基于最小方差的 *K*-means 用户聚类推荐算法

杨大鑫,王荣波,黄孝喜,谌志群
(杭州电子科技大学 计算机学院,浙江 杭州 310018)

摘 要:协同过滤推荐算法是一种传统的推荐技术,具有简单高效的特点,在实际中有广泛的应用,获得了大量研究者的青睐。虽然传统的协同过滤推荐算法在一定程度上缓解了用户当前所面临的信息超载问题,但其在处理大数据时存在的数据稀疏性和扩展性等问题却日益突出。于是,提出了一种基于最小方差的 *K* -means 用户聚类推荐算法。在缓解数据稀疏性方面,利用 Weighted Slope One 算法对初始用户—项目评分矩阵进行有效填充,降低了数据稀疏性;在提高算法扩展性方面,采用基于最小方差的 *K* -means 算法对用户评分数据进行聚类,将相似的用户聚到一起,减小目标用户的最近邻搜索空间,提高了算法扩展性。通过在 MovieLens 数据集上的对比实验,结果表明,相比于传统的协同过滤推荐算法,改进算法具有更高的推荐准确度。

关键词:信息过载;协同过滤算法;Weighted Slope One;最小方差; *K* -means 聚类
中图分类号:TP391 **文献标识码**:A **文章编号**:1673-629X(2018)01-0104-04
doi:10.3969/j.issn.1673-629X.2018.01.022

K-means User Clustering Recommendation Algorithm Based on Minimum Variance

YANG Da-xin, WANG Rong-bo, HUANG Xiao-xi, CHEN Zhi-qun
(School of Computer, Hangzhou Dianzi University, Hangzhou 310018, China)

Abstract: Collaborative filtering recommendation algorithm is a kind of traditional recommendation technology which is so simple and efficient with a wide range of applications that it has been favorite by a large number of researchers. Although the traditional collaborative filtering recommendation algorithm has alleviated the information overload faced by users to a certain extent, the data sparsity and expansibility in dealing with large data is becoming more and more prominent. For this, a *K* -means user clustering recommendation algorithm based on minimum variance is proposed. The Weighted Slope One algorithm is used to fill the initial user-item scoring matrix effectively, and the data sparsity is reduced. Then *K* -means algorithm based on minimum variance is adopted to carry out the user rating data clustering, with similar users clustered together to reduce the target user's nearest neighbor search space and improve its expansibility. The contrast experiments on MovieLens datasets show that the proposed algorithm has higher recommendation accuracy than the conventional collaborative filtering recommendation algorithm.

Key words: information overload; collaborative filtering algorithm; Weighted Slope One; minimum variance; *K* -means clustering

0 引 言

随着互联网和电子商务的飞速发展,在大量的商品信息面前,用户或消费者往往很难发现最需要或最合适的商品。在信息爆炸的时代,如何解决信息超载的问题,受到了越来越多的关注,并提出了多种推荐算法,主要包括基于规则的推荐算法[1]、基于模型的推荐算法[2]和协同过滤推荐算法[3]等。虽然协同过滤推荐算法在现实中有很多的应用[4-8],但是它在处理大数据时会存在稀疏性和扩展性等问题,导致推荐效果不理想。对研究人员有很大的挑战。其中,文献[9]提出了一种通过矩阵聚类的协作过滤算法,利用矩阵聚类算法对用户评分数据进行聚类,然后对聚类后的子矩阵进行协作过滤,提高了算法的推荐精度。文献[10]利用 BP 神经网络来预测用户对项目的评分,减

收稿日期:2017-01-05 **修回日期**:2017-05-10 **网络出版时间**:2017-09-27
基金项目:国家自然科学基金青年项目(61202281);教育部人文社会科学研究青年基金项目(10YJCZH052)
作者简介:杨大鑫(1992-),男,硕士研究生,研究方向为智能信息处理、推荐系统;王荣波,博士,副教授,硕士生导师,研究方向为自然语言处理、中文信息处理;黄孝喜,博士,硕士生导师,研究方向为自然语言处理、人工智能;谌志群,硕士,副教授,硕士生导师,研究方向为大规模网络文本的智能化处理、复杂网络及其应用。

小了数据集的稀疏性,提高了推荐系统的推荐质量。但是训练 BP 神经网络模型需要额外的时间花销。文献[11]利用奇异值对高维数据矩阵进行降维,使得评分矩阵变得密集,以此来减小数据矩阵的稀疏性。但是奇异值分解之后会导致数据丢失,在高维数据矩阵中,降维效果不是很理想。文献[12]对用户和项目两个维度进行联合聚类,通过对聚类信息的平滑处理和对用户未评分项目的预测,在数据稀疏的问题上进行改进,从而提高了推荐的质量。

针对传统的协同过滤推荐算法在处理大数据时存在的稀疏性、扩展性等问题,文中提出了一种基于最小方差的 K -means 用户聚类推荐算法。首先利用 Weighted Slope One 算法对数据矩阵中的未评分项进行预测,减小其稀疏性;然后通过基于最小方差的 K -means 算法对填充后的评分数据进行聚类,减少用户最近邻搜索空间,提高算法的扩展性;最后在目标用户所在的类中利用传统的基于用户的协同过滤进行推荐,生成最终的推荐结果。并通过与其他算法的对比验证该算法。

1　相关背景知识

1.1　用户相似性计算

基于用户的协同过滤推荐算法通过用户对项目的历史评分来度量用户之间的相似度。其中,计算用户相似度的方法有如下几种:皮尔逊(Pearson)相关系数法、余弦向量法和修正的余弦向量法等。文中采用余弦向量法来产生用户最近邻居,公式如下:

$$\mathrm{sim}(u,v)=\cos(\vec{u},\vec{v})=\frac{\vec{u}*\vec{v}}{\|\vec{u}\|*\|\vec{v}\|} \tag{1}$$

其中,$\vec{u}$ 和 $\vec{v}$ 分别为用户 u 和用户 v 对各个项目的评分;$\|\vec{u}\|$ 和 $\|\vec{v}\|$ 分别为用户 u 和用户 v 对项目评分向量的模。

1.2　项目评分预测

利用式(1)的相似性计算可以得到目标用户 u 的最近邻居。设目标用户 u 的最近邻居的集合为 N_u,则可以从目标用户 u 对最近邻居集合项目的评分中得到其对项目 i 的预测评分,公式如下:

$$P_{u,i}=\overline{R_u}+\frac{\sum_{n\in N_u}\sum_{n\in N_u}\mathrm{sim}(u,n)\times(R_{n,i}-\overline{R_n})}{\sum_{n\in N_u}|\mathrm{sim}(u,n)|} \tag{2}$$

其中,$\overline{R_u}$ 为目标用户 u 对已评分项的评分均值;$\overline{R_n}$ 为邻居 n 对已评分项的评分均值;$R_{n,i}$ 为邻居 n 对项目 i 的评分;$\mathrm{sim}(u,n)$ 为目标用户 u 和邻居 n 之间的相似度。

1.3　基于 Weighted Slope One 算法降低数据稀疏性

初始数据矩阵中的 0 元素经过 Weighted Slope One 算法处理后可以降低其稀疏性。因为在推荐系统的应用过程中,用户对项目的评分数通常会大于 2,所以为了平衡各个评分项目对于目标项目的影响,需要用到 Weighted Slope One 算法[13],它是 Slope One 算法[14]的一个递进算法。

首先,定义一个数据矩阵 $\boldsymbol{R}$,$R_{u,i}$ 为用户 u 对项目 Item_i 的评分;I_u 为用户 u 所有评分的项目;U_i 为对项目 Item_i 进行评分的用户集合;$U_{ij}=U_i\cap U_j$ 为对 Item_i 和 Item_j 两个项目都评过分的用户集合;Num_{ij} 为集合中元素的数目。

项目 Item_i 和 Item_j 之间评分的偏差可以由式(3)得到:

$$\mathrm{Dis}_{ij}=\frac{\sum_{u\in U_{ij}}(R_{ui}-R_{uj})}{\mathrm{Num}_{ij}} \tag{3}$$

最后,得到目标用户 user 对项目 Item_i 的预测评分:

$$\mathrm{Prediction}_{\mathrm{user},i}=\frac{\sum_{j\in I_{\mathrm{user}}}\mathrm{Num}_{ij}(\mathrm{Dis}_{ij}+R_{\mathrm{user}j})}{\sum_{j\in I_{\mathrm{user}}}\mathrm{Num}_{ij}} \tag{4}$$

1.4　*K*-means 算法

设待聚类的数据集为 $X=\{x_i\,|\,x_i\in R^P,i=1,2,\cdots,n\}$。$K$ 个聚类中心分别为 $M_1,M_2,\cdots,M_k$,于是用 $w_j(j=1,2,\cdots,k)$ 表示聚类的 k 个类别。

定义 1:两个数据对象间的欧氏距离为:

$$d(x_i,x_j)=\sqrt{(x_i-x_j)^{\mathrm{T}}(x_i-x_j)} \tag{5}$$

定义 2:同类别中数据对象的算术平均为:

$$M_j=\frac{1}{N_j}\sum_{x\in w_j}X \tag{6}$$

定义 3:聚类准则函数为:

$$E=\sum_{i=1}^{k}\sum_{j=1}^{n_j}d(X_j,Z_i) \tag{7}$$

在传统 K -means 算法中,初始簇心的选择是随机的,通过相似度计算,把数据集中的数据样本与最近的初始中心归为一类,不断重复这一过程,直到各个初始中心在某个精度范围内不变。具体算法步骤如下:

(1)在包含 N 个样本的 X 中随机选择 k 个样本数据作为初始簇心 $M_i(i=1,2,\cdots,k)$;

(2)利用式(5),计算 X 中各个样本数据 p 到 M_i 的距离 $d(p,M_i)$;

(3)找到每个样本数据 p 的最小的 $d(p,M_i)$,把 p 加入到与 M_i 相同的簇中;

(4)完成所有样本的遍历之后,通过式(6)重新计算 M_i 的值,作为新的簇心;

(5)重复步骤 2 ~ 4,直到目标准则函数 E 取值不再变化。

2 基于最小方差优化 K-means 初始簇心

在众多聚类算法中, K -means 算法十分典型,虽然实现起来简单方便,但也有一些弊端。首先, K 值是根据人的经验随机确定的,具有一定的盲目性,如果不了解要聚类的数据,那么给出合理的 K 值就会非常困难;其次,对初始簇心的选择也是随机的,不同的簇心会导致不同的聚类效果,如果选择了孤立点,则聚类结果会有很大的差异。

针对 K -means 算法存在的缺陷,许多研究者对其进行了优化[15-17]。文中则对初始聚类中心基于最小方差进行优化[18],在不同范围内选择 K 个方差最小的样本作为初始簇心。根据方差的定义,一个样本的方差越小,它附近的数据分布就会越密集,使得簇心的选取就会越客观,聚类结果越准确。具体选取步骤如下:

定义 4:样本 x_i 到所有样本距离的平均值为:

$$m_i = \frac{1}{n}\sum_{j=1}^{n} d(x_i, x_j) \tag{8}$$

定义 5:样本 x_i 的方差为:

$$\mathrm{var}_i = \frac{1}{n-1}\sum (d(x_i, x_j) - m_i)^2 \tag{9}$$

定义 6:数据集样本的平均距离为:

$$\mathrm{cmean} = \frac{2}{n(n+1)}\sum_{i=1}^{n}\sum_{j=1}^{i} d(x_i, x_j) \tag{10}$$

(1)利用式(5)、(8)、(9)计算出数据集 X 中各个样本的方差,然后在 X 中找到方差最小的样本 x_i',把它当作第一个簇心 M_i',添加到集合 C 中;

(2)利用式(10)计算出 X 中各个样本之间距离的平均值 cmean;

(3)在以 cmean 为半径的圆之外寻找另一个方差最小的样本,把它作为第二个簇心,添加到集合 C 中;

(4)重复上一步,在剩余的样本中不断寻找,找到 K 个簇心之后,算法结束。

通过上述步骤选取到的中心点紧密度极高,可以较好地反映样本的分布情况,具有一定的客观性,聚类结果更加稳定、可靠。

3 基于最小方差的 K -means 用户聚类推荐算法

该算法中每条数据主要由用户、项目、评分 3 部分组成。设目标用户为 u ,用户集合为 $U = \{u_1, u_2, \cdots, u_m\}$,基于最小方差优化后的 K -means 算法生成的用户集合表示为 $U = \{C_1, C_2, \cdots, C_k\}$ 。其中, k 为生成的簇类个数, C_k 表示第 k 个簇类。基于最小方差的 K -means 用户聚类推荐算法的描述步骤如下:

输入:用户-项目数据矩阵 $\boldsymbol{R}_{m\times n}$,聚类个数 k ;

输出: N 个推荐项目。

Step1:利用式(4)消除 $\boldsymbol{R}_{m\times n}$ 中的 0 元素,得到矩阵 $\boldsymbol{R}'_{m\times n}$;

Step2:以集合 C 中元素 $M_i'(i = 1, 2, \cdots, k)$ 为初始簇心,把 $\boldsymbol{R}'_{m\times n}$ 中的数据通过 K -means 算法分成 k 类;

Step3:利用式(5)计算 u 与 k 个簇心之间的相似度,然后把 u 加入到与其最相似的类中;

Step4:利用式(1)计算 u 与同类中其他用户的相似性,得到其最近邻居集 $N_{uj}(j = 1, 2, \cdots, m)$;

Step5:得到最近邻居集之后,可以根据它们对项目的评分,利用式(2)求得 u 对待推荐项目的预测分,从高到低排序之后,把前 N 个项目推荐给 u 。

4 实 验

4.1 数据集与评测标准

采用 MovieLens 数据集对算法的性能进行测试。数据集中包括 943 个用户对 1 682 部电影的 10 万多条评分,评分为 1 ~ 5 之间的整数。评分值越大说明用户对这部电影越喜欢,0 表示用户没有对该电影进行评分。根据实验需要,将其中的 80% 作为训练集,其余 20% 作为测试集。

平均绝对偏差(Mean Absolute Error, MAE)是一种常用的推荐质量度量方法,通过计算预测评分与实际真实评分之间的偏差来度量预测的准确性。MAE 越小,推荐精度越高。MAE 的计算公式如下:

$$\mathrm{MAE} = \frac{\sum_{i=1}^{N} |p_i - q_i|}{N} \tag{11}$$

其中, p_i 表示用户预测的评分; q_i 表示对应的实际评分。

4.2 实验结果与分析

实验首先对初始数据中的 0 元素进行有效消除,然后采用基于最小方差的 K -means 算法对处理后的数据进行聚类,最后在目标用户所在类中用传统的协同过滤算法输出最终的推荐结果。通过实验对传统的协同过滤算法、简单 K -means 用户聚类推荐算法和文中算法进行了对比。设定用户聚类数为 14,最近邻居的个数从 5 增加到 50,间隔为 5。实验结果如图 1 所示。

由图 1 可知,三种算法的 MAE 值都会随着目标用户最近邻个数的增加而降低,说明推荐的准确率可以随着目标用户的最近邻居数的增加而得到有效提高。而文中算法在不同邻居个数下的 MAE 值都是最低的,由此可见,与传统协同推荐算法和简单用户聚类推荐

算法相比,文中算法具有更好的推荐效果。

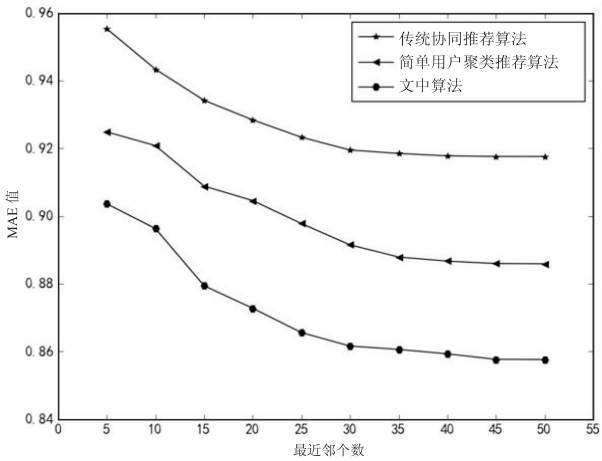

图 1　实验结果对比

在进行传统的推荐之前,文中算法由于对初始数据进行了填充,并且对用户数据基于最小方差进行了聚类,使得用户最近邻搜索空间更具客观性,选取到的最近邻居更加准确。实验结果表明,相比于传统的协同过滤推荐算法,文中算法的准确度更高。

5　结束语

从数据稀疏性和算法扩展性两方面进行改进,文中提出了一种基于最小方差的 K-means 用户聚类推荐算法。一方面,利用 Weighted Slope One 算法对初始数据矩阵进行填充,降低其稀疏性;另一方面,采用基于最小方差的 K-means 算法对填充后的数据进行聚类,提高算法的扩展性。实验结果表明,文中算法在一定程度上改善了数据稀疏性和算法扩展性,提高了算法的推荐质量。

参考文献:

[1] 陈华月,余　刚,朱征宇. 基于加权关联规则的用户关注项目推荐算法[J]. 计算机工程,2006,32(6):86-88.

[2] 伍杰华,朱岸青,蔡雪莲,等. 基于隐朴素贝叶斯模型的社会关系推荐[J]. 计算机应用研究,2014,31(5):1381-1384.

[3] SCHAFER J B,DAN F,JON H,et al. Collaborative filtering recommender systems[C]//The adaptive web:methods and strategies of web personalization,lecture notes in computer science. Berlin:Springer-Verlag,2007:291-324.

[4] 游　文,叶水生. 电子商务推荐系统中的协同过滤推荐[J]. 计算机技术与发展,2006,16(9):70-72.

[5] 曹一鸣. 基于协同过滤的个性化新闻推荐系统的研究与实现[D]. 北京:北京邮电大学,2013.

[6] KONSTAN J,MILLER B,MALTZ D,et al. GroupLens:applying collaborative filtering to usenet news[J]. Communications of the ACM,1997,40(3):77-87.

[7] PARK S T,PENNOCK D M. Applying collaborative filtering techniques to movie search for better ranking and browsing[C]//Proceedings of the 13th ACM SIGKDD international conference on knowledge discovery and data mining. New York:ACM,2007:550-559.

[8] 孟　佩,曹　菡,师　军. 基于 Softmax 回归模型的协同过滤算法研究与应用[J]. 计算机技术与发展,2016,26(12):153-155.

[9] 高凤荣,邢春晓,杜小勇,等. 基于矩阵聚类的协作过滤算法[J]. 华中科技大学学报:自然科学版,2005,33:257-260.

[10] 张　锋,常会友. 使用 BP 神经网络缓解协同过滤推荐算法的稀疏性问题[J]. 计算机研究与发展,2006,43(4):667-672.

[11] VOZALIS M G,MARGARITIS K G. Applying SVD on item-based filtering[C]//Proceedings of 5th international conference on intelligent systems design and applications. [s. l.]:IEEE,2005:464-469.

[12] 韦素云,肖静静,业　宁. 基于联合聚类平滑的协同过滤算法[J]. 计算机研究与发展,2013,50(S):163-169.

[13] 郑　丹,王名扬,陈广胜. 基于 Weighted-slope One 的用户聚类推荐算法研究[J]. 计算机技术与发展,2016,26(4):51-55.

[14] LEMIRE D,MACLACHLAN A. Slope one predictors for online rating-based collaborative filtering[C]//SIAM data mining. California:SIAM,2005:21-23.

[15] ERISOGLU M,CALIS N,SAKALLIOGLU S. A new algorithm for initial cluster centers in k-means algorithm[J]. Pattern Recognition Letters,2011,32(14):1701-1705.

[16] 汪　中,刘贵全,陈恩红. 一种优化初始中心点的 K-means 算法[J]. 模式识别与人工智能,2009,22(2):299-304.

[17] 杨善林,李永森,胡笑旋,等. K-means 算法中的 k 值优化问题研究[J]. 系统工程理论与实践,2006,26(2):97-101.

[18] 谢娟英,王艳娥. 最小方差优化初始聚类中心的 K-means 算法[J]. 计算机工程,2014,40(8):205-211.