Content-Aware Collaborative Music Recommendation Using Pre-trained Neural Networks

Dawen Liang DLIANG@EE.COLUMBIA.EDU
Minshu Zhan MZ2468@COLUMBIA.EDU
Daniel P. W. Ellis DPWE@EE.COLUMBIA.EDU
LabROSA, Department of Electrical Engineering, Columbia University, New York, NY, USA

Abstract

To incorporate content information into collaborative filtering methods, we train a neural network on semantic tagging information as a content model and use it as a prior in a collaborative filtering model. Such a system allows the user listening data to "speak for itself". The proposed system is evaluated on the Million Song Dataset and shows performance comparably better than the collaborative filtering approaches, in addition to favorable results in the cold-start case.

1. Introduction

Two primary approaches exist in recommendation: collaborative filtering and content-based methods. For music, the state-of-the-art recommendation results have been achieved by collaborative filtering methods, which requires only information on users' listening history.

On the other hand, modeling musical content for the purpose of taste prediction is difficult due to the structural complexity present in music data which is hard to capture by simple models. Deep learning has shown its power in various pattern recognition tasks with its capability of extracting hierarchical representations from raw data. In music recommendation, van den Oord et al. (2013) have experimented with neural networks on predicting the song latent representation from musical content.

It is natural to combine collaborative filtering and content models in recommendation to utilize different sources of information. A successful attempt is collaborative topic regression (Wang & Blei, 2011), which joins latent Dirichlet allocation (LDA) as a content model on article with collaborative filtering. Collaborative topic regression achieves good performance on scientific article recommendation.

Inspired by the works mentioned above, we create a content-aware collaborative music recommendation system. As the name suggests, the system has two components: the content model and the collaborative filtering model. To obtain a powerful content model, we pre-train a multi-layered neural network to predict semantic tags from vector-quantized acoustic feature. The output of the last hidden layer is treated as a higher-level representation of the musical content, which is used as a prior for the song latent representation in collaborative filtering. We evaluate our system on the Million Song Dataset and show competitive performance to the state-of-the-art system.

2. Proposed approach

Adopting the same structure as that of Wang & Blei (2011), our system consists of two components: a content model which is based on a pre-trained neural network and a collaborative filtering model based on matrix factorization.

2.1. Supervised pre-training

Inspired by the success of transfer learning in computer vision which exploits deep convolutional neural networks (Krizhevsky et al., 2012), in our system we pre-train a multi-layer neural network in a supervised semantic tagging prediction task and use it as the content model.

Our training data comes from Liang et al. (2014) which consists of $370K$ tracks from the Million Song Dataset and the pre-processed *last.fm* data with a vocabulary of 561 tags. The input to the network is ℓ_1-normalized vector-quantization (VQ) histograms.

We treat music tagging as a binary classification problem: For each tag, we make independent predictions on whether

the song is tagged with it or not. We fit the output of the network $f(\boldsymbol{x}_i) \in \mathbb{R}^{561}$ into logistic regression classifiers. Therefore, given tag labels $y_{it} \in \{-1, 1\}$ for song i and tag t, the network is trained to minimize the following loss:

$$\mathcal{L}_{\text{tag}} = \sum_{i,t} \log(1 + \exp(-y_{it} f_t(\boldsymbol{x}_i))$$

Here we use a network with three fully-connected hidden layers and ReLU activations with dropout. Each layer has 1,200 neurons. Stochastic gradient descent with mini-batch of size 100 is used with AdaGrad (Duchi et al., 2011) for adjusting the learning rate. We notice that both dropout and AdaGrad are crucial for getting the good performance.

2.2. Content-aware collaborative filtering

We can interpret the output of the last hidden layer $\boldsymbol{h}_i \in \mathbb{R}^{F_h}$ (here $F_h = 1200$) as a latent content representation of song i. Because of the way that the network is trained, this latent representation is supposed to be highly correlated to the semantic tags ("topics" of music). Therefore, we can take a similar approach to Wang & Blei (2011). The generative process for the proposed model is as follows:

- For user u, draw latent factor $\boldsymbol{\theta}_u \sim \mathcal{N}(0, \lambda_\theta^{-1} I_K)$.
- For song i, draw song latent factor:
$$\boldsymbol{\beta}_i \sim \mathcal{N}(W\boldsymbol{h}_i, \lambda_\beta^{-1} I_K).$$
- For each user-song pair (u, i), draw implicit feedback (whether user u listened to song i):
$$r_{ui} \sim \mathcal{N}(\boldsymbol{\theta}_u^T \boldsymbol{\beta}_i, c_{ui}^{-1}).$$

Here the weight matrix $W \in \mathbb{R}^{K \times F_h}$ transforms the learned content representation from the neural networks into the collaborative filtering latent space via $W\boldsymbol{h}_i$. We set the confidence c_{ui} following the same heuristic in Hu et al. (2008).

We want to emphasize that our proposed model is *content-aware* instead of *content-based*. Just like collaborative topic regression, our proposed model is still fundamentally based on collaborative filtering. The content model is only used as a prior and can be deviated if the model thinks it is necessary to explain the data.

For notational convenience, we define the concatenated user latent factors matrix $\Theta \stackrel{\triangle}{=} [\boldsymbol{\theta}_1 | \cdots | \boldsymbol{\theta}_U] \in \mathbb{R}^{K \times U}$ and song latent factors matrix $B \stackrel{\triangle}{=} [\boldsymbol{\beta}_1 | \cdots | \boldsymbol{\beta}_I] \in \mathbb{R}^{K \times I}$. We estimate the model parameters $\{\Theta, B, W\}$ via maximum *a posteriori*. The complete log-likelihood is written as:

$$\mathcal{L} = -\sum_{u,i} \frac{c_{ui}}{2} (r_{ui} - \boldsymbol{\theta}_u^T \boldsymbol{\beta}_i)^2 - \frac{\lambda_\theta}{2} \sum_u \boldsymbol{\theta}_u^T \boldsymbol{\theta}_u - \frac{\lambda_\beta}{2} \sum_i (\boldsymbol{\beta}_i - W\boldsymbol{h}_i)^T (\boldsymbol{\beta}_i - W\boldsymbol{h}_i)$$

Take the gradient of the complete log-likelihood with respect to the model parameters and set it to 0, we can obtain the following closed-form coordinate updates:

$$\boldsymbol{\theta}_u \leftarrow (BC_u B^T + \lambda_\theta I_K)^{-1} BC_u \boldsymbol{r}_u \quad (1)$$

$$\boldsymbol{\beta}_i \leftarrow (\Theta C_i \Theta^T + \lambda_\beta I_K)^{-1} (\Theta C_i \boldsymbol{r}_i + \lambda_\beta W \boldsymbol{h}_i) \quad (2)$$

$$W^T \leftarrow (H^T H + \lambda_W I_{F_h})^{-1} H^T B^T \quad (3)$$

where $C_u \in \mathbb{R}^{I \times I}$ is a diagonal matrix with c_{ui}, $i = 1, \cdots, I$ as its diagonal elements, and $\boldsymbol{r}_u \in \mathbb{R}^I$ is the feedback for user u. C_i and $\boldsymbol{r}_i$ are similarly defined. $H \in \mathbb{R}^{I \times F_h}$ is the concatenated output from the last hidden layer $[\boldsymbol{h}_1 | \cdots | \boldsymbol{h}_I]^T$. When updating W, we add a small ridge term λ_W to the diagonal of the matrix to regularize and avoid numerical problems when inverting.

After the model is trained, we can make *in-matrix* prediction by $\hat{r}_{ui} = \boldsymbol{\theta}_u \boldsymbol{\beta}_i$. Similar to collaborative topic regression, we can also make *out-of-matrix* prediction for songs that no one has listened to by only using the content $\hat{r}_{ui} = \boldsymbol{\theta}_u^T (W\boldsymbol{h}_i)$.

2.3. Connections to relevant work

Collaborative topic regression (Wang & Blei, 2011) The main difference that sets our method apart from collaborative topic regression is the content model. As a feature extractor, LDA can only produce linear factors due to its bilinear nature. On the other hand, multi-layer neural network used by in our system is capable of capturing the non-linearities in the feature space.

Deep content-based music recommendation (van den Oord et al., 2013) Our method is very similar to this approach, but we will point out two major differences:

First, the neural network is used for different purposes. We use it as a content feature extractor. The neural network in van den Oord et al. (2013) maps content directly to the latent factors learned from the weighted matrix factorization (Hu et al., 2008), and the resulting model is expected to operate similarly to collaborative filtering even when usage data is absent.

Secondly, the performance of van den Oord et al. (2013) is upper-bounded by that of Hu et al. (2008), as the neural network is trained to map content to the latent factors learned from the weighted matrix factorization. What we propose in this paper, on the other hand, uses content as an *addition* to the collaborative filtering, in a similar manner as the collaborative topic regression.

Acknowledgements

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Tesla K40 GPU used for this research.

References

Duchi, John, Hazan, Elad, and Singer, Yoram. Adaptive subgradient methods for online learning and stochastic optimization. *The Journal of Machine Learning Research*, 12:2121–2159, 2011.

Hu, Yifan, Koren, Yehuda, and Volinsky, Chris. Collaborative filtering for implicit feedback datasets. In *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*, pp. 263–272. IEEE, 2008.

Krizhevsky, Alex, Sutskever, Ilya, and Hinton, Geoffrey E. ImageNet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pp. 1097–1105, 2012.

Liang, Dawen, Paisley, John, and Ellis, Daniel P. W. Codebook-based scalable music tagging with Poisson matrix factorization. In *Proceedings of the International Society for Music Information Retrieval Conference*, pp. 167–172, 2014.

van den Oord, Aäron, Dieleman, Sander, and Schrauwen, Benjamin. Deep content-based music recommendation. In *Advances in Neural Information Processing Systems*, pp. 2643–2651, 2013.

Wang, Chong and Blei, David M. Collaborative topic modeling for recommending scientific articles. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 448–456. ACM, 2011.