Academic Journal of Computing & Information Science
ISSN 2616-5775 Vol. 3, Issue 1: 17-23, DOI: 10.25236/AJCIS.030102

Cross Attentive Collaborative Filtering for Recommender Systems

Rongjie Shan, Wenming Ma, Mingming Qi

School of Computer and Control Engineering, Yantai University, Yantai 264005, China.

ABSTRACT. *The information on the Internet is increasingly complicated, users can not use the search engine to accurately find effective information. But the emergence of the recommendation system can effectively alleviate this problem. Algorithms such as collaborative filtering have a large drop in recommended quality when user ratings are sparse. We propose a novel cross attentive collaborative filtering model. Our model can learn more about the potential relationships between users, items and ratings. To validate the validity of the algorithm model, experiments were performed on public datasets. The experimental results show that the mae and rmse index have decreased.*

KEYWORDS: *attention, sparsity, potential relationship, recommender systems*

1. Introduction

The development of Internet technology has led to the continuous growth of information data in network nodes, which has brought rich information to users and users can easily obtain massive amounts of knowledge data. However, the exponential growth of information data has created an information overload problem, and a large amount of redundant information makes it difficult for users to obtain valid information in a limited time. The advent of search engines allows users to search using algorithms such as keywords and fuzzy queries in the early stages of data growth, thereby alleviating the problem of information overload [1]. Since entering the era of big data, data engine search technology has become weaker. Because of the existence of various redundant information, when the user needs are not clear, the user cannot be satisfied.

In response to the existing problems, a personalized recommendation system emerges as the times require, and it has become an effective method to alleviate data overload problems. The recommendation system does not require the user to provide explicit demand information. Analyze existing user data through various recommendation algorithms, actively recommend information content similar to user interest, guide users to find points of interest. Personalized recommendations

ISSN 2616-5775 Vol. 3, Issue 1: 17-23, DOI: 10.25236/AJCIS.030102

rely on user behavior information [2,3]. When the information is insufficient, the algorithm is prone to large errors. Data sparsity is a major challenge for recommendation systems.

In this paper, we propose a cross attentive collaborative filtering score prediction model which mining potential relationships. The neural network's nonlinear, multi-input, self-learning data and other characteristics are used for training and learning. At the same time, the attentional mechanism is used to mine the potential scoring model, and then the accuracy of scoring prediction is improved.

2. Related work

In the study of the recommender system, the sparsity of the score record data and the cold start of the user and the item are issues of concern. In some traditional recommendation algorithms, the non-negative matrix factor (NMF) proposed by Lee and Seung uses matrix decomposition to find potential data features under non-negative constraints, and predicts users' interest in unknown items[4]. In [5], using the algorithm based on probability matrix decomposition, the scoring matrix is decomposed into potential feature matrices of users and items, and is not subject to non-negative conditions.

The two decomposed matrices are restored to the scoring matrix form, filling in the blank records. The idea of this type of algorithm fills the blank record by decomposing the scoring matrix into a potential feature factor vector. Matrix decomposition correlation methods are largely limited by the sparseness of the data[6]. Too sparse results in large algorithm errors and low data filling quality, which in turn affects the overall effect of the recommendation. Therefore, fully exploiting the potential scoring features and patterns of users and item matrices is an effective way to solve the scoring prediction problem.

3. Cross attentive collaborative filtering model

Attention mechanism has a wide range of applications in depth learning such as image processing, speech recognition and natural language understanding. Attention mechanisms look for key information in a global context, giving higher weight to critical information and forming "attention"[7]. For example, users have different levels of attention to different types of movies. Movies that like love and romantic categories have less attention to horror and suspense categories. It is unreasonable to assign the same weight to global score analysis[8]. The Attention Mechanism can be used to rationalize the distribution of weights for different performances and contribute to the improvement of predictions.

ISSN 2616-5775 Vol. 3, Issue 1: 17-23, DOI: 10.25236/AJCIS.030102

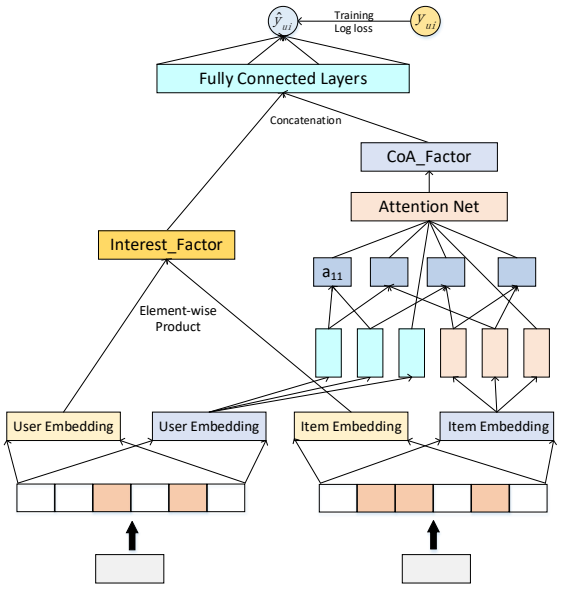

Figure. 1 Model Architecture

The model consists of two parts, Interest_Factor and CoA_Factor, where Interest Factor can handle potential scoring features, and CoA_Factor can make weight allocation more reasonable. Interest_Factor, which uses linear kernels to handle interactions between linear features, as defined below:

$$G = \mathbf{u}_i \odot \mathbf{v}_i \tag{1}$$

Where $\mathbf{u}_i$, $\mathbf{v}_i$ are the user and item vectors respectively. CoA_Factor, weighted using attention mechanism.For each id is unique, the Item-Embedding is formed into three copies, denoted as q_1, k_1, m_1 , and the User-Embedding copies into three, denoted as q_2, k_2, m_2 ,and $a_{11}, a_{12}, a_{21}, a_{22}$ is obtained by cross-multiplication of Item-Embedding and User-Embedding.

$$\begin{aligned} a_{11} &= \mathbf{k}_1 \odot \mathbf{q}_1, a_{12} = \mathbf{k}_2 \odot \mathbf{q}_1 \\ a_{21} &= \mathbf{k}_1 \odot \mathbf{q}_2, a_{22} = \mathbf{k}_2 \odot \mathbf{q}_2 \end{aligned} \tag{2}$$

Among them, "$\odot$"denotes the element-wise product of two matrices, the cross-point multiplication of User-Embedding and Item-Embedding, so that a vector with a correlation can obtain a larger value, and a non-correlated one appears as a smaller value.

The definition of Attention Net is as follows:

ISSN 2616-5775 Vol. 3, Issue 1: 17-23, DOI: 10.25236/AJCIS.030102

$$N=a_{out}(W_i^T\begin{bmatrix}M_1\\M_2\end{bmatrix}+b_i),$$

$$M_1 = A_{11} \odot m_1 + A_{12} \odot m_2,$$
$$M_2 = A_{21} \odot m_1 + A_{22} \odot m_2 \quad (3)$$
$$A_{11}, A_{12} = \frac{\exp(a_{11}, a_{12})}{\sum \exp(a_{11}, a_{12})},$$
$$A_{21}, A_{22} = \frac{\exp(a_{21}, a_{22})}{\sum \exp(a_{21}, a_{22})}$$

Attention Net layer uses the softmax function to calculate a_{11}, a_{12} at the same time. After the calculation, the result is split into the original dimension to get A_{11}, A_{12} .Then calculated in the same way to get A_{21}, A_{22} .M_1, M_2 are obtained by multiplying and summing $A_{11}, A_{12}, A_{21}, A_{22}$ with m_1 and m_2.

After processing, we have got Interest_Factor and CoA_Factor. Combine Interest_Factor and CoA_Factor, and then input to the fully connected neural network layer.

The deep neural network can be divided into three layers as a whole. The first layer can be divided into the input layer, the last layer can be divided into the output layer, and the middle network layer can be divided into the hidden layer. The hidden layer can be composed of one to multiple layers, and each layer can set a different number of neurons, which is the most complicated part. The connection method of the hidden layer can be fully connected, or dropout technology can be used to make partial connections between neurons, which has a strong expression ability combined with activation functions.

The definition of Multi-layer neural network is as follows:

$$z_1 = \phi_1(G, N) = \begin{bmatrix}G\\N\end{bmatrix},$$
$$\phi_2(z_1) = a_x(W_2^T z_1 + b_2),$$
..... (4)
$$\phi_k(z_{k-1}) = a_k(W_k^T z_{k-1} + b_k),$$
$$\hat{y} = \sigma(h^T \phi_k(z_{k-1}))$$

Where W_x, b_x, a_x denote the weight, bias and activation function for multi-layer neural network. σ is also the activation function, and then we can get the final predicted score $\hat{y}$ .We use MSE as the loss function, defined as follows:

ISSN 2616-5775 Vol. 3, Issue 1: 17-23, DOI: 10.25236/AJCIS.030102

$$Loss_{MSE} = \frac{1}{N}\sum_{t=1}^{N}(y-\hat{y})^2 \quad (5)$$

Among them, y is the actual score, $\hat{y}$ is the predicted score, and N is the set number of batches. The average of multiple records is used as the Loss of each training. As the number of trainings increases, Loss shows a downward trend, but too small Loss may be overfitting.

In neural networks, we use Adam as the gradient descent optimization algorithm, and learning rate set to 0.005. Adam iteratively updates the neural network weights based on the training data, which can replace the first-order optimization algorithm of the traditional stochastic gradient descent process. Adam is suitable for solving problems containing very noisy or sparse gradients. Hyperparameters can be explained intuitively, and basically require only a small amount of tuning parameters.

4. Experiment

The programming language used in this experiment is Python3.6, the deep learning framework is Keras+Tensorflows1.14, the operating system is Windows7, the clock speed is 3.20GHz and the memory is 8GB.

The experimental data was obtained from the Book-Crossing community with kind permission from Ron Hornbaker, contains 1,149,780 ratings about 271,379 books. The scoring was processed so that the scoring range was 1 to 5. The size of the dataset is controlled by choosing different numbers of people.

4.1 Evaluation Protocols

The scoring prediction indicators used in this paper are the root mean square error (RMSE) and the mean absolute error (MSE). The evaluation indicators are defined as follows:

$$\text{MAE} = \frac{\sum_{i \in T_E} | r_i - r_i^* |}{| T_E |} \quad (6)$$

$$\text{RMSE} = \sqrt{\frac{\sum_{i \in T_E} | r_i - r_i^* |^2}{| T_E |}} \quad (7)$$

Where $|T_E|$ represents the number of records in the test set, r_i represents the actual score, and r_i^* represents the predicted score for the i records.

In the experimental evaluation, the smaller the error value of MAE and RMSE, the higher the accuracy of the representative model and the better the effect.

ISSN 2616-5775 Vol. 3, Issue 1: 17-23, DOI: 10.25236/AJCIS.030102

4.2 Baselines

MF: Matrix factorization decomposes the sparse user rating matrix into two matrices, fills them with the form of multiplication and restores the original matrix dimensions, and learns the potential rating mode in the form of feature vectors.

GMF: Generalized matrix decomposition, which uses Embedding vectorization to represent users and items, learns the intrinsic relationship by dot multiplication, and uses gradient descent iterative training.

MLP: Multilayer Perceptron, which learning the potential relationship between users and items through neural network unit connections.

CACF: This method is proposed in this paper, using the attention mechanism.

4.3 Performance Comparison

We have performed several experiments, taking the mean as the final result of each algorithm. In the experiment, we used 30% of the data as the test set and 70% of the data as the training set.

For each experiment, the training set and test set are randomly generated in proportion, and the results are slightly different. We use the average of RMSE and MAE as the currently recommended results. Recommendation of books score prediction for different numbers of people. Here are the results of the experiment.

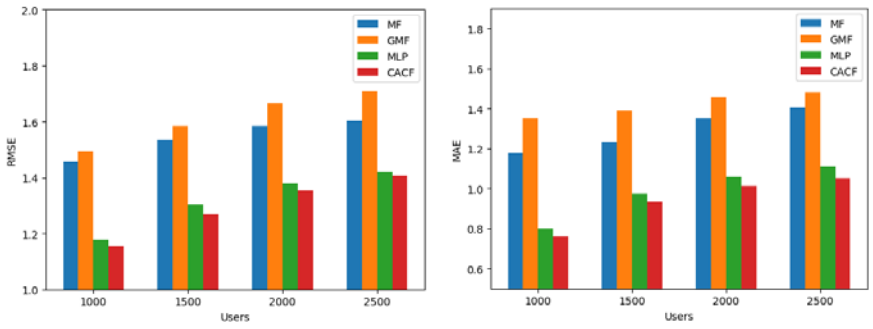

Figure. 2 Results of different methods

It can be seen from the figure that the algorithm model has better results in multiple sets of data, and the RMSE evaluation value is the lowest. From multiple sets of experimental data, as the number of recommended people increases, the sparsity gradually increases. Experimental results are reports Fig.2. Compared with other methods, our model has more advantages and lower scoring errors.

ISSN 2616-5775 Vol. 3, Issue 1: 17-23, DOI: 10.25236/AJCIS.030102

5. Conclusion

In this paper, we propose a cross attentive collaborative filtering model that predicts user ratings by using the attention mechanism and can alleviate the effects of data sparsity. The next research focuses on the further optimization of the model, and gradually matures the model to adapt to different data environments in combination with various aspects of attribute information, so as to be applied to actual engineering projects.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No. 61602399).

References

[1] J.Bobadilla.,F.Ortega.,Hernando,et al(2013).Recommender systems survey. Knowledge Based Systems 46,p.109-132.
[2] R.Burke(2012).Hybrid recommender systems:Survey and experiments.In User modeling and user-adapted interaction,p.331-370.
[3] Xiangnan He,LiziLi Zhao，Hanwang Zhang et al(2017).In Neural Collaborative Filtering.International World Wide Web Conference Committee.
[4] D.Lee,HS.Seung(2001).Algorithms for non-negative matrix factorization . In Advances in neural information provessing systems.p.556-560.
[5] P.Peng,L.Xiao,et al(2017).Recommendation algorithm based on uer trust and interest with probability matrix factorization,In International Conference on Advanced Cloud & Big Data.IEEE Computer Society,p.355-361.
[6] X. He,H. Zhang,M.-Y. Kan,and T.-S.Chua(2016).Fast matrix factorization for online recommendation with implicit feedback . In SIGIR.p.549-558.
[7] A.Vaswani,N.Siiazeer,N.Parmar,et al(2017). Attention ia all you need.In Advances in Neural Information Processing Systems,p.5998-6008
[8] N. Srivastava and R.Salakhutdinov(2012).Multimodal learning with deep boltzmann machines. In NIPS,p.2222-2230.