Proceedings of the 5th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/15/20230826

Deep learning methods used in movie recommendation systems

Lexi Liu

Chengdu Foreign Languages School, Xipu Street, Pidu District, Chengdu City, Sichuan Province, China

1502806106@qq.com

Abstract. As the amount of internet movie data grows rapidly, traditional movie recommendation systems face increasing challenges. They typically rely on statistical algorithms such as item-based or user-based collaborative filtering. However, these algorithms struggle to handle large-scale data and often fail to capture the complexity and contextual information of user behavior. Therefore, deep learning techniques have been widely applied to movie recommendation systems. This paper reviews movie recommendation algorithms based on traditional statistical models and introduces three main deep learning techniques: Artificial Neural Networks (ANN), Convolutional Neural Networks (CNN), and Recurrent Neural Networks (RNN). ANN can extract features at different levels of users and movies; CNN can capture features of movie posters and movie data to recommend similar movies; RNN can consider user historical behavior and contextual information to better understand user interests and demands. The application of these deep learning techniques can enhance the accuracy and user experience of movie recommendation systems. This paper also demonstrates the advantages and disadvantages of these models and their specific application methods in movie recommendation systems, and points out the direction for further development and improvement of deep learning models in this field.

Keywords: recommendation system, deep learning, artificial neural network, CNN, RNN.

1. Introduction

Machine learning plays an extremely important role in movie recommendation systems [1]. It can help improve the accuracy and personalization of recommendations. With the development of the internet and the digital entertainment industry, people are increasingly inclined to watch movies at home. However, it is not easy to find content that one likes from millions of movies and TV programs. This is why movie recommendation systems have become so important. Machine learning can learn data features from a large amount of user data such as viewing history and ratings, and use this as the basis for training models to better predict user preferences and recommend movies and TV programs that users may like. Machine learning can also help recommendation systems solve the cold start problem, where new users joining the system may not have enough personal preference data for accurate recommendations. In this case, machine learning can use some data such as age, gender, and geographic location to infer new users' preferences. Through machine learning, recommendation systems can recommend according to each user's specific needs, making recommendations more accurate,

DOI: 10.54254/2755-2721/15/20230826

personalized, and targeted. Therefore, the application of machine learning in movie recommendation systems is of great significance. It can help recommendation systems better serve users, improve user experience, and also promote the development of the digital entertainment industry [2].

Traditional machine learning was initially used in recommendation systems to predict user preferences based on pre-defined features and corresponding relationships. For movie recommendation systems, this involved analyzing user preferences and movie characteristics to recommend movies. However, traditional machine learning methods have limitations in effectively learning from growing massive internet data, resulting in limited accuracy of recommendation systems. The emergence of deep learning technology as a powerful tool for recommendation systems can be attributed to advancements in graphics cards and parallel computing. With its powerful parameter space, deep learning can effectively learn from massive internet data, allowing for the discovery of hidden features and complex relationships between data without the need for manual feature extraction [3-6]. Compared with traditional machine learning algorithms, deep learning can improve the accuracy of recommendation systems and reduce the time and effort required for manual intervention, ultimately enhancing personalization and real-time performance. In this article, we explore the application of different deep learning models for various types of movie data on the internet, and provide a summary of their advantages and disadvantages. Our analysis provides guidance for the future construction of movie recommendation systems, emphasizing the importance of deep learning in leveraging large amounts of data to better serve users and promote the development of the digital entertainment industry.

2. Traditional machine learning methods

Traditional machine learning algorithms have been widely used in early movie recommendation systems. These methods analyze the relationship between different users and movies by building statistical models, extracting sensitive features of different user groups for different types of movies, in order to better meet the recommendation needs of users.

In our research on traditional machine learning, the focus was mainly on comparing the similarity of different groups of data to achieve recommendation through machine learning methods. Here we introduce two key recommendation algorithms.

The first method is to use similarity calculation to classify different users or movies to achieve movie recommendations. This method can be divided into two aspects. In the case of a movie recommendation system, they are based on user-based collaborative filtering and item-based collaborative filtering. Both of them mainly require the user's rating data for different movies as input features for the recommendation system. The essence of user-based collaborative filtering is to identify a group of users who have similar preferences to the target user based on existing ratings. By analyzing the evaluation of this group for a given movie, the system can decide whether to recommend the movie to the target user. In item-based collaborative filtering, the system first finds the movies rated highly by the target user, compares their ratings with other movies, and then recommends similar movies to the target user based on their similarity. These similarity-based methods have high flexibility and can use different correlation metrics, such as Pearson correlation coefficient and Spearman correlation coefficient [7-9].

$$r = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2(Y_i-\bar{Y})^2}}$$

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n^3-n}$$

The second method involves using the K-Nearest Neighbor (KNN) algorithm to classify users or movies into categories [10,12]. The KNN algorithm is the simplest non-parametric supervised classification algorithm. In the KNN algorithm, we need to calculate the distance between the object to be classified and its neighboring nodes based on its features (such as Euclidean distance, Chebyshev

DOI: 10.54254/2755-2721/15/20230826

distance, etc.), and then select the majority category among its K neighbors as the target object's category. The advantage of this method over the first method is that when recommending movies, we only need to consider the category of users or movies without comparing the similarity of individuals one by one, which improves the efficiency of the recommendation system. At the same time, this algorithm does not require parameters, which means that the model does not make any assumptions about the data, reducing the impact of special cases on the recommendation system. However, the KNN algorithm faces problems such as high computational complexity and high spatial complexity. At the same time, the imbalance of samples can also affect the effectiveness of this method.

3. Deep learning

Against the backdrop of the huge amount of data in the Internet today, traditional machine learning algorithms have shown their limitations. On the one hand, a large amount of data does not significantly improve the accuracy of traditional machine learning. On the other hand, traditional machine learning algorithms cannot fully explore the rules behind the data. In the research process, the continuous development of neural network direction has led to the concept of deep learning, which can deeply analyze the inherent features and rules of the samples, and its performance is positively correlated with the size of the data. That is to say, under the background of big data, the performance of deep learning will be better.

4. Artificial neural network

Artificial neural networks (ANNs) are information processing systems composed of a large number of interconnected neurons, analogous to the neural systems that transmit information through synapses in biology. ANNs typically consist of an input layer, hidden layers, and an output layer. In general, the more complex the problem and the more variables involved, the more layers and neurons in the hidden layers are required [13,14]. When we input data into an ANN system, the data is processed by a parameter matrix in the hidden layer, and the signal transmission between neurons is simulated through a nonlinear activation function [14]. With numerous model parameters, ANNs can learn the hidden patterns in the data, greatly improving the accuracy of recommendation systems. To evaluate the performance of an ANN system, we typically use error functions such as RMSE. The training process involves adjusting the parameters to minimize the error function, which can be achieved through gradient descent. The specific process of gradient descent involves using the backpropagation algorithm to obtain the gradient of the error function, and then continuously adjusting each parameter in the opposite direction of the gradient until the error function converges. However, due to the large number of parameters required by ANNs, the high cost of adjusting parameters is inevitable, which limits the application scenarios of ANNs.

5. Convolutional neural network

Convolutional Neural Network (CNN) is a deep learning model mainly used for processing image data [15]. In a movie recommendation system, we can use CNN to extract features from movie screenshots, posters, and other image data to help the system obtain more information about movies. During the process of using CNN to process input information, the information goes through a series of layers including convolutional layer, ReLU layer, pooling layer, and fully connected neural network, and eventually outputs a result [16]. In the convolutional layer, a convolutional kernel is used as a set of weighted elements to perform a weighted sum with elements in the input information. The role of the convolutional layer is to represent local features of the input information with more concise and distinctive numbers. The larger the number obtained through the convolution process, the more correlation there is between the local feature and the given template. Using convolutional kernels to process information greatly reduces the number of parameters and computational complexity compared to fully connected artificial neural networks, which improves efficiency. Additionally, CNNs largely reduce data volume and preserve spatial information in images by utilizing convolutional and pooling layers, avoiding the limitation of one-dimensional representation of all information. This makes CNNs

DOI: 10.54254/2755-2721/15/20230826

perform well in image classification and other tasks. However, although CNNs reduce the number of parameters through sparse connections, they still have high computational requirements and require a large amount of training data to continuously adjust parameters, which reflects the high operating costs of CNNs.

6. Recurrent neural network

Movie ratings are an important source of information for recommending movies to users, and an objective and accurate rating can correct many misleading movie information. Usually, movie ratings are presented in the form of text, and our information has a chronological relationship that needs to be combined and read together to be properly understood. Therefore, people have proposed recurrent neural networks (RNNs) to connect and process the chronological information [17-20]. RNNs can predict user preferences based on their historical behavior (such as which movies they have watched before and their ratings), thus recommending movies to them. In addition, RNNs can model sequence data over time, so they can sort recommendations based on time and provide users with time-based recommendations. Recurrent neural networks process input with a chronological relationship in sequence. In each step, the information retained from the previous step is passed to the current moment. At the same time, another part of the information comes from the current input, and the output information is passed to the next moment. However, this processing method has the disadvantage of severe data loss when facing a large amount of input data. Because each input data can only be passed on to the next loop in a certain proportion, the initial input data may decay to a very small proportion during the processing. However, even so, it still requires a very large amount of training data to adjust the parameters, resulting in high training costs.

7. Discussion

With the advancement of deep learning technology, movie recommendation systems can now effectively utilize big data on the internet with minimal human intervention. Deep learning algorithms can autonomously analyze and utilize larger and more diverse data, while also extracting data features and correlations that are not easily observed by humans. This results in significantly improved utilization rate and recommendation accuracy of data. However, while the advantages of deep learning are evident, there are still certain shortcomings that need to be addressed.

First of all, the aforementioned deep learning models still have certain shortcomings. For example, in the application of convolutional neural networks, the original information undergoes highly abstract processing through multiple layers of CNN, which leads to information loss during the processing. To address this issue, scientists have proposed residual structures, which preserve some of the previous data during each step of processing, reducing data loss and avoiding the problem of gradient vanishing [5]. In the case of recurrent neural networks, severe data loss is also a barrier to their development, which led to the emergence of LSTM (Long Short-Term Memory) and GRU (Gated Recurrent Unit) [17-20]. LSTM demonstrates great advantages in processing long-term memory, but has the limitation of extremely high computational complexity; GRU is a simplified form of LSTM, with reduced computational complexity and can be considered as a compromise between the original RNN and LSTM. The recent popular ChatGPT model is based on the Transformer model, which can better adapt to information processing in the context of big data and can be widely applied to various scenarios, but still has the drawback of high computational complexity and training cost, as is common with deep learning models [21].

Secondly, data imbalance can lead to a decrease in the accuracy of the recommendation model. In the era of the Internet, data is inevitably biased due to the preferences of the user group, and it cannot guarantee an equal relationship in terms of data volume for each category. In the movie recommendation system, this will lead to insufficient data for specific categories of movies, and users with niche preferences may not get ideal recommendation results. To address this issue, we can supplement data through other means to avoid learning biases caused by data. We can also perform data augmentation on existing data, such as rotating and adding noise to image information, to increase the volume of data.

DOI: 10.54254/2755-2721/15/20230826

Data augmentation operations need to consider the application scenario of the model, that is, whether the accuracy requirements are strict [22]. In addition to processing data, we can also use transfer learning to strengthen the connection between multiple tasks. For example, when recommending movies to users with niche preferences, we can use the parameters previously used to recommend similar users as the basis for improving the model, and then make adjustments to reduce the error caused by insufficient data to a certain extent [23]. In addition, Generative Adversarial Networks (GAN) can also be used to mitigate the obstacles caused by data imbalance. In the GAN model, there are mainly two parts: the generator and the discriminator. The generator generates data and the discriminator judges whether the data is true or false, and affects the judgment to have stronger discrimination ability. The results of discrimination can in turn affect the parameters of the generator to generate more realistic data. The generator and discriminator are trained in turn, making the generator have a strong ability to create data. At this time, the data created by the generator can be used to solve the problem of data imbalance [24].

Finally, not all recommendation system models are better with increased complexity. Deep learning faces limitations due to the huge computational requirements, and the value of traditional machine learning cannot be denied even in cases with small data and clear features, despite the superiority of deep learning.

References

[1] Marappan, R. & Bhaskaran, S. Movie Recommendation System Modeling Using Machine Learning. Int. J. Math. Eng. Biol. Appl. Comput. 12–16 (2022).

[2] Goyani, M. & Chaurasiya, N. A Review of Movie Recommendation System: Limitations, Survey and Challenges. ELCVIA Electron. Lett. Comput. Vis. Image Anal. 19, 18–37 (2020).

[3] Deng, L. & Yu, D. Deep Learning: Methods and Applications. Found. Trends® Signal Process. 7, 197–387 (2014).

[4] Kamilaris, A. & Prenafeta-Boldú, F. X. Deep learning in agriculture: A survey. Comput. Electron. Agric. 147, 70–90 (2018).

[5] He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image Recognition. (2015).

[6] LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. Nature 521, 436–444 (2015).

[7] Myers, L. & Sirois, M. J. Spearman Correlation Coefficients, Differences between. in Encyclopedia of Statistical Sciences (John Wiley & Sons, Ltd, 2006). doi:10.1002/0471667196.ess5050.pub2.

[8] Benesty, J., Chen, J. & Huang, Y. On the Importance of the Pearson Correlation Coefficient in Noise Reduction. IEEE Trans. Audio Speech Lang. Process. 16, 757–765 (2008).

[9] Sedgwick, P. Pearson's correlation coefficient. BMJ 345, e4483 (2012).

[10] Ahuja, R., Solanki, A. & Nayyar, A. Movie Recommender System Using K-Means Clustering AND K-Nearest Neighbor. in 2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence) 263–268 (2019). doi:10.1109/CONFLUENCE.2019.8776969.

[11] Ahmed, M., Seraj, R. & Islam, S. M. S. The k-means Algorithm: A Comprehensive Survey and Performance Evaluation. Electronics 9, 1295 (2020).

[12] Guo, G., Wang, H., Bell, D., Bi, Y. & Greer, K. KNN Model-Based Approach in Classification. in On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE (eds. Meersman, R., Tari, Z. & Schmidt, D. C.) 986–996 (Springer, 2003). doi:10.1007/978-3-540-39964-3_62.

[13] Jain, A. K., Mao, J. & Mohiuddin, K. M. Artificial neural networks: a tutorial. Computer 29, 31–44 (1996).

[14] Krogh, A. What are artificial neural networks? Nat. Biotechnol. 26, 195–197 (2008).

[15] Li, Z., Liu, F., Yang, W., Peng, S. & Zhou, J. A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects. IEEE Trans. Neural Netw. Learn. Syst. 33, 6999–7019 (2022).

[16] Gu, J. et al. Recent advances in convolutional neural networks. Pattern Recognit. 77, 354–377

(2018).
[17] Staudemeyer, R. C. & Morris, E. R. Understanding LSTM -- a tutorial into Long Short-Term Memory Recurrent Neural Networks. Preprint at https://doi.org/10.48550/arXiv.1909.09586 (2019).
[18] Smagulova, K. & James, A. P. A survey on LSTM memristive neural network architectures and applications. Eur. Phys. J. Spec. Top. 228, 2313–2324 (2019).
[19] Chung, J., Gulcehre, C., Cho, K. & Bengio, Y. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. Preprint at https://doi.org/10.48550/arXiv.1412.3555 (2014).
[20] Zhao, R. et al. Machine Health Monitoring Using Local Feature-Based Gated Recurrent Unit Networks. IEEE Trans. Ind. Electron. 65, 1539–1548 (2018).
[21] Vaswani, A. et al. Attention Is All You Need. (2017) doi:10.48550/ARXIV.1706.03762.
[22] Chlap, P. et al. A review of medical image data augmentation techniques for deep learning applications. J. Med. Imaging Radiat. Oncol. 65, 545–563 (2021).
[23] A survey of transfer learning | Journal of Big Data | Full Text. https://journalofbigdata.springeropen.com/articles/10.1186/s40537-016-0043-6.
[24] Creswell, A. et al. Generative Adversarial Networks: An Overview. IEEE Signal Process. Mag. 35, 53–65 (2018).