文章编号：2096-1618(2018)02-0107-06

基于 RBM 模型的豆瓣小组推荐系统设计与实现

刘宇宁， 陶宏才
(西南交通大学信息科学与技术学院，四川 成都 611756)

摘要：将受限玻尔兹曼机(restricted boltzmann machine, RBM)模型应用于推荐领域已成为一个很有意义的研究方向。针对豆瓣小组，设计实现了一个基于 RBM 模型的推荐系统，该系统由数据层、模型层、评测层 3 部分组成。数据层通过选取“豆瓣达人”数据，一定程度上解决了数据稀疏问题。模型层利用对比散度(contrastive divergence, CD)算法进行学习。实验结果表明，在豆瓣小组数据集上，RBM 模型相较传统协同过滤算法具有更好的推荐效果。

关 键 词：豆瓣小组；推荐系统；限制玻尔兹曼机；对比散度算法
中图分类号：TP301.6 **文献标志码**：A
doi：10.16836/j.cnki.jcuit.2018.02.001

0 引言

随着互联网技术的发展，Web 社区层出不穷。以豆瓣小组为例，截至 2017 年第三季度，豆瓣网小组数量已达 60 万，注册用户超过 1 亿。小组数量激增为用户提供了更多的选择，但过多的小组却带来了“信息过载”的问题。推荐系统作为一种信息过滤的重要手段，是目前用来解决信息过载问题极具潜力的方法。

通常，解决推荐问题的方法有 3 种：传统的协同过滤算法、聚类模型及基于搜索的方法[1]。其中协同过滤算法应用最为广泛，但其面临数据稀疏问题[2]。同时，传统协同过滤方法采用浅层模型难以学习到隐藏在用户及项目当中的深层次特征。

近年来，深度学习在自然语言处理、图像识别等领域取得突破性进展[3]，同时也为推荐系统的研究创立了新方向。深度学习在推荐系统领域的应用最早可追溯至 Netflix 竞赛后半程异军突起的 RBM 模型[4]；Hu 等[5]在 RBM 模型基础上提出了多层 RBM，用来解决群组推荐问题；文献[6-7]阐述了如何将深度学习技术应用于音乐推荐领域；Liu 等[8]基于 RNN 推荐方法，研究位置信息与社交网络中行为预测的关系；Wang 等[9]采用基于 CNN 的推荐方法，通过融合图像信息解决兴趣点推荐问题；鉴于 RBM 模型在深度学习领域中占据的核心位置及其自身的良好特性[10]，将研究如何将 RBM 模型应用在豆瓣小组推荐当中，进而实现一个基于 RBM 模型的豆瓣小组推荐系统，并对推荐结果进行评测、分析。

收稿日期：2018-01-13
基金项目：国家自然科学基金资助项目(61505168)

1 RBM 模型原理

1.1 RBM 结构与主要参数

RBM 模型是一种生成式随机神经网络，如图 1 所示。v 代表可见层，用于表示输入数据，h 代表隐藏层，可理解为特征数据。可见层单元和隐藏层单元都是二元变量，其状态取{0,1}。W 表示两层之间的连接权重，a 和 b 分别表示可见层和隐藏层的偏置。虽然 RBM 模型所要表示的分布无法有效计算，但通过 Gibbs 采样(Gibbs sampling)可得到服从该分布的随机样本。Roux 等[11]证明，只要隐藏层单元数目足够，RBM 可拟合任意离散分布。

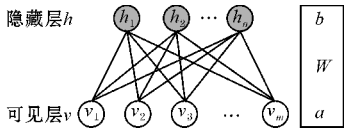

图 1 RBM 模型图表

RBM 模型的学习目标为最大程度拟合观测数据。对于一组输入数据而言，在未知其符合的概率分布时，很难对其进行学习。已有统计力学结论证明，任何概率分布都可以转换成基于能量的模型。因此，可以将学习数据的分布问题转换为求取一个能量模型的稳定状态，RBM 便是这样一种基于能量的模型。假设一个 RBM 模型中，用 v_i 表示第 i 个可见层单元的状态，用 h_j 表示第 j 个隐藏层单元的状态。对于一组给定的状态(v,h)，可将 RBM 作为一个系统所拥有的能量表示为

$$E(v,h;\theta) = -\sum_{ij} W_{ij} v_i h_j - \sum_i a_i v_i - \sum_j b_j h_j \quad (1)$$

其中，θ 是 RBM 的参数$\{W, a, b\}$。当该参数确定时，可根据 v 和 h 的联合配置能量，得到 v 和 h 的联合概率。希望最大化观测数据的似然函数得到 RBM 的参数，通过随机梯度下降进行求解。

1.2 权重学习方法

RBM 模型学习的关键是求出参数 θ 值，从而拟合给定的输入数据。通常采用 Hinton[12] 提出的对比散度(contrastive divergence, CD)算法进行学习，主要步骤描述如下：

在样本数据已知的情况下，随机初始化 W_{ij}，并通过可见层单元的状态 v_i 计算出隐藏层单元的状态 h_j，令 W_{ij} 的正向梯度表示为

$$pos(W_{ij}) = v_i \times h_j \quad (2)$$

以上过程称为 Positive-CD 算法阶段。同理，利用求得的隐藏层单元 h_j 反向计算可见层单元 v'_i，令 W_{ij} 的负向梯度表示为

$$neg(W_{ij}) = v'_i \times h_j \quad (3)$$

上述过程称为 Negative-CD 算法阶段。之后，更新权重，即：

$$W_{ij} = W_{ij} + L \times (pos(W_{ij}) - neg(W_{ij})) \quad (4)$$

其中，L 为学习率。一次模型训练过程即如上描述，实际训练时，定义最大训练次数 *epochs*，重复上述训练过程。

2 豆瓣小组推荐系统设计实现

2.1 系统框架

主要研究如何将 RBM 模型应用于推荐中，最终利用一个基于豆瓣小组的推荐系统展示研究结果。系统整体框架如图 2 所示，主要包含 3 个层次：数据层、模型层及评测层。数据层负责爬取豆瓣小组相关数据，并将数据输出至模型层进行训练。Covington 等[13] 在 YouTube 推荐系统的设计中提出，在为用户推荐少量视频前，先得到大量候选生成结果，然后利用特征工程对其进行排序，模型层借鉴该方法进行设计。依据模型训练结果得到候选生成数据，并融合其他特征进行权重计算，然后根据权重计算结果进行排名，最终输出推荐结果。评测层将利用不同的评测方法对推荐结果进行分析评价，得出结论。下文将详细介绍数据层与模型层的具体设计与实现，评测层的实现方法则放到实验评测部分阐述。

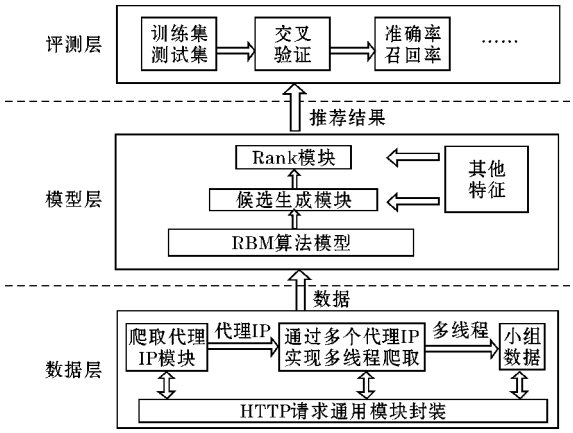

图 2 系统框架图

2.2 数据层的数据爬取

数据层使用爬虫程序对豆瓣小组相关数据进行爬取。其中，如何防止爬取时被封禁、提升爬取效率及内容质量问题成为爬虫程序设计的关键。

2.2.1 反爬虫策略

主要采用代理 IP 及爬取频率限制相结合的方式进行爬虫程序的设计。如图 2 的数据层所示，首先爬取到代理 IP 地址列表，接着，通过多个代理 IP 进行多线程爬取。其中，每个单独的爬虫线程需做好爬取频率控制，以防止因爬取过快而导致代理 IP 被豆瓣封禁。

2.2.2 豆瓣达人数据爬取

利用上述爬虫程序对小组数据进行爬取，除去无法爬取的小组(小组不存在、被解散或非公开小组等)，最终爬取到 417658 条小组数据。如图 3 所示，将所有小组按照已加入人数进行划分，主要分为：10 人以下、10 ~ 100 人、100 ~ 500 人、500 ~ 1000 人、1000 ~ 5000 人、5000 ~ 10000 人及 10000 人以上，在图中以不同的柱体表示。图 3 横轴代表豆瓣小组 ID 范围，此处共分为 7 个区段；纵轴代表某一 ID 范围内，不同柱体对应的人数。

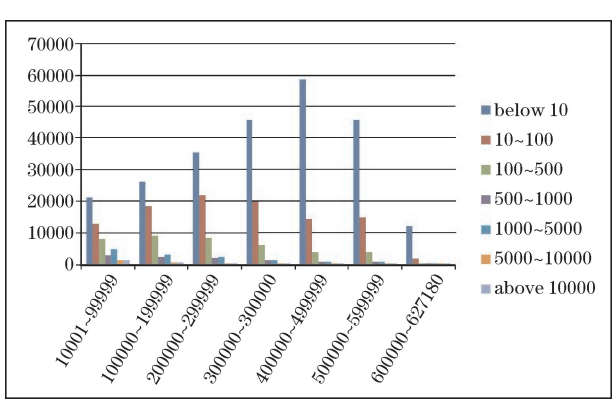

图 3 豆瓣小组 ID-人数分布情况

由图 3 可以明显地看到,大多数小组人数不足 10 人,而活跃小组只占所有小组的一小部分,这说明大多数小组质量较低。获得小组全量数据后,推荐系统还需要用户的行为数据。然而,截至 2017 年,豆瓣注册用户已超过 1 亿,爬取所有用户数据较为困难。

Amatriain 等[14]揭示了“专家用户”与一般用户在行为模式上的差异:“专家用户”对电影数据的评论更具参考价值,利用专家数据进行训练的模型在预测精度和推荐列表精度方面更具优势。同时,“专家用户”数据集的数据稀疏度要比全体用户数据集的稀疏程度低。受这一结论的启发,在训练算法模型时,只选取“专家用户”的行为数据,而不是大量“稀疏”的全体用户的行为数据。

对于豆瓣网来说,将“专家用户”定义为“豆瓣达人”,是豆瓣官方挑选的推荐给用户关注的对象。在豆瓣达人页面,一次只提供几位活跃用户,需要手动点击“换一换”才能展示其他“豆瓣达人”,如图 4 所示。此处的目标是穷举出所有“豆瓣达人”。

图 4　豆瓣达人

在穷举“豆瓣达人”的过程中,利用 Python 提供的 set 数据结构存储已获取到的“豆瓣达人”用户 ID,当获取新用户 ID 并向 set 中添加时,可保证用户 ID 不重复。在每次程序模拟“换一换”点击操作从而取到新用户 ID 前后,统计 set 的长度。同时,设定阈值 threshold,用来表示“换一换”操作前后 set 长度未发生变化的最大次数。初始时,将统计“换一换”操作前后 set 长度未发生变化的次数 count 置 0。通过不断调整阈值 threshold 的大小并观察获取到的用户 ID 数量,最终得出“豆瓣达人”的总数。

最终,选取 649 位豆瓣达人的行为数据。通过脚本统计,得到以下结果:649 位豆瓣达人总共加入了 27804 个小组。对这 27804 个小组按加入人数进行分段,并与所有小组已加入人数进行对比,得到如表 1 所示结果。其中,“10001 ~ 627180 num”代表所有小组成员数量情况,“active user group num”代表“豆瓣达人”加入的小组成员数量情况,“percentage”代表同一人数段“active user group num”与“10001 ~ 627180 num”的比值。由表 1 可以看出,人数较少的小组,豆瓣达人加入的比例也较小。随着小组加入人数规模上升,爬取到的豆瓣达人加入的小组占所有同等规模小组的比例也在上升。因此,爬取到的豆瓣达人的行为数据,能很好地包含豆瓣活跃小组的绝大多数,一定程度上,起到对豆瓣全量数据缩放的效果。于是,可将模型层输入的用户数据从接近 2 亿骤减至千人左右规模。

表 1　豆瓣达人加入的小组成员数与所有豆瓣小组成员数对比

option	10001 ~ 627180 num	active user group num	percentage/%
below 10	244752	588	0.24
10 ~ 100	104404	4130	3.96
100 ~ 500	39489	7407	18.76
500 ~ 1000	10149	3635	35.82
1000 ~ 5000	13056	7259	55.60
5000 ~ 10000	2763	2127	76.98
above 10000	3045	2542	83.48

2.3　模型层的推荐方法设计

主要介绍如何将 RBM 模型应用于小组推荐当中,并给出一些实践当中的具体优化方法。图 5 所示为利用 RBM 模型进行数据训练的通用流程。

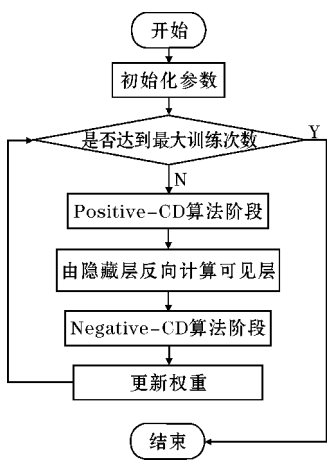

图 5　RBM 模型训练通用流程

初始化参数阶段,设定可见层及隐藏层的数目;利用正态分布函数初始化权重矩阵 $\boldsymbol{W}$;设定最大训练次数 epochs 及学习率 L;使用输入数据填充可见层单元等。

Positive-CD 算法阶段及 Negative-CD 算法阶段使用 sigmoid 激活函数,如式(5):

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{5}$$

在 Positive-CD 算法阶段,利用公式(2)计算 pos(W_{ij})时,通常 v_i 及 h_j 取值为 0 或 1。实际应用中,可利用公式(5)提供的激活函数对输入数据进行计算,作为 v_i 及 h_j 的实际替代值,Negative-CD 算法阶段亦如此。同时,当计算出来的权重 W_{ij} 较大时,通过添加正则项[15],对其进行惩罚。

在更新权重阶段,学习率 L 的选择很重要:L 过大,收敛速度变快,导致算法不稳定;而 L 过小虽可避免不稳定情况,但收敛速度将变慢。为解决这一矛盾,实践中通过增加动量项[16],使本次参数值的修改依赖上次参数值的修改。以权重 W_{ij} 为例,可采用式(6)对其进行更新:

$$W_{ij}^{(t+1)} = kW_{ij}^{(t)} + \in \frac{\partial L}{\partial W_{ij}^{(t)}} \tag{6}$$

其中,$W_{ij}^{(t+1)}$ 及 $W_{ij}^{(t)}$ 分别代表本次参数值与上一次的参数值。k 为动量项学习率,初始时,取0.5,随着模型训练趋于平稳,可取0.9。

训练完成后,隐藏层得出的结果即为模型求得的所有小组,去掉用户已加入小组,得到候选生成结果。再对其进行排序,得到最终推荐的小组。实际评测时,先利用训练集结果得出最佳参数组合,再利用最佳参数训练得到的模型,对测试集上的数据进行评测。

3 实验评测

3.1 评价指标

小组推荐问题可类比二分类问题,采用准确率、召回率对推荐结果进行评价。实验评测时,采用“10 折交叉验证法”拆分数据集为训练集和测试集。

将推荐给用户的小组记为 *Rec*,而用户在测试集上加入的小组记为 *Tes*,定义准确率和召回率计算公式。

准确率

$$Precision = \frac{\sum Rec \cap Tes}{\sum Rec} \tag{7}$$

召回率

$$Recall = \frac{\sum Rec \cap Tes}{\sum Tes} \tag{8}$$

然而,准确率和召回率在某些情况下是相互矛盾的,通常采用 *F*1 度量对两者进行加权调和平均。

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{9}$$

3.2 推荐结果评价

利用“10 折交叉验证法”对通过爬虫获取到的“豆瓣达人”数据进行拆分,在训练集上通过构建用户-小组矩阵,调整主要参数,进行模型训练,并在测试集上对模型进行评测。

以下展示训练 RBM 模型阶段调整主要参数后,*F*1 度量值的变化情况。该阶段使用训练集数据对模型进行训练,以得出适合模型的最佳参数组合。由于 RBM 模型中隐藏层单元的数目用于提取输入数据的特征,此处首先通过实验确定隐藏层单元数目的最佳值。

图 6 为训练次数 epochs 固定为 10[17],学习率 L 固定为 1 时,隐藏层单元数目的改变对 *F*1 度量值的影响。由图 6 可看出:随着隐藏层单元数目的不断增加,*F*1 度量值呈现上升趋势,设置为 30 时达到顶峰。随后,隐藏层数目的增加反而会使得 *F*1 度量值下降。这说明,隐藏层单元数目过少时,提取的特征数据不足;而过多时,则会影响模型的效果。因此,确定隐藏层单元最佳数为 30。

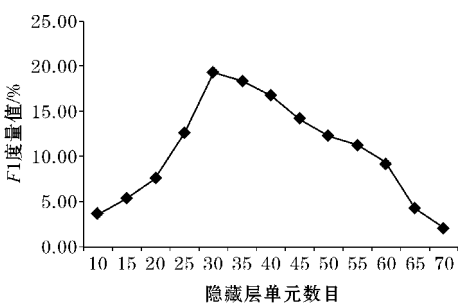

图 6 隐藏层单元数目对 *F*1 度量值的影响

在确定了隐藏层单元最佳数目后,接下来将以该最佳值作为隐藏层单元的固定值,通过改变训练次数 epochs 及学习率 L,分别判断其对 *F*1 度量值的影响。

图 7 展示了隐藏层单元数固定为 30,学习率 L 固定为 1 时,训练次数 epochs 的改变对 *F*1 度量值的影响。由图 7 可知,训练次数达到 50 次后,*F*1 基本稳定不变,故训练次数确定为 50。

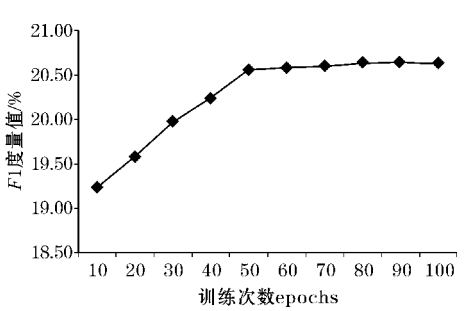

图 7 训练次数 epochs 对 *F*1 度量值的影响

同理,固定隐藏层单元数及训练次数,可得出学习率 L 的改变对 $F1$ 度量值的影响,如图 8 所示。

最终,通过不断调整模型参数的组合方式,得出适合豆瓣小组推荐的最佳参数,并利用最佳参数训练得到的模型在测试集上进行评测。

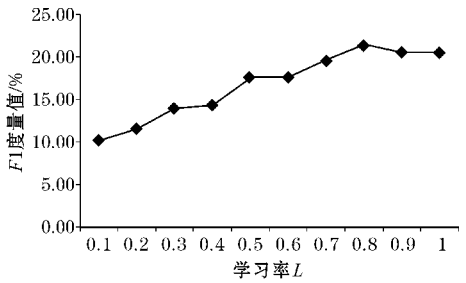

图 8　学习率 L 对 $F1$ 度量值的影响

同时,利用"10 折交叉验证法"对传统协同过滤算法(表中记为"CF")在该数据集下的表现进行评测,得出如表 2 所示结果。

表 2　传统协同过滤算法与 RBM 推荐效果对比

算法	准确率/%	召回率/%	$F1$ 度量值/%
CF	16.19	18.58	17.30
RBM	19.80	23.13	21.34

从表 2 可以看出,由于 RBM 模型能得到对于原始数据不同抽象程度的表示,因此可学习出更多的隐藏特征。在选择合适参数的情况下,相比于传统协同过滤采用的相似度计算方式,推荐效果更胜一筹。

4　结束语

传统协同过滤算法在向用户进行推荐时,侧重用户与用户之间或者项目与项目之间的相似度,往往倾向于提高与用户历史偏好的重合度。RBM 模型通过可见层与隐藏层之间的映射关系,试图挖掘用户及项目中更深层次的特征。实验结果表明,利用 RBM 模型构建的小组推荐系统相比传统协同过滤算法在各个维度的指标上表现更为突出。当然,尽管文中采用"豆瓣达人"数据作为评测的数据源,一定程度上缓解了数据稀疏性问题,但由实验结果可以看出,推荐准确率、召回率等指标仍不尽人意。未来,可考虑将 RBM 模型扩展至深度神经网络模型,以此进一步提升豆瓣小组的推荐效果。

参考文献:

[1] Linden G, Smith B, York J Amazon. Com Recommendations: Item-to-Item Collaborative Filtering [J]. IEEE Internet Computing, 2003, 7(1): 76-80.

[2] Georgiev K, Nakov P. A non-IID framework for collaborative filtering with restricted Boltzmann machines [C]. International Conference on International Conference on Machine Learning. JMLR. org, 2013: 1148-1156.

[3] Lecun Y, Bengio Y, Hinton G. Deep learning[J]. Nature, 2015, 521(7553): 436-444.

[4] Salakhutdinov R, Mnih A, Hinton G. Restricted Boltzmann machines for collaborative filtering[C]. International Conference on Machine Learning. ACM, 2007: 791-798.

[5] Hu L, Cao J, Xu G, et al. Deep modeling of group preferences for group-based recommendation [C]. Twenty-Eighth AAAI Conference on Artificial Intelligence. AAAI Press, 2014: 1861-1867.

[6] Wang X, Wang Y. Improving Content-based and Hybrid Music Recommendation using Deep Learning[C]. ACM International Conference on Multimedia. ACM, 2014: 627-636.

[7] Yon R. Music Personalization at Spotify[C]. ACM Conference on Recommender Systems. ACM, 2016: 373-373.

[8] Liu Q, Wu S, Wang L, et al. Predicting the next location: a recurrent model with spatial and temporal contexts[C]. Thirtieth AAAI Conference on Artificial Intelligence. AAAI Press, 2016: 194-200.

[9] Wang S, Wang Y, Tang J, et al. What Your Images Reveal: Exploiting Visual Contents for Point-of-Interest Recommendation [C]. The International Conference, 2017: 391-400.

[10] 张春霞,姬楠楠,王冠伟. 受限波尔兹曼机简介[J]. 工程数学学报, 2013(2): 159-173.

[11] Roux N L, Bengio Y. Representational power of restricted boltzmann machines and deep belief networks[J]. Neural Computation, 2008, 20(6): 1631.

[12] Hinton G E. Training products of experts by minimizing contrastive divergence [M]. MIT Press, 2002.

[13] Covington P, Adams J, Sargin E. Deep Neural Networks for YouTube Recommendations [C]. ACM Conference on Recommender Systems. ACM, 2016: 191-198.

[14] Amatriain X, Lathia N, Pujol J M, et al. The wisdom of the few: a collaborative filtering approach

based on expert opinions from the web[C]. International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2009:532-539.

[15] 罗恒. 基于协同过滤视角的受限玻尔兹曼机研究[D]. 上海:上海交通大学,2011.

[16] Geoffrey E. Hinton. A Practical Guide to Training Restricted Boltzmann Machines[J]. Momentum, 2012,9(1):599-619.

[17] 何洁月,马贝. 利用社交关系的实值条件受限玻尔兹曼机协同过滤推荐算法[J]. 计算机学报,2016(1):183-195.

Design and Implementation of the Recommendation System for Douban Group based on RBM Model

LIU Yu-ning, TAO Hong-cai

(School of Information Science & Technology, Southwest Jiaotong University, Chengdu 611756, China)

Abstract: Restricted Boltzmann Machine for recommendation has become one of the significant researches. In this paper, a recommendation system for Douban Group based on RBM model is designed and implemented. The system consists of three layers: data layer, model layer and evaluation layer. The data layer can solve the problem of data sparsity to a certain extent by selecting the data of "the Douban expert". The experimental results show that the RBM model rivals the traditional collaborative filtering algorithm by providing a better recommendation effect on the data set of the Douban Group.

Keywords: douban group; recommendation system; RBM; contrastive divergence