CUBC: A Generalized Representation Learning Method for User Behavioral Sequence

Yongqing Wang[1], Haopeng Zhang[1], Hao Gu[2], Lingling Yi[2], Huawei Shen[1] and Xueqi Cheng[1]
1. Institute of Computing Technology, Chinese Academy of Sciences, China
2. Weixin Group, Tencent Inc., China
1. wangyongqing,shenhuawei,gaojinhua,cxq@ict.ac.cn
2. chrisyi,nickgu@tencent.com

Abstract

The objective of user behavior coding is to learn high-level representation over behavioral sequence which can be fed into downstream tasks, e.g., profiling and recommendation. However, the existed approaches has low efficiency on extracting high-quality information from behavioral sequences. In this paper, a contrastive user behavior coding approach (CUBC) is proposed by maximizing context- and content-level mutual information between inputs, outputs and side information. Furthermore, a proposed content-level negative sampling is proved its effectiveness in estimating low bound of mutual information. The experimental results show great potentials to be applied in real applications.

1 Introduction

The increased online services facilitate users to access information on Internet, e.g., browsing webpages, chatting with others and shopping online, leaving tremendous sequential behavior records. In this way, the demand of learning a generalized representation from user behaviors has emerged in both industrial and academic circles [9]: multiple downstream learning tasks can be improved by the well-presented user behaviors, such as profiling and recommendation.

One foundational idea in sequence modeling is to learn compressed representations (following *encoder-decoder* way) that nonetheless can be used to reconstruct raw sequence [5]. The compressed representative codes can be learned in the form of autoencoders and generative models with constraint on lossless data distribution. However, such objective of representation learning [6] can merely afford to predict future, missing or contextual information in behavioral sequence, but failed in coding user characteristics, such as profiling. An alternative way is to introduce supervised information in sequence modeling, broadly applied in accurate modeling for specific learning tasks. The so-called *end-to-end* modeling methods attempt to directly learn the mappings from inputs to labels [7]. But learned representation of user behaviors can be hardly re-employed in other learning tasks. In this way, the remarkable implementation costs take additional system risks on business.

Different from sequence modeling in text, speech and images, learning a generalized representation of user behaviors prefers to capture both context- and content-level information. *Context-level information* is sensitive to temporal variation in one sequence, which can be used to predict future or missing behaviors [8]. Meanwhile, *content-level information* presents the sequence as whole, which can be generally applied to deduce stable characteristics relative to users [2]. Inspired by contrastive learning [1], in this paper, we attempt to learn high-level representation from user behavior records by mutual information maximization. The learned representation contains both context- and content-level information in sequences. Thus, the proposed framework consists of two objectives: 1) context-level mutual information maximization: the compact distributed vectors are learned by maximizing average mutual information between inputs and outputs, reflecting the temporal variation in a sequence. 2)

NeurIPS 2021 Workshop: Self-Supervised Learning - Theory and Practice.

content-level mutual information maximization: side information relative to user characteristics can be introduced to embed stable features in behavioral representation. Moreover, a *content-level negative sampling* is proposed to better approximate lower bound of mutual information when using Noise-Contrastive Estimation (NCE) [3] in training. The experimental results prove that the learned representation from our proposed method has better performance than other comparative models in multiple downstream tasks, including end-to-end approaches. Furthermore, the improvement from offline/online scenarios shows the great potentials of our proposed method in real applications.

2 Model

Firstly, we introduce some basic notations and definitions in user behavior coding. The original input $\{S^{(i)}\}_{i=1}^{N}$ is a collection of N sequential user behaviors, where a sequence $S = \{s_k | s_k \in \mathcal{T}\}_{k=1}^{M}$ records behaviors generated by one specific user. The symbols s and $\mathcal{T}$ denotes behavior token and its token set respectively, referring to the possible behaviors, e.g., browsing, chatting and shopping. Generally, the original behavior token s_k should be vectorized as input, i.e., $s_k \overset{\text{vec.}}{\rightarrow} x_k$, where the symbol x_k is an K_x-dimensional input vector corresponding to token s_k. Then the vectorized sequence can be presented as $\mathbf{x} = \{x_k | x_k \in \mathbb{R}^{K_x}\}_{k=1}^{M}$. The objective of user behavior coding is to generate high-quality representations of sequential user behaviors $\{y^{(i)} | y^{(i)} \in \mathbb{R}^{K_y}\}_{i=1}^{N}$, improving prediction or inference performance on kinds of potential learning tasks. Each $y^{(i)}$ presents a compressed vector from corresponding behavior sequence $S^{(i)}$ and the dimension is K_y.

Context-level Mutual Information Maximization: The objective of maximizing context-level mutual information is to learn the optimal mapping function f_θ: $\mathbf{x} \rightarrow y$ with parameters θ, that is,

$$I(X; Y) = D_{KL}(\mathbb{P}_{XY} || \mathbb{P}_X \otimes \mathbb{P}_Y),$$

where X and Y refer to the variables on input and output for one sequence, and mutual information is equal to Kullback-Leibler divergence between the joint distribution $\mathbb{P}_{XY}$ and the product of the marginal distribution $\mathbb{P}_X \otimes \mathbb{P}_Y$.

According to variational estimation on f-divergence, the mutual information $I(X; f_\theta(X))$ can be estimated by Jensen-Shannon divergence (JSD [4]), that is,

$$\mathcal{J}_{\theta,\omega}(X, f_\theta(X)) = \mathbb{E}_{\mathbb{P}_X}[-\text{sp}(-T_\omega(\mathbf{x}, f_\theta(\mathbf{x})))] - \mathbb{E}_{\mathbb{P}_{\tilde{X}} \otimes \mathbb{P}_X}[-\text{sp}(-T_\omega(\tilde{\mathbf{x}}, f_\theta(\mathbf{x})))], \tag{1}$$

where $\mathbf{x} \backsim \mathbb{P}_X$ and $\tilde{\mathbf{x}} \backsim \mathbb{P}_{\tilde{X}}$ are the sequences sampled from domain $\mathcal{X}$, the score function $T_\omega \in \mathbb{R}$ is parameterized by ω, and the $\text{sp}(z) = \log(1 + \exp(z))$ is a softplus function. The Equation 1 formalizes the lower bound of context-level mutual information. Moreover, the score function T_ω can be formalized as

$$T_\omega(\mathbf{x}, f_\theta(\mathbf{x})) = \frac{1}{M} \sum_{k=1}^{M} G_\omega(x_k, y),$$

where $x_k \in \mathbf{x}$, $y = f_\theta(\mathbf{x})$ and the function G_ω with parameters ω is a well-constructed neural network with concatenated inputs from vectorized token and sequence representation.

Content-level Mutual Information Maximization: Abundant user data offer us an opportunity to consider what if we introduce side information in contrastive learning so as to improve the quality of the learned representation from behavioral sequence. For example, information of user interests can encourage to understand deep interests in behavioral sequence so as to improve the quality of learned representation.

Similar to the formalization in context-level mutual information maximization, the optimization can be formalized as

$$\theta = \arg\max_\theta I(f_\theta(X); C), \tag{2}$$

where C is variable on side information. The objective can also be solved by maximizing JSD loss, where the score function T can be defined as

$$T_{\omega'}(f_\theta(\mathbf{x}), c_u) = G_{\omega'}(f_\theta(\mathbf{x}), c_u), \tag{3}$$

where c_u is a vector of side information corresponding to user u.

Content-level negative samples: A content-level negative sampling is proposed in order to better approximate lower bound of mutual information when using NCE to estimate JSD loss. Different

from negative sampling in token and context levels (e.g., CBOW and skip-gram), the faked behavioral sequences are directly drawn from the training dataset. For further simplifying negative sampling approach, we propose an efficient sampling strategy by shuffling sequences on mini-batch. The experimental results show that the new negative sampling strategy is appropriate for learning sequence representation than other sampling strategy on token and context level.

Model training: We choose *transformer* with multi-head attention to learn mapping function f_θ. The sequence representation is aggregated by attention mechanism. Finally, the integrated objective of our proposed model is to maximize the loss function as follows,

$$\mathbb{L}(\theta, \omega, \omega') = \mathcal{J}_{\theta,\omega}(X, f_\theta(X)) + \alpha \cdot \mathcal{J}_{\theta,\omega'}(f_\theta(X), C),$$

where α is hyper-parameter to balance context- and content-level mutual information functionalized in representation learning. The detail of learning framework is depicted in A.1.

3 Experiments

In this section, we conduct experiments to analyze prediction performance on our proposed learning framework in real applications, evaluating the effectiveness of our proposed method.

Datasets and Implementation: The evaluation is implemented on offline data from reading tracks of users on Wechat public subscription. We extract reading activities of users who live in one of the cities located in Guangzhou province, China, during June 1 to June 30, 2019. For better modeling the behavioral sequence, we only choose the top 48,150 popular public subscription accounts and filter out the sequences which contains less than 3 reading activities. The final dataset includes 687,192 users and its corresponding behavioral sequences. For sequence modeling, the token in sequences refers to account ID read by users. Besides, we use interests of users as side information in calculating content-level mutual information loss. The interests of users are obtained by users' reading preference in Wechat.

Comparative Methods: It is assuming that better representation can result in better prediction performance. To illustrate the performance of our proposed method, three typical downstream tasks are used for evaluation, including: gender prediction, age prediction and next token prediction, broadly used as benchmarks in business.

The compared methods include: 1) **Generative model:** The learned representation is generated by objective on sequence completion. Generally, y_M (the last output) is chosen for sequence representation; 2) **End-to-end model**: The model directly learn the mapping from input to downstream tasks. The architecture is same as our proposed model except loss function; 3) **Feature-based model:** The downstream models are directly constructed by side information.

All constructed features are fed into similar downstream models to evaluate the prediction performance on tasks. For gender and age predictions, two layer fully connected models are implemented. Meanwhile, a bilinear scoring function score$(y, x) = \sigma(y^T \mathbf{W} x)$ is used for next token prediction, where σ is a sigmoid function and $\mathbf{W}$ is a learnable parametric matrix. The highest score achieved by pair (y, x) refers to the predicted next token.

Table 1: Prediction performance on gender/age/next token prediction.

Model	Gender prediction				Age prediction				Next token prediction	
	Acc.	Precision	Recall	F1 score	Acc.	Precision	Recall	F1 score	nDCG@10	MRR
Generative (y_M)	0.6402	0.6236	0.7074	0.6629	0.6591	0.6912	0.5695	0.6245	0.6971	0.6805
End-to-end	0.7435	**0.7475**	0.7356	0.7415	0.7232	**0.7836**	0.6131	0.6879	0.9356	0.9067
Feature-based	**0.7465**	0.7381	0.7644	0.7510	0.6903	0.7072	0.6444	0.6743	0.5917	0.4783
Token-level NS	0.6751	0.6673	0.6987	0.6826	0.7228	0.7613	0.6452	0.6985	0.8680	0.7649
Context-level NS	0.5070	0.5038	**0.9449**	0.6572	0.5164	0.5074	**0.9690**	0.6660	0.1032	0.0430
CUBC (context)	0.7300	0.7296	0.7310	0.7303	0.7334	0.7407	0.7143	**0.7273**	**0.9692**	**0.9454**
CUBC (content)*	0.7349	0.7336	0.7379	0.7357	0.7226	0.7458	0.6715	0.7067	–	–
CUBC (context+content)	0.7463	0.7348	0.7710	**0.7525**	**0.7378**	0.7710	0.6730	0.7187	0.9672	0.9429

* The learned sequence representation is insensitive to position, thus it can hardly be applied into next token prediction.

Prediction Performance: Table 1 shows the experimental results. The symbols Token-level NS and Context-level NS present token and context-level negative sampling stategies used in CUBC. Meanwhile, the symbols CUBC (context), CUBC (content) and CUBC (context+content) refer to

the models with considering of context, content or context+content-level mutual information. It is obvious that both token- and context-level negative samples has limit performance in behavioral sequence learning. In all three tasks, the learned representations from CUBC outperform other comparative methods, even better than end-to-end model. Interestingly, feature-based model and CUBC (context+content) both perform better in gender prediction task. It indicates that high quality side information can optimize the learned representation. However, it is not equal to directly transform side information to learned representation. The proof is presented in A.4.1.

Offline and Online Evaluation: To evaluate performance in practical application, the learned sequence representation associated with other extracted features are also fed into offline tagging model for training and testing. Concatenated with learned representation from CUBC, the prediction accuracy is increased to 0.9305 (+0.81%), 0.9403 (+0.23%) and 0.9551 (+0.20%) respectively. Furthermore, we also take A/B test experiments on online game recommendation production. After concatenating the learned representation from CUBC, the registration rate is lifted by +0.11%, +0.51%, 0.82% and +1.11% in four games. The experimental results show the great potentials of our proposed method in real applications.

4 Conclusion

In this paper, we propose an efficient coding method to learn sequence representation from users behaviors. The proposed CUBC model maximizes context- and content-level mutual information between inputs, outputs and side information, and utilize a content-level negative sampling strategy, improving the quality of learned sequence representation.

References

[1] S. Arora, H. Khandeparkar, M. Khodak, O. Plevrakis, and N. Saunshi. A theoretical analysis of contrastive unsupervised representation learning. *ICLR*, 2019.

[2] C. Chen, S. Kim, H. Bui, R. Rossi, E. Koh, B. Kveton, and R. Bunescu. Predictive analysis by leveraging temporal user behavior and user embeddings. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 2175–2182, 2018.

[3] M. Gutmann and A. Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 297–304, 2010.

[4] R. D. Hjelm, A. Fedorov, S. Lavoie-Marchildon, K. Grewal, P. Bachman, A. Trischler, and Y. Bengio. Learning deep representations by mutual information estimation and maximization. *ICLR*, 2018.

[5] T. Mikolov, M. Karafiát, L. Burget, J. Černockỳ, and S. Khudanpur. Recurrent neural network based language model. In *Eleventh annual conference of the international speech communication association*, 2010.

[6] M. Pavlovski, J. Gligorijevic, I. Stojkovic, S. Agrawal, S. Komirishetty, D. Gligorijevic, N. Bhamidipati, and Z. Obradovic. Time-aware user embeddings as a service. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3194–3202, 2020.

[7] Q. Pi, W. Bian, G. Zhou, X. Zhu, and K. Gai. Practice on long sequential user behavior modeling for click-through rate prediction. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2671–2679, 2019.

[8] Y. Zhu, H. Li, Y. Liao, B. Wang, Z. Guan, H. Liu, and D. Cai. What to do next: Modeling user behaviors by time-lstm. In *IJCAI*, volume 17, pages 3602–3608, 2017.

[9] Y. Zhu, D. Xi, B. Song, F. Zhuang, S. Chen, X. Gu, and Q. He. Modeling users' behavior sequences with hierarchical explainable network for cross-domain fraud detection. In *Proceedings of The Web Conference 2020*, pages 928–938, 2020.