计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.200600109

基于显式反馈协同过滤算法的偏好与共性平衡

黄超然　甘咏诗
香港浸会大学计算机科学院　香港 999077

摘　要　基于显式反馈的协同过滤算法只存在3个变量,其相似度计算方法依赖用户评分数据的显式反馈行为,而未考虑现实推荐场景中存在的隐性因素影响[5],这决定了协同过滤算法被限制于挖掘用户及商品的偏好,而缺乏挖掘用户和商品共性的能力。对此,学术界提出了不同的创新想法以改进传统协同过滤算法,但大多数的改进是基于协同过滤的垂直改进,如向算法加入分类、聚类、时间序列等机制,即对算法结构进行改进而不对变量因素进行改进,因此仍然无法深入挖掘用户和商品的共性因素。文中提出水平改进方法,即协同过滤与回归加权平均(Collaborative Filtering & Regression Weighted Average,CRW),旨在保留协同过滤对偏好的计算,并通过树回归算法计算且挖掘出用户和商品的共性因素,对协同过滤的预测结果和回归预测结果进行加权平均,以平衡协同过滤偏好性强而共性弱的问题。实验结果表明在适当的加权系数 a 下,CRW 预测结果均方误差相比于单一的协同过滤和回归的预测结果均方误差有明显的降低,表明 CRW 具有更高的推荐精度。

关键词:协同过滤;显式反馈;树回归;推荐系统;偏好与共性
中图法分类号　TP391.3

Balance Between Preference and Universality Based on Explicit Feedback Collaborative Filtering

HUANG Chao-ran and GAN Yong-shi
Department of Computer Science, Hong Kong Baptist University, Hong Kong 999077, China

Abstract　Collaborative filtering (CF) based on explicit feedback only exists three variables, and its similarity computing method depends on the explicit feedback of user's rating data, but never considers the implicit factors existed in the real word's recommendation, which determines that CF is limited in mining the preference of users and items, but it lacks of the abilities of mining the universality of users and items[5]. Academia has proposed various of innovative ideas to improve the traditional CF, but most of improvements are vertical improvements for CF algorithm like adding the mechanisms of classification, clustering and time series to the algorithm, which improve the algorithm structure but barely improve the variable factors. Therefore, it still cannot mining the universality of users and items deeply. This paper proposes a horizontal improvement: Collaborative Filtering & Regression Weighted Average (CRW), intending to mine the universality of users and items through tree regression while keeping the preference of users and items through CF, and conducting weighted average between the predicting result of regression and CF, in order to balance the strength of preference and the weakness of universality of CF. Experiment result shows that with a proper weighting coefficient a, the mean square error of predicting result of CRW is distinctly lower than that of CF and regression, which shows CRW performs better than single CF and regression.

Keywords　Collaborative filtering, Explicit feedback, Tree regression, Recommended system, Preference and universality

1　基于显式反馈的协同过滤和树回归

1.1　基于用户以及商品的协同过滤

基于用户的协同过滤算法是通过用户的历史行为数据发现用户对商品的喜好,并对这些喜好进行度量和打分。算法根据不同用户对相同商品或内容的态度和偏好程度计算用户之间的关系。该算法在预测某一用户对某个商品的喜好程度时,首先根据所有用户的历史评价数据,利用相关系数算法如皮尔森相关系数计算用户之间的相关性,根据与该用户最相似的几个用户对该商品的评价[1],来预测该用户对该商品的评价。

基于商品的协同过滤预先根据所有用户的历史偏好数据计算物品之间的相似性,把与用户喜欢物品的相类似物品推荐给用户。用户 x 对商品 i 的预测评分的数学表达式为:

$$\hat{r}_{xi}=\frac{\sum_{j\in N(i;x)} s_{ij}\cdot r_{xj}}{\sum_{j\in N(i;x)} s_{ij}}$$

其中,s_{ij} 为商品 i 和 j 的相似度,r_{xj} 为用户 x 对商品 j 的评分,$N(i;x)$ 为用户 x 评价的一组与商品 i 相似的商品。

1.2　基于模型的协同过滤

由于随着用户和商品的数目增多,用户-商品-评分矩阵(简称评分矩阵)将非常庞大,且评分矩阵为稀疏矩阵,导致在实际场景中运用协同过滤的计算量通常非常巨大,因此学者

本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
通信作者:黄超然(triumph_huang@foxmail.com)

引入了基于分解的方法。主流的基于分解的方法将评分矩阵分解为2个矩阵，即用户相关矩阵和商品相关矩阵，这里被称为隐因子模型(Latent Factor Model)，通常利用梯度下降法或交替最小二乘法ALS(Alternating Least Squares)求出用户矩阵和商品矩阵的最优解，最终预测评分[3]。用户 x 对商品 i 的预测评分的数学表达式为：

$$\hat{r}_{xi}=Q_i \cdot P_x^{\mathrm{T}}$$

其中，Q_i 为商品矩阵中商品 i 的因子，P_x^{T} 为用户矩阵中用户 x 的因子。

1.3 基于显式反馈协同过滤的不足和改进

显式反馈指用户明确表示对物品喜好的行为，如用户对商品的评分、对商品表示喜欢或不喜欢等。算法通过观察到的所有用户给产品的打分，来推断每个用户的喜好并向用户推荐适合的商品。根据上文数学表达式得知，无论是基于用户的协同过滤、基于商品的协同过滤，还是基于模型的协同过滤，协同过滤算法的变量只有3个，分别为用户、商品、评分，这一条件决定了协同过滤能够很好地探测出用户对商品的偏好，但是，这一条件也限制了协同过滤挖掘用户和商品共性的能力。协同过滤更多的是探测某一特定用户与某一特定商品之间的相互偏好(如图1左侧)。

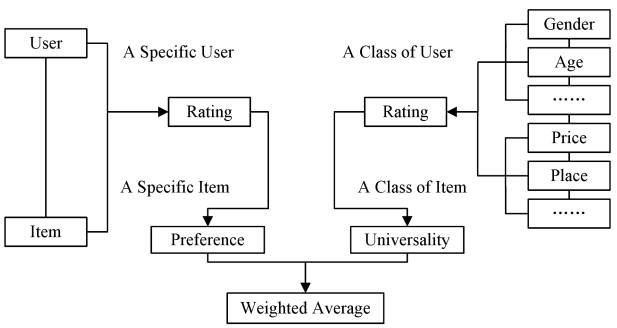

图1 加权平均偏好与共性

Fig. 1 Weighted average preference and universality

本文认为共性因素对预测的影响非常大，认为引入用户及商品的共性因子(如图1右侧)是必要的。本文引入共性因子的方法为，通过树模型回归算法来计算用户及商品的共性，最后以加权平均方法结合协同过滤的偏好得分和回归的共性得分(如图1)，这也是本文的核心思想——协同过滤与回归加权平均(Collaborative Filtering & Regression Weighted Average，CRW)。

1.4 树回归

本文认为，协同过滤算法的性质决定了其针对的是某一特定用户对某一特定商品的评价，而忽略了某一类人对某一类商品的普遍评价，因此本文认为挖掘用户和商品的共性特征，即把用户各个特征变量以及商品各个特征变量作为独立变量，利用回归算法进行共性的探测，能够弥补协同过滤缺乏共性的不足。由于输入变量为复杂的多变量数据，且非独立变量不为连续变量，因此本文引入树模型回归算法。树模型回归算法能够对多变量和非线性的数据进行建模，且对多变量非线性数据的拟合程度高[10]。本文引入决策回归树、随机森林和梯度提升回归树3种算法分别与协同过滤进行加权平均，以此进行实验测试，对比CRW与协同过滤和回归算法的均方误差。

2 CRW模型实验

2.1 实验目的与内容

对上文针对单一协同过滤算法无法探测用户和商品共性的问题，本文提出CRW以改进现有基于协同过滤的推荐算法。实验的目的为对比CRW与协同过滤和回归算法的均方误差。具体分为3步：1)分别建立协同过滤和回归模型；2)对两者的预测结果进行加权平均并最优化系数；3)对比CRW与两种算法的均方误差。

2.2 数据集

MovieLens数据集是美国GroupLens实验室从MovieLens网站采集的用户对电影评分的数据。用户对电影的评分区间为[0,5]，分值越大表示用户越喜欢这部电影，数据还包含用户特征以及电影特征。本文采用943名用户对1 682部电影的100 000个数据评分，采用3个用户特征(包括用户年龄、收入、职业)以及20个电影特征(包括电影发行年份，是否为动作类、冒险类等(哑变量))。

2.3 训练-测试集的划分

本文认为，训练集经模型训练之后，已达到较为拟合的水平，因此加权平均发生在训练集没有任何意义。加权平均应发生在测试集，以将协同过滤算法和回归算法共同拟合于测试集。因此本文摒弃传统训练-测试集的划分，把数据集划分为训练集、测试1集、测试2集。首先将数据集按8∶2的比例分为总训练集和测试2集，再将总训练集8∶2的比例划分为训练集和测试1集。利用训练集训练得到协同过滤模型和回归模型，通过测试1集求出最优系数 a 得到CRW模型，最后以测试2集测试CRW模型的拟合效果。

2.4 CRW模型

在得到协同过滤和回归模型之后，CRW模型的预测公式为：

$$\hat{r}_w=a\hat{r}_c+(1-a)\hat{r}_R$$

其中，a 为加权平均系数，$\hat{r}_c$ 为协同过滤预测结果，$\hat{r}_R$ 为回归预测结果。其中 a 的定义域为[0,1]，a 的最优解由迭代法求出。a 和$(1-a)$在CRW模型里代表偏好比重和共性比重，a 越大，偏好在模型中越重要，共性在模型中就越不重要，反之亦然。

2.5 评价指标

本文采用均方误差(Mean Square Error)作为实验的测试标准，表示为：

$$MSE=\frac{1}{n}\sum_{i=1}^{n}[f(x_i)-y_i]^2$$

其中，n 为样本总数，$f(x_i)$ 为算法对样本i的预测值，y_i 为样本i的实际值。MSE 值越小，表示算法预测的误差越小，推荐精度越高。

2.6 结果分析

本文采用基于ALS的协同过滤和3种回归算法(决策树回归DT、随机森林回归Random Forest、梯度提升回归树GBRT)进行实验。经过训练后得到协同过滤和回归模型，之后利用测试1集进行对 a 求解，在求解过程中发现 a 与 MSE 的关系曲线呈下凸形态(见图2)，这意味 MSE 最低时 a 不在0或者1的位置上而是在(0,1)区间内，意味着协同过滤预测结果和回归预测结果的加权平均的预测精度大于单一协同过

滤预测精度和单一回归预测精度。

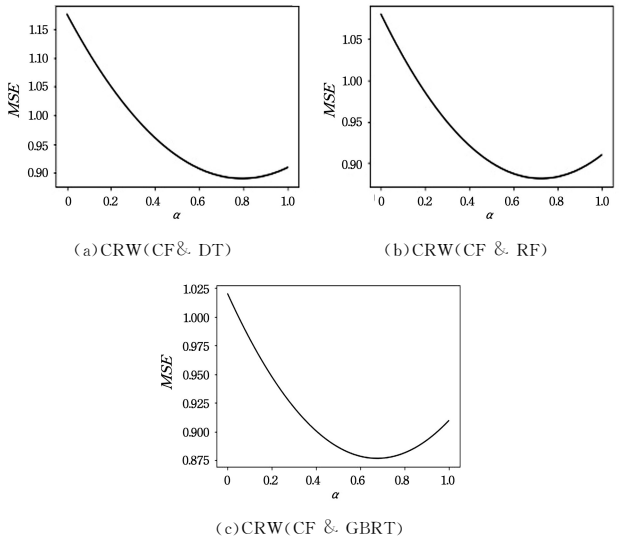

图 2 测试 1 集中 CRW 加权系数 a 与 *MSE* 的关系

Fig. 2 Relation between CRW weighted coefficient a and *MSE* in test set 1

由迭代法计算出 CRW(协同过滤，决策树回归)最的优 a 系数为 0.79，CRW(协同过滤，随机森林回归)的最优 a 系数为 0.72，CRW(协同过滤，梯度提升回归树)的最优 a 系数为 0.68。在 CRW 运用不同的回归算法，得出的 a 系数即偏好比重是不同的。最终经过测试 2 集评估得到 CRW(协同过滤，梯度提升回归树)的 *MSE* 最小，为 0.862(见表 1)，比单一的协同过滤算法预测的均方误差低了 3.274%，有效地提高了预测精度。

表 1 测试 2 集中 CRW 与单一协同过滤和回归算法的 *MSE* 对比

Table 1 Comparison of *MSE* between CRW and single CF and regression in test set 2

Algorithm	CF	Regression	CRW	Compare CRW and CF/%
CRW(CF,DT)	0.892	1.137	0.872	↓2.224
CRW(CF,RF)	0.892	1.064	0.865	↓2.952
CRW(CF,GBRT)	0.892	1.006	0.862	↓3.274
Mean	0.892	1.069	0.866	↓2.817

由表 1 得出，无论利用 3 种回归算法中的任意一种算法与协同过滤进行加权平均(使用最优系数 a)，其预测结果的误差都小于单一的协同过滤算法或者单一的回归算法，证明了 CRW 能够有效提升推荐预测的精度。

结束语 实验证明 CRW 相对于单一的协同过滤和单一的回归算法，其均方误差有明显的降低，意味着 CRW 较单一的协同过滤或回归有明显的精度提升。但由于需要同时建立两个模型，其时间复杂度也有所增加。优化算法结构，降低时间复杂度，将成为 CRW 未来的研究方向。

参 考 文 献

[1] JI Y M，LIKE，LIU S D，et al. Collaborative filtering recommendation algorithm based on interactive data classification[J/OL]. The Journal of China Universities of Posts and Telecommunications.[2020-06-16]. https://doi. org/10. 19682/j. cnki. 1005-8885. 2020. 0024.

[2] HU Y T，XIONG F，LU D Y，et al. Movie collaborative filtering with multiplex implicit feedbacks[J]. Neurocomputing，2020，398：485-494.

[3] WANG Z，JIAN P L，YONG L Z. Muhammad Hammad Memon. Bayesian pairwise learning to rank via one-class collaborative filtering[J]. Neurocomputing，2019，367：176-187.

[4] HAPER F M，KONSTAN J A. TheMovieLens Datasets：History and Context[J]. ACM Transactions on Interactive Intelligent System(TiiS)，2006，5(4)：1-19.

[5] DONG L Y，XIU G Y，MA J Q. Collaborative filtering algorithm based on weight adjustment and user preference [J]. Journal of Jilin University (Science Edition)，2020，58(3)：599-604.

[6] ZHANG W，CUI Y B，LI J，et al. ollaborative filtering recommendation based on clustering matrix approximation [J]. Operations Research and Management，2020，29(4)：171-178.

[7] DONG Y H，ZHU C Y. Collaborative filtering algorithm based on improved user attribute score [J]. Computer Engineering and Design，2020，41(2)：425-431.

[8] LU H，SHI Z B，LIU Z B. Collaborative filtering recommendation algorithm integrating user interest and rating difference [J]. Computer Engineering and Application，2020，56(7)：24-29.

[9] ZHAO W T，LU X. Collaborative filtering algorithm based on user characteristics and similar confidence [J]. Measurement and Control Technology，2019，38(8)：95-98，102.

[10] WU S. Research and Application of Recommendation Technology based on regression tree Model [D]. Nanjing：Nanjing University，2018.

HUANG Chao-ran，born in 1995，postgraduate. His main research interests include data mining and social network analysis.