

基于自编码机和聚类的混合推荐算法

赵 阳^{1,2}, 朱全银^{1,2}, 胡荣林², 瞿学新²

(1 南京工业大学 计算机科学与技术学院, 江苏 南京 211800;

2 淮阴工学院 计算机与软件工程系, 江苏 淮安 223003)

摘 要: 针对传统的协同过滤推荐算法在稀疏数据集上表现不佳的情况, 提出了一种将自编码器(Autoencoder)和聚类结合的混合推荐算法. 首先将用户项目评分数据和用户人口统计学数据作为自编码器的输入, 提取用户特征. 然后利用提取到的用户特征对用户聚类得到用户类别, 从而使近邻搜索范围减小. 接着通过计算平均绝对误差(MAE)寻找到适用于同一类别用户的推荐算法, 最后组合各类别上适用的推荐算法, 得到混合推荐模型. 推荐阶段, 计算目标用户类别, 并使用混合推荐模型得到推荐结果. 实验结果表明, 该算法可以有效的提高推荐质量.

关键词: 混合推荐; 协同过滤; 自编码器; 聚类; 平均绝对误差

中图分类号: TP391.9

文献标识码: A

文章编号: 1000-7180(2018)11-0052-05

Hybrid Recommendation Algorithm Based on Autoencoder and Clustering

ZHAO Yang^{1,2}, ZHU Quan-Yin^{1,2}, HU Rong-Lin², QU Xue-Xin²

(1 School of Computer Science and Technology, Nanjing Tech University, Nanjing 211800, China;

2 Faculty of Computer and Software Engineering, Huaiyin Institute of Technology, Huaian 223001, China)

Abstract: With the problem that traditional collaborative filtering recommendation algorithm does not recommend high quality on sparse data sets, a hybrid recommendation algorithm that based on autoencoder and clustering is proposed. First, the user's project rating data and user demographic data as the input data from the autoencoder to extract user characteristics. Then use the extracted user characteristics to cluster the users to obtain the user categories, so that the search range of the neighbor decreases. Then search for a recommendation algorithm for the same type of users by calculating mean absolute error (MAE), and get a hybrid recommendation model. Finally calculate the target user category and use the hybrid recommendation model to get the recommended results. Experimental results show that this algorithm can effectively improve the quality of recommendation.

Key words: hybrid recommender; collaborative filtering; autoencoder; clustering; mean absolute error

1 引言

随着信息技术不断地发展, 互联网上的信息发生了指数型的增长, 用户无法快速的在网上寻找到自己想要的信息, 这就是“信息过载”问题. 学术界和工业界都在不断地探索提高信息服务质量解决“信息过载”问题的方法, 由此诞生了个性化推荐技

术^[1]. 精确的个性化推荐服务不仅能够促进用户消费, 而且可以提升用户体验. 在所有个性化推荐技术中应用最成功使用最广泛的就是协同过滤推荐算法. 它的基本思想就是与目标用户相似的用户群体对商品或货物的评分也是与目标用户相似的. 协同过滤推荐算法首先通过相似度算法寻找与目标用户相似的用户, 然后根据相似用户对商品的评分计算

收稿日期: 2018-03-01; **修回日期:** 2018-03-15

基金项目: 江苏省科技厅 2015 年重点研发计划项目(BE2015127); 江苏高校自然科学基金(15KJB52004); 江苏省公共服务平台建设项目(16SYC-114); 淮安市重点研发计划(HAG201604)

目标用户对商品的评分,最后选择评分排名最高的前 N 项商品推荐给目标用户.

但是商品种类和用户人数的不断增加导致的用户评分矩阵稀疏问题,对推荐算法的准确性和时效性都造成了很大的影响. 为了提高推荐质量,国内外专家为了解决上述问题做出了许多尝试. 例如文献[2]根据用户历史行为数据预测每个用户对项目的偏好,通过融合评分和用户偏好提高了推荐准确性. 文献[3]提出了一种基于 SVD 与模糊聚类的协同过滤推荐算法,通过缩小搜索邻居范围和改进 Pearson 相关系数来提高推荐质量. 但是这两种算法都忽略了用户的人口统计学信息对用户相似度造成的影响. 文献[4-6]都结合了聚类技术改进协同过滤推荐算法. 但是并没有考虑到聚类后各类别用户的差异.

针对以上问题,本文先将用户评分数据和用户人口统计学数据结合,再使用自编码器提取用户特征,最后使用提取到的用户特征对用户聚类,从而使近邻搜索范围减小. 为了进一步提高推荐质量,本文采用了混合推荐策略,针对不同类别的用户使用不同的推荐算法. 实验结果表明,本算法有效的克服了数据稀疏性的影响,提高了推荐质量.

2 基本理论

2.1 自编码器

自编码器 (Autoencoder) 是一种无监督学习算法,它由三层网络组成,其中输入层与输出层神经元数量相等,隐藏层神经元数量少于输入输出层. 自编码器将输入的特征经过非线性变化到隐藏层上,通过这种方法,可以降低特征的维度,提高了计算效率^[7]. 自编码器结构如图 1 所示.

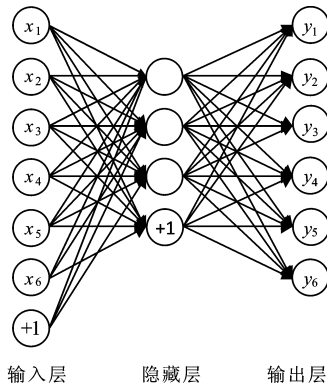


图 1 自编码器结构图

自编码器的训练分为编码和解码两个步骤. 其

中编码部分,数据 x 在输入层通过激活函数映射到隐藏层 h ,激活函数如式(1)所示:

$$h = f(x) = \frac{1}{1 + \exp(-(\omega_1 x + b_1))} \tag{1}$$

式中, ω_1 为编码部分的权值, b_1 为偏置单元.

解码部分为隐藏层 h 还原输入数据,还原得到的数据用 y 表示,那么解码公式可以如式(2)所示:

$$y = g(h) = \frac{1}{1 + \exp(-(\omega_2 h + b_2))} \tag{2}$$

式中, ω_2 为解码部分的权值, b_2 为偏置单元. 自编码机的目标函数如式(3)所示:

$$L = \frac{1}{2n} \|x - y\|^2 \tag{3}$$

式中, n 为输入的样本个数.

2.2 K-means++ 聚类算法

K-means 算法是最主流的聚类算法之一,然而它的初始中心是随机选择的^[8]. 随机选择初始中心可能会造成聚类的结果和数据的实际分布相差巨大. K-means++ 算法改进了初始中心的选择,使算法可以获得全局最优解. 由于 K-means++ 聚类算法的准确性高和速度快的特点,本文选择 K-means++ 聚类算法对用户进行聚类. K-means++ 选择初始中心点:

- (1) 从数据点中随机选择一个中心;
- (2) 对于每个数据点 x ,计算 x 和最近的中心之间的距离 $d(x)$;
- (3) 随机选择一个新的数据点作为一个新的中心,使用一个加权概率分布,其中一个点 x 被选择成与 $d(x)^2$ 的概率成比例;
- (4) 重复步骤(2)和(3),直到选定 k 个中心;
- (5) 选择了初始中心后,使用标准的 K-means 算法进行聚类.

2.3 协同过滤推荐算法

协同过滤算法是通过其他用户评分计算得到目标用户的推荐列表. 协同过滤算法首先使用相似度度量方法搜索与目标用户最相似的 K 个用户 (K 可以根据需求设定),然后根据搜索到的用户对目标用户未评分的项目评分,最后产生推荐列表^[9].

协同过滤推荐算法中,定义用户数据集为 $U = \{User_1, User_2, \dots, User_n\}$,项目数据集 $I = \{Item_1, Item_2, \dots, Item_m\}$,那么用户评分数据可以用一个 $n \times m$ 的矩阵表示,如表 1 所示.

表 1 用户评分矩阵

	Item ₁	...	Item _k	...	Item _m
User ₁	$r_{1,1}$	...	$r_{1,k}$	...	$r_{1,m}$
...	...	...	...	...	...
User _j	$r_{j,1}$	...	$r_{j,k}$	...	$r_{j,m}$
...	...	...	...	...	...
User _n	$r_{n,1}$	...	$r_{n,k}$	...	$r_{n,m}$

3 基于自编码机和聚类的混合推荐算法

基于自编码机和聚类的混合推荐算法可以结合多种推荐算法的优点,提升推荐性能. 本文选取两种典型的推荐方法: KNNBaseLine 算法和基于 SVD 的推荐算法. 首先使用自编码器对训练集中的用户特征矩阵降维, 其中用户特征矩阵包含用户对项目的评分和用户自身的年龄、性别、职业等人口统计学信息; 接着使用 K-means++ 算法对用户聚类, 得到 K 个类别的用户; 最后通过计算 MAE 值确认同一类别用户适用的推荐算法, 得到混合推荐模型.

3.1 相似度计算

皮尔逊相关系数的概念是来自于统计学中的线性回归模型, 它是衡量两组变量之间线性比密程度的无量纲指标, 它的取值范围为 $[-1, 1]$, 其中 -1 指的是负相关, 1 指的是正相关, 0 指的是不相关. 一般情况下, 我们首先要对数据进行处理, 将负相关和正相关结合, 这时它的取值范围为 $[0, 1]$, 当它的值越大时代表相关性越强^[10]. 当使用皮尔逊相关系数计算相似度时, 用户 u 和用户 v 的相似度计算公式 $\text{sim}(u, v)$ 如公式(4)所示:

$$\text{sim}(u, v) = \frac{\sum_{i \in I_{uv}} (R_{ui} - \bar{R}_u)(R_{vi} - \bar{R}_v)}{\sqrt{\sum_{i \in I_{uw}} (R_{ui} - \bar{R}_u)^2} \sqrt{\sum_{i \in I_{uv}} (R_{vi} - \bar{R}_v)^2}} \quad (4)$$

3.2 特征提取

在用户的人口统计学信息中年龄、性别、职业、收入等特征并不总是连续值, 有可能是一些离散值. 例如性别可分为“男”和“女”. 在构建用户特征矩阵时, 我们需要把这类特征数字化. 本文使用 one-hot 编码处理离散的用户特征, 其方法时使用 m 个二元特征来表示一个有 m 种可能值的离散特征, 例如用 $\{10\}$ 表示男性用户, $\{01\}$ 表示女性用户.

为了减小因为数据量级不同对用户特征提取造成的影响, 需要对用户数据进行标准化处理. 本文使用 z-score 算法对用户数据标准化处理, z-score 如式(5)所示.

$$x^* = \frac{x - \mu}{\sigma} \quad (5)$$

式中, μ 为所有用户数据均值; σ 为所有用户数据标准差. 使用自编码器提取用户特征的主要步骤如下:

输入: 用户评分数据、用户人口统计学数据、隐藏层维数

输出: 用户特征

(1) 将用户的评分数据和人口统计学数据结合, 以向量形式作为自编码器输入, 同时设定自编码器隐藏层维数;

(2) 利用公式(1)计算输入数据通过隐藏层到输出层的映射;

(3) 利用公式(3)计算输出数据与输入数据的偏差;

(4) 通过后向传播对网络参数进行调整;

(5) 重复步骤(2)(3)(4)直至目标函数收敛或者到达设定的迭代次数;

(6) 迭代停止时, 隐藏层的数据既为用户的特征编码.

3.3 训练模型

首先使用 K-means++ 算法对经过处理后的数据聚类, 得到 K 个类别的用户. 然后对每类别用户中使用 KNNBaseline 推荐算法测试每个用户类别的推荐效果. KNNBaseline 推荐算法的推荐方法如式(6)所示:

$$\hat{r}_{ui} = b_{ui} + \frac{\sum_{v \in N_i^k(u)} \text{sim}(u, v) * (r_{vi} - b_{vi})}{\sum_{v \in N_i^k(u)} \text{sim}(u, v)} \quad (6)$$

式中, b_{ui} 是 baseline 值, 计算公式如式(7) 所示:

$$b_{ui} = u + b_u + b_i \quad (7)$$

式(6) 中 u 是平均得分, b_u 是用户的偏置, b_i 是项目的偏置. 接着使用基于 SVD 的推荐算法. 基于 SVD 的推荐算法预估值如式(8) 所示:

$$\hat{r}_{ui} = u + b_u + b_i + q_i^T p_u \quad (8)$$

式中, 目标函数如公式(9) 所示:

$$\sum_{r_{ui} \in R_{\text{train}}} (r_{ui} - \hat{r}_{ui}) + \lambda(b_i^2 + b_u^2 + \|q_i\|^2 + \|p_u\|^2) \quad (9)$$

接着计算每个类别用户使用 KNNBaseline 算法和基于 SVD 推荐算法的 MAE 值, 得到最终混合推荐模型. 训练混合推荐模型给出推荐的主要步骤如下:

输入: 用户特征、聚类个数、待混合的推荐算法、待推荐的用户数据

输出:混合推荐模型、推荐结果

(1)使用 K-means++算法对用户特征聚类;

(2)利用公式(6)测试每个类别用户在 KNN-Baseline 算法上的表现,并计算 MAE 值;

(3)利用公式(8)测试每个类别用户在基于 SVD 的推荐算法上的表现,并计算 MAE 值;

(4)将各个类别表现最好的算法组合形成混合推荐模型;

(5)当推荐模型训练完成后,有推荐需求的用户首先被分类,然后使用该类别最适合的推荐算法,最后得到推荐列表.

4 实验结果及分析

4.1 实验数据和环境

实验数据集使用的是美国 Grouplens 项目组提供的 MovieLens 1M 数据.数据集包括 6 040 个用户对 3 952 部电影的 1 000 209 条评分数据.评分范围从 1 到 5,其中每个用户至少对 20 部以上的电影进行了评分.数据集包含了 3 个训练集(users. dat, ratings. dat, movies. dat),其中 users. dat 中包含了特征提取需要的用户人口统计学信息数据.实验计算机配置:CPU 为 Intel 的 Core I7,内存 16 G,操作系统为 Ubuntu16. 04,程序由 Python 语言实现.

4.2 评价标准

平均绝对误差指的是对目标用户来说,目标项目的预测评分和目标用户真实评分之间的偏差程度.当计算得到的 MAE 值越小,证明推荐质量性越高^[11].假设 $\{p_1, p_2, \dots, p_n\}$ 是待推荐目标的评分集合, $\{q_1, q_2, \dots, q_n\}$ 是真实项目的评分集合,那么 MAE 的定义如公式(10)所示:

$$MAE = \frac{\sum_{i=1}^n |p_i - q_i|}{n}$$

(10)

4.3 实验结果及分析

在试验中,为了更好的测量推荐算法的准确性,本文采用 5-折交叉验证法^[12].将每组数据分为 5 个子集,每个子集都将做一次测试集,其余时间作为训练集,交叉验证重复 5 次,并将 5 次测试的平均交叉验证值作为最终结果.

实验对比了基于自编码器与聚类的 KNNBase-Line 算法和传统的 KNNBaseLine 算法,实验结果如图 2 所示.

从图 2 可以看出,当 k 值较小时基于自编码器与聚类的 KNNBaseLine 算法的效果要明显优于传统的

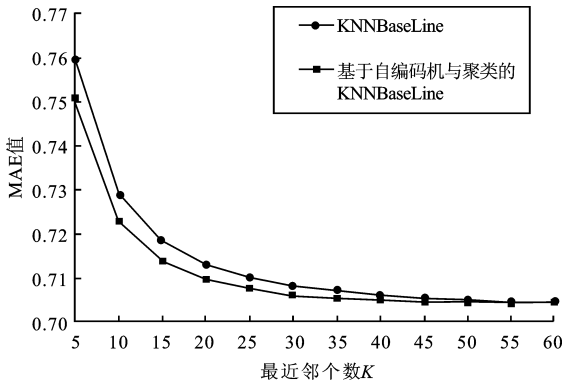


图 2 不同 K 值与 MAE 的关系

KNNBaseLine 算法.随着 k 值的不断增大两个算法的推荐效果均有提高但 MAE 值也趋于平缓.可以证明,将用户评分和用户人口统计学数据结合再通过自编码器提取用户特征能够提高推荐质量.

接下来,对比的是 KNNBaseLine 算法、基于自编码器与聚类的 KNNBaseLine 算法、基于 SVD 的推荐算法和基于自编码器与聚类的混合推荐算法.其中图 3 显示的是基于 SVD 的推荐算法在不同分解参数 k 下的性能,图 4 显示的是对比实验.

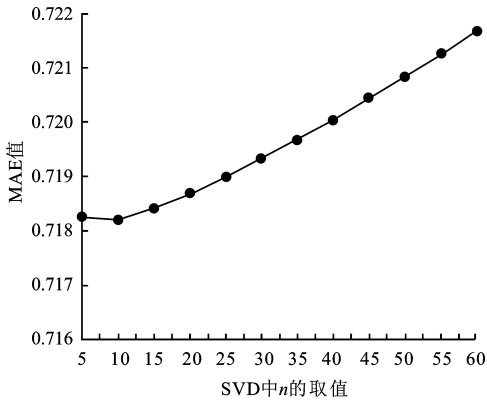


图 3 不同分解系数 n 与 MAE 的关系

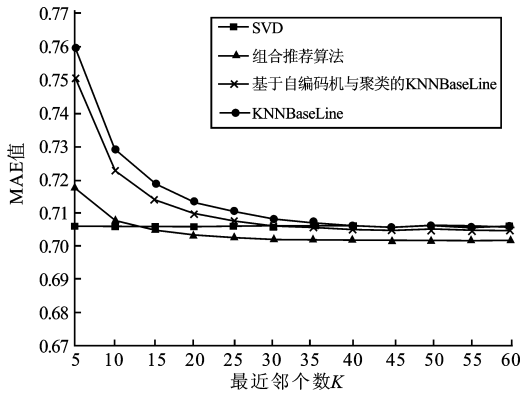


图 4 不同 K 值与 MAE 的关系

从图 3 可以看出,当 n 值取 10 时基于 SVD 的推荐算法效果最佳,当 n 值不断增加时推荐的效果越来越差.从图 4 可以看出,当最近邻个数 $k < 15$ 时,混合推荐算法的 MAE 值仅次于基于 SVD 的推荐算法.当 $k > 15$ 后,混合推荐算法性能明显优于其它三种算法.

5 结束语

针对传统的协同过滤推荐算法因为数据稀疏,导致推荐质量不高的问题.本文提出了一种基于自编码机和聚类的混合推荐算法.该算法首先利用自编码器提取用户特征,然后对用户聚类,最后针对不同类别的用户选择不同的推荐算法.实验结果表明,本文提出的混合推荐算法能够有效的提高推荐质量.接下来的研究将围绕以下两个方面:研究聚类个数对混合推荐造成的影响;寻找更好的算法混合策略.

参考文献:

- [1] Resnick P, Iacovou N, Suchak M, et al. Group Lens: an open architecture for collaborative filtering of net-news[C]// ACM Conference on Computer Supported Cooperative Work. Chapel Hill, NC, USA, ACM, 1994:175-186.
- [2] 王志虎,黄曼莹.基于用户历史行为的协同过滤推荐算法[J].微电子学与计算机,2017,34(5):132-136.
- [3] 林建辉,严宣辉,黄波.基于 SVD 与模糊聚类的协同过滤推荐算法[J].计算机系统应用,2016,25(11):156-163.
- [4] 朱东郡,李敬兆,谭大禹,等.基于标签聚类和兴趣划分的协同过滤推荐算法[J].计算机工程,2017,43(11):146-151.
- [5] Hu R, Dou W, Liu J. ClubCF: a clustering-based collaborative filtering approach for big data application [J]. Emerging Topics in Computing IEEE Transactions on, 2014, 2(3):302-313.

- [6] Wen J, Zhou W. An improved item-based collaborative filtering algorithm based on clustering method[J]. Journal of Computational Information Systems, 2012, 8(2):571-578.
- [7] Li X, She J. Collaborative variational autoencoder for recommender systems[C]// The ACM SIGKDD International Conference. Nova Scotia, Canada, ACM, 2017:305-314.
- [8] David Arthur, Sergei Vassilvitskii. K-Means++: the advantages of careful seeding[J]. Society for Industrial and Applied Mathematics, 2007, 11(6):1027-1035.
- [9] Breese John S, Heckerman, David Kadie Carl. Empirical analysis of predictive algorithms for collaborative filtering [J]. Uncertainty in Artificial Intelligence, 2013, 98 (7): 43-52.
- [10] Shang M S, Lü L, Zeng W, et al. Relevance is more significant than correlation: Information filtering on sparse data[J]. Europhysics Letters, 2010, 88 (6): 68008-68012.
- [11] 朱郁筱,吕琳媛.推荐系统评价指标综述[J].电子科技大学学报,2012,41(2):163-175.
- [12] Shang S, Hui Y, Hui P, et al. Privacy preserving recommendation system based on groups[J]. Computer Science, 2013(129): 407-409.

作者简介:

赵阳 男,(1992-),硕士研究生.研究方向为数据挖掘、推荐系统.

朱全银(通讯作者) 男,(1966-),教授,硕士生导师.研究方向为智能信息处理、接口与通信、数据挖掘等.

E-mail:hyitzqy@126.com.

胡荣林 男,(1978-),博士,副教授,硕士生导师.研究方向为物联网应用技术.

瞿学新 男,(1992-),硕士研究生.研究方向为数据挖掘、机器学习.