Biomedical Research 2018; Special Issue: S377-S379 *ISSN 0970-938X*
www.biomedres.info

Toward an improved collaborative filtering algorithm for omics data adjusted by double factors.

Li-Ping Li*, Guang-Li Xu, Wen-Xia Ding

Department of Engineering, Hubei Communications Technical College, China University of Geoscience, PR China

Abstract

A comprehensive recommendation algorithm adjusted by double factor based on improved Particle Swarm Optimization (PSO) and K-means was proposed to further improve algorithm performance in the high throughput omics data filtering. It uses User Behavior Factor (UBF) to adjust similarity. Meanwhile, it also introduces Global Supplement Factor (GSF) to adjust parameters in the adjacent phase selection and supplement items. The experiment shows that the improved algorithm can achieve good efficiency and recommendation accuracy. The application of this algorithm in biomarker feature set filtering has also been evaluated in this study.

Keywords: Particle swarm optimization, User behavior factor, Global supplement factor.

Accepted on August 28, 2017

Introduction

Personalised medicine is the revolutionising contribution of advanced health care research. With the evolution of advanced techniques like artificial intelligence technologies, proteomic analysis etc., accurate disease prediction becomes possible. The "omics data sets" enabled the identification of exact biomarkers associated with specific phenotype expression. With the emergence of researches in the personalized medicine filed and promotion of machine learning techniques, more and more well-known commercial and social networking sites used the recommender systems [1]. It is proved that the users actually need systems that fully exploit features and effectively push useful information. Information overload caused by dump of large amount of content will greatly decrease information usage efficiency, so that users have no way to select valuable information [2]. Many statistical and research institutions have arrived at the conclusion that the usage of recommender system bring users benefit and enhance of attention, the growth rate of which is generally high [3]. In summary, the personalized recommendation of recommender system plays a significant role in guiding customers concern about goods and enhancing satisfactory degree [4]. The filtering method's helps to identify the more related data subset with informative features thus improving prediction accuracy [5].

Biomarker identification and selection tools enable great feature selection properties and increased discriminative power (A filter-based feature selection approach for identifying potential biomarkers for lung cancer). In order to improve performance of recommender system and recommendation quality for biomarker feature set filtering, a comprehensive recommendation algorithm combining UBF and GSF is proposed in the paper.

UBF and GSF

The paper proposes a similarity parameter to adjust deviation of similarity. If the user number of two different items generating behavior records is not high, even the number of co-scoring items is small, it is of valuable. In order to avoid wrong punishment caused by too less evaluation number of user, another threshold is set as the UBF.

If the user number of co-scoring on two goods is less than α and the user number of individual scoring larger than a certain amount, the following similarity parameter is used to modify the computed similarity.

$$sim(i,j) = \frac{(UBF \cdot \alpha + n \cdot (1 - UBF))}{\alpha} \cdot sim(i,j) \rightarrow (1)$$

where α is the threshold of co-scoring user number, which can be determined by size of whole dataset and average common user number. The UBF is the adjust proportion to determine the adjustment value.

In order not to affect similarity and ensure effectiveness of UBF, the value of UBF is set 0.9 in the paper after tests. Here, n is the number of co-scoring. When the value is less than α, the factor is used for adjustment.

In order to ensure several items with the largest similarity been introduced to the nearest selection scope, the paper introduces GSF. After selected K items from searching space, compute its average similarity x and obtain USF using the formula:

$$\begin{matrix} if\ (C_{\max} < C_i \& \lambda \le K/10) \\ then\ \lambda = \lambda + 1 \end{matrix} \rightarrow (2)$$

where $\lambda=0,\ i \in K$ is the condition to execute the formula; C_i is the absolute value of difference between some item and the

mean; C_{max} is the difference between item with largest similarity and absolute value.

Finally, λ deviated items are removed out and λ items whose similarity larger than C_{max}-$\overline{C}$ from searching space are added to final neighbors.

Algorithm Design

The paper uses improved PSO and K-means mixed algorithm for item clustering. As to user scoring matrix A(m, n), it contains scoring results of m users on n items. The planned clustering number is s and k particles are used. Finally, s clustering and centers can be obtained.

Step 1: Mark the set of n items in scoring matrix A(m, n) as $I=\{I_1,I_2,\ldots,I_n\}$, which is also the set of samples. Then, the population is initialized. Each particle is s × m dimensional vector. Randomly select s cluster centers where particle located and classify each sample to a cluster. Randomly initialize particle and set different velocity. After k times same processes, generate total k particles. The current position of the i^{th} particle is pPos(i) and the optimal individual position as pBest(i). The population optimal position is the best one of all individuals, marked as gBest.

Step 2: Calculate fitness values of the particles in turn.

Step 3: Traverse k particles. If the fitness of current position is better than the original one, replace the optimal position pBest(i) with the current position pPos(i).

Step 4: Traverse k particle. If the fitness of current position is better than that of whole population, replace the global optimal position gBest with the current position pPos(i).

Step 5: Update position and velocity according to formulas of PSO algorithm.

Step 6: Generate optimal extrapolate particle with the formula.

Step 7: Carry out small-scale mutation operation.

Step 8: Compute global fitness variance. If it is less than the threshold, determine the particle population arrive at convergence. Select several optimal particles to carry out local searching with K-means algorithm to bypass trap of early convergence. If the maximum value of traverse has not been achieved or the position is not good, return to Step 8.

In the nearest neighbors searching, the threshold can be adjusted constantly. After calculated nearest neighbors of the target item and obtained forecast score, compute the Mean Absolute Error (MAE) by comparing to score in the test set.

Experiment and Result Analysis

In the dataset selection, the MovieLens of GroupLens is used. The interesting of users can be reflected by different score on different movie. The user behavior is divided randomly in accordance with uniform distribution to generate several test sets and training sets. Where, the training set accounts for 80% and remaining for test.

In this experiment, the neighbor number also increase from 10 with interval 5. The MAE value is recorded as shown in Figure 1. In case of less neighbor number about 10 to 15, the value is close. With increase of neighbor number, the value of different level reduces at different degree. The MAE of UBF-GSF obtains better recommendation effect compared to item-based CF algorithm. When the neighbor number increases from 10 to 15, the previously close value now achieves some obvious advantages. In the subsequent that neighbor numbers continue increase, the difference of MAE between improved mixed algorithm and UBF-GSF also increases gradually. It can be concluded that the introduction of two factors significantly improve the recommendation effect.

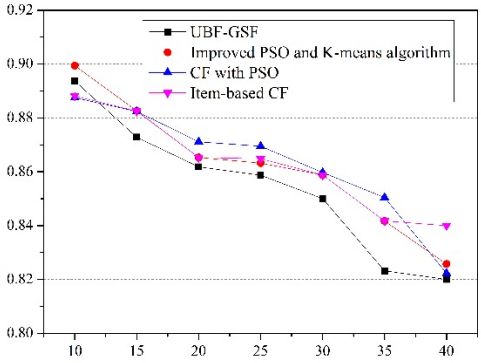

Figure 1. *MAE line chart.*

Conclusion

In this work, a novel CF algorithm integrates clustering and adjust factors is proposed. In the clustering phase, it uses improved PSO and K-means mixed algorithm. Then, UBF and GSF factors are introduced from improving algorithm performance. The experiment also proves its validity. However, the computation process of the algorithm is complex and computation is large. It also needs improvement in aspects of improving recommendation quality and accelerating recommendation.

References

1. Aditya PH, Budi I, Munajat Q. A comparative analysis of memory-based and model-based collaborative filtering on the implementation of recommender system for E-commerce in Indonesia: A case study PT X. International Conference on Advanced Computer Science and Information Systems (ICACSIS), Malang, 2016.
2. Bhajantri LB, Nalini N, Rathod SH. Collaborative filtering technique based recommendation in ubiquitous commerce. International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT), Davangere, 2015.
3. Yingtong D, Yang H, Deng X. A Survey of Collaborative Filtering Algorithms for Social Recommender Systems.

12th International Conference on Semantics, Knowledge and Grids (SKG), Beijing, 2016.

4. Hasan M, Ahmed S, Malik MAI, Ahmed S. A comprehensive approach towards user-based collaborative filtering recommender system. International Workshop on Computational Intelligence (IWCI), Dhaka, 2016.
5. Assawamakin A, Prueksaaroon S, Kulawonganunchai S, Shaw PJ, Varavithya V, Ruangrajitpakorn T, Tongsima S. Biomarker selection and classification of "-omics" data using a two-step bayes classification framework. Biomed Res Int 2013; 2013: 148014.

***Correspondence to**

Li-Ping Li

Department of Engineering

Hubei Communications Technical College

China University of Geoscience

PR China