Dual gated attentive-autoencoder for content-aware Recommendation

Ming Jin¹, Jun Li²,* , Chenyang Li²

¹,²School of computer and artificial intelligence, Wenzhou University, Wenzhou , Zhejiang, China

*Corresponding author e-mail:omama@wzu.edu.cn

Abstract. In recent years, content-aware recommendation of personalized products using automatic encoders has become the mainstream technology in the field of personalized recommendation systems. But the technology still faces the following main problems: 1. The automatic encoder structure ignores the inherent duality of the model and separates the training encoder from the decoder. 2. The automatic encoder structure makes use of implicit feedback data and the difficulty of combining heterogeneous data. 3. The traditional content-aware recommendation based on the auto-encoder ignores the neighboring merchandise information of the merchandise. In order to solve these problems, this paper proposes a dual gated attentive-autoencoder based on the door attention mechanism. The model is compared to the current optimal outcome model on the three recommended system public datasets. The experimental results show that the recall rate and the normalized damage cumulative gain of the auto-encoder content perception recommendation based on the dual-gate attention mechanism are better than the current optimal result model.

1. Introduction

With the rapid growth in the use of Internet services and mobile devices, it makes it easier for people to access online products and multimedia content such as movies and articles. People have multiple options for online products and multimedia content, but with this growth trend, people have increased the difficulty of choice. Personalized recommendations are especially important in everyday life.

Although the current personalized recommendation system method has achieved good results, there are some limitations: 1. The traditional method [1] does not consider the description of a commodity between different words when learning the hidden layer representation of the word vector. The importance of it. Treating information words and other words in the same way may result in an incomplete understanding of the product information by the user. 2. The traditional method [1] [2] combines the hidden layer representation from the heterogeneous information by using the regularization term with weights. This approach does not ensure that the hidden layer representation of the data is completely extracted from the heterogeneous data and triggers tedious hyperparameter adjustments. 3. The traditional method ignores the relationship information between commodities, such as the movie of the same movie genre or the mutual reference information between articles. Closely related items are likely to share the same subject or have similar attributes. Therefore, exploring the user's preferences for similar items of the product also facilitates inferring the user's
preference for the item. 4. When the traditional method uses the automatic encoder structure to extract the hidden layer representation of the commodity information, since the model structure duality attribute exists between the encoder and the decoder in the automatic encoder structure, the conventional method often ignores the relationship between the encoder and the decoder. The structure of the dual property, separately training the encoder and decoder structure.

In order to solve the above problems, this paper proposes a content-aware recommendation model based on Gated Attentive Dual-Autoencoder (GADAE). GADAE consists of four modules: the word attention mechanism module, the proximity attention mechanism module, the neural gate structure, and the dual stack automatic encoder. Inspired by the dual learning mechanism, this paper designs a stacking autoencoder as a dual stacking autoencoder, and proposes a stacking autoencoder and decoder as a dual closed loop. The strategy gradient method is used to train the encoder and the decoder simultaneously, so that the encoder and the decoder are sharing feedback information. This paper evaluates the dual auto-encoder of the door attention mechanism proposed in this paper on three real-world datasets, and proves the good recommendation performance of the proposed model.

2. Dual automatic encoder based on door attention mechanism

The dual auto-encoder architecture based on the gate attention mechanism proposed in this paper is shown in Figure 2. This paper designs the main encoder, the neural gate structure, and the dual decoder as a dual closed loop. This section first introduces the dual autoencoder learned from the input binary user product rating. Then introduce the word attention mechanism module. Then the paper introduces the proposed neural gate structure module. Finally, this section introduces the loss function and training process of the proximity attention mechanism module and model.

2.1 Dual automatic encoder

This paper proposes to use the stacking autoencoder to model from the implicit data of the user-commodity to extract the hidden layer feature information in the implicit data. The conventional stacked autoencoder does not consider the structural dual information between the encoder and the decoder, separately trains the encoder and the decoder, and the feedback signal between the encoder and the decoder is not shared. This paper proposes a new dual-stack automatic encoder structure, in which the primary encoder and the dual decoder structure are designed as a dual closed loop. Then, the primary encoder and the dual decoder are jointly trained to share the feedback signals of the primary encoder and the dual decoder, and the performance of the dual-layer automatic encoder to extract the hidden layer feature information in the implicit data is improved.

As shown in the blue module of Figure 1, the binary user product rating $e_i \in \mathbb{R}^n$ is first encoded using the primary encoder to the hidden layer representation $e'_i$ of the commodity rating.

The main encoder encode is expressed as:

$$
\text{The main encoder encode}:
\begin{align*}
    e'_i &= a_1(W_e e_i + b_1) \\
    e''_i &= a_2(W_2 e'_i + b_2)
\end{align*}
$$
The decoding of the dual decoder is expressed as:

\[
\begin{align*}
\tilde{r}_i &= a_i(W_i \tilde{e}_i^r + b_i) \\
\tilde{u}_i &= a_u(W_u \tilde{e}_i^u + b_u)
\end{align*}
\] (2)

The subscript \(i\) in \(\tilde{e}_i^r\) represents the specific item \(i\) information, and the superscript \(r\) represents the hidden layer representation encoded in the binary commodity rating. \(W_i \in \mathbb{R}^{v \times m}\), \(W_u \in \mathbb{R}^{v \times m}\), \(W_u \in \mathbb{R}^{m \times v}\) represents a weight matrix. \(m\) represents the number of users, \(v\) represents the dimension of the first hidden layer, and \(v\) represents the bottleneck layer dimension.

### 2.2 Word attention mechanism module

The traditional method of embedding commodity information ignores the importance of describing the same commodity words. This paper proposes a module of word attention mechanism based on the sequence of commodity descriptors. As shown in the purple module in Figure 2, the word attention mechanism module proposed in this paper adaptively selects information words with different importance for the commodity description information.

The word attention mechanism module proposed in this paper is inspired by the Transformer model in machine translation [4], using the multi-dimensional attention mechanism to learn the hidden layer information of the commodity description text, instead of using complex cyclic neural networks and convolutional neural networks. Because in practical applications, users prefer product information that can be summarized by a few words, rather than the sequence relationship of word time.

The purpose of using the word attention mechanism module is to assign the importance of different word embedding from word embedding \(E_i\) The Word Attention Mechanism module calculates the word attention mechanism module weights from the input word embedding \(E_i\) using the two-layer neural network of vanilla attention mechanism:

\[
e_i = \text{softmax}(w_i^T \tanh(W_i E_i + b_i))
\] (3)

\(w_i \in \mathbb{R}^r, W_i \in \mathbb{R}^{r \times r}, b_i \in \mathbb{R}^r\). Here the softmax function ensures that the sum of the weights is 1, and then we calculate the sum of the weights of the word embedding \(E_i\):

\[
E_i^w = \sum_{i \neq j} e_{ij}^w e_j
\] (4)

The use of the word attention mechanism module to calculate the importance value for a single word embedding results in a single content that would cause the recommendation system model to recommend the merchandise content. If \(e_i\) is a single-dimension merchandise information, the recommendation system performance is degraded. Therefore, this paper proposes a multi-dimensional \(e_i\) word attention mechanism module to calculate word embedding weights, and multi-dimensional \(e_i\) can contain different aspects of product information. Assume that \(c_i\) is the attention feature extracted from word embedding. This paper expands \(W_i \in \mathbb{R}^{c \times c}\), using the multi-dimensional attention mechanism to calculate the weight of the word attention mechanism module:

\[
e_i = \text{softmax}(w_i^T \tanh(W_i E_i + b_i) + b_i)
\] (5)

\(e_i \in \mathbb{R}^{c \times c}\) is the attention weight matrix, \(b_i \in \mathbb{R}^c\) is the deviation term. Through the word embedding extracted by the multi-dimensional attention mechanism, we can get the characteristic matrix representation of the commodity content item:

\[
E_i^c = E_i E_i^T
\] (6)

Finally, the neural layer is used to obtain the hidden layer feature representation of the commodity content feature matrix:

\[
e_i^c = c_i(E_i^T w_i)
\] (7)
2.3 Neural gate structure module

The neural gate structure module designed in this paper is to fuse the user commodity rating vector and the commodity content description information vector into two kinds of heterogeneous data. Combining the product rating information and the product content description information, the recommendation system is better for the marked product data.

As shown in FIG 2, the neural gate structure G is defined to merge the commodity rating vector $e_i^c$ with the commodity content description information vector $e_i^c$, and is proposed to be calculated using the sigmod function:

$$G = \text{sig mod}(W_{g1}e_i^c + W_{g2}e_i^c + b_g)$$

(8)

$W_{g1} \in \mathbb{R}^{m \times r}, W_{g2} \in \mathbb{R}^{m \times r}, b_g \in \mathbb{R}$ is the neural gate layer parameter. The fusion of commodity rating information and product description information is calculated as:

$$e_i^c = G \odot e_i^c + (1 - G) \odot e_i^c$$

(9)

Where $\odot$ represents the product of the elements, and through the neural gate structure module, we obtain the fusion feature vector of the commodity rating information and the commodity content description information.

2.4 Proximity attention mechanism module

After obtaining the fusion feature vector 1 of the commodity rating information and the commodity content description information, the article considers the commodity feature information similar to the commodity fusion feature vector 2, that is, the adjacent commodity feature information. In real life, such as movies, the same film, the papers cited by the papers have neighboring information. In the past recommendation system, it often ignores the neighboring commodity information of the commodity. This paper proposes the proximity attention mechanism module to extract the feature information between the commodity and the adjacent commodity.

First, we define that the neighboring item of the item $i$ is $n_i$, and the neighboring item $n_i$ can select the item information similarity of the item from the user binary rating to select the adjacent item information of the item $i$. The hidden layer feature representation $e_i^n$ of the adjacent commodity is $n_i$ is calculated as:

$$e_i^n = \sum_{j \in n_i} \text{soft max}(\tanh(e_i^T W^n e_j^o)) e_i^o, \forall j \in n_i$$

(10)

$W^n \in \mathbb{R}^{r \times r}$ is the adjacent attention mechanism module layer parameter. In order to simultaneously extract the adjacent item of the item $i$ into $n_i$ pieces of information, the dual decoder calculates as:

$$\hat{a}_i = a_i(W^n a_i(W^n e_i^o + b_i) + W_a a_i(W^n e_i^o + b_i) + b_i)$$

(11)

2.5 Model loss function and model training

The model modeling in this paper is based on implicit data. In order to better model the user's preference information and implicit data information, this paper inserts a confidence matrix into the square loss function, and designs the model loss of the dual automatic encoder with the door attention mechanism. The function is:

$$L_{GAE} = \sum_{i=1}^{n} \sum_{j=1}^{m} ||C_{i,j}(D_{i,j} - \hat{D}_{i,j})||_2^2 + ||C^T(T - \hat{T}^T)||_F^2$$

(12)

Where $\odot$ represents the product of the element and $||\cdot||_F$ is the F-norm of the matrix. The confidence matrix is defined as:

$$C_{i,j} = \begin{cases} 
\rho & \text{if } D_{i,j} = 1 \\
1 & \text{else} 
\end{cases}$$

(13)

$C \in \mathbb{R}^{m \times n}$. The objective function of the dual autoencoder for designing the door attention mechanism is:

$$L = L_{GAE} + \lambda(||W^1||_F^2 + ||W^2||_F^2)$$

(14)
3. Experiment setup

3.1 Experimental data

Experimental Data Sets This paper selects data sets of three standard implicit feedback data from the real world: CiteULike-a [5], movielens-20M [6], Amazon-Books [7]. These three data sets select different areas with different sparsity. The experimental data set statistics are shown in Table 1.

| data sets      | users  | items  | rates    | words | density   |
|----------------|--------|--------|----------|-------|-----------|
| CiteULike-a    | 5511   | 16980  | 204986   | 8000  | 0.217%    |
| movielens-20M  | 138493 | 18307  | 1997049  | 12397 | 0.788%    |
| Amazon-Books   | 65476  | 41264  | 1947765  | 27584 | 0.072%    |

3.2 Experimental evaluation indicators

Experimental evaluation indicators This paper uses the recall rate (Recall) and ranking indicators to represent the normalized damage cumulative gain (NDCG). Among them, Recall means that in the recommendation system, it is equivalent to a multi-interest query, that is, each user is a query word, and then returns the Top k commodity associated with each query word, that is, returns the Topk commodity of interest to each user. percentage. The NDCG indicates that the score of each recommendation result relevance is accumulated as the score of the entire recommendation list (list).

3.3 Comparative experimental model

In order to verify the validity of the dual automatic encoder (GADAE) based on the gate attention mechanism proposed in this paper, this paper designs and compares the three types of recommendation system models.

The first experimental comparison model is a traditional recommendation system model based on implicit feedback. 1. Weighted Regularized Matrix Factorization (WRMF) [3], WRMF uses implicit binary feedback as a binary instance, then considers all entries in the user-item interaction matrix (including unobserved entry). WRMF has a confidence value to control the weight of the positive and negative entries. 2. Collaborative Denoising Autoencoder (CDAE) [4], CDAE model uses the noise reduction automatic encoder DAE to learn the hidden layer feature representation from the implicit feedback information.

The second experimental comparison is a recommendation system model based on the word bag model. 3. Collaborative Variational Autoencoder (CVAE) [1], CVAE is a generated implicit variable model that uses the variational automatic encoder model to generate commodity content and user ratings. 4. Collaborative Metric Learning Model (CMLM) [8], CMLM uses metric learning strategies to learn commodity features.

The third comparison experiment is a recommendation system model based on word sequences. 5. Convolutional Matrix Factorization (ConvMF) [9], The ConvMF model extracts commodity content information features by using convolutional neural networks and probability matrix decomposition techniques. 6. Joint Representation Learning Model (JRLM) [10], JRLM is a framework model for joint learning of Top-N recommendation system features.

3.4 Experimental environment and experimental settings

This paper first sets up the model GADAE proposed in this paper, sets the commodity content description word embedding maximum length to 300, sets the grid search hyperparameter for the four data sets to \([m,100,50,100, m]\), and sets the CiteULike-a dataset grid search threshold \(\rho = 5\), movielens- The 20M dataset grid search threshold \(\rho = 20\), the Amazon-Books and Amazon-CDs dataset grid search thresholds are \(\rho = 15\) and \(\rho = 20\), respectively. The network learning rate is set to 0.01 and the batch size is set to 1024.

Experimental comparison models: WRMF, CDAE, CVAE, CMLM, ConvMF and JRLM, first set the hidden variable dimension of all models to 50. The WRMF, CDAE model experimental setup is consistent with its original paper. For the CVAE model, in order to achieve better performance, this paper sets experimental parameters \(\lambda_0 = 0.1, \lambda_c = 10, \lambda_c = 0.01\) For CMLM, this article sets the marginal
parameter $m = 2$ and sets the experimental parameter to $\lambda_1 = 0.1, \lambda_2 = 1$. For the model ConvMF, the experimental setup of this paper is consistent with the original paper. For JRLM, this article sets the batch learning size to 64.

4. Experimental results and analysis

The results of the recall rate (Recall) and the normalized damage cumulative gain (NDCG) of the dual automatic encoder (GADAE) and experimental comparison models WRMF, CDAE, CVAE, CMLM, ConvMF and JRLM based on the gate attention mechanism proposed in this paper. As shown in Fig. 2, Fig. 3 shows the recall rate (Recall) and normalized damage cumulative gain (NDCG) results of several models in the data set CiteULike-a.

![Figure 2 Recall on CiteULike-a data set](image)

![Figure 3 NDCG on CiteULike-a data set](image)

As shown in Fig. 4, Fig. 5 shows the recall rate (Recall) and normalized damage cumulative gain (NDCG) results of several models in the data set movielens-20M.

![Figure 4 Recall on movielens-20M data set](image)

![Figure 5 NDCG on movielens-20M data set](image)

As shown in Figure 6, Figure 7 shows the recall (Recall) and normalized damage cumulative gain (NDCG) results for several models in the dataset Amazon-Books.

![Figure 6 Recall on Amazon-Books data set](image)

![Figure 7 NDCG on Amazon-Books data set](image)

From the comparison experimental results of the three sets of data sets, we can see that the dual automatic encoder (GADAE) based on the gate attention mechanism is compared with the experimental comparison model models WRMF, CDAE, CVAE, CMLM, ConvMF and JRLM, in the recall rate Recall. (k=5,10,15,20) Under the evaluation, GADAE achieved better results than the 6 sets of comparative models. Under the loss cumulative gain (NDCG) evaluation, except in the data set movielens-20M, GADAE is slightly lower than the JRLM model and the CMLM model under the NDCG (k=20) evaluation, and the other two sets of data sets are in the NDCG (k =5,10,15,20) GADAE is superior to the other 6 sets of comparison models. GADAE's recall rate (Recall) and normalized damage cumulative gain (NDCG) results outperform the word-based recommendation.
system models ConvMF and JRLM, although ConvMF and JRLM use the sentence vector-based doc2vec model to model and use Convolutional neural network structure, but these two methods have neglected the user attention rating and the adjacent attention feature information of the product content. The proximity attention mechanism module proposed in this paper combines the commodity content adjacent product information to highlight the information words. GADAE's recall rate (Recall) and normalized damage cumulative gain (NDCG) results are superior to the recommendation system model CVAE, CMLM based on the word bag model, thinking that CVAE, CMLM is using the word bag vector instead of embedding the product content. Using the attention mechanism module of the word vector, this makes the information between the commodity feature words independent. The GADAE model proposed in this paper uses the word embedding technique and the attention mechanism module of the word vector to consider the importance of describing the same commodity words. GADAE performance is significantly better than the traditional implicit feedback-based recommendation system model WRMF, CDAE. This is a heterogeneous data that is superior to the neural gate structure proposed in this paper to fuse user binary ratings and commodity content information. Finally, GADAE's indicators in Recall and NDCG are superior to autoencoder-based architectures: CDAE, CVAE, which is due to the design of the encoder and decoder structure in the automatic encoder as a dual closed loop, while training The encoder and decoder consider the structural dual and feedback signals between the encoder and the decoder.

5. Conclusion
The dual automatic encoder (GADAE) based on the gate attention mechanism proposed in this paper is applied to content-aware recommendation. In this paper, we design a dual-closed dual-stack automatic encoder, combined with neural gate structure, word attention mechanism and proximity attention mechanism. Using the dual closed-loop automatic encoder to learn the hidden layer feature representation of commodity information, using the neural layer to fuse user binary rating and commodity content heterogeneous data, using the word attention mechanism and proximity attention mechanism to learn commodity information words and neighboring goods feature. The experimental results verify that the proposed GADAE has better performance than the current latest model.

Reference
[1] Li X, She J. Collaborative variational autoencoder for recommender systems[C]//Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2017: 305-314.
[2] Kim D, Park C, Oh J, et al. Convolutional matrix factorization for document context-aware recommendation[C]//Proceedings of the 10th ACM Conference on Recommender Systems. ACM, 2016: 233-240.
[3] Hu Y , Koren Y , Volinsky C . Collaborative Filtering for Implicit Feedback Datasets[C]//2008 Eighth IEEE International Conference on Data Mining. IEEE, 2009.
[4] Wu Y, DuBois C, Zheng A X, et al. Collaborative denoising auto-encoders for top-n recommender systems[C]//Proceedings of the Ninth ACM International Conference on Web Search and Data Mining. ACM, 2016: 153-162.
[5] Wang C, Blei D M. Collaborative topic modeling for recommending scientific articles[C]//Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2011: 448-456.
[6] Harper F M, Konstan J A. The movielens datasets: History and context[J]. Acm transactions on interactive intelligent systems (tiis), 2016, 5(4): 19.
[7] He R, McAuley J. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering[C]//proceedings of the 25th international conference on world wide web. International World Wide Web Conferences Steering Committee, 2016: 507-517.
[8] Hsieh C K, Yang L, Cui Y, et al. Collaborative metric learning[C]//Proceedings of the 26th international conference on world wide web. International World Wide Web Conferences Steering Committee, 2017: 193-201.
[9] Kim D, Park C, Oh J, et al. Convolutional matrix factorization for document context-aware recommendation[C]//Proceedings of the 10th ACM Conference on Recommender Systems. ACM, 2016: 233-240.

[10] Zhang Y, Ai Q, Chen X, et al. Joint representation learning for top-n recommendation with heterogeneous information sources[C]//Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. ACM, 2017: 1449-1458.