DOI:10.16707/j.cnki.fjpc.2016.01.001

# 一种神经网络与用户偏好相结合的协同过滤推荐算法

周高云，张雅云，陈冬隐，程红举
(福州大学数学与计算机科学学院　福建　福州　350108)

【摘　要】随着互联网服务和产品的大量增加，高效、可靠的推荐系统变得越来越重要。协同过滤是应用最广泛的推荐算法之一。传统的基于用户的协同过滤仅考虑共同评分的物品来计算用户的相似度，容易忽略用户之间的相关性。本文采用一种基于神经网络和用户偏好的协同过滤方法，首先通过聚类算法解决稀疏性问题，其次通过神经网络与用户偏好相结合来学习用户之间的相关性，最后利用训练后的多层神经网络来计算用户对物品的评分情况，来有效提高推荐效果。本文采用 EachMovie 数据来进行验证，并和现有的一些方法进行比较。实验结果表明，所推荐的协同过滤推荐算法在准确率、召回率、F1 三个指标上均优于传统的方法，推荐效果比传统的方法好。

【关键词】推荐算法；神经网络；用户偏好；聚类；协同过滤

## 0 引言

在当今信息爆炸的时代，如何快速获取用户所需要的信息变得越来越重要。推荐系统的出现能够很好的解决这样一个问题，推荐系统能够向用户提供符合他们偏好的物品。为了提高用户的满意度，一个可靠、高效的推荐系统变得越来越重要。协同过滤(基于领域的算法)是应用最广泛的推荐算法之一，它主要是利用行为的相似度计算兴趣的相似度[1]。协同过滤主要有两大类，一类是找到与目标用户兴趣相似的用户集合，根据相似用户集来向用户推荐用户没有听说过的物品，称为基于用户(user-based)的协同过滤(UBCF)[2]；另一类是计算物品之间的相似度，根据物品的相似度和用户的历史行为为用户进行推荐，称为基于物品(item-based)的协同过滤(IBCF)[3]。

协同过滤有两种传统的处理方法：k-NN 方法和关联规则方法[4-5]。虽然这两种方法能在一定程度上为用户推荐感兴趣的物品，但是他们的推荐效果不是很理想，性能不高，其主要原因是受到它们本身模型限制。k-NN 方法假设数据属性之间相对独立，但通常情况下并非如此；相关规则只能有限的表示数据之间复杂关系[6-7]。

聚类[8]算法可以减少复杂操作所需要处理的数据规模，降低数据稀疏性。用户偏好能够很好的反应物品的流行度。神经网络[9]能够有效的学习用户之间的复杂关系，隐藏节点能够提高性能，同时神经网络也很容易添加各种信息作为输入层特征向量的一部分来进一步提高模型的学习性能。传统协同过滤方法由于存在稀疏问题或只能有限的表达数据之间关系，导致推荐效果不好。针对传统方法存在的不足，本文采用一种用户偏好与神经网络模型(P-MLP)相结合的方法，该方法首先对用户聚类，降低数据稀疏性问题，其次利用神经网络与用户偏好相结合来有效学习用户之间复杂关系。实验结果表明，本文方法在准确率、召回率、F1 三个指标上均优于传统的方法，推荐效果比传统的方法好。

## 1 问题描述

对于一个给定的目标用户 $u_t$ 和其他用户集 $U=\{u_1,u_2,\cdots,u_n\}$，假设用户 $u_t$ 的历史行为 $H=\{I_1,I_2,\cdots,I_m\}$，其中 $I_i$ $(1\leqslant i\leqslant n)$ 表示用户 $u_t$ 对物品的评分。对于一个目标物品集合 $T=\{t_1,t_2,\cdots,t_n\}$，本文旨在将目标物品集 $T$ 中用户 $u_t$ 喜欢的物品推荐给用户 $u_t$。

### 1.1 用户偏好

根据原始的用户-物品评分数据集，可以得到用户分布矩阵，用户被分到三个子集中：(1) 用户不喜欢这个物品 (不喜欢)；(2) 用户对这个物品喜欢程度一般(一般喜欢)；(3) 用户很喜欢这个物品(很喜欢)。表 1 是一个 $n\times3$ 的用户分布矩阵，$n$ 表示物品集的数量，$N_{ij}$ 表示喜欢(不喜欢、一般喜欢)物品 $i$ 的人数，其中 $1\leqslant j\leqslant3$。

表 1　用户分布矩阵

| *Item* | 不喜欢 | 一般喜欢 | 很喜欢 |
|---|---|---|---|
| $I_1$ | $N_{I1}$ | $N_{I2}$ | $N_{I3}$ |
| … | … | … | … |
| $I_i$ | $N_{i1}$ | $N_{i2}$ | $N_{i3}$ |
| $I_j$ | $N_{j1}$ | $N_{j2}$ | $N_{j3}$ |
| $I_n$ | $N_{n1}$ | $N_{n2}$ | $N_{n3}$ |

表 2 是一个概率矩阵，通过表 1 计算得到，$P_{ij}$ 表示所有用户中选择集合 $j$ 所占的概率，$i$ 表示物品。$P_{ij}$ 可由下面公式计算得到：

$$P_{ij}=\frac{N_{ij}}{\sum_{j=1}^{3}N_{ij}}$$

表 2　概率矩阵

| *Item* | 不喜欢 | 一般喜欢 | 很喜欢 |
|---|---|---|---|
| $I_1$ | $P_{I1}$ | $P_{I2}$ | $P_{I3}$ |
| … | … | … | … |
| $I_i$ | $P_{i1}$ | $P_{i2}$ | $P_{i3}$ |
| $I_j$ | $P_{j1}$ | $P_{j2}$ | $P_{j3}$ |
| $I_n$ | $P_{n1}$ | $P_{n2}$ | $P_{n3}$ |

### 1.2 三层感知机模型

三层感知机(MLP)[8]主要分为输入层、隐藏层、输出层，其中输入层到隐藏层是一个全连接的层，隐藏层到输出层是一个分类器 softmax 回归(逻辑回归)。可以采用下述模型：

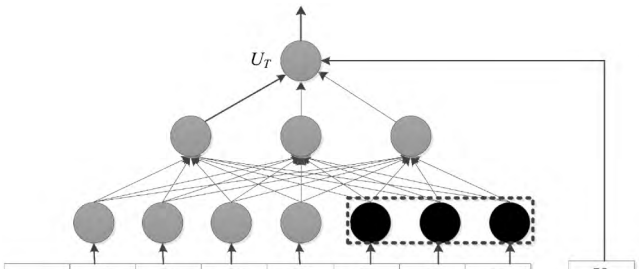
$$\begin{cases}s(x)=\tanh(x)=\dfrac{e^{x}-e^{-x}}{e^{x}+e^{-x}}\\ f(x)=G(B_2+W_2(s(W_1x+B_1)))\end{cases}$$

基金项目 福建省自然科学基金(2011J01345) 福州大学发展基金(2008-XQ-23) 福建省科技创新平台计划项目(2009J1007)

式(2)中 $B_1$ 表示输入层到隐藏层的偏置，$W_1$ 表示输入层到隐藏层的权重（也叫链接系数），$B_2$ 表示隐藏层到输出层的偏置，$W_2$ 表示隐藏层到输出层的权重，函数 $G$（采用 softmax 分类器）和 $s$ 是两个激活函数。

1.3 信息集成

当数据集中的评分信息不足的时候，就会存在稀疏问题，导致模型性能不高。因此，为了进一步提高性能，解决稀疏问题，在三层感知机模型（MLP）的输入层输入向量中加入用户偏好（P-MLP）。如图 2 所示：




| | $U_1$ | $U_2$ | $U_3$ | $U_4$ | $P_{i1}$ | $P_{i2}$ | $P_{i3}$ | $U_T$ |
|---|---|---|---|---|---|---|---|---|
| $I_1$ | 0.6 | 0.2 | 0.2 | 0.0 | 0.38 | 0.178 | 0.442 | 1.0 |
| $I_2$ | 0.2 | 0.6 | 0.2 | 1.0 | 0.15 | 0.575 | 0.276 | 0.6 |
| $I_3$ | 0.6 | 1.0 | 0.2 | -0.6 | 0.414 | 0.127 | 0.459 | -0.6 |
| $I_T$ | 0.2 | 0.4 | 1.0 | 0.6 | 0.372 | 0.19 | 0.437 | ? |

图 2 P-MLP 模型

其中 $P_i$ 表示用户对物品 $i$ 的偏好概率，$U=\{u_1,u_2,\cdots,u_n\}$ 表示与目标用户相似度最高的用户集，$n$ 表示选取的相似用户个数。

2 算法描述

聚类能够降低数据的稀疏性，神经网络能够有效的学习用户之间复杂关系，用户偏好能够反映物品流行度。因此通过对用户聚类，在神经网络的输入层中加入用户偏好，就能实现降低数据的稀疏性和有效的学习用户之间复杂关系。

具体实现时，需要先对模型进行训练，才能对目标用户行为进行预测。具体算法步骤如下：

step 1 使用 k-means 方法对用户聚类并得到行为相似的用户；

step 2 计算用户的偏好，得到概率矩阵；

step 3 对数据集按 4:1 分成训练集和测试集；

step 4 得到训练集和测试集中每个用户的特征向量 Vfinal；

step 5 使用训练集对 P-MLP 模型进行训练，得到最终模型；

step 6 根据训练好的模型对测试集进行预测。

对于一个用户 $u$ 和物品 $i$，step 1 得到 $U=\{u_1,u_2,\cdots,u_m\}$ 和 $Vsim=[r_{1i},r_{2i},\cdots,r_{mi}]$，其中 $m$ 表示选取相似用户的数量，$U$ 表示相似用户集，$r_{mi}$ 表示用户 $u_m$ 对物品 $i$ 的评分。step 2 得到用户对物品 $i$ 的偏好向量 $P_i=\left[\frac{m_{1i}}{n},\frac{m_{2i}}{n},\frac{m_{3i}}{n}\right]$，其中 $n$ 表示对物品 $i$ 评过分的用户总人数，$m_{1i}$ 表示喜欢物品 $i$ 的人数，$m_{2i}$ 表示物品 $i$ 喜欢程度一般的人数，$m_{3i}$ 表示不喜欢物品 $i$ 的人数。通过 step 1 和 step 2 可以得到 step 4 中每个用户最终的特征向量 $Vfinal=\left[r_{1i},r_{2i},\cdots,r_{mi},\frac{m_{1i}}{n},\frac{m_{2i}}{n},\frac{m_{3i}}{n}\right]$。

3 实验与结果分析

在这部分，本文将从准确度（precision），召回率（recall），F1 三个指标对比分析本文方法和传统方法。首先介绍实验使用的数据集以及对数据集的处理，其次介绍相关性能指标，最后分析使用不同聚类 K 值对性能影响、比较加入偏好的三层感知模型（P-MLP）与单纯的三层感知模型（MLP）、P-MLP 与 K-NN 和关联规则的比较。

3.1 数据集

本文实验使用 EachMovie 数据集。该数据集包含来自 72196 个用户对 1628 部电影的 2,811,983 个评分。电影打分有六个不同的值，从 0.0 到 1.0 的分值，分值越高表示电影越受用户喜欢。本文选择为 100 部以上电影评过分的前 1000 个用户作为实验数据集，然后将数据集随机分为训练集和测试集，训练集和测试的比例是 4:1。在实验中，为了更有效的训练本文的方法，本文对数据集进行了转换（如表 3 所示），对于用户没有过行为的电影本文设定评分值为 0.0。一般情况下处理没有作用过的缺省数据是比较困难的，通过对评分的转换，能够方便本文方法的训练。在训练 MLP 模型时，本文仅考虑目标用户评分大于 0.7 或小于 0.3，用户对物品评分大于 0.7 被认为是喜欢该物品，小于 0.3 被认为是不喜欢该物品。

表 3 评分转换

| 原始评分 | 0.0 | 0.2 | 0.4 | 0.6 | 0.8 | 1.0 | 没有行为 |
|---|---|---|---|---|---|---|---|
| 转换后 | -1.0 | -0.6 | -0.2 | 0.2 | 0.6 | 1.0 | 0.0 |

3.2 衡量标准

本文使用三个不同的指标来评估本文方法与传统方法的性能[10]。

用户间的相似度本文使用改进余弦相似度计算得到相似度最高的前 $n$ 个用户，$n$ 表示参照用户的数量。使用三个不同的指标来评估方法性能：准确度（precision），召回率（recall），F1。

令 $R(u)$ 是根据用户在训练集上的行为给用户做出的推荐列表，而 $T(u)$ 是用户在测试集上的行为列表，$U$ 是用户集合，那么推荐结果的召回率定义为：

$$\text{Recall}=\frac{\sum_{u\in U}|R(u)\cap T(u)|}{\sum_{u\in U}|T(u)|}$$

推荐结果的准确率定义为：

$$\text{Precision}=\frac{\sum_{u\in U}|R(u)\cap T(u)|}{\sum_{u\in U}|R(u)|}$$

推荐结果的 F1 定义为：

$$\text{F1}=\frac{2*\text{Recall}*\text{Precision}}{\text{Recall}+\text{Precision}}$$

3.3 聚类K 值的影响

使用 K-means 对用户聚类，聚类的数量 $K$ 值对整体的结果也有很大影响。实验结果表明，在其他条件相同的情况下（相同的 P-MLP 方法，相似度最高用户数量 $n$=10），不同的 $K$ 值在三个指标上面都会有差异，随着 $K$ 值的增大，三个指标先由小到大然后再由大到小，结果如表 4 所示：

表 4 不同 K 值性能比较 (%)

| | K = 5 | K = 10 | K = 15 |
|---|---|---|---|
| 准确率 | 75.4 | 86.0 | 78.7 |
| 召回率 | 56.2 | 62.3 | 57.4 |
| F1 | 64.4 | 72.25 | 66.38 |

3.4 P–MLP 与 MLP 比较

MLP 是三层感知模型，P–MLP 是加入用户偏好的三层感知模型，在实验中其他条件都相同（聚类的数量 K=10，相似度最高用户数量 $n$=10），P–MLP 的整体性能都比 MLP 模型性能高，推荐结果的准确率、召回率、F1 都是 P–MLP 模型高。两个模型方法的性能如表 5 所示：

表 5 MLP 与 P-MLP 性能比较 (%)

| | MLP | P–MLP |
|---|---|---|
| 准确率 | 81.1 | 86.0 |
| 召回率 | 60.2 | 62.3 |
| F1 | 69.1 | 72.25 |

3.5 P–MLP 与传统方法性能比较

在这部分中，将比较本文的方法与现有的协同过滤方法。为了比较的公平性，训练集和测试集都是一样的。数据集选择在 EachMovie 数据集中为 100 部以上电影评过分的前 1000 个用户。数据集按 4:1 随机分成训练集和测试集。表 6 展示了 P–MLP 方法与 KNN、关联规则的比较结果，其中 P–MLP 方法分别使用了聚类数量 $K$ = 5、$K$ = 10、$K$ = 15 三个模型方法。通过这个表可以看出，本文的 P–MLP 方法与现有的方法在性能上有很大的提高，其中在 P–MLP 模型方法中，当聚类 K = 10 的时候，总体的性能最高。准确率、召回率、F1 都最大。

表 6 性能比较 (%)

| | KNN | 关联规则 | P–MLP (K = 5) | P–MLP (K = 10) | P–MLP (K = 15) |
|---|---|---|---|---|---|
| 准确率 | 62.3 | 73.4 | 75.4 | 86.0 | 78.7 |
| 召回率 | 53.2 | 56.3 | 56.2 | 62.3 | 57.4 |
| F1 | 57.4 | 63.72 | 64.4 | 72.25 | 66.38 |

4 结论

本文主要针对传统协同过滤算法中存在的稀疏性问题和用户之间复杂关系表达问题，采用了一种神经网络与用户偏好相结合的协同过滤推荐算法，该方法首先对用户进行聚类，然后采用神经网络与用户偏好结合训练得到模型，最后利用训练后的多层神经网络来预测用户对物品的评分情况。实验结果表明，本文采用的方法与传统的方法（KNN、关联规则）比较，在准确度（precision），召回率（recall），F1 三个指标上有较大的提高。

参考文献：

[1] Shih Y Y, Liu D R. Product recommendation approaches: Collaborative filtering via customer lifetime value and customer demands [J]. Expert Systems with Applications, 2008, 35(1): 350-360.

[2] Sarwar, B.M., Karypis, G., Konstan, J.A. and Ried, J.: Item-based Collaborative Filtering Recommender Algorithms. Accepted for publication at the WWW10 Conference (2001).

[3] Linden G, Smith B, York J. Amazon. com recommendations: Item-to-item collaborative filtering [J]. Internet Computing, IEEE, 2003, 7 (1): 76-80.

[4] Lin W, Alvarez S A, Ruiz C. Collaborative recommendation via adaptive association rule mining [J]. Data Mining and Knowledge Discovery, 2000, 6: 83-105.

[5] Sarwar B, Karypis G, Konstan J, et al. Analysis of recommendation algorithms for e-commerce [C] //Proceedings of the 2nd ACM conference on Electronic commerce. ACM, 2000: 158-167.

[6] Herlocker J L, Konstan J A, Borchers A, et al. An algorithmic framework for performing collaborative filtering [C] //Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 1999: 230-237.

[7] Claypool M, Gokhale A, Miranda T, et al. Combining content-based and collaborative filters in an online newspaper [C] //Proceedings of ACM SIGIR workshop on recommender systems. 1999, 60.

[8] 李秀娟. KNN 分类算法研究 [J]. 科技信息, 2009 (31): 81-81.

[9] Junlin Z, Heng C, Tongwen H, et al. A Distributional Representation Model For Collaborative Filtering [J]. arXiv preprint arXiv:1502.04163, 2015.

[10] 推荐系统实践 [M]. 人民邮电出版社, 2012.

作者简介：

周高云(1989 - )，男，硕士研究生 通讯联系人 程红举(1975 - )，男，博士，副教授。

（上接第 102 页）

可以与现有网络管理系统良好集成，具有强大的生命力，减少资源浪费，提高了网络的可靠性和可扩展性，简化网络管理操作人员和计算机网络维护人员的工作。

（三）更加智能化

随着信息化建设的不断发展，计算机网络管理技术逐步转向了智能化。计算机网络管理工作的复杂程度以及专业技术含量在不断变高，网络管理员除了要具备扎实的专业知识外，还应具有丰富的网络管理经验。智能化的计算机网络管理技术可以支持系统功能的自动调整，防止某些网络资源性能的下降，这将极大地适应网络管理瞬变性、动态性和实时性的特点。计算机网络的智能化管理还能促进网络管理功能的优化，大大提高网络资源的利用率。按照科学性、可持续性的方式，不断提高计算机网络管理技术的水平，促进计算机网络管理技术智能化和开放式的融合发展。

四、结语

计算机网络管理技术的发展和革新是无时无刻不在进行的，促进计算机网络管理技术朝着更加层次化、集成化、智能化方向发展，使计算机网络管理技术的应用价值更加显著，进一步为经济效益及社会效益的实现奠定坚实的基础。

参考文献：

[1] 田海宇，计算机网络管理技术探析 [J].科学导报，2015(13)

[2] 杨常勇、张鲲，浅谈如何加强计算机网络管理技术创新应用 [J].信息工程，2013.5

[3] 李超宇，计算机网络管理技术的研究 [J].科技传播，2013.3（上）

[4] 孙培岩，计算机网络管理技术探索 [J].产业与科技论坛，2015(20)