diversa lunghezza.

I soggetti in totale erano 457, 236 femmine e 221 maschi. 366 dall’università cinese di cui 44 esperti di musica e gli altri no.

Sono stati invitati poi 47 esperti di inglese per annotare il dataset.

Ogni brano ha ricevuto un totale di almeno 10 annotazioni incluse una da un esperto di musica e una da un esperto di inglese.

Le electrodermal activities sono prese in modo continuo con un sampling rate di 50 Hz. È stata sviluppata una applicazione desktop per annotare i brani con una sliding area per acquisire i dati dinamicamente su una scala da 1 a 9 con un sampling rate di 2 Hz. Viene chiesta poi una annotazione generale di tutto il brano. Poi viene fatto ascoltare di nuovo il brano per fare una annotazione di tipo Valence-Arousal.

I soggetti impiegano 50 min circa per le annotazioni, fatto firmare una liberatoria e fatto vedere un video su come fare le annotazioni con anche un momento di training per rendere famigliare il fatto delle annotazioni durante i brani.

Per staccare da un brano ad un altro è stata aggiunta una procedura di relax di 10 secondi in cui si chiedeva al soggetto di rilassarsi con la musica di *In my Life* di *Kevin Kern* come fatto in [9].

Il trasduttore EDA era [BSL-SS3LA](https://www.biopac.com/product/eda-finger-transducer-bsl/) posto sulle dita della mano non dominante.

Dopo il settaggio iniziale ad ogni persona viene chiesto di fare l’annotazione di 20 brani uno dei quali duplicato per garantire l’alto livello dei dati come fatto in [4] (i dati vengono considerati validi solo se il bias tra le clip duplicate è entro lo 0.25 nello spazio VA).

A screenshot of a cell phone

Description automatically generatedAlla fine vengono pagati circa 6€.

In conclusione, sono stati sottoposti all’esperimento 457 persone di cui accettate 401 (87.7%).

La distribuzione delle annotazioni statiche è più concentrata nel primo quadrante del VA.

Initial Orientation Time (IOT) esprime il fatto che le persone che devono annotare hanno un tempo iniziale in cui danno delle risposte poco significative, secondo [14] solo dopo 8 s di IOT si può considerare valida una annotazione per il Valence, mentre si passa a 12 secondi per l’Arousal.

In [16] nel 2013 e poi [2] nel 2017 hanno visto che si iniziano ad avere delle risposte accettabili dopo il decimo secondo. Per questo hanno deciso di scartare le annotazioni per i primi 15 secondi dalle annotazioni dinamiche.

Per valutare la consistenza delle annotazioni hanno usato Cronbach’s come definito in [13].

Le annotazioni hanno un sampling rate di 2 Hz e le annotazioni vengono processate secondo la formula (come fatto anche in [2]):

Dove:

* indica il soggetto.
* indica l’istante di tempo.
* la label annotata dal soggetto al tempo .
* la media di tutte le label dinamiche annotate dal soggetto .
* la media di tutte le label per quella canzone da tutti i soggetti.

A screenshot of a cell phone

Description automatically generatedLa tabella indica la media (pesata tra i brani) e la deviazione standard della Cronbach’s del dataset PMEmo:

Il dataset viene poi usato per fare automatic emotion recognition che cerca di predire le emozioni in modo statico e dinamico di un brano.

Per MER non c’è ancora un modello, ma è utile l’INTERSPEECH ComParE (Computational Paralinguistics Evaluation) del 2013 [15] che ha proposto un set di features per dare detection. In questo set ci sono descrittori di basso livello (LLDs, MFFCs, energy, etc.). ComParE ha applicato anche funzioni statistiche come media e momenti a LLDs.

Questo porta ad un set di 6373 dimensioni che funziona bene per music mood regression.

PMEmo utilizza tutte le 6373 dimensioni di features per il riconoscimento a livello statico, ma vengono poi estratte solo 260 dimensioni per il riconoscimento dinamico (per ridurre il calcolo computazionale).

L’estrazione delle features è fatto con openSMILE in [6] e non vengono implementati altri metodi di estrazione di features così da dimostrare la capacità di PMEmo per il task di MER.

Hanno lavorato con due metodi per classificare le emozioni in VA:

* Multivariate Linear Regression (MLR)
* Support Vector Regressio (SVR)

Per tutti i brani hanno fatto training e testing dei classificatori usanto le 6373-dimension features separando le label statiche di Valence e Arousal e .

Il modello statico è:

Dove:

* è il numero dei brani.
* è il numero del brano.
* è il set di features della -esima canzone.
* è il valore di Valence o Arousal per la -esima canzone.

Ispirati da [21] hanno deciso di organizzare il dynamic MER model creando un classificatore globale con 6373 livelli per suddividere il valor medio delle emozioni dinamiche e nel frattempo usando una segmentazione a 260 livelli per features locali per trovare le fluttuazioni nel tempo.

Il modello globale:

Il modello locale:

La dynamic emotion:

Dove:

* è il numero dei brani.
* è il numero del brano.
* è il set globale.
* è il numero di timestamps nella -esima canzone.
* è una matrice di 260 colonne e righe che rappresentano le features locali del brano.
* è la media delle emozioni calcolate dinamicamente.
* è la fluttuazione ad ogni timestamp.

Prima di fare il regression model hanno fatto una resize delle annotazioni in per essere di miglior comprensione.

Performance di MER statica:

il task statico è di predire l’emozione generale di tutto il brano musicale rappresentato da un singolo valore nel piano VA. Per fare training il PMEmo dataset è stato diviso in 11 parti, 10 che costituiscono il training set e uno usato per testare il modello.

Sono stati usati per metriche di valutazione Root Mean Square Error (RMSE) e Pearson Correlation Coefficient () calcolati individualmente per Valence e Arousal.

A screenshot of a cell phone

Description automatically generatedEsempio su due run, run 1 basato su MLR, trained sulle 10 parti del modello, run 2 basato su SVR con Radial Basis kernel function, sempre trained sulle 10 parti del modello:

Performance di MER dinamica:

Essendo che gli estratti di brani del dataset PMEmo sono state fatte due operazioni:

* Global-scale operation: per ogni brano, estrarre solo una feature globale e mapparla in una emozione globale.
* Local-scale operation: per ogni brano, dividere in segmenti da un secondo con 50% di overlap (con 2 Hz di fluctuation frequency) e poi estraendo le features lovali da ogni segmento proiettandolo nel mood space.

A screenshot of a cell phone

Description automatically generatedIl fatto della divisione del dataset per fare training è fatta come nei calcoli della performance statica. La valutazione è:

L’errore maggiore è nella scala globale di RMSE, consistentemente al lavoro di [21] e i valori di nella scala locale non sono ottimali, il che limita le performance del dynamic emotion recognition. Questo fa si che non si riesca a riconoscere la variazione immediatamente dell’emozione nel breve tempo.

Predizione delle emozioni usando EDA:

loro hanno fatto dynamic MERcon EDA che varia nel tempo. È stato applicato in primis un LPF a 0.6 Hz per togliere il rumore, poi la skin electric conductance è stata scalata in z-score come:

Dove:

* è il vettore.
* è la media di .
* è la deviazione standard di .

A screenshot of a cell phone

Description automatically generatedHanno quindi fatto il training e testing MLR e SVR modelli con pre-processed EDA data:

Valori bassi di RMSE indicano la correlazione tra perceived emotion e evoked emotion.