Research on recommendation algorithm based on collaborative filtering of fusion model

Yichen Wang

Northeastern University, Shenyang, China

Corresponding author and e-mail: Yichen Wang, 1801676@stu.neu.edu.cn

Abstract. Based on the facts nowadays, recommendation systems are widely used, the focus of this research is how to effectively improve the accuracy and adaptability of such systems. The research proposes a collaborative filtering model algorithm based on the fusion model, which combines five classical algorithms: Singular value decomposition, Knn-Baseline, K-Means, Non-negativistic matrix factorization and SlopeOne algorithm. By combining the outputs of five models, and then carrying out regression and fusion, a collaborative filtering algorithm is obtained. The model can effectively integrate the advantages of the five models, and can carry out experimental validation on the datasets of Movielens-1M and Movielens-100k. The experimental results are more accurate than applying the five algorithms individually, and the regression model has good adaptability and predictability.

1. Introduction
With the rapid development of the Internet, information has also presented explosive growth. Users expect to get personalized content in e-commerce, entertainment and social media, but information overload makes it a requirement that cannot be met easily. To provide users personalized content in the age of the information explosion, a large number of researchers, scholars and related enterprises have carried out studies on this problem and recommendation technology has been rapidly developed.

Current recommendation algorithms mainly include collaborative filtering recommendation, deep-learning-based recommendation, mixed recommendation, etc. Among these, collaborative filtering recommendation is a widely used type of recommendation algorithm [1]. Its core idea is to recommend among users with similar interests. Collaborative filtering algorithms can be roughly divided into three categories. The first is memory-based collaborative filtering, which is used to recommend to the target user an item which he hasn’t interacted with but which a highly similar has interacted with, mainly by calculating the similarity between the users [2]. The second is item-based recommendation, which mainly calculates the similarity between the target item and items with which the user has interacted to determine whether the target item should be recommended to the user [3]. The last is model-based collaborative filtering, which restores known ratings through various models and predicts the unknown rating. Due to the rapid development of deep learning in recent years, model-based collaborative filtering algorithms have become one of the most effective algorithms.

Model fusion is one of the most effective methods to improve the accuracy of weak classification models. In this paper, model fusion is used to integrate singular value decomposition (SVD) [4], knn-baseline, K-Means [5], non-negative matrix factorization (NMF) [6] and SlopeOne algorithm [7] which are all widely used. The fusion method designed in this paper is regression fusion. Regression models
can enhance the matching ability of recommendation systems in different data sets. Although its accuracy is not as good as the classification model, it can increase the accuracy and adaptability of the model. In this paper, average fusion, linear fusion and non-linear fusion are also tried, and good improvement is achieved. Among them, gradient boosting decision tree (GBDT) fusion with non-linear fusion is 12% lower than the worst K-Means algorithm among the five algorithms tested, and it showed a good improvement.

2. Related work

2.1. Singular value decomposition

SVD is one of the most important matrix decomposition techniques in linear algebra. The algorithm also belongs to the improved version of matrix decomposition algorithms. The improvement is that the method of matrix decomposition becomes singular value decomposition, and then the next steps are basically the same, while the gradient descent method is used to fit continuously. In the recommendation system, the SVD-based matrix decomposition algorithm has consistently good properties and performance.

2.2. K-Means

The k-means algorithm is one of the most common clustering algorithms. The algorithm is widely used in large-scale data mining because of its simple idea, low difficulty and good clustering effect. The basic idea of the algorithm is:

First, the user needs to determine how many categories the objects to be classified will be divided into; that is, the size of the class k value. Then k objects are selected randomly from the object to be classified as the initial center of k clusters, the distance is calculated from each object to be classified to the center of each cluster, and the objects are divided into their nearest clusters. Based on the initial clustering results, the cluster center of each cluster is re-calculated and the objects are re-divided, and the process is iterated until the clustering results no longer change.

2.3. Collaborative filtering algorithm based on baseline

To assess whether a strategy is good or not, a baseline of comparison should be established to follow up on the improvement of the algorithm’s effectiveness. Here, we can simply model the recommendation algorithm as a baseline. We can suppose that our training data are: <user, item, rate> triad, where user is user id, item is item id (item can be a movie on the MovieLens, a book on the Amazon, or a keyword on the Baidu keyword tool); rating is user’s rating on item; the real rating of the user u on item i is recorded as r_ui; and the estimated rating of the baseline model is bu_i. Thus, the model is as follows:

\[ \hat{b}_{ui} = \mu + b_u + b_i \]  

Where \( \mu \) is the average of all known ratings, \( b_u \) is the deviation of the user’s rating from the average, and \( b_i \) is the deviation of the item rated relative to the average, which can reflect the popularity, for example, of the movie. \( \hat{b}_{ui} \) is an estimate of the rating of the user u on the item i for the baseline model.

2.4. Non-negativistic matrix factorization

NMF (Non-negativistic matrix factorization), for any given non-negative matrix V, can find a non-negative matrix W and a non-negative matrix H, which meets the condition \( V = WH \), thus decomposing a non-negative matrix into the product of two non-negative matrices. Therefore, this algorithm is also a kind of matrix decomposition algorithm. Each column of the matrix V represents an observation, and each row represents a feature; the matrix W is called a base matrix, and the matrix H is called a coefficient matrix or a weight matrix. By replacing the original matrix with the coefficient matrix H, we
can reduce the dimension of the original matrix and get the dimension-reducing matrix of the data feature, so as to reduce the storage space.

2.5. SlopeOne
The SlopeOne algorithm is a very simple collaborative filtering algorithm. The main ideas are as follows: If the user \( u \) rated the item \( j \), then he is going to rate the item \( i \). We only need to calculate the average difference of ratings rated by a group of people who rating the item \( i \) and the item \( j \) at the same time; then, we can calculate the rating of the user \( u \) on the item \( i \) based on the difference of ratings. Of course, there are many items \( j \). There may be fewer people who rate items, and there are more people who rate items. So it’s clear that the item rated by most people should have the largest percentage of the rating.

3. Design of the collaborative filtering algorithm based on the fusion model
In this paper, a collaborative filtering algorithm is designed, based on the fusion model. Model fusion is a good model optimization method at present. Model fusion involves training multiple models and then integrating them into one model according to a certain method. It is easy to understand, simple to get and effective. In this paper, five classical algorithms, SVD, knn-baseline, K-Means, NMF and SlopeOne, are trained and fused. The fusion model can extract multi-dimensional features and effectively increase the accuracy and adaptability of the model, as shown in Figure 1.

![Figure 1. Flowchart of first round training data.](image)

After the new input is obtained, the new input is fitted with the label of the train set again, and the fitting method can select different complex fitting models. To meet the needs of different scenes, this paper selects the average fusion model, the linear fusion model and the non-linear fusion model, which can effectively fit the second model and improve its precision, as shown in Figure 2.
The composite model can be represented as follows

$$\text{Label} = f([g_1(x), g_2(x), g_3(x), g_4(x), g_5(x)])$$  \hspace{1cm} (2)

The inputs of $g_1(x)$, $g_2(x)$, $g_3(x)$, $g_4(x)$, $g_5(x)$ are the inputs of the recommendation system. The results are the results of five different model-based recommendation systems. $[g_1(x), g_2(x), g_3(x), g_4(x), g_5(x)]$ labels a vector consisting of the five results. $f(x)$ is the fusion model, and finally ‘Label’ is the output of the composite model and is also the predictive value. As shown in Figure 3, an input feature goes through the fusion model and comes up with the final result.

**Figure 2.** Fusion model design.

**Figure 3.** Predicting process of user data.
3.1. Average fusion method
For the regression model, the average fusion method usually refers to the weighted average method; that is, the weighted average method is applied to the regression output results obtained by the five models to get new output, which is the value of the average fusion method. The advantages of the average fusion model are that it is fast and adapts to many scenes. When the errors of the five models are distributed positively or negatively, the average fusion method has high fusion efficiency, which can effectively reduce errors and improve fusion precision.

3.2. Linear fusion method
For the regression model, if the inputs of multiple weak regression models are not distributed positively or negatively, the average fusion method may not meet the needs of fusion, and more complex fusion methods are needed. The linear fusion method is a simple regression fusion. The most common is the ordinary least square method, commonly used in statistics. However, the ordinary least square method requires strong linear correlation between the data. For data without such strong linear correlation, the effect may be less. The linear regressions in this paper are performed by using ElasticNet regression, ridge regression and Lasso regression. Lasso technique is to add the L1 regular expression to the penalty function; ridge regression is to add the L2 regular expression; and ElasticNet regression uses L1 and L2 regular expressions at the same time, which is equivalent to a mixture of the first two methods. L1 regularization is the sum of the absolute values of each element in the weight vector, L2 regularization refers to the sum of squares of each element in the weight vector, and then calculates the square root.

3.3. Linear fusion method
In fact, among most data, linear correlation data are not frequent; most data may be more non-linearly correlated. In this paper, the decision tree series model is used for GBDT fusion. The decision tree can be used as either classification or regression. In this paper, the GBDT algorithm of the decision tree series is used for fusion. GBDT [8] is one of the best algorithms to fit a real distribution in the traditional machine learning algorithm, and GBDT is quite good. It can be used for classification or regression and it can be used to screen features.

4. Experiment and result analysis
4.1. Experimental settings
To test the effect of the model, two real data sets from different fields are used: data sets of MovieLens-1M and MovieLens-100k. The MovieLens data set is a ratings data set that has been collected and provided by GroupLens Research from the MovieLens website. The MovieLens-100K data set was released in April 1998, and the stable benchmark data set was from 100,000 ratings on 1,700 movies by 1,000 users. The MovieLens-1M movie data set, released in February 2003, is a 1-million data set from 6,000 users on 4,000 movies. The training set: verification set = 3:1. Because the training process is too long, this training all uses the server to carry on the training.

4.2. Experimental results
To ensure the randomness of the training data, cross-validation is not adopted in this training, and the training set is always selected randomly: It is trained under the test set = 3:1, the MovieLens-100k data set is trained 10 times for all models, and the best one is displayed, as shown in Table 1. To display conveniently, the comparison results of MAE and RMSE of the algorithms of the MovieLens-100k are all stated to three decimal places. The simulation results are shown in Figure 4.
As we can see from Table 1, the accuracy and RMSE of Avg-fusion, Lasso-fusion, Elastic-fusion and GBDT-fusion are all better than the five classical models, of which Avg-fusion is lower than Lasso-fusion, but the accuracy and RMSE of Lasso-fusion are worse than Elastic-fusion, and GBDT-fusion is the best in the fusion models.

MovieLens-1M data set trains all models 5 times, and the best one is shown, as shown in Table 2. To display conveniently, the comparison results of MAE and RMSE of the MovieLens-100k are all stated to three decimal places. The simulation results are shown in Figure 5.
Table 2. Experimental results of movielens-1M.

| Model       | MAE  | RMSE |
|-------------|------|------|
| K-Means     | 0.738| 0.928|
| SVD         | 0.681| 0.869|
| SlopeOne    | 0.711| 0.903|
| NMF         | 0.725| 0.917|
| KNNBaseline | 0.736| 0.891|
| Avg-fusion  | 0.665| 0.856|
| Lasso-fusion| 0.642| 0.844|
| Ridge-fusion| 0.639| 0.838|
| Elastic-fusion| 0.635| 0.832|
| GBDT-fusion | 0.618| 0.817|

Figure 5. Experimental results of movielens-1M.

Compared with Table 1 and Table 2, in Table 1, the best performance is KNN-Baseline with the lowest RMSE, 0.934, but in Table 2, the best performance is SVD. It shows that SVD has better adaptability and accuracy in a large data set, and is therefore suitable for large data sets. It shows that KNN-Baseline algorithm has better adaptability and accuracy in small data sets, and it is therefore suitable for small data sets. The accuracy and RMSE of Avg-fusion, Lasso-fusion, Elastic-fusion and GBDT-fusion are all better than the five classical models, with Avg-fusion lower than Lasso-fusion, which means that the inputs of the five models have a certain linear correlation, but the accuracy and RMSE of Lasso-fusion are worse than Elastic-fusion, which shows that the inputs and labels of the five models also have a strong non-linear correlation. GBDT-fusion is the best in the fusion models, which indicates that the non-linear correlation of the models is strong.
5. Conclusions
In this paper, a collaborative filtering algorithm based on the fusion model is designed, which combines SVD, knnbaseline, k-means, NMF algorithm and SlopeOne algorithm, experiments with average fusion, linear fusion and nonlinear fusion, and makes some improvements. The results show that GBDT fusion among the non-linear fusions is 12% better than the worst K-Means algorithm among the five algorithms. Moreover, the regression model able to match different data sets has good adaptability in prediction. In later stages, we can try to apply a classification model to the outputs of the five models and the corresponding rating labels, to match the appropriate data set, and improve the accuracy of the recommendation model from another angle.

References
[1] Koren Y, Bell R. Advances in collaborative filtering [M]// Recommender systems handbook. Springer, Boston, MA, 2015: 77-118.
[2] Resnick P, Iacovou N, Suchak M, et al. GroupLens: an open architecture for collaborative filtering of netnews [C]//Proceedings of the 1994 ACM conference on Computer supported cooperative work. ACM, 1994: 175-186.
[3] Sarwar B M, Karypis G, Konstan J A, et al. Item-based collaborative filtering recommendation algorithms [J]. Www, 2001, 1: 285-295.
[4] Zhang Xirui, Sang Maodong, Du Yigang, Lin Muqing, Zhu Lei. Application and Performance Analysis of Singular Value Decomposition Filter in Contrast-enhanced Ultrasound [J/OL]. Applied Acoustics: 1-10 [2020-11-08]. http://kns.cnki.net/kcms/detail/11.2121.O4.20200813.0928.002.html.
[5] Dong Jiashun, Wang Xingdong, Li Dianjie, Tang Bo, Li Zhen. A Visual Detection Method for Steel Tube Surface Defects Based on Improved K-means Algorithm [J]. Journal of Wuhan University of Science and Technology, 2020, 43 (06): 439-446.
[6] Wang Yinsong, Sun Tianshu. Actuator Failure Status Assessment Based on Two-class NMF Networks [J/ OL]. Journal of System Simulation: 1-10 [2020-11-08]. http://kns.cnki.net/kcms/detail/11.3092.V.20200831.1712.005.html.
[7] Shi Xiaoxi. Research on Mixed Recommendation Algorithm Based on Slope One Algorithm [D]. East China Normal University, 2020.
[8] Xu Yongrui, Zuo Fengkai, Zhu Xinshan, Li Shuoishi, Liu Hongrui, Sun Biao. Research on Load Forecasting of Improved GBDT Algorithm [J/ OL]. Proceedings of the CSU-EPSSA: 1-8 [2020-11-08]. https://doi.org/10.19635/j.cnki.cusu-epsa.000618.