Deep content-based music recommendation

A¨aron van den Oord, Sander Dieleman, Benjamin Schrauwen

Park, Jun Sep

Goal

- Latent factor model for recommendation
- The latent factors from music audio when they cannot be obtained from usage data (implicit feedback)

Challenge – traditional method

- Collaborative filtering approach

- Cold Start Problem
    - Usage Data is scarce
    - Necessity of rating information (explicit feedback)

Latent Factor

- Implicit feedback (Taste Prole Subset contains play counts per song and per user.)

- WMF to learn latent factor representations of all users and items
  (Weighted Matrix Factorization)

- Preference vector : p_{ui} (u : user , i : song)
- Assume the user enjoys the song : $p_{ui} = 1$

- Confidence issue : $c_{ui} = 1 + \alpha \log(1 + \epsilon^{-1} r_{ui})$
  (α , ϵ are hyperparameters)

WMF objective function

- $\min_{x*y*} \sum_{u,i} c_{ui}(p_{ui} - x_u^T y_i)^2 + \lambda(\sum_u ||x_u||^2 + \sum_i ||y_i||^2)$

→ MSE(mean square error) → L2 regularization

- λ : regularization parameter
- x_u : latent factor vector for user u
- y_i : latent factor vector for song i

- ALS(alternating least square) optimization method

  → to take all possible combinations into account (versus gradient descent)

Predicting (music audio → latent factors)

- Time series (audio signal) → vector of real numbers
- Converting ~ Regression Problem

- 2 Methods
  - BOW (bag-of-words) : input to a classifier / regressor
  - CNN : Convolutional neural network
    (state-of-the-art speech recognition)

CNN (convolutional neural network)

- Ingredients of the Network
    1. ReLU (vanishing problem)
    2. Theano library (GPU acceleration)
    3. MSD training data

- MFCCs : (method) extracting the characteristics of sound
    → analyzing the spectrum of intervals by dividing a short term

Objective function for real value

- MSE

$$\min_{\theta} \sum ||y_i - y_i'||^2$$

- WPE (weighted prediction error)

$$\min_{\theta} \sum_{u,i} c_{ui}(p_{ui} - x_u^T y_i')^2$$

θ : model parameter
y_i : latent factor vector for song i
y_i' : prediction by the model

Dataset

- MSD (million song dataset)
  - precomputed audio feature
  - 1 million contemporary songs
  - Taste Profile Subset (play counts : over 380,000 songs)
  - Last.fm (tagging for over 500,000 songs)

- 7digital.com
  - 29 seconds audio clips for 99% of the dataset

Experiment 1 - quantitative

<< MLR >>

- MLR(Multiple Linear regression)
- input : bag-of-words representation

<< linear MLP >>

- linear regression using latent factor vectors with 400 dimensions
- input : bag-of-words representation

<< CNN >>

- CNN trained on log-scaled , latent factor vectors with 50 dimensions
  - minimizing MSE
  - minimizing WPE

Experiment 1 - quantitative

- Initial experiments
  - the 9,330 most popular songs & listening data for only 20,000 users
  - 1,881 songs for testing

- other experiments, we used all available data
  - 382,410 songs and 1 million users in total
  - 46,728 songs for testing

- evaluation index
  - mean average precision (mAP)
  - area under the ROC curve (AUC)

Experiment 1 - quantitative

Model	mAP	AUC
MLR	0.01801	0.60608
linear regression	0.02389	0.63518
MLP	0.02536	0.64611
CNN with MSE	0.05016	0.70987
CNN with WPE	0.04323	0.70101

Table 2: Results for all considered models on a subset of the dataset containing only the 9,330 most popular songs, and listening data for 20,000 users.

Based on subset of the dataset

Model	mAP	AUC
random	0.00015	0.49935
linear regression	0.00101	0.64522
CNN with MSE	0.00672	0.77192
upper bound	0.23278	0.96070

Table 3: Results for linear regression on a bag-of-words representation of the audio signals, and a convolutional neural network trained with the MSE objective, on the full dataset (382,410 songs and 1 million users). Also shown are the scores achieved when the latent factor vectors are randomized, and when they are learned from usage data using WMF (upper bound).

Based on audio signal

Upper bound
→ latent factor vectors are obtained from usage data

Experiment 1 - quantitative

- Interestingly, the WPE objective does not result in improved performance.

  → latent factor vectors for popular songs, at the expense of all other songs.

- Limitation: we are unable to predict the popularity of the songs, which considerably affects the AUC and mAP scores.

Experiment 2 - qualitative

- Comparing
  - cosine similarity : the predicted usage patterns
  - WMF : latent factors (50 dimensions)

- Result
  - mostly different
  - quite reasonable in the sense

Query	Most similar tracks (WMF)	Most similar tracks (predicted)
Jonas Brothers - Hold On	Jonas Brothers - Games Miley Cyrus - G.N.O. (Girl's Night Out) Miley Cyrus - Girls Just Wanna Have Fun Jonas Brothers - Year 3000 Jonas Brothers - BB Good	Jonas Brothers - Video Girl Jonas Brothers - Games New Found Glory - My Friends Over You My Chemical Romance - Thank You For The Venom My Chemical Romance - Teenagers
Beyoncé - Speechless	Beyoncé - Gift From Virgo Beyonce - Daddy Rihanna / J-Status - Crazy Little Thing Called Love Beyoncé - Dangerously In Love Rihanna - Haunted	Daniel Bedingfield - If You're Not The One Rihanna - Haunted Alejandro Sanz - Siempre Es De Noche Madonna - Miles Away Lil Wayne / Shanell - American Star
Coldplay - I Ran Away	Coldplay - Careful Where You Stand Coldplay - The Goldrush Coldplay - X & Y Coldplay - Square One Jonas Brothers - BB Good	Arcade Fire - Keep The Car Running M83 - You Appearing Angus & Julia Stone - Hollywood Bon Iver - Creature Fear Coldplay - The Goldrush
Daft Punk - Rock'n Roll	Daft Punk - Short Circuit Daft Punk - Nightvision Daft Punk - Too Long (Gonzales Version) Daft Punk - Aerodynamite Daft Punk - One More Time / Aerodynamic	Boys Noize - Shine Shine Boys Noize - Lava Lava Flying Lotus - Pet Monster Shotglass LCD Soundsystem - One Touch Justice - One Minute To Midnight

Table 4: A few songs and their closest matches in terms of usage patterns, using latent factors obtained with WMF and using latent factors predicted by a convolutional neural network.

Experiment 2 - qualitative

- Visualization
  - distribution of prediction (two dimensions using t-SNE)

- Result
  - Well clustered
  - Appealing to the same audience

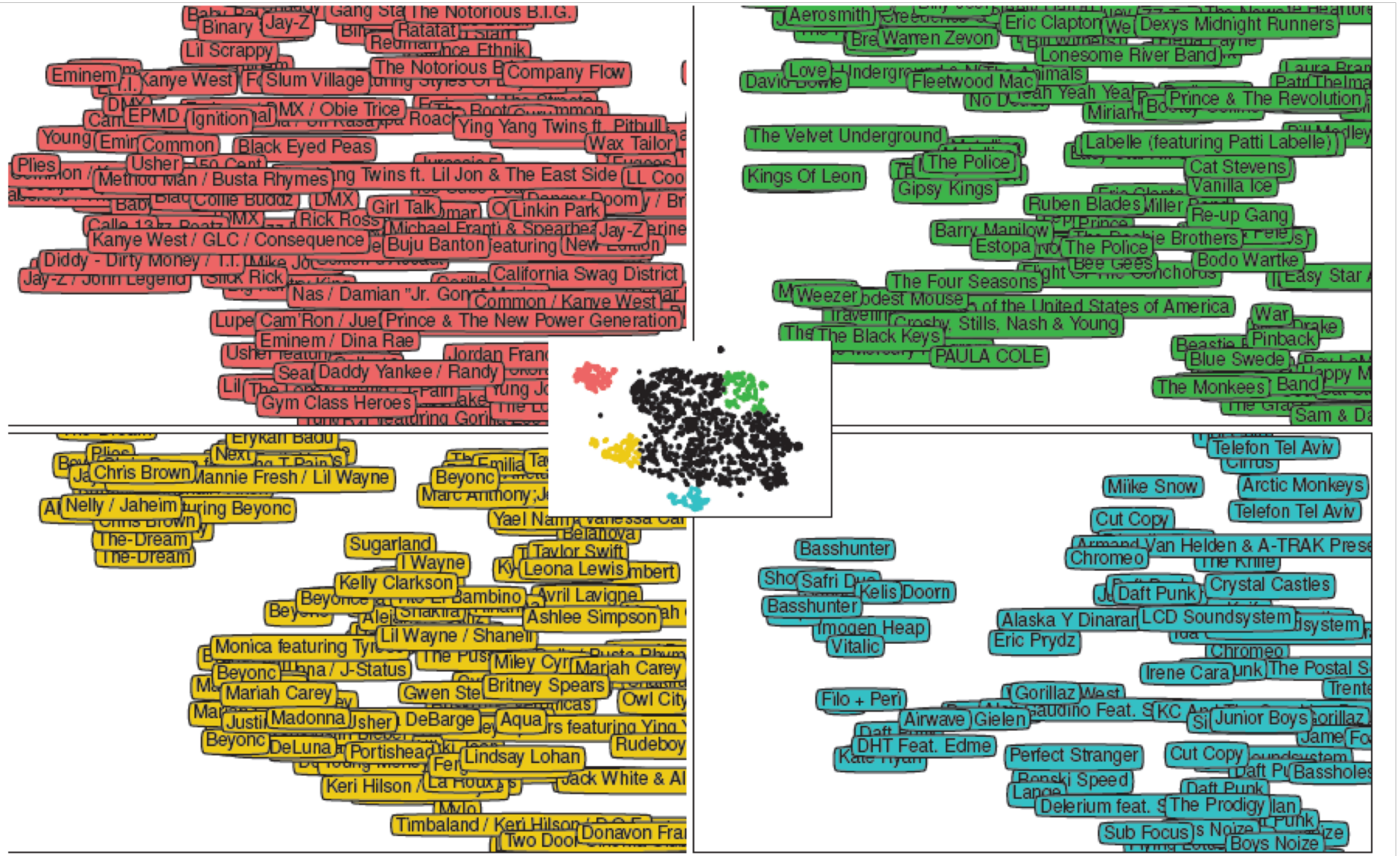

Figure 1: t-SNE visualization of the distribution of predicted usage patterns, using latent factors predicted from audio. A few close-ups show artists whose songs are projected in specific areas. We can discern hip-hop (red), rock (green), pop (yellow) and electronic music (blue). This figure is best viewed in color.

Result

- Predicting latent factors from music audio is a viable method for recommending new and unpopular music.
- CNN significantly outperforming the traditional approaches.
- Prediction from audio signals seem to be sensible.