基于注意力机制和知识蒸馏的电影评分预测

刘 彤 于思洁 倪维健*
(山东科技大学计算机科学与工程学院 山东 青岛 266590)

摘 要 针对简单的因子分解机模型(Factorization Machines, FM)对于高阶交互的时间复杂度高和神经网络解决复杂问题尺寸过大问题,以电影评分预测为例,提出一种注意力机制和知识蒸馏的深度网络预测模型(Knowledge Distillation Attention Deep Network,K-ADN)。结合注意力网络区分交互特征的重要度而得到注意力值,利用深度神经网络(Deep Neural Networks,DNN)处理高阶特征组合,建立神经网络模型作为教师模型,从知识蒸馏技术出发,以教师模型确保精确度,以学生模型精简模型尺寸,以求获得更有效的评分预测结果。以豆瓣电影为数据来源进行的实验结果表明,该模型预测的精确度有所提高,通过知识蒸馏后参数量减少 86%。

关键词 电影评分 深度神经网络 注意力网络 评分预测 知识蒸馏

中图分类号 TP391 **文献标志码** A DOI:10.3969/j.issn.1000-386x.2024.07.009

MOVIE RATING PREDICTION BASED ON ATTENTION MECHANISM AND KNOWLEDGE DISTILLATION

Liu Tong Yu Sijie Ni Weijian*
(*College of Computer Science and Engineering, Shandong University of Science and Technology, Qingdao 266590, Shandong, China*)

Abstract In view of the high time complexity of high-order interaction and the large size of complex problem solved by neural network in the simple factorization machines (FM), a knowledge distillation attention deep network (K-ADN) model is proposed based on the movie score prediction as an example. Combined with the attention network to distinguish the importance of interactive features, the attention value was obtained. Deep neural networks (DNN) were used to deal with the combination of high-order features, and the neural network model was established as the teacher model. Starting from the knowledge distillation technology, the teacher model was used to ensure the accuracy, and the student model was used to simplify the model size, so as to obtain more effective scoring prediction results. The experimental results based on Douban movie show that the accuracy of the model is improved, and the parameters are reduced by 86% after knowledge distillation.

Keywords Movie rating Deep neural network Attention network Scoring prediction Knowledge distillation

0 引 言

当代人们的休闲娱乐方式多样,电影就是其中最主要、最简单的途径之一。电影业蓬勃发展,各类型电影争奇斗艳,即满足了观众的精神文化需求,又推动着经济的不断繁荣发展。然而,电影产业仍然受到各方面因素的限制,如电影自身制作团队的人员、技术配备、投资商的运营及市场或者大众的偏好等。多元化因素影响下的电影行业往往会出现高投资低评分的现象。为解决这一问题,可以根据已知的导演、编剧、演员等班底组成、电影类型等较为可靠的参考信息,通过预测的方式来分析该电影上映后的整体评分情况。

传统的评分预测算法,例如贝叶斯算法、协同过滤算法,虽极大地促进了信息推荐领域的发展,但仍有其局限性。机器学习在推荐系统中有很广泛的应用[1]。

收稿日期:2021 - 03 - 17。国家自然科学基金项目(71704096,61602278);青岛社科规划项目(QDSKL2001117)。**刘彤**,副教授,主研领域:机器学习与信息检索,数据挖掘,知识图谱。**于思洁**,硕士生。**倪维健**,教授。

薄玲玲[2]提出了基于因子分解机(FM)的电影推荐方法。基于因子分解机和聚类技术,利用已选用户的电影评级预测未被选择的用户对新电影的评级,实现对新电影的推荐。黄东晋等[3]提出基于混合特征的预测模型,采用支持向量机算法(Support Vector Machine, SVM)初步训练预测评分。将该评分作为一维新特征和电影特征信息一起通过随机森林算法训练预测最终评分。Kim 等[4]采用 DeepFM 模型处理 API 服务的多维 QoS 属性的交互层。提取包括共现次数在内的特征组合,预测 API 服务的质量得分。

近几年,以深度学习为代表的表征学习方法在自然语言处理(Natural Language Processing, NLP)[5]、图像分析(Image Analysis, IA)[6]和人脸识别(Face Recognition, FR)[7]等领域得到普遍关注。Basu[8]提出了基于 Levenberg-Marquardt 反向传播算法的电影分级预测方法。深度学习技术简化了数据集的数据分析、降维等过程,直接传递至网络,节约了大量的时间和避免资源的浪费。

谷歌机器翻译团队于 2017 年 6 月提出的自注意力(Attention)机制,近几年在众多领域都得到了广泛应用[9]。它可以有效地探索到用户与项目之间的非线性、非平凡关系,进一步了解用户需求和项目特点,从而更好地实现用户与项目之间的交互推荐。Wang 等[10]提出了一种将潜在因素模型与评论相结合的层次注意模型,利用潜在因子模型的因子向量来引导注意网络,并将因子向量与评论学习到的特征表示相结合,对重要词汇和信息性评论进行集中预测。

然而,电影行业突飞猛进的同时也积累了大量且丰富的电影数据,这就要求特征工程阶段更加简化、快速和精确。另一方面,深度学习模型也趋向复杂,网络深度逐步加深的同时,模型参数量也相应增加。知识蒸馏[11]为当下解决此类问题提供了一种研究方向。通过引入教师网络和学生网络的概念将复杂、学习能力强的教师网络中学习的知识蒸馏到简单、参数量小的学生网络。Saputra 等[12]通过注意模仿损失(Attentional Imitation Loss, AIL)和通过注意暗示训练(Attention Hint Training, AHT)方法学习教师的中间表征时,将此信心分数注入主要训练任务。

本文针对以上相关知识,提出集成知识蒸馏和注意力机制的电影评分预测模型(K-ADN)。该模型构建了结合注意力的深度网络模型,从教师模型角度提高预测性能,进而指导学生模型学习网络结构的中间层特征。同时,知识蒸馏兼顾学生模型规模与预测准确性,可以有效精简学生模型,同时提高评分预测的实时性。

1　模型设计

本文方法在深度网络的基础上引入注意力机制的思想构建教师网络,以浅层神经网络为学生模型。模型结构如图 1 所示。

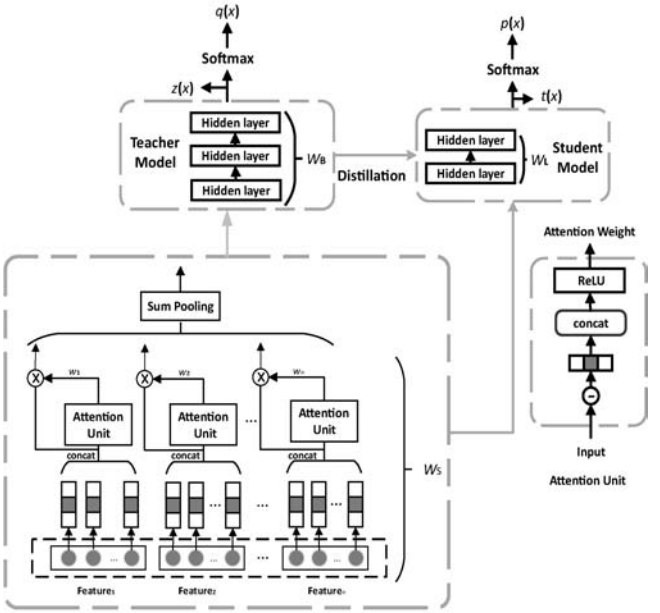

图 1　K-ADN 模型结构

在电影评分预测问题中,电影本身的特征是至关重要的,在建模电影表示时发挥着关键作用。K-ADN 模型不会通过使用相同的向量来表达所有电影的不同特征,而是通过局部激活单元,结合注意力网络区分交互特征的重要度而得到注意力值,进行自适应地调整电影表示。然后,建立神经网络模型作为教师模型,利用深度神经网络(DNN)处理高阶特征组合,最后从知识蒸馏技术出发,将教师模型的知识集成到一个简单的学生模型上,两部分同时训练。学生网络的整个学习过程由教师网络指导。学生网络不仅可以从目标输出中学习,而且可以从具有更多学习能力的复杂模型提供的训练过程中学习。让学生模型同时拟合预测结果和决策层的网络结构表示,缩小两个网络的性能差距,以教师模型确保精确度,以学生模型精简模型尺寸。

1.1　特征表示

通过编码将数据转化为高维稀疏二进制特征。在数学上,第 i 个特征组的编码向量被形式化为 $\boldsymbol{t}_i \in \mathbf{R}^{K_i}$, K_i 表示特征组 i 的维数。$\boldsymbol{t}_i[j]$ 表示第 j 个元素,且 $\boldsymbol{t}_i[j] \in \{0,1\}$, $\sum_{j=1}^{K_i} \boldsymbol{t}_i[j] = k$。$k=1$ 的向量 $\boldsymbol{t}_i$ 表示 one-hot 编码,$k>1$ 表示 multi-hot 编码。一个电影实例可以表示为 $\boldsymbol{x} = [\boldsymbol{t}_1^{\mathrm{T}}, \boldsymbol{t}_2^{\mathrm{T}}, \cdots, \boldsymbol{t}_M^{\mathrm{T}}]^{\mathrm{T}}$,其中 M 是特征组的数目, $\sum_{i=1}^{M} K_i =$

K, K 是整个特征空间的维数。由此,以电影类型和国家为例,电影特征实例可以表示为:

$$\underbrace{[0,\cdots,1,\cdots,1,\cdots,0]}_{T_{ype}=\{剧情,传记\}}\quad\underbrace{[0,\cdots,1,\cdots,0]}_{C_{ountry}=\{美国\}}$$

1.2 注意力网络

针对不同的电影,不同的属性特征的权重是不同的。例如科幻电影,可能更侧重于技术大国,剧情片可能更侧重于演员特征。同时,每部电影的同一特征包含的条目数量也不相同。因此,增加池化层(Pooling),Pooling 使用 sum polling 操作。以此得到固定长度的 Embedding 向量,抽象表示电影特征。

在 K-ADN 场景中,引入了一个新设计的局部激活单元,对于不同的电影来说,需要进行自适应地调整电影表示。在 Embedding 层到池化层得到 attention 表示时,给电影特征赋予不同的权重,完成局部激活。如式(1)所示。

$$\boldsymbol{v}_U(A)=f(\boldsymbol{v}_A,\boldsymbol{e}_1,\boldsymbol{e}_2,\cdots,\boldsymbol{e}_H)=\sum_{j=1}^{H}a(\boldsymbol{e}_j,v_A)\boldsymbol{e}_j=\sum_{j=1}^{H}w_j\boldsymbol{e}_j \tag{1}$$

式中:$\boldsymbol{e}_1,\boldsymbol{e}_2,\cdots,\boldsymbol{e}_H$是特征 U 的嵌入向量列表,长度为 H;$\boldsymbol{v}_A$是目标电影 A 的嵌入向量,$\boldsymbol{v}_U(A)$是经过局部激活后得到的目标电影向量表示;$a(\cdot)$是一个前馈网络,其输出得到激活权重;w_j是属性特征影响着每个电影评分的权重,通过 Activation Unit 计算得出。

使用 PReLU 激活函数[13],它只增加了极少量的参数,一方面并未大幅度增加网络的计算量;另一方面也控制了过拟合的可能。其表达式如式(2)所示。

$$f(y_i)=\begin{cases}y_i & y_i>0\\ a_iy_i & y_i\leqslant 0\end{cases} \tag{2}$$

式中:a_i为一个可学习的参数,用于控制 $y_i\leqslant 0$ 时激活函数的斜率。

采用自适应正则的方式防止过拟合。根据特征出现的频率,自适应地调整其正则化的强度:给予频率高的特征较小的正则化强度;反之,给予频率低的特征较大的正则化强度。

$$I_i=\begin{cases}1 & \exists(x_j,y_j)\in B,\text{s.t. }[x_j]_i\neq 0\\ 0 & 其他\end{cases} \tag{3}$$

$$w_i\leftarrow w_i\leftarrow\eta\Big[\frac{1}{b}\sum_{(x_j,y_j)\in B}^{H}\frac{\partial L(f(x),y_j)}{\partial w_i}+\lambda\frac{1}{n_i}w_iI_i\Big]$$

式中:B 表示小批量样本,大小为 b;n_i表示特征出现的频率;I_i是非零特征的样本。

1.3 知识蒸馏

知识蒸馏分为两部分:教师模型和学生模型。教师模型采用复杂的网络结构来实现电影评分的精确预测,教师网络的多层感知器(Muti-Layer Perception,MLP)隐层更深,神经元更多。但结构复杂、层数很深的深度模型不能很好地满足电影评分预测响应时间的限制,为了在减少响应时间的前提下还能保证很好的预测性能,用教师网络指导 MLP 隐层较浅、神经元数目较少的学生网络,两个网络共享底层特征嵌入层,但 MLP 部分参数是私有的。它们都有各自的特定层,用于对电影评分进行学习和预测,从而得到更好的训练效果。

分别用 $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 表示网络结构输入的电影特征和电影评分真值标签。令 L 表示轻量网络,即学生网络,参数由两部分组成:共享层中的参数 $\boldsymbol{W}_S$和用于预测的轻量级网络中的参数 $\boldsymbol{W}_L$。其 softmax 输出形式为:

$$p(\boldsymbol{x})=\text{softmax}(l(\boldsymbol{x})) \tag{4}$$

式中:$l(\boldsymbol{x})$表示学生网络中从输入到 softmax 之前的 logits 的映射。令 B 表示教师网络,包含底层共享参数 $\boldsymbol{W}_S$和特定的权重参数 $\boldsymbol{W}_B$,以获得最终的输出。其 softmax 输出形式为:

$$q(\boldsymbol{x})=\text{softmax}(z(\boldsymbol{x})) \tag{5}$$

式中:$z(x)$表示教师网络中从输入到 softmax 之前的 logits 的映射。

为了使学生网络的训练更接近于真实的电影评分标签 y,且更近似于教师网络学习的知识具有更多的表示能力,本文在训练目标中加入监督损失,以便将知识从教师网络传递到学生网络。此时,该模型的目标函数定义如下:

$$L(\boldsymbol{x};\boldsymbol{W}_S,\boldsymbol{W}_L,\boldsymbol{W}_B)=H(\boldsymbol{y},p(\boldsymbol{x}))+H(\boldsymbol{y},q(\boldsymbol{x}))+\lambda\|l(\boldsymbol{x})-z(\boldsymbol{x})\|^2 \tag{6}$$

式中:前两项 $H(m,n)=-\sum_i m_i\log n_i$ 为交叉熵损失,来学习真实的电影评分;第三项为监督损失;λ 是平衡交叉熵和监督损失的超参数。通过这种类似正则化的方法使得两个网络参数接近,达到知识传递的目的。

2 实验与结果分析

2.1 数据描述

本文数据来源于豆瓣网[14],选择的对象为上映时间跨度为 2007 年 1 月 1 日至 2019 年 12 月 31 日范围内的电影。通过网络爬虫进行爬取,在选择研究对象的同时处理上述范围内符合以下条件的电影:

(1) 重复上映的电影如《小森林》,分别于 2014 年 8 月 30 日和 2015 年 2 月 14 日上映。此类电影保留最新上映的信息。

(2) 类型为“歌舞”的范围内包含少量演唱会、春节联欢晚会等不属于电影范畴的条目,予以剔除。

(3) 缺少部分标签的电影,如 2017 年上映的《与死神共舞之夜》缺少主演信息,此类电影予以剔除。

(4) 当评价人数少于 30 时,网站不显示该电影评分。此类缺少评分项的电影予以删除。

最终共筛选得到 9 160 部电影及其信息,建立得到一个电影资料库。

部分电影数据如图 2 所示。图中序号为 8306 的电影 ID 为 2062678;电影名称为《对话尼克松》;导演姓名为朗·霍华德;编剧是皮特·摩根;演员有多个,包括弗兰克·兰格拉、麦克·辛和山姆·洛克威尔等人,在表格中用"|"分隔;电影类型为剧情片和传记片;制片国家为美国和英国;语言为英语;电影片长为 122 分钟;上映日期为 2008 年 12 月 5 日;豆瓣首页评分为 8 分;评价人数为 13 047 人次;影评数量为 143;短评数量为 2 148。

图 2　电影信息库部分示例

2.2　评价指标

为了评价预测模型的预测效果,采用平均绝对误差(MAE)和均方根绝对误差(RMSE)两个评价指标进行分析。公式如下:

$$L_{\mathrm{MAE}}^{\mathrm{K\text{-}DAN}} = \frac{1}{n}\sum_{i=1}^{n}|o_i - p_i| \tag{7}$$

$$L_{\mathrm{RMSE}}^{\mathrm{K\text{-}DAN}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(o_i - p_i)^2} \tag{8}$$

式中:o_i表示观测值(observation),即电影的真实评分;p_i表示预测值(prediction value),分别是电影的预测得分;n 是预测的电影的数量。

2.3　实验结果及分析

2.3.1　不同模型结果对比

本文实验平台搭建在 Ubuntu 上,深度网络模型基于 PyTorch 实现。蒸馏过程中学习速度设置为 1E-5,Adam 作为优化器,batch-size 为 32。教师模型全连接网络共三个隐藏层,隐藏节点数目分别为 1 024、512 和 256,学生模型两个隐藏层的隐藏节点分别为 256、64。在豆瓣数据集上训练 50 轮,每轮在验证集上验证 2 次,取验证结果最佳的模型在测试集上进行测试。

实验对比 SVM 模型、FM 模型、DeepFM 模型及改进后的 K-ADN 模型在豆瓣数据集上的运行结果,且增加消融实验对比分析注意力部分和知识蒸馏部分的结果。关于电影评分预测的问题实验结果如表 1、图 3 - 图 4 所示。

表 1　不同模型对比实验结果

指标	SVM	FM	DeepFM	AT	KD	K-ADN
MAE	1.112	0.801	0.703	0.652	0.695	0.611
RMSE	1.380	1.085	0.862	0.798	0.854	0.750

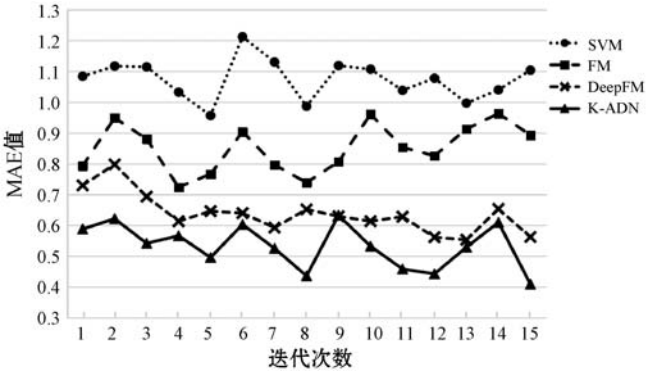

图 3　各模型的 MAE 指标对比

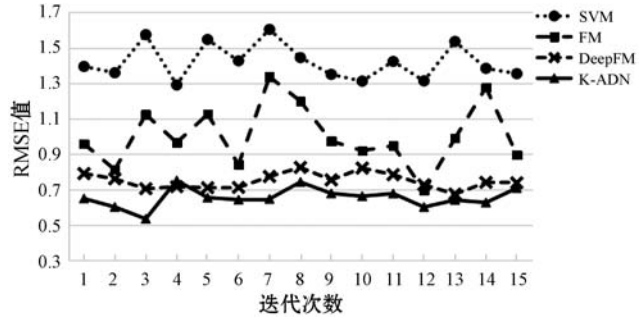

图 4　各模型的 RMSE 指标对比

表 1 展示了不同模型的评价指标均值,且加入消融实验对比仅改进注意力部分和仅改进蒸馏部分的结果,其中:AT 表示仅加入注意力的结果;KD 表示仅加入知识蒸馏的结果。可以看出,改进后的 K-ADN 算法的 MAE 值与 RMSE 值均明显低于其余基准模型,MAE 均值约为 0.618,较前者下降幅度为 12.09%,RMSE 均值约为 0.67,较前者下降幅度为 10.6%。AT 部分的结果较深度网络有所降低,直接对多个 Embedding 向量进行等权的 sum-pooling 会带来信息的丢失,且相对重要的 Embedding 向量也无法完全突出自己所包含的信息。本文采取了 weighted-sum pooling 让模型更加关注有用的信息,区分交互特征的重要度有助于对电影评分预测的建模。KD 部分结果与 Attention 部分相差无几,教师模型在特征 Embedding 层和 MLP 层间可加入不同方法的特征组合功能,体现出教师模型较强的模型表达和泛化能力,而蒸馏过程也增强 student 的泛化能力,使其更接近教师模型的效果。

图 3 和图 4 分别截取了 15 次迭代结果绘制成图,更为直观地展示了 K-ADN 模型与其他三种模型的对比效果。虽然模型中有个例结果逊于其他模型,但整体结果较其他模型有很大提升。

图 5 展示了新模型拟合的预测值与豆瓣分数之间

的对比,可以看出,各个电影的两分数之间差距较小。其中:《真相至上》的差值最大,为 0.84;《飓风营救 2》的差值最小,为 0.28,拟合效果较好。

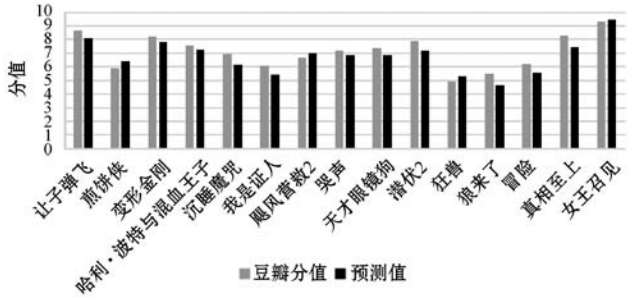

图 5 部分豆瓣电影分值对比

2.3.2 教师模型与学生模型实验对比

此部分实验以深度神经网络为基础构建教师模型,将其部分知识蒸馏到以轻量网络为基础的学生模型上,两部分同时训练,使得学生网络同时可以学到教师网络的最终输出和训练过程。从表 2 可知,学生模型大大降低了电影评分预测模型的大小,相比于教师模型参数量减少 86%,同时,运行时间也减少为原来的 24%,有利于在电影特征数据量少的情况下,用更轻量的网络模型进行评分预测。

表 2 教师模型与蒸馏后学生模型对比实验结果

模型	模型大小/KB	运行时间/s
教师模型	1 319 757	941
学生模型	184 765	225

3 结 语

本文从电影自身信息出发来预测电影客观评分,提出基于知识蒸馏和注意力机制的电影评分预测模型。通过对比三种基准算法与 K-ADN 算法且增加消融实验来检验评分预测模型的性能,以豆瓣电影网站提供的电影信息及评分作为数据来源。对 2007 年至 2019 年的 9 160 部不同类型的电影进行实证研究。结果显示,在评分预测模型中,K-ADN 改进模型表现出良好的效果,在保证准确率的同时也降低了参数量,可以有效用作预测数据源。

近年来,各种文本情感分析的技术愈发成熟,Xie 等[15]从电影情节入手,结合 NLP 算法来分析某电影受关注程度。未来工作可从情感分析方面着手,融入电影摘要信息,提取情感特征,侧面辅助更进一步精确地预测电影评分。

参 考 文 献

[1] Ni W J, Liu T, Zeng Q T, et al. Robust factorization machines for credit default prediction[C]//Pacific Rim International Conference on Artificial Intelligence, 2018: 941 - 953.

[2] 薄玲玲. 一种基于因子分解机和主动学习的新电影推荐方法[D]. 西安:西北大学,2018.

[3] 黄东晋,耿晓云,李娜,等. 基于混合特征的电影评分预测系统[J]. 计算机技术与发展,2020,30(12):136 - 141.

[4] Kim Y J, Cheong Y G, Lee J H. Prediction of a movie's success from plot summaries using deep learning models [C]//2nd Workshop on Storytelling, 2019: 127 - 135.

[5] Strubell E, Ganesh A, Mccallum A. Energy and policy considerations for deep learning in NLP [C]//57th Annual Meeting of the Association for Computational Linguistics, 2019: 53 - 55.

[6] Litjens G, Kooi T, Bejnordi B E, et al. A survey on deep learning in medical image analysis[J]. Medical Image Analysis, 2017, 42: 13 - 16.

[7] Sun Y, Wang X G, Tang X O, et al. Deep learning face representation by joint identification-verification [C]//27th International Conference on Neural Information Processing Systems, 2014: 1988 - 1996.

[8] Basu S. Movie rating prediction system based on opinion mining and artificial neural networks [C]//International Conference on Advanced Computing Networking and Informatics, 2019: 41 - 47.

[9] Liu T, Yin X R, Ni W J. Next basket recommendation model based on attribute-aware multi-level attention[J]. IEEE Access, 2020, 8: 153872 - 153880.

[10] Wang X C, Liu H T, Wang P Y, et al. Neural review rating prediction with hierarchical attentions and latent factors [C]//International Conference on Database Systems for Advanced Applications, 2019: 363 - 367.

[11] Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network[J]. Computer Science, 2015, 14(7): 38 - 39.

[12] Saputra M R, Gusmao P, Almalioglu Y, et al. Distilling knowledge from a deep pose regressor network[C]//EEE/CVF International Conference on Computer Vision, 2019: 263 - 272.

[13] He K M, Zhang X Y, Ren S Q, et al. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification[C]//IEEE International Conference on Computer Version, 2015: 1026 - 1034.

[14] 豆瓣电影网[EB/OL]. [2021 - 03 - 10]. https://movie.douban.com.

[15] Xie H, Wang H, Zhao C, et al. Movie score prediction model based on movie plots[C]//Data Science: 5th International Conference of Pioneering Computer Scientists, Engineers and Educators, 2019: 626 - 634.