Research on the Combination of Probability Statistics and Recommendation Algorithm

Yejun Wu, Hongli Yang

Nanjing Institute of Technology, Nanjing, 211167, Jiangsu, China

wyj852507@sina.com, yanghongli1016@163.com

Keywords: Probability Statistics; Recommendation Algorithm; Data Processing; Assessment Index

Abstract: With the continuous development of information technology, the emergence of massive data makes recommendation algorithm become the core technology to solve the problem of user information acquisition. However, the traditional recommendation algorithm has some limitations. This article focuses on the fusion of probability statistics and recommendation algorithm. By expounding the basic theory of probability statistics and recommendation algorithm, this article deeply analyzes the application of probability statistics in data preprocessing, model construction and result assessment of recommendation algorithm, and discusses the advantages and challenges brought by the combination of them. Combining probability statistics with recommendation algorithm can improve the accuracy of recommendation algorithm, enhance its adaptability to complex data, and expand the diversity of recommendation. However, in this process, it also faces many challenges, such as the increase of computational complexity, the inconsistency between probability hypothesis and actual situation, and privacy protection. This study provides a new idea for the optimization of recommendation algorithm, promotes the application of probability and statistics theory in this field, and has important theoretical reference value and guiding significance for related research and practice.

1. Introduction

With the rapid development of information technology, the amount of data carried by the Internet has exploded. Faced with massive information, it is difficult for users to obtain the content they need efficiently [1]. As the key technology to solve this problem, recommendation algorithm has been widely used in many fields, such as e-commerce platform, social media, online music and video. It provides personalized information recommendation service for users, thus enhancing the user experience and increasing the stickiness and commercial value of platform users [2]. The traditional recommendation algorithm has some limitations in the face of complex and changeable data and diverse user needs [3]. Content-based recommendation algorithm mainly makes recommendations based on the characteristics of items, but fails to capture the dynamic changes of users' interests; Collaborative filtering recommendation algorithm faces the problems of data sparsity and cold start, which affects the accuracy and stability of recommendation effect.

Probability statistics, as a discipline to study the statistical regularity of random phenomena, provides strong theoretical support for solving the above dilemma of recommendation algorithm [4]. Many theories and methods in probability statistics can effectively model and analyze the uncertainty and randomness of data [5]. By integrating the theory of probability and statistics into the recommendation algorithm, the data can be cleaned, feature extracted and selected more scientifically in the data preprocessing stage, and the data quality can be improved. In the process of model construction, the model structure and parameters are optimized by means of probability distribution and statistical inference to enhance the adaptability of recommendation algorithm to complex data; In the assessment of recommendation results, the performance of recommendation algorithm is comprehensively and objectively evaluated quantitatively with the help of probability and statistical indicators [6].

In view of this, it is of great practical significance to explore the integration of probability

  DOI: 10.25236/icetmr.2025.004

statistics and recommendation algorithm. This research can open up a new path and provide a new method for the optimization of recommendation algorithm, thus improving the performance of recommendation system and enhancing user satisfaction. It is also helpful to broaden the application scope of probability and statistics theory and promote the development of interdisciplinary research. This article aims to explore the combination mode, advantages and challenges of probability statistics and recommendation algorithm in an all-round way, and provide theoretical reference and guidance for research and practice in related fields.

2. Probability statistics and recommendation algorithm theory

Probability theory focuses on the quantitative law of random phenomena, in which random variables are an important tool to describe random phenomena. When analyzing the user's preference for goods, the user's score can be regarded as a random variable, and its value reflects the different preferences of users [7]. Probability distribution further describes the probability of random variables taking different values, such as normal distribution and Poisson distribution, and different distributions are suitable for different types of data characteristics. Statistics focuses on how to collect, sort out and analyze data. Statistics such as mean and variance can help us understand the concentration trend and dispersion degree of data. For example, by calculating the average value of a group of users' ratings, we can know the overall assessment level, and the variance can reflect the fluctuation of ratings. Correlation analysis is used to study the degree of correlation between variables, and in the recommendation algorithm, we can find out the potential relationship between different commodities or users.

There are also many types of recommendation algorithms. Content-based recommendation algorithm makes recommendations according to the characteristics of items themselves. In music recommendation, it can recommend related songs for users who like similar music characteristics according to the characteristics of song style, singer, release time and so on [8]. Collaborative filtering recommendation algorithm is based on user behavior data, looking for other users with similar interests to the target user, and then recommending items that similar users like. On the e-commerce platform, by analyzing the user's purchase records, it finds a group of users with similar purchase behaviors, and recommends products purchased by this group for the target users but not yet purchased by the target users [9]. These algorithms have their own limitations. Content-based recommendation algorithm is difficult to capture the dynamic changes of users' interests, and collaborative filtering recommendation algorithm is easily affected by data sparsity, resulting in cold start.

3. Application mode of probability statistics in recommendation algorithm

Probability statistics plays a vital and indispensable role in the actual operation of recommendation algorithm. Its application covers many key links such as data preprocessing, model construction and recommendation result assessment.

(1) Data preprocessing

As the first step of recommendation algorithm, data preprocessing aims at improving data quality and laying a solid foundation for subsequent model training. In this link, the probability statistical method plays an extremely important role. When dealing with user behavior data, the interference of data noise and outliers is common. Using the concepts of mean, median and standard deviation in statistics, we can identify and deal with these abnormal data. Assuming that there is a group of users' scoring data for commodities, a reasonable threshold range can be set by calculating the mean and standard deviation of the scores, and the scores beyond this range can be corrected or eliminated as abnormal values. In feature extraction, correlation analysis in probability statistics can help screen out features closely related to recommended targets.

(2) Model construction link

Probability statistics provides rich and solid theoretical support for the construction of recommendation algorithm model. Taking the common collaborative filtering recommendation

algorithm as an example, when calculating the similarity between users, methods such as cosine similarity or Pearson correlation coefficient in probability statistics can be used. By analyzing user behavior data, these methods measure the similarity between users from the perspective of probability statistics, and then recommend products that similar users like for target users.

Probability distribution hypothesis is a common means in constructing forecasting model. It is assumed that users' ratings of goods obey normal distribution. Based on this assumption, model parameters can be determined by using maximum likelihood estimation and other methods, so as to predict users' preference for unrated goods. Bayesian network is also a powerful model building tool based on probability statistics. It can describe the complex dependence structure between data through the conditional probability relationship between nodes, so as to make recommendation prediction more accurately. Table 1 shows the comparison of recommended model parameters based on probability statistics:

Table 1: Comparison of Parameters in Recommendation Models Based on Probability and Statistics

Recommendation Model	Probability and Statistics Method	Key Parameters	Parameter Functions
Cosine Similarity-Based Collaborative Filtering	Cosine similarity calculation	Similarity threshold	Determines the range of similar users
Prediction Model Based on Normal Distribution Assumption	Maximum likelihood estimation	Mean, standard deviation	Describes the rating distribution and predicts ratings
Bayesian Network Model	Conditional probability calculation	Conditional probability tables	Characterizes data dependencies and makes recommendation predictions

(3) Recommendation result assessment link

The assessment of recommendation results is a key step to measure the performance of recommendation algorithms. Probability statistics provides a series of quantitative assessment indexes, which help us evaluate the merits of recommendation algorithms comprehensively and objectively. Accuracy and recall are two commonly used indicators. Accuracy indicates the proportion of related items in the recommendation results, and recall rate measures the proportion of all related items recommended. In a movie recommendation system, if 10 movies are recommended, and 6 of them are of interest to users, the accuracy rate is 60%. However, if there are 20 movies that users are interested in, the recommended 6 movies are only part of them, and the recall rate is 30%.

In addition, the MSE (Mean Squared Error) is also an important assessment index, which is used to measure the degree of error between the predicted score and the actual score. By calculating the average of the sum of squares of the difference between the predicted score and the actual score, MSE can reflect the prediction accuracy of the recommended algorithm. Lower MSE value means that the prediction result of the recommendation algorithm is closer to the actual situation. These probabilistic statistical indicators complement each other, which provides a strong basis for the optimization of recommendation algorithm.

4. Advantages and challenges of combining probability statistics with recommendation algorithm

(1) The advantages of combination

The combination of probability statistics and recommendation algorithm brings many obvious advantages to the recommendation system, but at the same time, it inevitably faces a series of challenges. Correlation analysis and probability distribution model in probability statistics can deeply explore the potential relations and laws behind the data. Taking e-commerce recommendation as an example, by analyzing the data such as user's purchase history and browsing records, the user's preferences are accurately located by using probability statistics method, and then the products that meet their actual needs are pushed for users. For example, Bayesian inference is

used to update the prior knowledge of users' interests according to their past behaviors, which makes the recommendation more suitable for users' actual needs.

The data in reality often has the characteristics of high dimension, high noise and sparse data. Dimension reduction methods in probability statistics, such as principal component analysis, can reduce the data dimension and noise interference while retaining key information, so that the recommendation algorithm can handle complex data better. Aiming at the problem of sparse data, matrix decomposition technology based on probability statistics can effectively fill the missing data and improve the recommendation effect. Probabilistic statistical methods can find hidden association patterns in data, which are not limited to the common user-object relationship. See Figure 1 for comparison of performance advantages of recommendation algorithm before and after combining probability statistics:

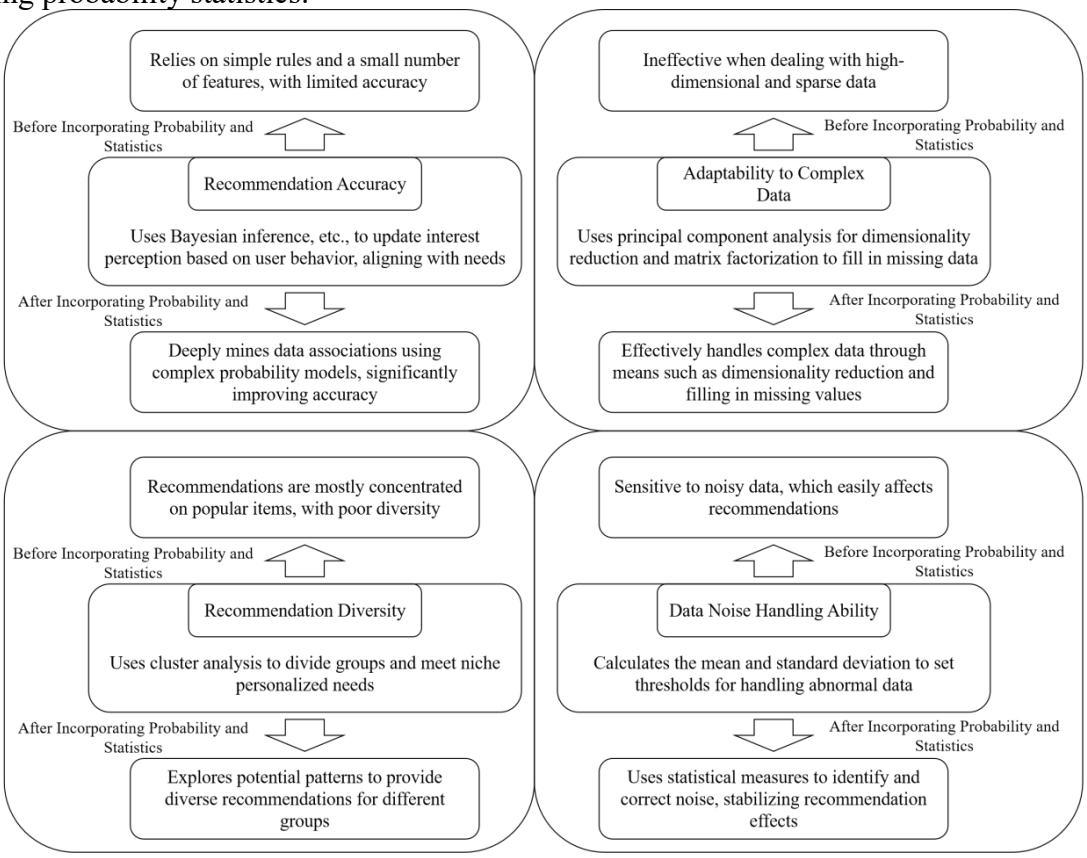

Figure 1 Comparison of performance advantages of recommendation algorithms before and after combining probability statistics

(2) Challenges

Although the combination of probability statistics and recommendation algorithm has many advantages, it also faces a series of thorny challenges. It is a big problem to increase the computational complexity of probability statistical model under high-dimensional data. With the increasing data dimension, the computational complexity of probabilistic statistical models, such as Bayesian networks and complex probability graph models, increases exponentially. This requires a lot of computing resources, and it will also lead to a longer response time of the recommendation algorithm, which will affect the user experience. It is also a common problem that the probability hypothesis does not match the actual data distribution. In the process of building the model, we often make some assumptions about the probability distribution of the data, such as the assumption that the user's score obeys the normal distribution. However, the distribution characteristics of the actual data may be extremely complex and inconsistent with the pre-assumption, which will lead to model deviation and further reduce the accuracy of the recommendation algorithm.

The contradiction between data privacy protection and the demand of probability and statistical analysis can not be ignored. Probability statistical analysis usually needs a lot of user data to ensure

the accuracy and reliability of the results, but user data involves personal privacy. How to carry out effective probability statistical analysis on the premise of protecting users' privacy has become an urgent problem. For example, although differential privacy technology can protect data privacy to a certain extent, it may affect the availability of data, and then affect the effect of probability statistical analysis.

5. Conclusions

In today's information explosion era, recommendation algorithms are widely used in various fields. However, the limitations of traditional recommendation algorithms urge us to seek new optimization approaches, which are strongly supported by probability statistics. This article deeply discusses the combination of probability statistics and recommendation algorithm, and comprehensively analyzes its theoretical basis, application mode, advantages and challenges.

By combing the basic theories of probability statistics and recommendation algorithm, we can know that the random variables, probability distribution, mean variance and other theories in probability statistics are closely connected with the principles of content-based and collaborative filtering, which lays a solid foundation for the combination of the two. In the application mode, probability statistics runs through the whole process of data preprocessing, model construction and recommendation result assessment of recommendation algorithm. Used for cleaning data and screening features during data preprocessing; In model construction, it helps to calculate similarity and determine model parameters; The assessment of recommendation results provides quantitative indicators such as accuracy, recall and mean square error, which comprehensively improves the performance of recommendation algorithm.

The combination of probability statistics and recommendation algorithm has obvious advantages, which can improve the accuracy of recommendation results, make recommendations more suitable for users' actual needs, enhance the adaptability to complex data, effectively deal with problems such as high-dimensional and sparse data, and meet the individual needs of different users. However, it should not be ignored that the combination process also faces many challenges, such as the increase of computational complexity of probability statistical model under high-dimensional data, the inconsistency between probability hypothesis and actual data distribution, and the contradiction between data privacy protection and analysis requirements. In the future, related research should focus on solving these challenges, further explore the innovative integration of probability statistics and recommendation algorithms, promote the optimization and development of recommendation systems in more scenarios, and provide users with better and more efficient personalized recommendation services.

Acknowledgements

The authors acknowledge the Cross-Curriculum Project of "Jiebang Guashuai" (Open Bidding for Key Projects) at Tianyinhu Institute of Science and Technology Innovation, Nanjing Institute of Technology:“ The Topic of University Mathematics and Physics Comprehensive Course” (NO:2025TKJA01).

References

[1] Li Huaxiaoyang, Xu Qing, Wang Zhuoning. A Point-of-Interest Recommendation Algorithm Integrating User Spatial Behavior Characteristics[J]. Journal of Geomatics Science and Technology, 2024, 40(6): 654-660.

[2] Li Jianfeng, Chen Hailong, Zhai Jun. A Personalized Recommendation Algorithm Under Complete Cold Start[J]. Computer Engineering and Design, 2024, 45(8): 2329-2335.

[3] Tian Renjie, Jing Mingli, Jiao Long. A Graph Contrastive Learning Recommendation Algorithm Based on Hybrid Negative Sampling[J]. Journal of Computer Applications, 2025, 45(4): 1053-1060.

[4] Guo Yuhan, Zhu Rushu. A Dynamic Pick-Up Point Recommendation Algorithm Based on Multi-Mode Deep Forest and Iterative Kuhn-Munkres[J]. Application Research of Computers, 2024, 41(12): 3634-3644.

[5] Ge Han, Ren Shumin, Zhang Hongliang. An Intelligent Recommendation Algorithm for Network Association Information Considering Data Sparsity[J]. Computer Simulation, 2024, 41(10): 311-316.

[6] Yan Mengmeng, Wang Haitao, He Jianfeng, Chen Xing. A Sequential Recommendation Algorithm Integrating Hierarchical Attention Mechanism and User Dynamic Preferences[J]. Mini-Micro Systems, 2024, 45(3): 621-628.

[7] Fan Jiaobei, Qian Yuhua, Peng Fu. A Citation Recommendation Algorithm Integrating Multi-Level Interactive Attention[J]. Mini-Micro Systems, 2023, 44(12): 2656-2662.

[8] Li Hui, Zheng Shanhong, Wang Guochun. A Social Recommendation Algorithm Based on Distributed and Graph Attention Mechanisms[J]. Computer Engineering and Design, 2024, 45(11): 3463-3470.

[9] Xia Xiang, Liu Jiang, Ni Feng. A Recommendation Algorithm Based on Implicit Feedback and Weighted User Preferences[J]. Computer Technology and Development, 2024, 34(3): 140-146.