文章编号：1003-0077(2018)11-0086-11

ACMF：基于卷积注意力模型的评分预测研究

商 齐[1]，曾碧卿[1,2]，王盛玉[1]，周才东[1]，曾 锋[1]

（1. 华南师范大学 计算机学院，广东 广州 510631；
2. 华南师范大学 软件学院，广东 佛山 528225）

摘 要：评分数据稀疏是影响评分预测的主要因素之一。为了解决数据稀疏问题，一些推荐模型利用辅助信息改善评分预测的准确率。然而大多数推荐模型缺乏对辅助信息的深入理解，因此还有很大的提升空间。鉴于卷积神经网络在特征提取方面和注意力机制在特征选择方面的突出表现，该文提出一种融合卷积注意力神经网络（Attention Convolutional Neural Network，ACNN）的概率矩阵分解模型：基于卷积注意力的矩阵分解（Attention Convolutional Model based Matrix Factorization，ACMF），该模型首先使用词嵌入将高维、稀疏的词向量压缩成低维、稠密的特征向量；接着，通过局部注意力层和卷积层学习评论文档的特征；然后，利用用户和物品的潜在模型生成评分预测矩阵；最后计算评分矩阵的均方根误差。在 ML-100k、ML-1m、ML-10m、Amazon 数据集上的实验结果表明，与当前取得最好预测准确率的 PHD 模型相比，ACMF 模型在预测准确率上分别提高了 3.57%、1.25%、0.37%和 0.16%。

关键词：卷积神经网络；注意力机制；评分预测
中图分类号：TP391　　**文献标识码**：A

ACMF：Rating Prediction Based on Attention Convolutional Model

SHANG Qi[1], ZENG Biqing[1,2], WANG Shengyu[1], ZHOU Caidong[1], ZENG Feng[1]
(1. School of Computer, South China Normal University, Guangzhou, Guangdong 510631, China;
2. School of Software, South China Normal University, Foshan, Guangdong 528225, China)

Abstract: The sparseness of rating data is one of the main factors that affect the recommender models prediction. To exploit the advantage of convolutional neural networks in feature extraction and attention mechanism in feature selection, a probability matrix factorization model(PMF) with attention convolutional neural network(ACNN) is proposed as attention convolutional model based matrix factorization(ACMF). Firstly, the ACMF model compresses the high dimensional and sparse word vectors into low dimensional and dense feature vectors through word embedding technique. Then, it uses the local attention layer and convolutional layer to learn the feature of review document, and utilizes the user and item's latent models to reconstruct the rating prediction matrix. Finally, the loss function is set as the root-mean-square error of rating matrix. Compared with the best prediction model PHD, the ACMF model increases the accuracy rate on ML-100k, ML-1m, ML-10m and Amazon datasets by 3.57%, 1.25%, 0.37% and 0.16%, respectively.

Keywords: convolutional neural network; attention mechanism; rating prediction

0 引言

用户物品评分数据持续增长的同时也带来了数据稀疏问题，而数据稀疏使得传统协同过滤方法的评分预测准确率下降[1]。为了提高预测的准确率，文献[2-7]在评分预测任务中引入辅助信息（包括用户的人口统计学、社交网络和物品的评论文本等），例如，文献[2]提出协同主题回归模型（Collaborative Topic Regression，CTR），该模型属于概率图

收稿日期：2018-01-31　**定稿日期**：2018-03-31
基金项目：国家自然科学基金(61503143)；华南师范大学研究生创新计划项目(2016lkxm59)

模型(probability graph model)，模型整合了主题模型(topic model)、潜在狄利克雷分布[8](Latent Dirichlet Allocation，LDA)和概率矩阵分解[9](Probability Matrix Factorization，PMF)；文献[3,5,10-11]堆叠降噪自动编码器(Stacked Denoising Autoencoder，SDAE)提取物品的描述文档(如评论和摘要)特征以提高评分预测的精度；文献[12-13]将辅助信息按照用户和物品分别进行聚类，然后用卷积神经网络和注意力机制提取文本信息的特征。

然而，现有的模型通常使用词袋模型提取文本的特征，而词袋模型不考虑词的相互关系、词的上下文以及词序的影响。例如，“The paper is about CNN，not RNN.”和“The paper is about RNN，not CNN.”，LDA 和 SDAE 将文档看作由互不相同的词构成的词袋，忽略词序和上下文信息的影响，导致 LDA 和 SDAE 无法准确理解文档，进而降低评分预测的准确率。

为克服词袋模型的缺陷，文献[11,14-15]使用卷积神经网络提取文档的上下文特征，但是输入到 CNN 的文档未经过关键词筛选，存在大量的非关键词，而 CNN 又无法自动判别输入文档中哪些词为关键词(对评分预测至关重要)，导致 CNN 缺乏关键信息的提取能力。另外，CNN 在文档处理中往往忽略长度为 1 的卷积核，导致 CNN 缺乏单一词特征的提取。因此，本文提出一种有效的特征提取模型 ACMF，该模型可有效提取物品描述文档的上下文信息和单一词特征，并克服数据稀疏问题，从而提高评分预测的准确率。ACMF 模型的核心思想是将 CNN 模型和注意力机制整合到 PMF 框架中，使整合后的模型能够有效利用协同信息和上下文信息完成评分预测任务。为了验证 ACMF 模型的评分预测能力，本文在稀疏度不同的四个数据集上进行实验，实验结果表明，在数据稀疏和稠密的条件下，ACMF 模型的预测准确率均超过了 R-ConvMF、aSDAE 和 PHD 模型。

本文的主要贡献总结如下：

① 突破了词袋模型的限制，提出 ACMF 模型，该模型充分考虑词序和词的上下文因素，对不同的物品赋予差异化的高斯噪声，可有效提取文档的特征，并用于评分预测；

② ACMF 模型将 ACNN 整合到 PMF 框架下，并利用相关的正则化参数平衡评分信息和物品描述文档信息，从而减少数据稀疏带来的不利影响；

③ 在 ML-100k、ML-1m、ML-10m、Amazon 数据集上定性和定量分析了 ACMF 模型的实验效果，ACMF 模型比 PHD 模型的预测准确率分别提高了 3.57%、1.25%、0.37%和 0.16%。

1　相关工作

本节简要回顾矩阵分解、注意力机制和卷积神经网络的基本概念和相关研究工作。

1.1　矩阵分解

矩阵分解是一种基于模型的协同过滤方法，该方法首先在共享的潜在空间中找到用户和物品的潜在模型 U 和 V，然后计算用户潜在模型和物品潜在模型的内积，最后将内积的结果作为预测的评分矩阵[16]。假设有 N 个用户，M 个物品，用户评分矩阵 $R\in\mathbb{R}^{N\times M}$。在矩阵分解中用户 i 和物品 j 分别表示成 k 维向量，$u_i,v_j\in\mathbb{R}^k$。矩阵 R 的评分项 r_{ij} 近似地等于用户潜在向量和物品潜在向量的内积。通常采用最小化损失函数 L 来训练潜在模型，采用 L_2 范式作为正则项，防止模型过拟合，如式(1)所示。

$$L=\sum_{i}^{N}\sum_{j}^{M}I_{ij}\ (r_{ij}-u_i^{\mathrm{T}}v_j)^2+\lambda_u\sum_{i}^{N}\ \|u_i\|^2+\lambda_v\sum_{j}^{M}\ \|v_j\|^2 \tag{1}$$

其中，I 表示元素为 0 或 1 的方阵，I_{ij} 是指示函数，如果用户 i 对物品 j 评过分，则 $I_{ij}=1$，否则 $I_{ij}=0$。

1.2　注意力机制

注意力机制已成功应用于自然语言处理、图像处理、数据挖掘等多个领域。在文本挖掘中，注意力机制有利于模型更有效地发现和构建文档特征，使模型在训练时有选择地进行特征提取。文献[17]将注意力机制用于图片描述；在机器翻译领域[18-19]，注意力机制改进了原有的 encoder-decoder 翻译模型；Yin 等[20]提出一种基于注意力机制的卷积神经网络，将其用于句子对建模任务中；Wang 等[21]利用基于多层注意力机制的卷积神经网络进行句子关系分类；这些方法的成功说明注意力机制与 CNN 结合的有效性。

1.3 卷积神经网络

卷积神经网络最早用于图像识别问题[22]，现已在信息检索和数据挖掘等多个领域成功应用。尽管推荐系统和图像识别的任务不同，但CNN模型经过改造可用于推荐系统领域，比如CNN音乐推荐[23]，通过CNN分析歌曲的音频，并根据音频的潜在模型对歌曲进行评分预测，但是该CNN模型仅适用于音频的推荐，不适合文档的处理。文献[24-25]使用CNN将图像的特征融入推荐系统，利用多媒体进行特征的融合可以有效提升推荐的效果。

2 基于卷积注意力的矩阵分解(ACMF)

如图1所示，本文提出融合卷积神经网络和注意力机制的矩阵分解模型：ACMF。该模型包括两个子模块：PMF和ACNN。其中，U和V分别表示用户和物品的潜在模型，W表示ACNN网络所有的权值和偏置项的统称，R为用户—物品评分矩阵，σ^2表示方差，X为物品的描述文档，i、j和k分别为三个实线框内的元素所用到的角标。图1左侧虚线框表示PMF，右侧虚线框表示ACNN，ACMF的目标是将ACNN整合到PMF框架中，并通过PMF和ACNN联系物品的评分和描述文档信息用于评分预测。下面依次介绍ACMF模型的两个子模块以及模型的潜在变量优化。

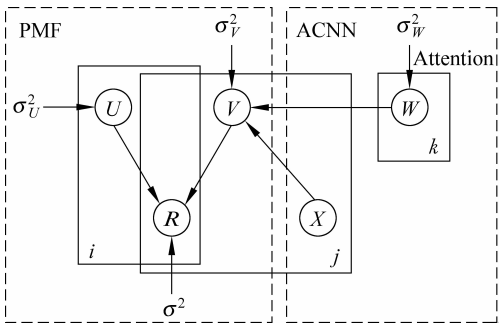

图1 ACMF模型的结构

(1) ACNN的网络结构。

ACNN根据物品的描述文档，生成物品的潜在模型，该网络的结构如下：

嵌入层：将序列化后的物品描述文档进行词嵌入操作，得到文档的特征表示。

局部注意力层：在嵌入层和卷积层之间添加局部注意力层，使ACNN网络在训练的过程中有选择地进行词特征提取，降低无关词对评分预测的影响。

卷积层：卷积层增加长度为1的卷积核，以便提取局部注意力层的单一词特征。

ACNN的池化层和全连接层与CNN模型相同。

(2) 概率矩阵分解。

PMF模型以概率生成的角度解释用户和物品的潜在模型。

(3) ACMF模型中潜在变量的优化。

PHD[11]为ACMF模型的主要对比模型，为保证实验的公平性，本文采用与PHD相同的优化方法：最大后验概率(Maximum A Posteriori, MAP)估计，优化ACMF模型的潜在变量。

2.1 ACNN的网络结构

ACNN网络的目标是生成物品的潜在模型，网络结构如图2所示，主要包括5层：嵌入层、注意力层、卷积层、池化层和输出层。

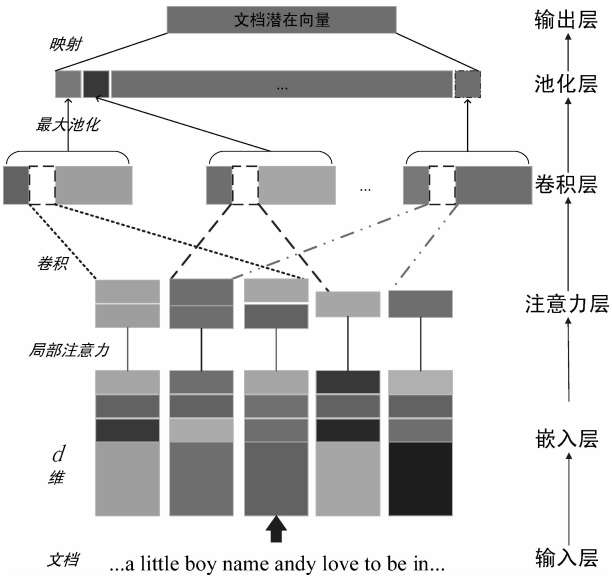

图2 ACNN网络结构图

2.1.1 嵌入层

假设文档包含T个词，每个词对应的词向量维度均为d。其中，词向量的初始化通过预训练的词嵌入模型Glove[26]完成，然后，通过神经网络的优化过程进一步训练词向量。记长度为T的文档对应的词嵌入矩阵为D，则$D\in\mathbb{R}^{d\times T}$，如式(2)所示。

$$D=(x_1,x_2,\cdots,x_T),0\leqslant T\leqslant 300 \quad (2)$$

2.1.2 局部注意力层

注意力机制分为局部注意力机制和全局注意力

机制[12]。局部注意力侧重用户的偏好和物品的属性特征，全局注意力侧重文档的语义特征。在评分预测中，用户的偏好和物品的属性更为重要，因此本文使用局部注意力机制。

在 ACMF 模型中，局部注意力模块通过滑动窗口来获取文本序列的注意力得分，用以表示各中心词的信息量大小。假设 x_t 为滑动窗口的中心词，窗口的长度为 w，利用参数矩阵 $W^1_{l\text{-}att}$ 和偏置项 $b^1_{l\text{-}att}$ 为每个词计算注意力得分，方法如下：

$$X_{l\text{-}att,t}=(x_{t+\frac{-w+1}{2}},x_{t+\frac{-w+3}{2}},\cdots,x_t,\cdots,x_{t+\frac{w-1}{2}})^{\mathrm{T}} \tag{3}$$

$$\mathrm{score}(t)=g(X_{l\text{-}att,t}*W^1_{l\text{-}att,t}+b^1_{l\text{-}att,t}),t\in[1,T] \tag{4}$$

其中，$*$ 表示符号两边的矩阵对应位置的元素相乘然后相加的操作。激励函数 $g(\cdot)$ 为 sigmoid 函数。score(t)为注意力得分，也是第 t 个词嵌入的权重。对于第 t 个词，词嵌入的权重序列为：

$$\hat{X}_{l\text{-}att,t}=(\hat{x}_t,\hat{x}_{t+1},\cdots,\hat{x}_{t+w-1})^{\mathrm{T}},t\in[1,T] \tag{5}$$

整个文档的词嵌入权重序列 $\hat{X}_{l\text{-}att}$ 为：

$$\hat{X}_{l\text{-}att}=(\hat{x}_1,\hat{x}_2,\cdots,\hat{x}_t,\cdots,\hat{x}_T)^T \tag{6}$$

其中，$\hat{x}_t=\mathrm{score}(t)x_t,t\in[1,T]$

2.1.3　卷积层

卷积层用于词的上下文特征提取，如式(7)所示，记第 j 个卷积核 $W^j_c\in\mathbb{R}^{d\times ws}$ 提取的特征为 $c^j_t\in\mathbb{R}^1$，该卷积核的长度 ws 决定被卷积的词数：

$$c^j_t=f(W^j_c*(\hat{X}_{l\text{-}att})_{(:,t:(t+ws-1))}+b^j_c) \tag{7}$$

其中，$*$ 表示卷积操作，$b^j_c\in\mathbb{R}^1$ 是 W^j_c 的偏置，f 表示非线性的激励函数 ReLU。

经过卷积核 $W^j_c(c\in\{1,5\},j=1,2,\cdots,n_c)$ 的文档对应的上下文特征向量为 $c^j\in\mathbb{R}^{T-ws+1}$，如式(8)所示。

$$c^j=[c^j_1,c^j_2,\cdots,c^j_t,\cdots,c^j_{T-ws+1}] \tag{8}$$

2.1.4　池化层

本文使用最大池化将文档的表示向量降低至 n_c 维：$d_f=[\max(c^1),\max(c^2),\cdots,\max(c^j),\cdots,\max(c^{n_c})]$

其中，c^j 是由第 j 个共享权值矩阵 W^j_c 提取到的上下文特征向量，该特征向量的长度为 $T-ws+1$。

2.1.5　输出层

输出层将上一层输出的高级特征用于评分预测，因此本文将 d_f 映射到 k 维空间中(用户或物品的潜在空间)，得到的文档潜在向量如式(9)所示。

$$s=\tanh(W_{f_2}\{\tanh(W_{f_1}d_f+b_{f_1})\}+b_{f_2}) \tag{9}$$

其中，$W_{f_1}\in\mathbb{R}^{f\times n_c}$，$W_{f_2}\in\mathbb{R}^{k\times f}$ 为映射矩阵，$b_{f_1}\in\mathbb{R}^f$，$b_{f_2}\in\mathbb{R}^k$ 分别为 W_{f_1} 和 W_{f_2} 的偏置向量，结果 $s\in\mathbb{R}^k$ 用于计算预测的评分。

回顾从输入层到输出层的整个过程，ACNN 可以看作一个函数，该函数以原始文档作为输入，返回潜在向量作为文档的输出。如式(10)所示。

$$s_j=\mathrm{acnn}_W(X_j) \tag{10}$$

其中，X_j 记为物品 j 的原始文档，s_j 记为物品 j 的文档潜在向量，为避免混淆，将所有权重和偏置变量记为 W。

2.2　基于卷积注意力矩阵分解的概率模型

PMF 的目标是找到用户和物品的潜在模型 U 和 V，然后用 $U^{\mathrm{T}}V$ 重建评分矩阵 R，其中 $U\in\mathbb{R}^{k\times N}$，$V\in\mathbb{R}^{k\times M}$。从概率的角度来看，已知评分的条件分布[9]表示如式(11)所示。

$$p(R\mid U,V,\sigma^2)=\prod_i^N\prod_j^M N(r_{ij}\mid u_i^{\mathrm{T}}v_j,\sigma^2)^{I_{ij}} \tag{11}$$

其中，$N(x|\mu,\sigma^2)$ 是均值为 μ 方差为 σ^2 的高斯正态分布的概率密度函数。

对于用户潜在模型，本文使用零均值球面高斯先验生成，方差为 σ^2_U。

$$p(U\mid\sigma^2_U)=\prod_i^N N(u_i\mid 0,\sigma^2_U I) \tag{12}$$

对于物品潜在模型，本文从三个变量中生成：

① ACNN 网络的内部参数 W；

② 变量 X_j 表示物品 j 的文档；

③ 变量 ϵ 作为高斯噪声，帮助 ACMF 模型更深入地优化物品潜在模型中的评分。

文献[27]的实验结果表明，单个物品的评分数量和高斯噪声的方差成反比。因此，物品潜在模型可以通过下面的方程得到：

$$v_j=acnn_W(X_j)+\epsilon_j \tag{13}$$

$$\epsilon_j\sim N\left(0,\frac{\sigma^2_V}{h(n_j)}I\right) \tag{14}$$

其中，n_j 为物品 j 拥有的评分数，$h(\cdot)$ 为平方根函数，物品潜在模型的条件分布为：

$$p(V\mid W,X,\sigma^2_V)=\prod_j^M N\left(v_j\mid\mathrm{acnn}_W(X_j),\frac{\sigma^2_V}{h(n_j)}I\right) \tag{15}$$

其中，X 是物品的描述文档集合，文档潜在向量 $\mathrm{acnn}_W(X_j)$ 和物品的高斯噪声 $\frac{\sigma^2_V}{h(n_j)}I$ 分别作为

高斯分布的均值和方差，在ACNN和PMF模型之间起着桥梁的作用，将物品的描述文档和评分紧密联系起来。

对于W中的权值w_k，使用零均值球面高斯先验计算，如式(16)所示。

$$p(W \mid \sigma_W^2) = \prod_k^{|w_k|} N(w_k \mid 0, \sigma_W^2) \tag{16}$$

将变量U、V、W的后验概率估计整合在一起，如式(17)所示。

$$p(U,V,W \mid R,X,\sigma^2,\sigma_U^2,\sigma_V^2,\sigma_W^2) \approx p(R \mid U,V,\sigma^2)p(U \mid \sigma_U^2)p(V \mid W,X,\sigma_V^2)p(W \mid \sigma_W^2) \tag{17}$$

2.3 优化方法

本文采用最大后验概率估计优化潜在变量U、V和W，如式(18)所示。

$$\max_{U,V,W} p(U,V,W \mid R,X,\sigma^2,\sigma_U^2,\sigma_V^2,\sigma_W^2) = \max_{U,V,W} [p(R \mid U,V,\sigma^2)p(U \mid \sigma_U^2) p(V \mid W,X,\sigma_V^2)p(W \mid \sigma_W^2)] \tag{18}$$

对式(18)取负对数，最大后验估计转化为求损失函数的最小值，如式(19)所示。

$$L(U,V,W) = \sum_i^N \sum_j^M \frac{I_{ij}}{2}(r_{ij} - u_i^{\mathrm{T}} v_j)^2 + \frac{\lambda_U}{2}\sum_i^N \| u_i \|^2 + \frac{\lambda_V}{2}\sum_j^M h(n_j) \| v_j - \mathrm{acnn}_W(X_j) \|^2 + \frac{\lambda_W}{2}\sum_k^{|w_k|} \| w_k \|^2 \tag{19}$$

其中，λ_U等于σ^2/σ_U^2，λ_V等于σ^2/σ_V^2，λ_W等于σ^2/σ_W^2(与对比模型保持一致)。为了最小化L，本文采用与PHD相同的坐标下降法优化潜在变量，对U(或V)进行优化时，将式(19)中W和V(或U)看作临时常量，式(19)变为关于U(或V)的二次函数，而参数U(或V)的优化可简化为L关于u_i(或v_j)的微分，如式(20)、式(21)所示。

$$u_i \leftarrow (VI_iV^{\mathrm{T}} + \lambda_U I_K)^{-1} VR_i \tag{20}$$

$$v_j \leftarrow (UI_jU^T + h(n_j)\lambda_V I_K)^{-1} (UR_j + h(n_j)\lambda_V \mathrm{acnn}_W(X_j)) \tag{21}$$

其中，$I_i \in R^{M\times M}$、$I_j \in R^{N\times N}$和$I_K \in R^{k\times k}$均为对角矩阵，I_i的对角元素为I_{ij}，$j=1,2,\cdots,M$，R_i为用户i的评分向量，其分量为$(r_{ij})_{j=1}^M$。对于物品j，R_j的定义分别与R_i类似。式(20)和式(21)分别显示了用户和物品的潜在向量u_i，v_j的更新过程，其中λ_U和λ_V为平衡参数，用于平衡评分信息和描述文档。

然而，W不能通过解析解的方式优化，因为W与ACNN中的特征密切相关，但U和V作为临时常量时，式(19)中L可以变为加权误差平方函数，如式(22)所示。

$$\varepsilon(W) = \frac{\lambda_V}{2}\sum_j^M h(n_j) \| v_j - \mathrm{acnn}_W(X_j) \|^2 + \frac{\lambda_W}{2}\sum_k^{|w_k|} \| w_k \|^2 + \mathrm{constant} \tag{22}$$

于是，可使用反向传播算法优化参数W，并重复整个优化过程，直至ε收敛或达到预先定义的迭代次数，如式(23)所示。

$$\nabla_{w_k}\varepsilon(W) = -\lambda_V \sum_j^M h(n_j)(v_j - \nabla_{w_k}\mathrm{acnn}_W(X_j)) + \lambda_W w_k \tag{23}$$

如算法1(ACMF)所示，为防止ACMF模型过拟合，重复整个优化过程，直到模型在验证集上满足提前停止条件。通过优化参数U、V和W，最后就可以预测用户关于物品的未知评分，如式(24)所示。

$$r_{ij} \approx E[r_{ij} \mid u_i^{\mathrm{T}} v_j, \sigma^2] = u_i^{\mathrm{T}} v_j = u_i^{\mathrm{T}}(\mathrm{acnn}_W(X_j) + \epsilon_j) \tag{24}$$

2.4 时间复杂度分析

算法1 ACMF

已知：R：评分矩阵，X：物品的描述文档
目标：优化潜在因子U、V和W

```
随机初始化U和W
for j≤M do
    通过 v_j←acnn_W(X_j)初始化V
end for
repeat
    for i≤N do
        update: u_i←(VI_iV^T+λ_U I_K)^{-1}VR_i
    end for
    for j≤M do
        update:
    v_j←(UI_jU^T+h(n_j)λ_V I_K)^{-1}(UR_j+h(n_j)λ_V acnn_W(X_j))
    end for
    repeat
        for j≤M do
            通过公式(23)更新W
        end for
    until 收敛
until 在验证集上满足提前停止条件
```

对于每一轮训练(epoch)，更新用户和物品潜在

变量的时间复杂度为 $O(k^2 n_R + k^3 N + k^3 M)$，其中 n_R 表示已知评分项(非空项)的数量，k 为用户(或物品)潜在向量的维度。ACNN 网络更新权重和偏置的时间复杂度为 $O(n_c \cdot d \cdot T \cdot M)$。一轮训练过程的总时间复杂度，如式(25)所示。

$$O(k^2 n_R + k^3 N + k^3 M + n_c \cdot d \cdot T \cdot M) \quad (25)$$

3　实验

3.1　数据集

为了验证本文提出的模型的评分预测能力，本文使用 MovieLens① 和 Amazon② 数据集，数据集包括物品的描述文档和用户-物品评分数据(分值的范围为[1,5])，Amazon 和 MovieLens 的物品描述文档分别从 Amazon 和 IMDB③ 官网获取。对物品描述文档的预处理流程与文献[14]类似，步骤如下：

① 设置最大原始文档长度为 300(大于 300 的删掉)；

② 去除停用词；

③ 为每个词计算 TF-IDF 的值；

④ 去除语料库特有的停用词，即文档频率高于 0.5 的词；

⑤ 挑出频率最高的前 8 000 个不同的词作为词汇表；

⑥ 将原始文档中没有出现在词汇表里的词去掉。

对于评分数据，从评分数据集中去除没有描述文档的物品。对于 Amazon 评分数据，去除评分少于 6 项的用户得到 AIV-6，最后经过统计得到表 1。与 MovieLens 评分数据相比，AIV-6 评分数据更加稀疏。

表 1　四个数据集的数据统计

数据集	用户数	物品数	评分数	稠密度/%
ML-100k	943	1 546	94 808	6.503
ML-1m	6 040	3 544	993 482	4.641
ML-10m	69 878	10 073	9 945 875	1.413
AIV-6	5 072	10 843	48 836	0.089

3.2　对比模型

本文对比的模型如下：

① NMF[28]：非负矩阵分解是在矩阵分解的基础上对分解完成的矩阵加上非负的限制条件；

② PMF[9]：概率矩阵分解是标准的评分预测模型，该模型从概率的角度解释评分预测任务；

③ SVD[29]：奇异值分解是标准的评分预测模型，该模型仅将评分信息用于协同过滤；

④ R-ConvMF[14]：卷积矩阵分解通过预训练的词嵌入模型增强卷积神经网络的评分预测能力。

⑤ aSDAE[10]：通过堆叠降噪自动编码器分别添加用户和物品的辅助信息，提高评分预测的准确率。

⑥ PHD[11]：PHD 模型为 R-ConvMF 和 aSDAE 模型的融合，R-ConvMF 和 aSDAE 分别用于处理物品的描述文档和用户的辅助信息，最后结合两者以提高评分预测能力；

⑦ ACMF：ACMF 是本文提出的模型。

其中，传统推荐模型 NMF、PMF 和 SVD 采用 Surprise④ 辅助实现。

3.3　评估指标

本文选取均方根误差(RMSE)作为模型的评估指标。将数据集随机切分为训练集(80%)、验证集(10%)和测试集(10%)，训练集中每个用户或每个物品都至少有一项评分。比如测试集的均方根误差计算如式(26)所示。

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i,j}^{N,M} (r_{ij} - \hat{r}_{ij})^2}{n_R * 10\%}} \quad (26)$$

为了增强实验结果的可靠性，本文对于每一组实验从数据集切分开始，每组重复 5 次，最后取 5 次实验的平均值作为最终的实验结果。

3.4　实验环境和参数设置

实验环境：大数据服务器，参数为 Tesla K40m，12CPU(Intel(R)，Xeon(R)，1899.994MHz)，2GPU(K40，11439MiB 内存)。

参数设置：ACMF 的训练采用基于 Adam 优化器的批梯度下降法，每次训练 128 个样本；用户(或物品)潜在向量的维度 k 设为 50，用[0,1]区间的值随机初始化 U 和 V，各模型在 ML-100k、ML-1m、

① https://grouplens.org/datasets/movielens/
② http://jmcauley.ucsd.edu/data/amazon/
③ http://www.imdb.com/
④ https://github.com/NicolasHug/Surprise/

ML-10m、AIV-6 数据集上的正则化参数 λ_U 和 λ_V 设置如表 2 所示。

表 2 λ_U 和 λ_V 的参数设置

模型	ML-100k		ML-1m		ML-10m		AIV-6	
	λ_U	λ_V	λ_U	λ_V	λ_U	λ_V	λ_U	λ_V
NMF	0.06	0.06	0.06	0.06	0.08	0.08	0.06	0.06
PMF	0.01	0.01	0.01	0.01	0.01	0.01	0.02	0.02
SVD	0.005	0.005	0.01	0.01	0.01	0.01	0.02	0.02
R-ConvMF	2	210	3	250	2	210	250	1
aSDAE	2	210	3	250	2	210	250	1
PHD	2	210	3	250	2	210	250	1
ACMF	5	50	2	210	50	10	1	40

对于 ACNN 网络，本文作如下设置：

① 初始化词的潜在向量：通过 Glove 初始化词的潜在因子，维度为 200；

② 局部注意力层的窗口长度为 5；

③ 在卷积层使用长度为 1 和 5 的卷积核，两种卷积核的数量均为 50 个；

④ 卷积层的激励函数选择 ReLU，可以避免反向传播过程中的梯度消失问题；

⑤ 使用两层 Dropout 防止 ACNN 网络过拟合，并且设置丢失率为 0.4 和 0.2。

3.5 实验结果

实验说明：R-ConvMF、aSDAE、PHD 和 ACMF 模型中均使用了物品的辅助信息，但 R-ConvMF 和 ACMF 没有添加用户的辅助信息。

在训练集比例为 80%的 MovieLens 数据集上，各模型的测试均方根误差结果如表 3 所示。

表 3 不同模型在四个数据集上的平均 RMSE

模型	数据集			
	ML-100k	ML-1m	ML-10m	AIV-6
NMF	0.97400	0.91250	0.86912	1.21990
PMF	0.95220	0.88900	0.82703	1.05570
SVD	0.94530	0.86840	0.79521	1.05450
R-ConvMF	0.94474	0.86648	0.79774	1.66675
aSDAE	0.94418	0.86296	0.78434	1.30362
PHD	0.92963	0.84742	0.79002	1.04249
ACMF-G	0.919 25	0.855 56	0.786 55	1.050 41
ACMF	**0.896 47**	**0.836 79**	**0.781 42**	**1.040 83**
提升幅度/%	3.57	1.25	0.37	0.16

表 3 中，“提升幅度”表示 ACMF 与当前取得最小预测误差的对比模型相比，预测准确率提升的幅度。表 3 中，ACMF-G 表示使用全局注意力机制的 ACMF 模型，与使用局部注意力机制的 ACMF 相比，ACMF-G 在 ML-100k、ML-1m、ML-10m、AIV-6 数据集上的均方根误差均大于 ACMF，这是因为全局注意力机制关注的是长文本的特征，对长文本的所有词都分配相应的权重，无关的词也会获得相应的权重，从而降低了重点词的权重，干扰模型的特征选择，因此不能很好地突出对评分预测产生影响的词，使得传入卷积层的词没有经过筛选，降低了特征的质量，进而导致模型的预测准确率降低。而局部注意力机制更关注当前窗口中最重要的词，对其分配较大的权重，有利于锁定评分相关的词，最终使得 ACMF 模型取得较好的预测准确率。与 PHD 相比，ACMF 在 ML-100k、ML-1m、ML-10m、AIV-6 数据集上的提升幅度分别为 3.57%、1.25%、0.37%和 0.16%，说明深入理解物品的描述文档比简单添加用户辅助信息在提升评分预测准确率方面更重要。与传统推荐模型 NMF、PMF 和 SVD 相比，ACMF 模型在 ML-100k、ML-1m、ML-10m、

AIV-6 数据集上的提升幅度分别为 5.17%、3.64%、1.73%和 1.30%，说明 ACMF 模型比传统模型的特征提取能力强，深入理解物品的辅助信息对评分预测有效。

在 AIV-6 数据集上，模型 ACMF 与 R-ConvMF 均未添加用户辅助信息，但与 R-ConvMF 相比，ACMF 在 ML-100k、ML-1m、ML-10m 和 AIV-6 数据集上分别提升了：5.11%、3.43%、2.05%、37.6%，在 AIV-6 数据集上尤为明显，说明在数据集极其稀疏时，ACNN 网络的局部注意力层可以有效改善物品描述文档的特征选择和特征提取过程，促进模型深入理解物品的描述文档，进而生成高质量的物品潜在模型，同时减轻数据稀疏对评分预测的影响。

在 AIV-6 数据集上，PHD（或 aSDAE）和 R-ConvMF 对比，说明在数据极其稀疏时，加入用户的辅助信息可有效改善评分预测的精度。

与传统推荐模型 NMF、PMF 和 SVD 相比，R-ConvMF 和 aSDAE 在 AIV-6 数据集上预测误差较大，原因是 R-ConvMF 和 aSDAE 模型缺乏文档特征的选择和提取能力，aSDAE 利用降噪自动编码器处理用户和物品的文档信息，R-ConvMF 则利用 CNN 模型提取物品描述文档的特征，但输入到 aSDAE 和 CNN 模型的物品描述文档未经过关键词筛选流程，存在大量的非关键词，而 aSDAE 和 CNN 均无法自动判别输入文档中哪些词为关键词，因而影响了物品潜在模型的生成，进而影响评分预测的准确率；另一方面，CNN 在处理文档时往往忽略长度为 1 的卷积核，导致模型缺乏对单一词特征的提取能力，不能深入理解物品的描述文档，无法将物品文档特征准确地映射到物品潜在空间中，也就无法构建高效的物品潜在模型，使得 CNN 模型难以适应数据稀疏场景。

图 3 和表 4 为不同模型在稠密度不同的 ML-1m 数据集上的平均均方根误差。由图 3 和表 4 可知，与其他对比模型相比，ACMF 模型在 ML-1m 数据集上生成的潜在模型更为高效。从图 3 可以看出，当数据由稀疏逐渐变得稠密时，各模型的预测能力都有一定程度的提升，说明数据稀疏问题制约了模型的预测精度。表 4 中 ACMF 模型与 PHD 相比，当训练集比例从 20%增长到 80%时，ACMF 相比 PHD 的提升幅度从 0.18%增长到 1.25%，说明当数据稠密时，ACMF 模型可以产生更精确的潜在模型，证明 ACMF 将 CNN 和注意力机制整合到 PMF 框架的方式有效。

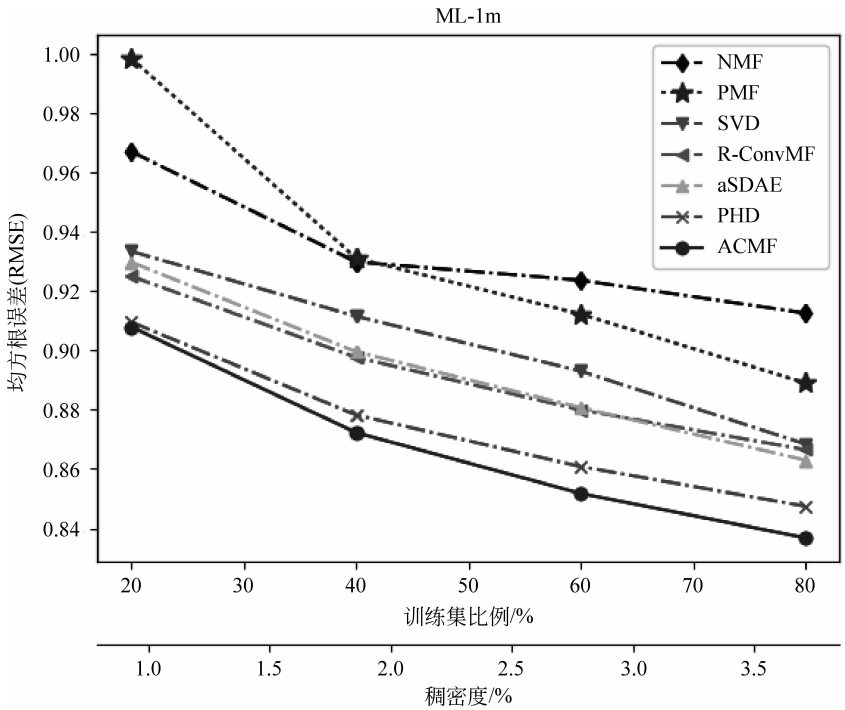

图 3　不同模型在 ML-1m 数据集上的平均 RMSE

表 4 不同模型在 ML-1m 数据集上的平均 RMSE

模型	训练集占整个数据集的比例(稠密度)			
	20%	40%	60%	80%
	(0.93%)	(1.86%)	(2.78%)	(3.71%)
NMF	0.966 90	0.929 70	0.923 60	0.912 50
PMF	0.998 30	0.931 00	0.912 20	0.889 00
SVD	0.933 30	0.911 40	0.893 00	0.868 40
R-ConvMF	0.924 89	0.897 56	0.879 90	0.866 48
aSDAE	0.929 61	0.899 48	0.880 67	0.862 96
PHD	0.909 47	0.878 14	0.860 82	0.847 42
ACMF	**0.907 82**	**0.872 20**	**0.851 77**	**0.836 79**
提升幅度	0.18%	0.68%	1.05%	1.25%

表 5 显示了 λ_U 和 λ_V 对评分预测的影响。其中,ML-100k、ML-1m、ML-10m 和 AIV-6 数据集的训练集比例均为 80%。通过网格搜索发现,(λ_U,λ_V)在 ML-100k、ML-1m、ML-10m 和 AIV-6 数据集上分别取(5,50)、(2,210)、(50,10)、(1,40)时,ACMF 模型的预测准确率最高,说明合适的(λ_U,λ_V)使得评分信息和物品描述文档信息分别映射到用户和物品的潜在空间,并在评分信息和物品描述文档信息之间达到平衡,从而提高 ACMF 模型的评分预测准确率。

表 5 λ_U 和 λ_V 对评分预测的影响

稠密度	6.503%	ML-100k							
λ_U	2	2	2.5	5	5	3	3	4	6
λ_V	50	210	250	50	45	60	48	60	60
RMSE	0.902 62	0.904 92	0.912 35	**0.896 47**	0.897 51	0.897 97	0.898 42	0.898 42	0.902 70
稠密度	4.641%	ML-1m							
λ_U	2	2	2	2	3	3	3	1	1
λ_V	200	210	220	250	150	250	300	150	100
RMSE	0.837 99	**0.836 79**	0.837 67	0.837 98	0.837 42	0.842 28	0.845 58	0.860 14	0.873 02
稠密度	1.413%	ML-10m							
λ_U	15	10	10	50	60	50	50	60	45
λ_V	80	60	50	10	10	5	15	15	15
RMSE	0.790 08	0.783 54	0.784 71	**0.781 42**	0.782 23	0.786 26	0.783 41	0.785 85	0.782 26
稠密度	0.089%	AIV-6							
λ_U	250	100	1	1	5	1	1	1	0.1
λ_V	1	1	100	50	50	10	30	40	50
RMSE	1.639 40	1.259 78	1.061 76	1.042 76	1.133 90	1.102 74	1.043 69	**1.040 83**	1.133 53

3.6 时间复杂度

本文所提出的 ACMF 模型和各对比模型的训练时间复杂度如表 6 所示,其中 p 为用户辅助信息的二进制向量维度,k_1 为堆叠降噪自动编码器的第一层的输出维度。对于传统模型,NMF 和 SVD 的时间复杂度由文献[30]得出。对于深度学习模型,PHD 模型的时间复杂度由 aSDAE、R-ConvMF 模型的权值更新和概率矩阵分解三部分组成,其中,aSDAE 的时间复杂度参考文献[3]计算得来。由表 6 可知,PHD 模型的时间复杂度大于 ACMF,这是因为 PHD 模型融合了多个模型,因此模型的训练参数增多,增加了模型的时间复杂度。

表 6 时间复杂度对比

模型	时间复杂度
NMF	$O(kn_R)$
PMF	$O(k^2 n_R + k^3 N + k^3 M)$

续表

模型	时间复杂度
SVD	$O(N^2M+M^3)$
R-ConvMF	$O(k^2n_R+k^3N+k^3M+n_c \cdot d \cdot T \cdot M)$
aSDAE	$O(k^2N+k^3+k_1NM+pk_1N)$
PHD	$O(k^2N+k^3+k_1NM+pk_1N+k^2n_R+k^3N+k^3M+n_c \cdot d \cdot T \cdot M)$
ACMF	$O(k^2n_R+k^3N+k^3M+n_c \cdot d \cdot T \cdot M)$

4　总结与展望

本文提出混合协同过滤模型 ACMF，将卷积注意力神经网络 ACNN 整合到 PMF 框架下。PMF 用于处理用户评分信息，ACNN 处理物品的文档信息，最终在 PMF 框架下实现对接，完成矩阵分解。与 CNN 结构不同，ACNN 在嵌入层和卷积层之间添加局部注意力层，增强了 CNN 模型的特征选择能力和单一词特征的提取能力。实验结果表明，ACMF 可以从用户-物品评分和物品的描述文档中分别学习到用户和物品的潜在模型，并在评分信息和文档信息之间达到一种平衡，从而提高了评分预测的准确率。

虽然 ACMF 模型在四个数据集上均取得了较高的预测准确率，但模型缺少对用户辅助信息的特征提取，下一步将探索用户的辅助信息对评分预测的影响，然后设计新的模型，进一步提升评分预测的准确率。

参考文献

[1] Herlocker J L, et al. Evaluating collaborative filtering recommender systems[J]. ACM Transactions on Information Systems (TOIS), 2004, 22(1): 5-53.

[2] Wang C, Blei D M. Collaborative topic modeling for recommending scientific articles[C]//Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2011: 448-456.

[3] Wang H, Wang N, Yeung D Y. Collaborative deep learning for recommender systems[C]//Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2015: 1235-1244.

[4] McAuley J, Leskovec J. Hidden factors and hidden topics: Understanding rating dimensions with review text[C]//Proceedings of the 7th ACM Conference on Recommender Systems. ACM, 2013: 165-172.

[5] Li S, Kawale J, Fu Y. Deep collaborative filtering via marginalized denoising auto-encoder[C]//Proceedings of the 24th ACM International on Conference on Information and Knowledge Management. ACM, 2015: 811-820.

[6] Ling G, Lyu M R, King I. Ratings meet reviews: A combined approach to recommend[C]//Proceedings of the 8th ACM Conference on Recommender Systems. ACM, 2014: 105-112.

[7] 马春平，陈文亮. 基于评论主题分析的评分预测方法研究[J]. 中文信息学报，2017，31(2)：204-211.

[8] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet allocation[J]. Journal of Machine Learning Research, 2003, 3(Jan): 993-1022.

[9] Mnih A, Salakhutdinov R. Probabilistic matrix factorization[C]//Advances in Neural Information Processing Systems. 2008: 1257-1264.

[10] Dong X, et al. A hybrid collaborative filtering model with deep structure for recommender systems[C]//AAAI, 2017: 1309-1315.

[11] Liu J, Wang D, Ding Y. PHD: A probabilistic model of hybrid deep collaborative filtering for recommender systems[C]//Proceedings of the Ninth Asian Conference on Machine Learning, PMLR 77, 2017: 224-239.

[12] Seo S, et al. Interpretable convolutional neural networks with dual local and global attention for review rating prediction[C]//Eleventh ACM Conference on Recommender Systems. ACM, 2017: 297-305.

[13] Seo S, et al. Representation learning of users and items for review rating prediction using attention-based convolutional neural network[C]//3rd International Workshop on Machine Learning Methods for Recommender Systems (MLRec)(SDM'17), 2017.

[14] Kim D, et al. Deep hybrid recommender systems via exploiting document context and statistics of items[J]. Information Sciences, 2017, 417: 72-87.

[15] Kim D, et al. Convolutional matrix factorization for document context-aware recommendation[C]//Proceedings of the 10th ACM Conference on Recommender Systems. ACM, 2016: 233-240.

[16] Koren Y. Factorization meets the neighborhood: A multifaceted collaborative filtering model[C]//Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2008: 426-434.

[17] Xu K, et al. Show, attend and tell: Neural image caption generation with visual attention[C]//International Conference on Machine Learning, 2015: 2048-

2057.

[18] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate [J]. arXiv preprint arXiv: 1409.0473, 2014.

[19] Luong M T, Pham H, Manning C D. Effective approaches to attention-based neural machine translation [J]. arXiv preprint arXiv: 1508.04025, 2015.

[20] Yin W, et al. ABCNN: Attention-based convolutional neural network for modeling sentence pairs[J]. Computer Science, 2015.

[21] Wang Y, et al. Attention-based LSTM for aspect-level sentiment classification[C]//Conference on Empirical Methods in Natural Language Processing, 2016: 606-615.

[22] Lécun Y, et al. Gradient-based learning applied to document recognition[C]//Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[23] Oord A V D, Dieleman S, Schrauwen B. Deep content-based music recommendation[J]. Advances in Neural Information Processing Systems, 2013(26): 2643-2651.

[24] He R, Mcauley J. VBPR: Visual Bayesian personalized ranking from implicit feedback [C]//Thirtith AAAI Conference on Artificial Intelligence. AAAI Press, 2016: 144-150.

[25] Zhang F, et al. Collaborative knowledge base embedding for recommender Systems[C]//Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2016: 353-362.

[26] Pennington J, Socher R, Manning C. Glove: Global vectors for word representation[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014: 1532-1543.

[27] Salakhutdinov R, Mnih A. Bayesian probabilistic matrix factorization using Markov chain Monte Carlo [C]//Proceedings of the International Conference on Machine Learning. ACM, 2008: 880-887.

[28] Zhang S, et al. Learning from incomplete ratings using non-negative matrix factorization[C]//Proceedings of the Siam International Conference on Data Mining, April 20-22, 2006, Bethesda, Md, Usa. DBLP, 2006: 549-553.

[29] Sarwar B, et al. Application of dimensionality reduction in recommender systems[J]. Acm Webkdd Workshop, 2000.

[30] George T, Merugu S. A scalable collaborative Filtering framework based on co-clustering[C]//Proceedings of the IEEE International Conference on Data Mining. IEEE, 2005: 625-628.

商齐(1994—),硕士研究生,主要研究领域为自然语言处理、推荐系统。
E-mail: 1162276945@qq.com

王盛玉(1992—),硕士研究生,主要研究领域为自然语言处理、情感分析。
E-mail: wangshengyu_1992@163.com

曾碧卿(1969—),通信作者,博士,教授,主要研究领域为人工智能与自然语言处理。
E-mail: zengbiqing0528@163.com