Emotion recognition in Music

Anurag Sharma
Shivam Khandelwal

November 5, 2015

Outline

Music Emotion Recognition

What is Emotion in Music ?

- Music plays an important role in human society

Music Emotion Recognition

What is Emotion in Music ?

- Music plays an important role in human society
- Has ability to affect our mood and elicit emotions

Music Emotion Recognition

What is Emotion in Music ?

- Music plays an important role in human society
- Has ability to affect our mood and elicit emotions
- Sentiments and feeling induced while listening to the music is termed as emotion in music

Applications of MER

GNOOSIC DISCOVER NEW MUSIC

Why is MER important ?

- Emotion retrieval finds application in music recommender systems

Applications of MER

GNOOSIC DISCOVER NEW MUSIC

Why is MER important ?

- Emotion retrieval finds application in music recommender systems
- Useful for auto tagging of huge online music libraries

Applications of MER

GNOOSIC DISCOVER NEW MUSIC

Why is MER important ?

- Emotion retrieval finds application in music recommender systems
- Useful for auto tagging of huge online music libraries
- Used in social media information retrieval and sentiment analysis

Applications of MER

GNOOSIC DISCOVER NEW MUSIC

Why is MER important ?

- Emotion retrieval finds application in music recommender systems
- Useful for auto tagging of huge online music libraries
- Used in social media information retrieval and sentiment analysis
- Can also find applications in niche areas like Music therapy

Difficulties in MER

Why is it difficult?

- Less consensus on what emotions does a music elicit

Difficulties in MER

Why is it difficult?

- Less consensus on what emotions does a music elicit
- Perception variance among subjects is high

Difficulties in MER

Why is it difficult?

- Less consensus on what emotions does a music elicit
- Perception variance among subjects is high
- Quantification of emotion is also a difficult problem

Difficulties in MER

Why is it difficult?

- Less consensus on what emotions does a music elicit
- Perception variance among subjects is high
- Quantification of emotion is also a difficult problem
- Less knowledge about the ground truth and about what acoustic or signal processing features to use

Outline

Representation of Music Emotion

Categorical Description

Subjects are asked to choose from a set of predefined words. These words are then used for grouping and assigning tags to the soundtrack

Representation of Music Emotion

Categorical Description

Subjects are asked to choose from a set of predefined words. These words are then used for grouping and assigning tags to the soundtrack

Sample words

Clusters	Mood Adjectives
Cluster 1	aggressive, fiery, tense, intense
Cluster 2	passionate, rousing, boisterous, rowdy
Cluster 3	rollicking, cheerful, fun, sweet, amiable
Cluster 4	poignant, bittersweet, autumnal, brooding
Cluster 5	humorous, silly, quirky, whimsical, witty, wry

Some sample sound tracks

Multidimensional Description

- A set of predefined dimensions is taken and subjects are asked to assign a value in a given a range to each dimension based on the soundtrack

Some sample sound tracks

Multidimensional Description

- A set of predefined dimensions is taken and subjects are asked to assign a value in a given a range to each dimension based on the soundtrack
- Arousal and Valence are popular dimension for music description which we also use in our study

Some sample sound tracks

Valence - Arousal Dimensions

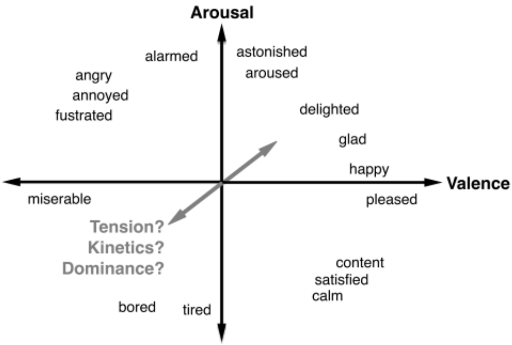


Figure - Describes the common emotions felt and corresponding AV values.

Outline

Problem Introduction

Problem at hand

Given a song clip and associated feature vector predict the arousal and valence values for the song

Problem Introduction

Problem at hand

Given a song clip and associated feature vector predict the arousal and valence values for the song

Continuous prediction

- Each song clip is of length 30 seconds and is divided in intervals of length 0.5 seconds each
- Predict A-V values for each of the 60 intervals for the sound clip

Mathematical description

The above described task can be modeled as a regression problem

Mathematical description

The above described task can be modeled as a regression problem
At some time instance t_i

- Let Y_i be an arousal or valence value for a song

Mathematical description

The above described task can be modeled as a regression problem
At some time instance t_i

- Let Y_i be an arousal or valence value for a song
- Now if $X_1, X_2, \ldots X_n$ are n features for the song at t_i

Mathematical description

The above described task can be modeled as a regression problem
At some time instance t_i

- Let Y_i be an arousal or valence value for a song
- Now if $X_1, X_2, \ldots X_n$ are n features for the song at t_i
- We want to model $Y_i \sim f(X_1, X_2, \ldots X_n)$ for some function f

Dataset description

Training dataset Description

- 744 music clips of 30 second length each
- Each clip partitioned into $0.5s$ interval (hence, 60 time instances)
- For instance t_i, we have Y_i and X_i where,
  1. Y_i = Arousal/Valence annotation for each song at t_i
  2. X_i = Data matrix with 6000 features for each song at t_i

Dataset description

Training dataset Description

- 744 music clips of 30 second length each
- Each clip partitioned into 0.5s interval (hence, 60 time instances)
- For instance t_i, we have Y_i and X_i where,
  1. Y_i = Arousal/Valence annotation for each song at t_i
  2. X_i = Data matrix with 6000 features for each song at t_i

Five low level features

RMS energy, MFCC, CHROMA, voicing probability, fundamental frequency

Data Visualization

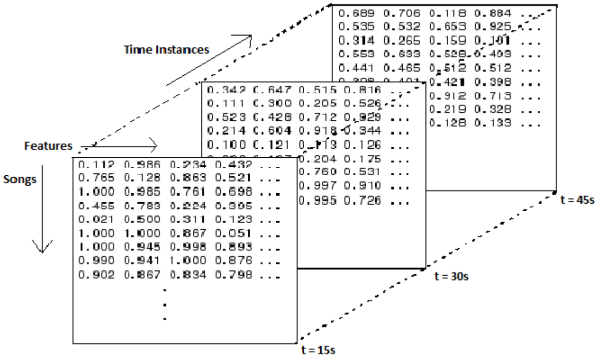


Figure: Visualization of data matrix X_i

Problem Challenges

Challenge A

- Large feature dimensions compared to the number of observations (8 GB of text files in total)

Problem Challenges

Challenge *A*

- Large feature dimensions compared to the number of observations (8 GB of text files in total)
- Many redundant features in the dataset

Problem Challenges

Challenge *A*

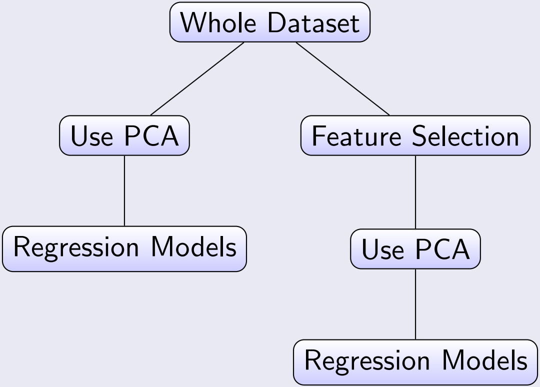
- Large feature dimensions compared to the number of observations (8 GB of text files in total)
- Many redundant features in the dataset

Challenge *B*

- Prediction of continuous Arousal - Valence values

Experiment setup

Two different approaches



Feature Selection

Dimensionality Reduction

- Removed near zero variance variables
- Removed correlated variables from the dataset
- Used top 50% correlated variables with y
- Reduced feature dimension to 1500 from 6000 (2 GB of text files)

Methodology

Assumptions

Independence assumption for each time interval

Methodology

Assumptions

Independence assumption for each time interval

Method

- Different model is trained for each time interval
- Gaussian filtering is done to incorporate temporal information

Models used

Linear Models

- Started off with **multiple linear regression**
- Since $n << p$, used **lasso** and **elastic net** models to penalize predictors

Models used

Linear Models

- Started off with **multiple linear regression**
- Since $n << p$, used **lasso** and **elastic net** models to penalize predictors

Non-linear Models

- To check whether given data was non-linear
- Used **SVR** with different kernels and **Random Forest**
- Most widely used methods in MER

Outline

Evaluation Criteria used in Literature

- The correlation between predicted A/V values and the ground truth is used for evaluation
- The Average correlation of 1000 songs in test dataset is reported for each model

Correlation Values for whole dataset

Using PCA of the whole data set for Arousal and Valence

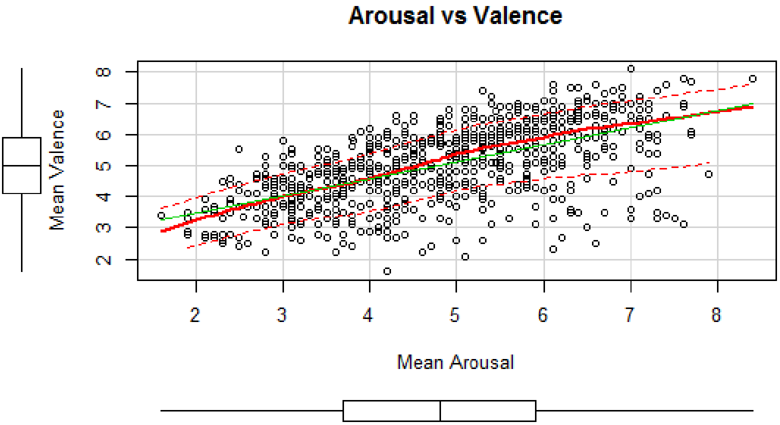
Methods (with PCA)	Correlation Values	
	Arousal	Valence
Multiple Linear Regression	0.082 ± 0.220	0.015 ± 0.071
Lasso Regression	0.096 ± 0.045	0.023 ± 0.011
Elastic Nets	0.103 ± 0.092	0.047 ± 0.059
Random Forest	**0.155 ± 0.113**	**0.077 ± 0.045**
SVR		
polynomial	0.098 ± 0.112	0.022 ± 0.054
radial	0.106 ± 0.215	0.034 ± 0.021
Baseline	0.050 ± 0.430	-0.020 ± 0.590

Correlation Values for reduced dataset

Using PCA of the reduced data set for Arousal and Valence

	Correlation Values	
Methods (with PCA)	Arousal	Valence
Multiple Linear Regression	0.103 ± 0.17	0.019 ± 0.04
Lasso Regression	0.196 ± 0.095	0.038 ± 0.015
Elastic Nets	0.215 ± 0.108	0.059 ± 0.021
Random Forest	0.209 ± 0.081	**0.094 ± 0.033**
SVR		
polynomial	0.098 ± 0.812	0.022 ± 0.054
radial	**0.223 ± 0.076**	0.074 ± 0.029
Baseline	0.050 ± 0.430	-0.020 ± 0.590

Mean plot for V-A



Correlation Values

Correlation between Arousal and Valence

Correlation is 0.557

This shows that we can use one for predicting the other and and the results are in line with the intuition.

Correlation Values

Correlation between Arousal and Valence

Correlation is 0.557

This shows that we can use one for predicting the other and and the results are in line with the intuition.

Using PCA of the reduced data set and Arousal for Valence

Method	Correlation
SVR	0.068 ± 0.047
SVR (smoothing)	0.096 ± 0.038
Random Forest	0.091 ± 0.039
Random Forest (smoothing)	$\mathbf{0.126} \pm \mathbf{0.031}$

Outline

Insights and future direction of work

Insights

- Non-linear models work best for A/V value prediction
- The strong dependence b/w valence and arousal can be exploited for prediction
- Gaussian smoothing greatly improves the correlation value – Emotions don't change fast

Insights and future direction of work

Insights

- Non-linear models work best for A/V value prediction
- The strong dependence b/w valence and arousal can be exploited for prediction
- Gaussian smoothing greatly improves the correlation value – Emotions don't change fast

Future Direction of work

- Partial Least Square regression is also used in similar problem previously
- Using other methods to incorporate temporal information
- Use predicted continuous A-V values to predict static A-V value

Thank you!