2020 13th International Symposium on Computational Intelligence and Design (ISCID) | 978-1-7281-8446-3/20/$31.00 ©2020 IEEE | DOI: 10.1109/ISCID51228.2020.00076

Application of improved k-means k-nearest neighbor algorithm in the movie recommendation system

Chang Cai[1], Li Wang [2]
College of Computer and Software Engineering,
University of Science and Technology Liaoning
Anshan, China
E-mail: 425073943@qq.com

***Abstract*—In this paper, we propose a clustering and reclassification method for movie recommendation. We use the improved K-means algorithm to cluster according to the scores of similar users, Firstly, the elbow function is used to estimate the number of clusters, and the elbow method is used to determine the K value. Then, the K-means algorithm of the maximum and minimum distance method is used to select the initial cluster center, and finally the cluster and cluster center are obtained. According to the similarity between the test data of the user's rating and user's personal information and the clustering center, they are divided into the cluster to which they belong, and the sample set in the cluster is used as the training set for K-nearest neighbor classification.**

Keywords: K-nearest neighbor algorithm; K-means algorithm; elbow method; maximum-minimum distance

I. INTRODUCTION

With the growth of big data generation across various fields, information overload is becoming a critical problem [1]. The recommendation system predicts the content that users are interested in by analyzing the historical behavior of customers[2].The key to the recommendation is to accurately obtain user preferences, which can help users find movies of interest and at the same time recommend good movies that are not easy to be found.

This paper combines improved K-means algorithm and K nearest neighbor algorithm to recommend movies of interest to users. The core idea of the K-nearest neighbor algorithm is to calculate the distance between the unlabeled sample and each sample in the data set on the test data and training data space, and decide based on the K nearest neighbors vote. However, the K-nearest neighbor classification calculation is huge, So consider using the K-means clustering algorithm to preprocess the data before K nearest neighbor classification. The K-means clustering algorithm has a great dependence on the K value and the selection of the initial clustering center [3]. By using the elbow method to select the K value, the running time and the number of runs can be reduced. The basic K-means algorithm is sensitive to the initial clustering centers. In the case of choosing inappropriate initial clustering centers, it may lead to undesirable clustering [4-5]. In this paper, the maximum-minimum distance method is applied to the K-means algorithm to select the initial clustering center. The improved K-means algorithm will solve the problem of the unsatisfactory clustering effect and improve the recommendation performance.

Literature[6] designed an improved mixed recommendation, the improved LeaderRank method is used to enhance the connectivity of the network, thereby increasing the convergence speed of the algorithm. Literature [7] proposed using canopy K-means clustering algorithm to cluster users, and then improve the recommendation algorithm by improving the user similarity calculation method. A new kernel k-means clustering method is proposed in reference [8]. The improvement of the recommendation algorithm and K-means algorithm has been studied, but none of them has made good progress.

II. RELATED BASIS

A. K-means clustering algorithm

Clustering is a process of classifying data into different classes and has become an important tool in data mining[9]. The selection of the initial clustering center will also significantly affect the clustering results [10]. The basic idea of the K-means algorithm is: randomly select K data points from the data set as the initial cluster center vector and set the initial K clusters as null values. Calculate the distance between each data point and each randomly selected initial cluster center vector. According to the principle of minimum distance, assign the data to the center vector closest to them, reduce the objective function value, and mark the assigned cluster center to Belonging to the cluster. Calculate the average of all data points in each cluster, and update the cluster average as the new cluster center. Repeat the operation until the cluster centers of the clusters do not change, obtain K clusters and cluster centers.

B. Determination of K value

In reference [11],the problem of manual input of K value in traditional K-means clustering algorithm, a method to obtain K value automatically is proposed. In this paper, the elbow method is used to estimate the number of clusters K of the data, and a K value and the initial cluster center of each category are set before clustering. When there is no specified K value in the K-means clustering algorithm, the optimal solution of K-means parameters is to minimize the cost function, the cost function is the sum of the degree of distortion of each class, the degree of distortion of each class is equal to the sum of the squared distances from each variable point to the center of its class, and the compactness of the members in the class is proportional to the degree of distortion of the class. The closer the members of the class

2473-3547/20/$31.00 ©2020 IEEE
DOI 10.1109/ISCID51228.2020.00076

are to each other, the smaller the degree of distortion, the value corresponding to the position where the improvement effect of the distortion degree decreases the most is the elbow.

The core of the elbow method is SSE (sum of the squared errors),

$$SSE = \sum_{i=1}^{k} \sum_{p \in C_i} \left| p - m_i \right|^2 \quad (1)$$

Among them, C_i is the ith cluster, p is the sample point in C_i, mi is the centroid of C_i (the mean of all samples in C_i), and SSE is the clustering error of all samples, representing the quality of the clustering effect. The more the number of clusters K, the more accurate the data division.

III. IMPROVED ALGORITHM

A. Improved K-means Algorithm

The traditional k-means clustering algorithm is a typical algorithm for solving clustering problems[12]. This paper proposes to use the maximum-minimum distance algorithm to improve the traditional K-means algorithm, which is a trial-based algorithm in the field of pattern recognition[13]. In the traditional K-means algorithm, all initial clustering centers are randomly selected, and the clustering effect is very unstable. However, the maximum-minimum distance algorithm only needs to select one initial cluster center randomly, and the remaining initial cluster centers are calculated according to the Euclidean distance criterion. Using this algorithm to select the initial cluster centers can reduce the number of clustering iterations. At the same time, it can avoid the clustering centers appearing close.

Improved K-means algorithm flow:

(1) Use the elbow method to determine the K value of the number of clusters, and set the initial K clusters to null values.

(2) Randomly select an initial cluster center Z_1 from the test data set $X=\{X_1, X_2...X_n\}$.

(3) Calculate the distance from each data point to Z_1, select the data point with the largest distance from Z_1 as the second initial cluster center Z_2.

$$\mathrm{d}_{ij} = \left\| X_i - Z_j \right\| (i = 1,2...n; j = 1,2...k) \quad (2)$$

(4) Calculate the distances from the remaining data points to Z1 and Z2, and find the minimum value of the distances among them (i ⩽ K).

$$d_i = \min[d_{i1}, d_{i2}](i = 1,2...n) \quad (3)$$

$$W = \theta \times \left\| Z_1 - Z_2 \right\| (\theta \text{ is the selected n scale factor}) \quad (4)$$

(5) Calculate the maximum distance value from the known minimum distance, and its corresponding data point is used as the ith (i ⩽ K) initial cluster center. When i> K or d_l<W, the initial cluster center selection is over, go to (6), otherwise repeat the above process.

$$d_l = \max\left[\min\left[d_{i1}, d_{i2} ... d_{ik}\right]\right] > W \quad (5)$$

(6) Calculate the distance from each data point $X_i(i=1,2...n)$ to each initial cluster center $Z_i(i=1,2...k)$, according to the principle of minimum distance, assign the data to The nearest cluster center to them.

(7) Mark the allocated Xi to the cluster z_i (i=1,2...k) to which it belongs. Calculate the average vector of all data points in each cluster, update the cluster average vector as the new cluster center, and repeat (5) and (6).

(8) After many iterations of calculation, until convergence, the final clustering result K clusters $Z=\{Z_1, Z_2, ..., Z_k\}$ and the cluster center $z=\{z_1, z_2... z_k\}$.

B. Improved K-means' K nearest neighbor algorithm

The K-nearest neighbor algorithm is very popular in data mining and recommendation systems [14-15]. However, the traditional K-nearest neighbor method needs to set an optimal K value for each test sample to perform K-nearest neighbor classification. When faced with a huge amount of data, the calculation is very time-consuming. This paper proposes that the improved K-means K-nearest neighbor algorithm can solve these problems of the K-nearest neighbor algorithm at the same time. The optimal K value has been selected when K-means clustering user rating data and the time cost for K-nearest neighbor classification is reduced.

After the improved K-means algorithm is used to cluster the training set of movie ratings, K clusters, and cluster centers of each cluster are obtained. Calculate the distance between the user to be classified and the cluster center of each cluster and the data in the cluster where the cluster center with the smallest distance is located is used as the training set of users waiting to be classified. In the new training set, calculate the distance between each user waiting to be classified and the training data. Find the K users closest to the users waiting to be classified, and take the category of the K users with the most categories as the category of the users waiting to be classified. Then recommend movies with higher ratings in this category to users. The specific steps to improve the algorithm are as follows:

(1) According to the improved K-means algorithm, the clustering results K clusters $Z=\{Z_1, Z_2 ...Z_k\}$ are obtained.

(2) Calculate the distance between the cluster center of each cluster and the user to be classified $u=\{u_1,u_2...u_n\}$ after clustering, according to the minimum distance principle, select the cluster where the cluster center with the smallest distance is located

$$\sum_{i=1}^{\mathrm{n}} \min_{j \in \{1,2,...,k\}} \left\| u_i - z_j \right\|^2 \quad (6)$$

(3) Take the data in the cluster as the new training set Y, and find the K nearest neighbor subsets Y_{u_k} closest to the user to be classified.

(4) According to Y_{u_k}, the calculation formula for calculating the membership degree of user u in each cluster is:

$$S_{u-Z_i} = \sum_{u_p \in Y_{u_k}} \sqrt{\sum_{i=1}^{n} \left(x_{ij} - x_{ip}\right)^2} \cdot \delta\left(u_p, Z_i\right) \quad (7)$$

Where $\delta\left(u_p, Z_i\right)=\begin{cases}1, d_p \in Z_i\\ 0, d_p \notin Z_i\end{cases}$ is the indicator function.

(5) Determine the category of u according to the classify u categories decision function of the user u to be classified:

$$C_u = \arg\max(S_u - z_i) \tag{8}$$

(6) Repeat the operation until all users to be classified have completed the classification.

C. Application of Improved Algorithm in Movie Recommendation

The whole system is divided into two parts: training and testing.

Training part: The K-means algorithm in this article is a user clustering algorithm based on the similarity of movie ratings. First, obtain the user's rating data for the movie, and randomly select a user rating of the movie that has been watched as the first initial clustering center. Then calculate the similarity between each user and the first initial cluster center user based on the Euclidean distance, and select the user with the largest distance from the first initial cluster center as the second initial cluster center. Repeat the operation and select the remaining users as the initial clustering center, and the user cluster is formed according to the minimum distance between the user rating data and the initial clustering center. The K-means algorithm needs to determine the K value in advance. This paper uses the elbow method to estimate the movie rating data and obtains the number of user clusters.

Testing part: When the test user enters the recommendation system, the target user's rating of the movie as a data point, calculate the distance to the cluster center of the cluster, divide the target user into the closest cluster, and use the users in the cluster as a new training set. In the new movie scoring training set, k nearest neighbor(KNN) users with high similarity are searched to form a set of nearest neighbor user set. According to the actual ratings of the K user's neighbors for the movies that have been watched, the target user's rating for the not watched movies is predicted, and the scores are sorted, and the list of movies recommended to users is sorted according to the ratings. The specific flow chart is as follows:

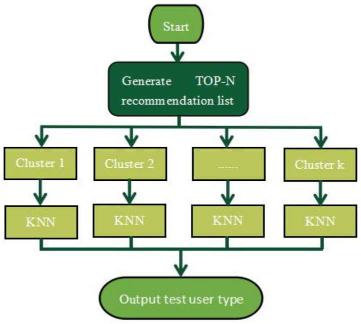
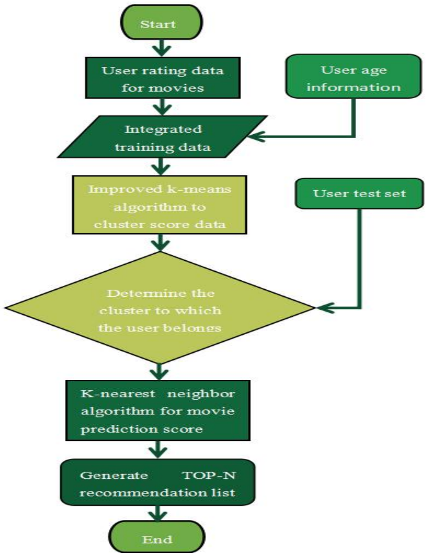


Figure 1. Improved K-means K-nearest neighbor algorithm flowchart



Start
User rating data for movies
User age information
Integrated training data
Improved k-means algorithm to cluster score data
User test set
Determine the cluster to which the user belongs
K-nearest neighbor algorithm for movie prediction score
Generate TOP-N recommendation list
End

Figure 2. System flow chart

IV. Experimental Results and Analysis

A. Experimental Environment and Data

The algorithm in this paper is implemented by python. The operating environment is Dual-Core Intel Core i5 CPU, the main frequency is 2.3GHz, memory is 8G, and the operating system is Macos64 bit system.

To verify the performance of the improved algorithm, we used the real Movie Lens data set. In the experiment, the data set is randomly divided into a testing set and training set according to the proportion of 20% and 80%. Determine user preferences through the user's movie ratings in the training set. According to the results of the training set, the movie recommendation is made to the users in the test set.

B. Model training and evaluation

In this paper, the film recommendation results are measured by three analysis indexes: accuracy, recall, F1 comprehensive evaluation accuracy, and the harmonic average of recall rate.

$$\text{Precision} = \frac{TP+TN}{TP+TN+FP+FN} \tag{9}$$

$$\text{Recall} = \frac{TP}{TP+FN} \tag{10}$$

$$F1 = \frac{2\times Precision\times \mathrm{Re}\,call}{Precision+\mathrm{Re}\,call} \tag{11}$$

TN means that no unnecessary movies are recommended, TP means movies that need to be actually recommended, FP refers to the movies that do not need to be actually recommended, FN refers to the movies that are not recommended, the accuracy rate indicates whether the user's favorite movie is predicted correctly,the recall rate represents the ratio of the movie correctly recommended to user u to the maximum number of hits of the recommended movie.

The experiment compares and analyzes the improved algorithm proposed in this article with CF-TP(Collaborative filtering algorithm based on time factor and user rating characteristics) and KNN in terms of accuracy, recall, and F1 value(Harmonic mean of comprehensive evaluation accuracy and recall rate).

TABLE I. ACCURACY

Recommended number of movies	*CF- TP*	*KNN*	*Improved KNN algorithm*
10	0.162	0.181	0.47
20	0.15	0.167	0.416
40	0.145	0.16	0.38
70	0.131	0.153	0.354
100	0.12	0.129	0.342

TABLE II. F1 METRICS

Recommended number of movies	*CF- TP*	*KNN*	*Improved KNN algorithm*
10	0.075	0.095	0.169
20	0.086	0.1	0.178
40	0.092	0.107	0.287
70	0.1	0.122	0.313
100	0.102	0.116	0.328

TABLE III. RECALL

Recommended number of movies	*CF- TP*	*KNN*	*Improved KNN algorithm*
10	0.049	0.064	0.103
20	0.06	0.072	0.116
40	0.067	0.08	0.23
70	0.081	0.102	0.28
100	0.089	0.105	0.315

From the experimental results of the data in the above three tables, the degree of optimization of the recommended results can be seen intuitively. Compared with the other two algorithms, the improved algorithm has significantly improved the accuracy of recommendation results. At the same time, the recall rate has also improved.

V. CONCLUSION

In this paper, we propose to use the elbow method to determine the K value and use the maximum and minimum distance method to optimize the K-means algorithm, taking into account the user's age information in the clustering data, it is finally applied to the K-nearest neighbor algorithm to recommend movies to the user. Experiment with the improved algorithm to evaluate the recommended performance of the algorithm.

REFERENCES

[1] J. Zhang, Y. Wang, Z. Yuan and Q. Jin, Personalized real-time movie recommendation system: Practical prototype and evaluation,Tsinghua Science and Technology, vol. 25, no. 2, pp. 180-191, April 2020, doi: 10.26599/TST.2018.9010118.

[2] X. Geng, J. P. Liu, Review of recommendation algorithms in data mining [J], Computer knowledge and technology, vol. 8, no. 19, pp:4691-4696, 2012.

[3] Shehroz S. Khan, Amir Ahmad, Cluster center initialization algorithm for K-means clustering, vol. 25, no. 11, pp. 1293-1302, 2004, doi:10.1016/j.patrec.2004.04.007.

[4] Kamran Rezaei, Hassan Rezaei, HFSMOOK-Means: An Improved K -Means Algorithm Using Hesitant Fuzzy Sets and Multi-objective Optimization[J], Arabian Journal for Science and Engineering, vol. 45, no. 8, PP 6241-6257, 2020 doi:10.1007/s13369-020-04620.

[5] B. J. Zhou, Y. Z. Tao, B. Ji, Optimization of initial clustering center based on K-means minimization of sum of squares of errors, Computer engineering and Application, vol. 54, no. 15, pp. 48-52, 2018.

[6] R. Lai, T. Wang and Y. Chen, Improved hybrid recommendation with user similarity for adult learners, The Journal of Engineering, vol. 2019, no. 11, pp. 8193-8197, 11 2019, doi: 10.1049/joe.2018.5353.

[7] J. Tao, J. Zhou and J. Gan, Recommendation Algorithm for Minority Cultural Resources Based on MapReduce, 2019 IEEE International Conference on Computer Science and Educational Informatization (CSEI), Kunming, China, 2019, pp:149-152, doi: 10.1109/CSEI47661.2019.8938905.

[8] Y. Zhang, J. Lu, F. Liu, et al., Does deep learning help topic extraction? A kernel k-means clustering method with word embedding, vol. 12, no. 4, pp. 1099-1117, 2018, doi:10.1016/j.joi.2018.09.004.

[9] Tingxuan Wang, Junyao Gao, An Improved K-Means Algorithm Based on Kurtosis Test, 2019 3rd International Conference on Artificial Intelligence, Automation and Control Technologies (AIACT), China, vol. 1267, no. 1, pp:25- 27,2019.

[10] R. H. Jiao, S. L. Liu, W. Wu, B. Y. Lin, Incremental kernel fuzzy c-means with optimizing cluster center initialization and delivery[J], vol. 45, no.8, PP 1273-1291, 2016.

[11] Y. Q. Xie, R. M. Fang, A K-means Clustering Algorithm for Automatically Obtaining K Value,2018 3rd International Conference on Electrical, Control and Automation Engineering (ECAE 2018), DEStech Publications, 2018, PP:135-139.

[12] Y. Z. Guo, An Improved Parallelization of K-means Algorithm based on HADOOP, Application of Computer Network and Information Technology, vol 1187, no 4, pp:1884-2021, May 8,2019, doi: 10.1088/1742-6596/1187/4/042029.

[13] J. Zhou, Z. Y. Xiong, Y. F. Zhang, Multi center clustering algorithm based on maximum minimum distance method[J], Computer application, vol. 2006, no. 6, pp:1425-1427.

[14] X. Y. Li, Optimal neighbor parameters in k-nearest neighbor collaborative filtering recommendation algorithm[J], Computer and digital engineering, vol. 46, no. 8, pp: 1525-1528+1619, 2018.

[15] S. C. Zhang, X. L. Li, X.F. Zhu, Efficient KNN Classification With Different Numbers of Nearest Neighbors, IEEE transactions on neural networks and learning systems, vol. 29, no. 5, MAY 2018.