计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications [doi: 10.15888/j.cnki.csa.009494]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

融合社交利益与图注意力网络的同伴互评分数预测①

杨　群, 訾玲玲, 丛　鑫

(重庆师范大学 计算机与信息科学学院, 重庆 401331)
通信作者: 訾玲玲, E-mail: lingling19812004@126.com

摘　要: 在同伴互评过程中, 评估者会因为战略性评估而导致评估分数不准确. 本文考虑了评估者之间的社交利益关系, 提出了一种融合社交利益与图注意力网络的同伴互评分数预测方法 GAT-SIROAN. 该方法由表示评估者与解决方案关系的加权网络 SIROAN 以及用来预测同伴互评分数的图注意力网络 GAT 构成. 在 SIROAN 中使用 ITSA 方法定义了评估者的两个特征: 自我评分能力和同伴评分能力, 并通过比较这两个特征来获取评估者之间的社交利益因子和关系. 在分数预测环节, 为了考虑每个节点的重要性, 使用自注意力机制来计算节点的注意力系数, 以此来提高预测能力. 采用最小化其均方根误差来学习网络的参数, 从而获取更准确的同伴互评预测分数. GAT-SIROAN 在真实数据集上与平均值、中位数、PeerRank、RankwithTA 以及 GCN-SOAN 这 5 个基线方法进行了对比实验, 结果表明 GAT-SIROAN 在 RMSE 指标上均优于基线方法.

关键词: 同伴互评; 社交利益; 加权网络; 图注意力网络; 分数预测

引用格式: 杨群,訾玲玲,丛鑫.融合社交利益与图注意力网络的同伴互评分数预测.计算机系统应用. http://www.c-s-a.org.cn/1003-3254/9494.html

Prediction of Peer Evaluation Scores by Integrating Social Benefits and Graph Attention Network

YANG Qun, ZI Ling-Ling, CONG Xin

(College of Computer and Information Science, Chongqing Normal University, Chongqing 401331, China)

Abstract: During peer evaluation, evaluators may give inaccurate evaluation scores as a result of strategic evaluation. Taking into account the evaluators' social interest (SI) relations, this study proposes a prediction method named graph attention network-social interest relation-oriented attention network (GAT-SIROAN) that integrates SI and the GAT. This method consists of a weighted network SIROAN that represents the evaluators' relations with the solutions and a GAT that is used to predict peer evaluation scores. In the SIROAN, the interrupted time-series analysis (ITSA) method is applied to define the evaluators' two characteristics: the self-evaluation ability and the peer evaluation ability, and these two characteristics are compared to obtain the SI factors and relations among the evaluators. In the score prediction stage, considering the importance of each node, this study uses a self-attention mechanism to calculate the attention coefficients at the nodes, thereby improving the prediction ability. Network parameters are learned by minimizing the root mean square error (RMSE) to obtain more accurate predicted peer evaluation scores. The GAT-SIROAN method is compared experimentally with five baseline methods, namely, the mean, median, PeerRank, RankwithTA, and GCN-SOIN methods, on real datasets. The results show that the GAT-SIROAN method outperforms all the above baseline methods in the RMSE.

Key words: peer evaluation; social benefits; weighted network; graph attention network (GAT); score prediction

① 基金项目: 重庆市教育科学规划重点课题 (K22YE205098); 重庆师范大学博士启动基金/人才引进项目 (21XLB030, 21XLB029)
收稿时间: 2023-11-07; 修改时间: 2023-12-11; 采用时间: 2024-01-05; csa 在线出版时间: 2024-04-01

在线教育课堂正逐渐兴起, 源于西方的 MOOC (massive open online courses) 也迅速遍布全球, MOOC 学习作为一种新型学习模式, 吸引了成千上万的学习者. 随着学习者人数的增多, 带来一个矛盾, 即庞大的学习者群体与紧缺的教师资源之间的矛盾[1]. 以 MOOCs edX 为例, 它曾经开设的“Circuits and Electronics” (6.002x) 课程有 155 000 名学习者注册, 教学团队仅有 12 名[2]. 由于学习者人数众多, 教师或助教无法为所有的解决方案进行打分, 因此迫切需要提出一种可扩展的替代传统学生评估的方法. 而同伴互评成为解决这一问题的实用解决方法, 即将解决方案分发给学习者的同伴, 由学习者的同伴进行解决方案的评估, 评估完成后将结果反馈给学习者. 同伴互评不仅减轻了教师批改解决方案的压力, 而且有助于加深学习者的理解[3].

同伴互评中的评估者通常根据课程教师提供的准则或基准对其他学习者提交的解决方案进行评估, 其中解决方案的最终得分是分数的某种聚合, 这种聚合通常是评估者给出的同伴互评分数的中值或平均值. 但由于评估者自身的知识储备不足以及缺乏准确评估他人的动机, 或是出于保护自身利益而做出的战略性评估 (如评估者为了获取更高的分数而恶意降低被评估者的评分或评估者受人际关系的影响给予被评估者不应有的高分等评估行为), 导致同伴互评的准确性和可靠性降低, 得出不公平或不准确的分数. 因此, 关于同伴互评的一个直接问题是同伴互评分数的准确性. 为了得到更准确的同伴互评分数, 许多学者对在线学习环境下同伴互评的真实分数估计展开了研究, 其中基于加权聚合的估计方法常用于评估者给予解决方案的反馈内容为分数的情况. 具体来说, 基于加权聚合的估计方法会因为评估者的评估准确性和可靠性去补偿在同伴互评过程中引入的故意或无意的偏见, 从而估计得出同伴互评的真实分数.

现有的基于加权聚合的估计方法大多基于这样的假设, 评估者的评估准确性通过评估者评估或提交的解决方案的质量来衡量. 而衡量评估者的评估准确性的方法主要分为两类, 一类是仅通过学生评估者自身提交或评估的解决方案来进行衡量, 没有使用教师或助教评估者评估的解决方案作为对比[4–6]; 另一类则是使用了教师或助教评估者评估的解决方案作为对比[7,8]来进行衡量.

然而, 在上述关于评估者评估准确性的考虑中仍然存在一些限制, 现有的考虑评估者评估准确性的因素大多与评估者的评估行为有关, 缺少对同伴互评过程中由于社交利益而存在的战略性评估的研究. 例如, 评估者对自身解决方案的评估与对同伴解决方案的评估之间的差异会在同伴互评过程中引入偏见, 而这种偏见并没有被考虑到影响评估者评估准确性的因素中去.

文献 [9]提出了一种由社会所有权评估网络 SOAN (social-ownership-assessment network) 和图卷积网络 GCN (graph convolution network) 构成的半监督聚合方法 GCN-SOAN. 在该机制中没有任何特定的或限制性的归纳偏见, 并且适应于各种模式的同伴互评. SOAN 中存在评估关系、社会关系以及所有权关系. 评估关系与所有权关系可以表示为节点类型为评估者与解决方案的加权图, 在评估关系中评估者与解决方案之间的边为同伴互评的分数, 在所有权关系中评估者与解决方案之间的边为评估者评估和解决的解决方案的数量. 社会关系则表现为节点类型为评估者的单类型图, 评估者之间的边展现了评估者之间的社交利益关系, 但并没有给出衡量社交利益关系的方法. 文中将在图卷积的过程中聚合邻居节点信息的操作视为同伴互评分数的加权聚合, 使用教师评估作为基本事实, 通过最小化其均方根误差来学习网络的参数, 以此来预测同伴互评的分数. 但 GCN 在聚合邻居节点时使用固定的邻居权重进行信息聚合, 无法有效地区分不同邻居节点的重要性.

本文在文献 [9]的工作上提出了 ITSA (is there a strategic assessment) 方法用于获取社交利益关系, 在 ITSA 方法中定义了评估者具有两个特征: 自我评分能力以及同伴评分能力, 通过比较两者的差异获得社交利益关系和社交利益因子, 创建了社交利益关系所有权评估网络 SIROAN (social interest relationship ownership assessment network), 同伴评分能力的计算方法则参考了文献 [8]中评估者的评分能力的计算方法. 并引入了图注意力网络 (graph attention network, GAT)[10], 在邻域聚合过程中使用自适应的注意力机制来计算节点的注意力系数, 允许每个节点对其邻居节点分配不同的权重, 考虑每个节点的重要性, 以此来聚合同伴互评的分数. 使用教师评分成绩作为基本事实, 通过最小化其均方根误差学习网络的参数, 来预测同伴互评分数. 我们的方法考虑了评估者之间的社交利益关系, 考虑了评估者如何相互影响和相互作用, 在一定程度上阻止了评估

者之间的战略性评估. 并且在聚合过程中考虑了注意力系数, 每个节点可以根据与其他节点的关系进行不同程度的注意, 使得在聚合过程中更加关注相关节点的特征, 降低对无关节点的依赖, 从而提高了预测能力, 能够得到更准确的同伴互评预测分数.

1 相关工作

1.1 同伴互评的真实分数估计

在同伴互评的真实分数估计方面已有许多研究工作, 根据评估者对解决方案评估内容的是质量的排名还是分数的高低, 同伴互评的真实分数估计方法可以分为序数 (ordinal) 估计方法和基数 (cardinal) 估计方法. 在序数估计方法中, 评估者对解决方案的评估内容是质量高低的排名, 主要利用矩阵分解[11]、模糊决策[12]、贝叶斯[13]等方法来估计同伴互评解决方案的质量. 在基数估计方法中, 评估者对解决方案的评估内容则是分数. 当前流行的基数估计方法有基于概率图模型的估计方法以及基于加权聚合的估计方法. 基于概率图模型的估计方法的核心思想是把与解决方案存在一定联系的真实分数、评估者的可靠性及偏见、互评分数视作显性或隐性的随机变量, 并建模为服从一定概率分布的模型, 然后基于显性的互评分数来推断隐性随机变量的值. 现有的概率图模型有 PG1–PG7[14–16]. 基于加权聚合的估计方法定义了评估者的准确性和信任度, 并根据它们的差异赋予不同的权重, 然后聚合评估者对同一解决方案的真实分数[17]. 基于加权聚合的估计方法将在第 1.2 节进行详细阐述.

1.2 基于加权聚合的估计方法

在基于加权聚合的估计方法中评估者的准确性和信任度与其提交或评估的解决方案质量有关. 例如, 文献 [4]提出的 Voncouver 算法, 该算法将不同评估者对同一份解决方案的评估差异视作评估者的评分准确性, 然后赋予准确性高的评估者更高的权重, 最后加权聚合得到该解决方案的一致分数. 文献 [6]则假设评估者提交的解决方案与其评估能力有关, 受到 Google 的网页排序 PageRank 算法[18]的启发提出了另一种迭代加权算法 PeerRank[6], 对每一份提交解决方案的多个同伴互评分数进行加权求和. 文献 [5]则将评估者的学习参与度视作影响评估解决方案真实分数准确性的因素, 赋予学习参与度高的评估者更高的权重. 而这种参与度只能通过线上课堂来获取, 对于传统课堂可以采用电子签到的形式来获取评估者的学习参与度[19]. 文献 [7]则在 PeerRank 基础上提出了 RankwithTA, RankwithTA 使每个学习者获得的分数取决于他们提交的解决方案的质量, 以及他们的审查和评估工作的质量, 以激励评估者正确评分, 并且通过由教师或助教对一些解决方案的评估来进行外部校准, 以提供准确性的基础. 在 SSPA 方法中, 文献 [8]使用了教师评估的分数作为基本事实, 提出了通过对比学生与教师的直接相似性 (如果学生与教师存在至少两份共同的评估) 与间接相似性 (如果学生与教师没有共同的评估) 来衡量学生的评分能力, 最后加权聚合同伴互评分数来推断出每个学生的最后得分. 文献 [20]提出了一种基于图的信任传播方法, 该方法把评估者、解决方案视作节点, 评估关系视作边, 将其建模为节点权重为评估者可靠性以及解决方案质量的加权图. 然后提出了基于图结构的解决方案分数更新策略以及评估者可靠性的传播策略, 以此推断解决方案真实分数以及评估者的评分可靠性. 此外, 文献 [21]把学生的学习行为以及评语视作评估者的可靠性, 然后量化学生的评分可靠性作为权值, 对他们给出的评估分数进行加权聚合, 最后得到解决方案的真实分数. 文献 [22]则使用了教师或助教对历史解决方案的评估纠正学生的偏见, 再从评语内容中提取学生的仔细度, 学生的评估仔细度越高, 其给出的评估越值得信赖, 则给该学生所打的分数赋予更高的权重以提升对解决方案真实分数估计的准确性. 文献 [9]将评估者与解决方案建模为图表示模型 SOAN, 并引入了图卷积神经网络来预测同伴互评分数. SOAN 模型为更广泛的同伴互评图神经网络方法的研究提供了根基, 本文在 SOAN 的基础上考虑了评估者之间的社交利益与其解决方案的相关性, 提出了衡量同伴互评之间社交利益关系的方法 ITSA, 构建了 SIROAN, 使用 GAT 来预测同伴互评分数, 并将该方法称为 GAT-SIROAN.

2 GAT-SIROAN 方法

本文提出的 GAT-SIROAN 方法用于预测同伴互评分数, GAT-SIROAN 由多关系加权网络 SIROAN 以及图注意力网络 GAT 两个模块组成. GAT-SIROAN 的目标是从嘈杂的同伴互评中预测基本事实评估.

GAT-SIROAN 方法如图 1 所示, 分为多关系网络 SIROAN 的构建和同伴互评分数预测两个阶段.

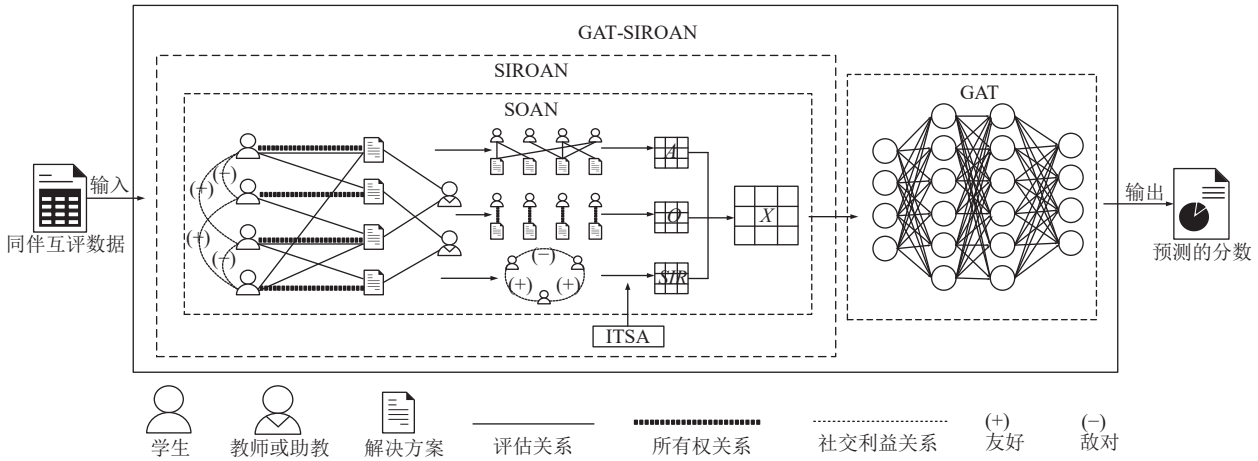

图 1 GAT-SIROAN 方法图

多关系网络 SIROAN 的构建: 给定同伴互评数据提取出评估者与解决方案之间的评估关系、所有权关系以及评估者与评估者之间的社交利益关系. 将提取出来关系转换为评估矩阵 A、所有权矩阵 O、社交利益关系矩阵 SIR, 由这 3 个关系矩阵初始化矩阵 $X=\begin{bmatrix} SIR & O+A \\ (O+A)^{\mathrm{T}} & 0_m \end{bmatrix}+I$, 作为第 2 阶段同伴互评分数预测的输入, 其中 T 为转置运算符, 0_m为$m\times m$值为 0 的矩阵.

同伴互评分数预测: 将节点特征初始化为十维全为 1 的向量, 同时将 X 输入到 GAT 中进行同伴互评分数的预测. 预测过程中在节点执行自我注意力机制去考虑每个节点的重要性, 最终获得同伴互评的预测分数.

SIROAN 由评估关系、所有权关系以及社交利益关系构成. 评估关系由采用评估分数构成评估矩阵 A 表示. 所有权关系使用所有权矩阵 O 表示, 该矩阵反映了解决方案所有权的分配情况, 其中解决方案的数量通过提交或评估来确定. 社交利益关系由社交利益关系矩阵 SIR 来表示, 采用社交利益因子构成, 其中社交利益因子由 ITSA 方法计算得出.

2.1 本文的主要符号及含义

本文的主要符号及其含义如表 1 所示.

评估矩阵 A: 一组有 n 个学生, 不同类型的解决方案被随机分配给任意的学生, 学生对解决方案的评估分数则收集在评估矩阵 A 中. 其中$A_{I_{g\text{-}t}\leftarrow\alpha}$表示学生$\alpha$对来自 g 组类型为 t 的解决方案的评估分数. $I_{g\text{-}t}$表示解决方案来自 g 组且类型为 t. 教师或助教对学生提交的解决方案的子集进行评估, $A_{I_{g\text{-}t}\leftarrow ta}$表示教师或助教 ta 对来自 g 组类型为 t 的解决方案的评估分数. 评估矩阵可以建模为有向加权图, 节点表示评估者以及解决方案, 有向边具有权值$A_{I_{g\text{-}t}\leftarrow\alpha}$表示评估者$\alpha$对来自 g 组类型为 t 的解决方案$I_{g\text{-}t}\leftarrow\alpha$的评估分数, I表示解决方案的集合, P表示评估者的集合包括教师助教以及学生, T表示评估者为教师或助教的集合, ξ表示评估者为学生的集合, 其中$T\subset P$, $\xi\subset P$, $T\not\subset\xi$, $\xi\not\subset T$.

表 1 本文的主要符号及其含义

符号	概念
A	评估矩阵
I	解决方案集合
P	评估者的集合
T	评估者为教师或助教的集合
ξ	评估者为学生的集合
g	分组
t	解决方案类型
$I_{g\text{-}t}$	解决方案来自g组且类型为t
$A_{I_{g\text{-}t}\leftarrow\alpha}$	学生α对来自g组类型为t的解决方案的评分
$A_{I_{g\text{-}t}\leftarrow ta}$	助教ta对来自g组的类型为t的解决方案的评分
ta	教师或助教
SIR	社交利益关系矩阵
sir	社交利益因子
O	所有权矩阵
A_T	助教评分集合
A_ξ	学生评分集合
SA	自我评分能力
PA	同伴评分能力
S_D	直接相似性
S_I	间接相似性
Q	相似性链

所有权矩阵 O: 在所有权矩阵 O 中, 使用 1 表示某一解决方案是由评估者α提交或评估的, 若是评估者α没有提交或评估某一解决方案则用 0 表示. 例如评估

者α提交了解决方案$I_{g1\text{-}t1}$则用 1 表示, 评估者α评估了解决方案$I_{g2\text{-}t2}$用 1 表示, 若是评估者α没有评估解决方案$I_{g2\text{-}t2}$则用 0 表示.

社交利益关系矩阵 *SIR*: *SIR* 社交利益关系矩阵收集了同伴互评之间的社交利益因子, 而社交利益因子则使用本文提出的 ITSA 方法计算得出.

评估矩阵, 所有权矩阵以及社交利益关系矩阵都从 SIROAN 提取出.

2.2 ITSA 方法

ITSA 的基本思想如下: 每个评估者拥有自我评分能力 (即评估自己解决方案的能力); 以及同伴评分能力 (即评估他人解决方案的能力). 自我评分能力被表征为每个评估者对解决方案评估的认知度, 同伴评分能力则被表征为每个评估者对被评估者解决方案的认可度. 通过比较自我评分能力和同伴评分能力得到社交利益因子和社交利益关系, ITSA 方法如图 2 所示.

每个学生的自我评分能力通过自我评估分数与教师或助教对该解决方案的评估分数进行相似度比较得出, 这种相似性通过皮尔逊系数计算 (如果评估者自我评估的解决方案与教师评估的解决方案大于 2 份) 或余弦相似性计算 (如果评估者自我评估的解决方案与教师评估的解决方案小于等于 2 份), 如式 (1) 所示:

$$SA=\begin{cases}\dfrac{\sum\limits_{I}\left(A^{\alpha}_{I_{g\text{-}t}\leftarrow\alpha}-\bar{A}_I\right)\left(A_{I_{g\text{-}t}\leftarrow T}-\bar{A}_I\right)}{\sum\limits_{I}\sqrt{\left(A^{\alpha}_{I_{g\text{-}t}\leftarrow\alpha}-\bar{A}_I\right)^2}\sqrt{\left(A_{Ig\text{-}t\leftarrow T}-\bar{A}_I\right)^2}}, & I\in(2,+\infty)\\ \dfrac{\sum\limits_{I}\left(A^{\alpha}_{Ig\text{-}t\leftarrow\alpha}A_{Ig\text{-}t\leftarrow T}\right)}{\sqrt{\sum\limits_{I}\left(A^{\alpha}_{I_{g\text{-}t}\leftarrow\alpha}-\bar{A}_I\right)^2}\sqrt{\sum\limits_{I}\left(A_{I_{g\text{-}t}\leftarrow T}-\bar{A}_I\right)^2}}, & I\in(0,2]\end{cases} \tag{1}$$

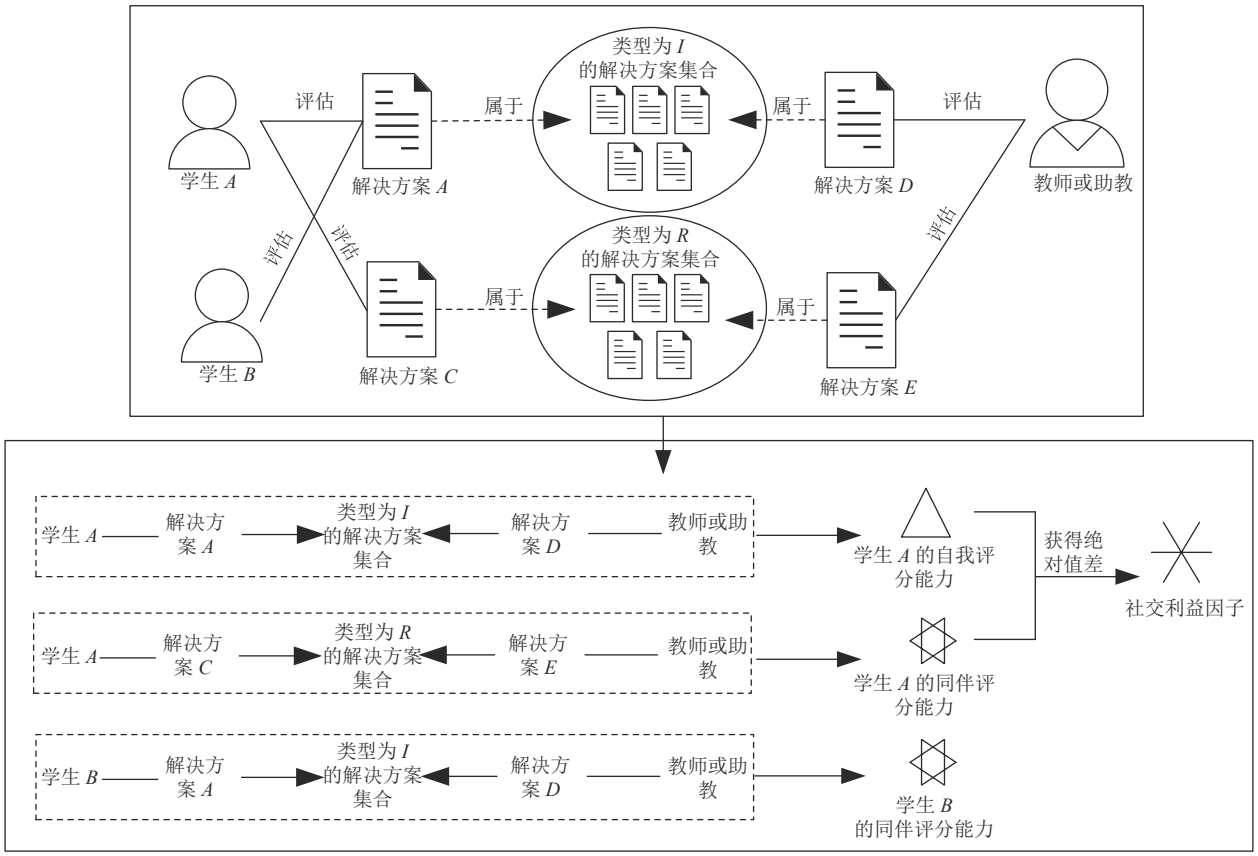

图 2 ITSA 方法

而同伴评分能力则参考了文献 [8]所提出的 SSPA 方案中衡量评估者评分能力的方法, 通过学生和教师或助教之间的相似性来衡量, 这种相似性是直接的 (学生与教师或助教评估了同一解决方案) 或间接的 (学生与教师或助教没有评估同一解决方案), 如式 (2) 所示:

$$PA=\begin{cases}S_D(\alpha,\beta)\\ S_I(\alpha,T)\end{cases} \tag{2}$$

直接相似性是指学生与教师或助教至少评估了两

份共同的解决方案, 相似值是他们评估在共同作业中的相似性的平均值. 如式 (3) 所示:

$$S_D = \sum_N sim(A^p_{I_{g\text{-}t}\leftarrow\alpha}, A^p_{I_{g\text{-}t}\leftarrow\beta})/N_{\alpha,\beta} \tag{3}$$

其中, $N_{\alpha,\beta}$表示学生α与学生β评估的解决方案的集合.

$$sim(A^p_{I_{g\text{-}t}\leftarrow\alpha}, A^p_{I_{g\text{-}t}\leftarrow\beta}) = 1 - |A^p_{I_{g\text{-}t}\leftarrow\alpha} - A^p_{I_{g\text{-}t}\leftarrow\beta}|/\lambda \tag{4}$$

间接相似性指学生与教师或助教之间没有共同评估的解决方案, 考虑一系列的评估者来计算这种相似性, 将其作为一种传递性度量, 每个相似性链 Q 都有直接相似性. 间接相似性定义为:

$$S_I = \max_{q\in Q(\alpha,t)} \prod S_D(\alpha,\beta) \tag{5}$$

通过对自我评分能力与同伴评分能力的比较得到该评估者对被评估者的社交利益关系, 社交利益因子则定义为自我评分能力绝对值与同伴评分能力绝对值的差值, 如式 (6) 所示:

$$sir = |SA| - |PA| \tag{6}$$

将计算得出的 sir 收集在 SIR 中.

2.3 图注意力网络

通过 ITSA 方法完成 3 个关系矩阵的构建, 然后初始化矩阵 X, 将 X 矩阵和节点特征$\vec{h}$输入到 GAT 中, $\vec{h}$被初始化为 10 维全为 1 的向量.

为了获得足够的表达能力, 将输入特征转化为更高层次的特征, 将节点特征应用一次权重化矩阵W, 作一次可学习的线性变化得到$W\vec{h}$. 考虑到在领域聚合时每个节点重要性的不同, 在节点上执行自我注意力机制, 将更高层次的特征$W\vec{h}$与一个可学习的参数矩阵α相乘, 来计算注意力系数$e_{I_{g\text{-}t}C}$ ($e_{I_{g\text{-}t}C}$表示作业节点$I_{g\text{-}t}$与评估者节点之间的注意力系数). 将计算得出的注意力系数经过一个 Softmax 函数进行归一化操作得到最后的结果$\alpha_{I_{g\text{-}t}C}$, 最后通过激活函数得到评估者节点的最终特征表示$W\vec{h}'$.

图 3 给出了计算获得评估者 C 节点新特征的过程, 评估者节点 C 与作业节点$I_{g1\text{-}t1}$相连接, 节点$I_{g1\text{-}t1}$与节点 C 经过一次可学习的线性变化后的特征为: $W\vec{h}_{I_{g1\text{-}t1}}$、$W\vec{h}_C$. 将变化后的特征与一个可学习的参数矩阵$\alpha$相乘后得到注意力系数: $e_{I_{g1\text{-}t1}C}$, 通过 Softmax 函数进行归一化操作后得到: $\alpha_{I_{g1\text{-}t1}C}$. 最后通过一个激活函数后得到节点 C 的最终特征$W\vec{h}'_C$.

本文将通过最小化其预测的均方根误差来学习 GAT-SIROAN 的参数:

$$L(\theta|\varsigma, D) = \frac{1}{|D|}\sum_{I_{g\text{-}t}\leftarrow\xi=1}^{|D|} (A_\xi - f(I_{g\text{-}t}\leftarrow\xi = 1|\theta,\varsigma))^2 \tag{7}$$

其中, $|D|$是训练数据集中的项数, $f(I_{g\text{-}t}\leftarrow\xi|\theta,\varsigma)$是解决方案$I_{g\text{-}t}\leftarrow\xi$的 GAT-SIROAN 的估值. 可以通过梯度下降的优化技术来最小化损失函数.

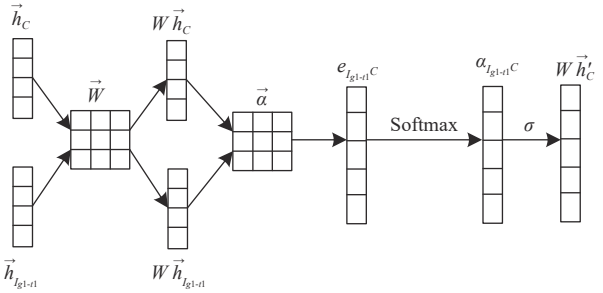

图 3 通过注意力机制得到的评估者 C 节点新的特征

本文采用了自注意力机制来计算节点注意力系数, 考虑每个节点的重要性, 以此来聚合同伴互评的成绩, 同样使用教师评分成绩作为基本事实, 通过最小化其均方根误差来学习网络的参数, 最终预测解决方案的同伴互评分数, GAT-SIROAN 阻止了评估者之间的战略性评估, 并且在聚合过程中考虑的注意力系数, 提高了预测能力, 能够得到更准确的评分.

3 实验分析

3.1 数据集

本文采用了文献 [23]提供的同伴互评数据集, 其中包含了 219 名学生在 6 种类型解决方案上的同伴互评分数、自我评估分数以及教师或助教评估的分数. 其中每种类型解决方案包括 2–5 个练习题, 本文选择了练习题数量大于 2 的解决方案类型, 一共有 4 个类型的解决方案, 即类型 1、类型 2、类型 3、类型 4. 所有学生被分成 79 组, 每组学生人数不同, 组内学生数量最多 3 人. 此外, 所有解决方案都由 6 名助教中的一名评估. 学生和助教都被提供了详细评估标准, 包括需要注意的错误以及如何分配分数. 将所有解决方案视为集合I, 并将解决方案的集合细分为来自哪个组的哪种类型的解决方案即$I_{g\text{-}t}$, 其中$I_{g\text{-}t}\in I$. 学生评估者视为集合ξ, 助教评估者视为集合T, 所有评估者视为集合P, 其中$\xi\subset P$, $T\subset P$. 与文献 [9]一样, 将同伴互评分数、自我评估分数以及教师或助教评估的分数

定在了 [0, 1]的范围内, 分别在 4 个类型的解决方案上进行了实验, 在实验过程中使用了评估关系以及社交利益关系, 并且区分了是否添加自我评估分数. 这些数据的结果汇总统计如表 2 所示.

表 2 真实世界同伴评分数据集的汇总统计

汇总项	类型1	类型2	类型3	类型4
解决方案数	225	308	380	237
同伴互评数量	965	620	1 889	1 133
自我评分数量	469	755	890	531

3.2 实验设置与评价指标

本文基于 PyTorch[24]和 PyTorch geometry[25]实现了 GAT-SIROAN. 在所有的实验中, 使用了两个嵌入层, 嵌入维数为 64, 并没有使用多头注意力机制, 使用了单头注意力机制. 将指数线性单元 (ELU) 作为所有隐层的激活函数. 使用 Adam 优化器[26]和 0.02 的学习速率训练了 1 500 个周期. 将值为 1 的向量初始化节点嵌入. 蒙特卡罗交叉验证[27]用于验证, 训练-测试分割比为 1:4, 即有 20% 的数据用于训练, 其余用于测试. 为了使结果更加可靠, 分别在 4 次随机分割上运行 GAT-SIROAN, 并报告这些分割的平均误差.

为了评估预测性能, 本文使用的同伴互评真实分数估计中常用的评价指标均方根误差 RMSE (root mean square error). RMSE 是一种常见的用于衡量预测值与真实值之间差异的指标, 它是将预测误差的平方和取平均后开方得到的. RMSE 可以用来评估模型的性能, 其中较小的 RMSE 值表示模型的预测较准确, 而较大的 RMSE 值则表示预测误差较大.

3.3 参数分析

网络层数与预测性能之间的关系是一个复杂且依赖于特定任务和数据集的问题, 一般来说, 增加网络的层数可以提供更多的非线性表示能力, 从而有可能提高模型的预测性能. 因此, 本文对在不同网络层数下模型所获得的性能进行了对比实验. 如图 4 所示, 本文选择了 RMSE 指标来显示 GAT-SIROAN 的性能是如何随网络层数的数量变化而变化的. 可以看到随着层数的增加均方根误差并没有太大的变化. 而产生这个现象的原因与 GAT 模型的中的“过渡平滑”问题有关[28].

在 X 矩阵的初始化中, 除了本文提出的使用 ITSA 得出的社交利益关系矩阵 *SIR* 外, 所有权矩阵 O 也是一个重要的参数. 因此本文还探讨了所有权矩阵 O 的加入是否能进一步提高 GAT-SIROAN 的性能, 如图 5 所示, 所有权矩阵 O 的加入, 模型在各个类型的解决方案上的 RMSE 提高, 性能降低. 当 X 矩阵中只存在评估矩阵 A 与社交利益关系矩阵 *SIR* 时, 模型性能最佳, 进一步说明了 ITSA 方法的有效性.

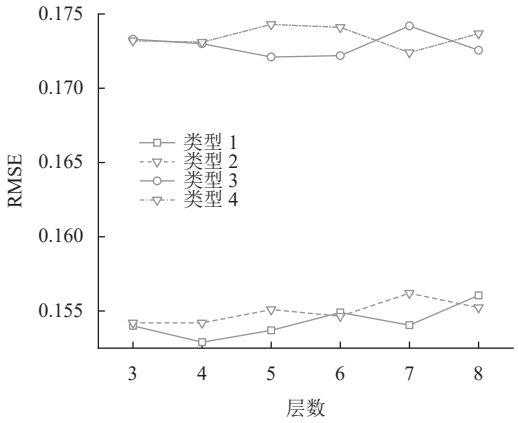

图 4 GAT-SIROAN 在不同网络层数下性能的展示

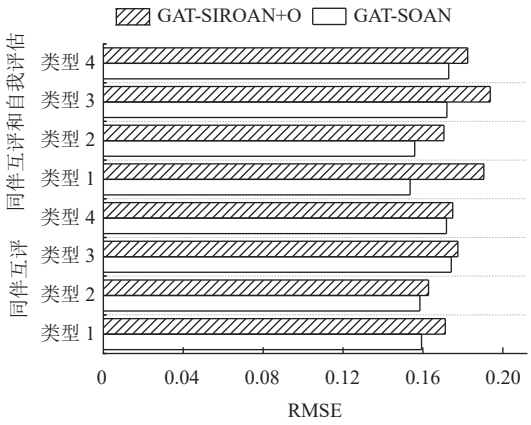

图 5 GAT-SIROAN 在是否添加所有权矩阵 O 后的性能展示

3.4 消融实验

为了验证本文提出 GAT-SIROAN 方法中各模块的有效性, 采用以下两种消融方案进行实验: (1) 使用 SIROAN 与两层 GCN 结合; (2) 使用 SOAN 与两层 GAT 结合. 评价指标采用均方根误差 RMSE.

如图 6 所示, 对于不同类型解决方案的数据集, 使用 SIROAN 在经过两层 GCN 运行后, 很明显 GCN-SIROAN 在 8 个数据集上的 RMSE 都低于 5 个基线方法. GCN-SIROAN 在 4 类解决方案中, 仅有同伴互评的情况下, 性能上比表现最好的 GCN-SOAN 提高了 0.97%、0.47%、0.33%、0.6%. 在添加了自我评分后, 在性能上提高了 1.15%、0.5%、0.11%、0.25%. 说明加入 ITSA 方法后 SIROAN 能在一定程度上提升同伴

互评分数预测的准确性. 图 7 展示了使用 SOAN 在经过两层 GAT 后在数据集上运行的结果. 仅存在同伴互评的情况下, GAT-SOAN 在类型 1、类型 3、类型 4 这 3 个数据集上的 RMSE 均比其他基线方法低, 在性能上提高了 1.06%、0.48%、0.21%. 在添加了自我评估后, GAT-SOAN 在类型 1、类型 3 两个数据集上的 RMSE 都低于其他基线方法, 在性能上提高了 0.21%、0.4%. GAT 也在一定程度上提高了预测能力.

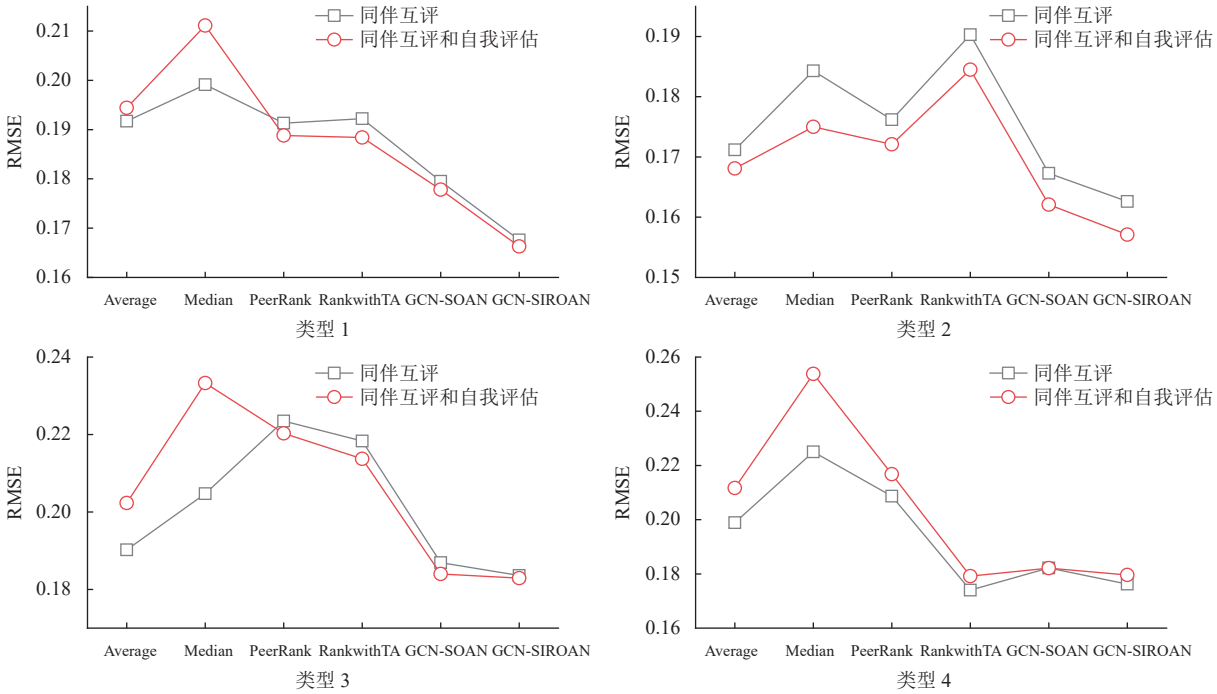

图 6 SIROAN 经过两层 GCN 后在真实数据集上运行的结果

3.5 基线实验

为了证明 GAT-SIROAN 优势, 将 GAT-SIROAN 与 5 种基线方法进行了比较, 使用了两个无监督的同伴互评方法: 平均值 (所有同伴互评分数的平均值) 以及中位数 (所有同伴互评分数的中位数), 使用了两个机器学习的同伴互评方法: 没有使用教师或助教评估的分数作为校准的 PeerRank[6]以及使用了教师或助教评估的分数作为校准的 RankwithTA[7], 还使用了深度学习的同伴互评方法 GCN-SOAN[9].

表 3 展示了 GAT-SIROAN 以及其他 5 个基线方法平均值、中位数、PeerRank[6]、RankWithTA[7]、GCN-SOAN[9]的性能. 如表 3 所示, 在无监督方法中平均值表现得最好, 相较于中位数, 在 4 类解决方案中平均值的 RMSE 都低于中位数, 性能高于中位数. 而 GAT-SIROAN 在 4 类解决方案中, 不论是否添加自我评估的分数, RMSE 都低于平均值, 性能上分别提升了 3.24%、1.31%、1.6%、2.71%、4.08%、1.22%、3.03%、3.92%. 在机器学习方法中, GAT-SIROAN 在仅有同伴互评分数的情况下, 在类型 1、类型 2 两个数据集上的 RMSE 都低于表现最好的 PeerRank[6], 在性能上分别提升了 3.2%、1.81%, 在类型 3、类型 4 两个数据集上的 RMSE 低于 RankwithTA[7], 性能上分别提升了 4.41%、0.22%. 添加了自我评估分数后, 在类型 1、类型 3、类型 4 这 3 个数据集上的 RMSE 低于 RankwithTA[7], 性能上提高了 3.48%、4.17%、0.67%, 在类型 2 上的 RMSE 低于 PeerRank[6], 性能提高了 1.62%. 在深度学习方法中, 不论是否添加自我评估分数, 在 4 类解决方案上 GAT-SIROAN 的 RMSE 都比 GCN-SOAN[9]低, 性能上分别提升了 2.02%、0.92%、1.27%、1.04%、2.42%、0.62%、1.2%、0.92%. 在所有基线中, GAT-SIROAN 在仅有同伴互评分数的情况下, 在类型 1、类型 2、类型 3 这 3 个数据集上的 RMSE 比基线方法中表现最优的 GCN-SOAN[9]低, 性能上提高了 2.02% 、0.92%、1.27%, 在类型 4 数据集上的 RMSE 比基线方法中表现最优的 RankwithTA[7]低, 性能上提高了 0.22%. 添加了自我评估分数后, 在类型 1、类型 2、类型 3 这 3 个数据集上的 RMSE 比基线方法中表现最优的 GCN-SOAN[9]低, 性能上提升了 2.42%、0.62%、

1.2%, 在类型 4 数据集上的 RMSE 比基线方法中表现最优的 RankwithTA[7]低, 性能上提升了 0.67%. 这些结果表明, 与这些基线方法相比, 本文提出的 GAT-SIROAN 方法在同伴互评真实分数估计的评估指标方面取得了最好的性能. 使用 ITSA 方法改进后的 SIROAN 在一定程度上阻止了评估者的战略性评估, 而在聚合过程中使用自适应的注意力机制能更好地捕捉节点之间的关系, 能使预测分数更接近基本事实.

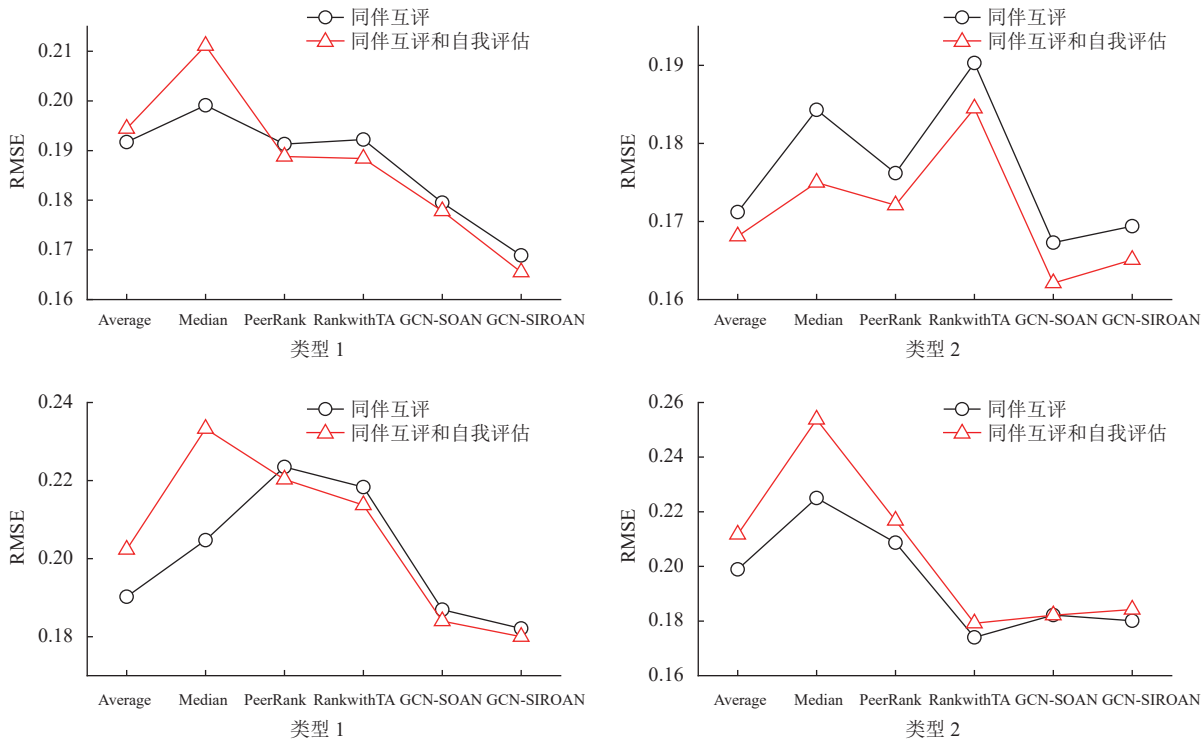

图 7 SOAN 经过两层 GAT 后在真实数据集上运行的结果

表 3 在真实数据集上的各种方法的均方根误差

方法	同伴互评				同伴互评和自我评估			
	类型1	类型2	类型3	类型4	类型1	类型2	类型3	类型4
平均值	0.1917	0.1712	0.1902	0.1989	0.1944	0.1681	0.2023	0.2117
中位数	0.1991	0.1843	0.2047	0.2250	0.2111	0.1750	0.2333	0.2538
PeerRank	0.1913	0.1762	0.2235	0.2087	0.1888	0.1721	0.2203	0.2168
RankwithTA	0.1922	0.1903	0.2183	0.1740	0.1884	0.1845	0.2137	0.1792
GCN-SOAN	0.1795	0.1673	0.1869	0.1822	0.1778	0.1621	0.1840	0.1821
GAT-SIROAN	0.1593	0.1581	0.1742	0.1718	0.1536	0.1559	0.1720	0.1725

4 结论与展望

在本文中, 提出了一种融合社交利益与图注意力网络的同伴互评分数预测方法 GAT-SIROAN, 该方法考虑了同伴互评过程中的战略性评估行为, 提出了获取同伴互评过程中社交利益关系的方法 ITSA, ITSA 通过对比自我评分能力以及同伴评分能力计算得出了社交利益因子从而得出互评过程中的社交利益关系, 以此创建了社交利益关系矩阵 *SIR*. 并将评估关系、所有权关系以及社交利益关系建模为加权网络 SIROAN. 利用 SIROAN 引入了 GAT 来预测同伴互评分数. GAT-SIROAN 阻止了评估者之间的战略性评估, 并且在聚合过程中考虑的注意力系数, 提高了预测能力, 能够得到更准确的评分. 在真实数据集上的实验结果表明, GAT-SIROAN 方法比现有的同伴互评方法表现得更好.

在未来会在社交利益关系矩阵初始化方面[9]以及针对关于浅层网络的过渡平滑现象进行研究[28], 以提高网络的性能.

参考文献

1 魏顺平. 在线学习自动评价模式构建与应用研究. 中国远

程教育, 2015(3): 38–45.
2 Breslow L, Pritchard D E, Deboer J, *et al*. Studying learning in the worldwide classroom research into edX's first MOOC. Research & Practice in Assessment, 2013, 8: 13–25.
3 Sadler PM, Good E. The impact of self- and peer-grading on student learning. Educational Assessment, 2006, 11(1): 1–31.
4 de Alfaro L, Shavlovsky M. CrowdGrader: A tool for crowdsourcing the evaluation of homework assignments. Proceedings of the 45th ACM Technical Symposium on Computer Science Education. Atlanta: ACM, 2014. 415–420.
5 García-Martínez C, Cerezo R, Bermúdez M, *et al*. Improving essay peer grading accuracy in massive open online courses using personalized weights from student's engagement and performance. Journal of Computer Assisted Learning, 2019, 35(1): 110–120. [doi: 10.1111/jcal.12316]
6 Walsh T. The PeerRank method for peer assessment. Proceedings of the 21st European Conference on Artificial Intelligence. Prague: IOS Press, 2014. 909–914.
7 Fang H, Wang YF, Jin Q, *et al*. RankwithTA: A robust and accurate peer grading mechanism for MOOCs. Proceedings of the 6th IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE). Hong Kong: IEEE, 2017. 497–502.
8 Wang YF, Fang H, Jin Q, *et al*. SSPA: An effective semi-supervised peer assessment method for large scale MOOCs. Interactive Learning Environments, 2022, 30(1): 158–176. [doi: 10.1080/10494820.2019.1648299]
9 Namanloo AA, Thorpe J, Salehi-Abari A. Improving peer assessment with graph convolutional networks. arXiv: 2111.04466, 2021.
10 Veličković P, Cucurull G, Casanova A, *et al*. Graph attention networks. arXiv:1710.10903, 2017.
11 Díez J, Luaces O, Alonso-Betanzos A, *et al*. Peer assessment in MOOCs using preference learning via matrix factorization. Proceedings of the 2013 NIPS Workshop on Data Driven Education. Lake Tahoe: NIPS, 2013.
12 Capuano N, Caballé S, Percannella G, *et al*. FOPA-MC: Fuzzy multi-criteria group decision making for peer assessment. Soft Computing, 2020, 24(23): 17679–17692. [doi: 10.1007/s00500-020-05155-5]
13 Waters AE, Tinapple D, Baraniuk RG. BayesRank: A Bayesian approach to ranked peer grading. Proceedings of the 2nd ACM Conference on Learning @ Scale. Vancouver: ACM, 2015. 177–183. [doi: 10.1145/2724660.2724672]
14 Piech C, Huang J, Chen ZH, *et al*. Tuned models of peer assessment in MOOCs. Proceedings of the 6th International Conference on Educational Data Mining. Memphis: EDM, 2013. 153–160.
15 Mi F, Yeung DY. Probabilistic graphical models for boosting cardinal and ordinal peer grading in MOOCs. Proceedings of the 29th AAAI Conference on Artificial Intelligence. Austin: AAAI, 2015. 454–460.
16 Wang TQ, LI Q, Gao J, *et al*. Improving peer assessment accuracy by incorporating relative peer grades. Proceedings of the 12th International Educational Data Mining Society. Montréal: IEDMS, 2019.
17 许嘉, 刘静, 于戈, 等. 面向在线教育的同伴互评技术综述. 计算机应用, 2022, 42(12): 3913–3923.
18 Page L, Brin S, Motwani R, *et al*. The PageRank citation ranking: Bringing order to the Web. https://courses.washington.edu/ir2010/readings/page.pdf. (1998-01-29).
19 刘俊岭, 李婷, 孙焕良, 等. 利用电子签到数据预测课程成绩. 计算机科学与探索, 2018, 12(6): 908–917. [doi: 10.3778/j.issn.1673-9418.1710070]
20 Darvishi A, Khosravi H, Sadiq S. Employing peer review to evaluate the quality of student generated content at scale: A trust propagation approach. Proceedings of the 8th ACM Conference on Learning @ Scale. New York: ACM, 2021. 139–150.
21 Li P, Yin ZR, Li FY. Quality control method for peer assessment system based on multi-dimensional information. Proceedings of the 17th International Conference on Web Information Systems and Applications. Guangzhou: Springer, 2020. 184–193.
22 Yuan Z, Downey D C. Practical methods for semi-automated peer grading in a classroom setting. Proceeding of the 28th ACM Conference on User Modeling, Adaptation and Personalization. Genoa: ACM, 2020. 363–367.
23 Sajjadi MSM, Alamgir M, von Luxburg U. Peer grading in a course on algorithms and data structures: Machine learning algorithms do not improve over simple baselines. Proceedings of the 3rd ACM Conference on Learning @ Scale. Edinburgh: ACM, 2016. 369–378.
24 Paszke A, Gross S, Chintala S, *et al*. Automatic differentiation in PyTorch. Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017). Long Beach: NIPS, 2017.
25 Fey M, Lenssen JE. Fast graph representation learning with PyTorch geometric. arXiv:1903.02428, 2019.
26 Kingma DP, Ba J. Adam: A method for stochastic optimization. Proceedings of the 3rd International Conference on Learning Representations. San Diego: ICLR, 2015.
27 Xu QS, Liang YZ. Monte Carlo cross validation. Chemometrics and Intelligent Laboratory Systems, 2001, 56(1): 1–11. [doi: 10.1016/S0169-7439(00)00122-2]
28 Chen M, Wei ZW, Huang ZF, *et al*. Simple and deep graph convolutional networks. Proceedings of the 37th International Conference on Machine Learning. PMLR, 2020. 1725–1735.

(校对责编: 孙君艳)