基于聚类的协作学习分组方法

祁天龙[1],任美睿[1,2],赵建宇[1],郭龙江[1,2]
(1. 陕西师范大学 计算机科学学院,陕西 西安 710062;
2. 陕西师范大学 现代教学技术教育部重点实验室,陕西 西安 710062)

摘 要:协作学习能够促进在线学习平台中学习者之间的沟通交流。同一小组内学习者共同协作完成既定任务过程中,不仅可以巩固已有知识,也能通过互相学习,获得新知识和新技能,在提高个人表现的同时,增加学习兴趣,从而有效地降低辍学率。近年来已提出了很多协作学习分组方法。然而,现有分组方法没有兼顾主题意愿、学习时间规律和小组规模这三个对小组沟通效率有重要影响的因素。该文依据学习者的主题意愿预分组,然后依据学习时间规律迭代地调用聚类算法将学习者划分到满足上下限的小组中,结果表明,上述方法形成的协作学习小组在满意度和时间重合度上均优于 IFST 和随机分组方法。最后,以 XuetangX 平台上的 1 754 名学习者为实验对象进行协作学习分组,实验结果表明,形成的小组有充分的协作学习时间,指派的主题能够很好地满足学习者的意愿,且各个小组之间成员数均衡。

关键词:协作学习;在线学习;分组方法;聚类;满意度;时间重合度
中图分类号:TP391.7 **文献标识码**:A **文章编号**:1673-629X(2023)06-0189-05
doi:10.3969/j.issn.1673-629X.2023.06.028

Collaborative Learning Grouping Method Based on Clustering

QI Tian-long[1],REN Mei-rui[1,2],ZHAO Jian-yu[1],GUO Long-jiang[1,2]
(1. School of Computer Science,Shaanxi Normal University,Xi'an 710062,China;
2. Key Laboratory of Modern Teaching Technology of Ministry of Education,Shaanxi Normal University,
Xi'an 710062,China)

Abstract:Collaborative learning can promote communication between learners on the online learning platform. When learners in the same group collaborate to complete the established task,they can not only consolidate the existing knowledge,but also acquire new knowledge and new skills through mutual learning. While improving individual performance,they can increase their interest in learning,thus effectively reducing the dropout rate. In recent years,many collaborative learning grouping methods have been proposed. However,the existing grouping methods do not take into account the three factors which have important influence on the efficiency of group communication:topic willingness,learning time rule and group size. According to the learners' topic willingness,they are pre-grouped,and then the clustering algorithm is iteratively called according to the learning time rule to divide the learners into groups that meet the upper and lower limits. It is showed that the formed collaborative learning groups are better than IFST and random grouping methods in terms of satisfaction and time coincidence. Finally,1 754 learners on the XuetangX platform are taken as the experimental objects for collaborative learning grouping. The experimental results show that the formed groups have sufficient collaborative learning time,the assigned topics can well meet the wishes of learners,and the size of each group is balanced.

Key words:collaborative learning;online learning;grouping method;clustering;satisfaction;time coincidence

0 引 言

近年来,在线学习平台取得了长足的发展,据统计,中国目前有 12 500 门左右的在线课程,学习人数已超过 2 亿人次[1]。然而,在线学习平台的课程完成率仅有不到 5%[2],限制了平台的发展。协作学习是以小组为单位的学习方式,对同一个主题感兴趣的学习者组成一个小组,组员间可以一起探讨主题,相互帮助,共同完成学习目标[3]。协作学习的过程中,不仅使

收稿日期:2022-08-12 **修回日期**:2022-12-14
基金项目:国家自然科学基金(61977044);教育部第二批新工科研究与实践项目(E-RGZN20201045);教育部 2021 年第二批产学合作协同育人项目(202102591018)
作者简介:祁天龙(1997-),男,硕士研究生,研究方向为教育大数据;通讯作者:任美睿(1972-),女,博士,副教授,CCF 会员(31901M),研究方向为移动计算、教育信息化、大数据分析与处理。

学习者的能力得以提升,还促进了组员之间的交流,从而有效地降低辍学率。

构建协作学习小组是协作学习中的关键问题,已有不少研究者提出了多种不同的分组方法。Andrejczuk 等[4]分组时考虑了学习者的性别、个性、能力和小组规模。Nand 等[5]以形成技能均衡的小组为目的,依据学习者的技能偏好和水平,使用萤火虫算法进行分组。

Flores-Parra 等[6]考虑了学习者在小组中可能的角色分工,使用社交网络的方法分组。桑治平等[7]依据学习者的兴趣、学习动机、知识水平等构建小组。潘芳等[8]考虑了在线学习者的性格、学习目标、风格、动机、认知水平等因素,使用了 Multi-Agent 分组。李浩君等[9]结合混合遗传算法提出了在线学习环境中基于任务驱动的协作学习小组构建方法。

聚类算法由于其复杂度低,可解释性强,近年来出现了基于聚类的协作学习分组方法。Akbar 等[10]考虑了小组规模和主题意愿,结合 HK-Means 算法(一种基于 K-Means 的改进算法)提出了改善学习者团队的形成方法(Improving Formation of Student Teams, IFST),取得了较好的实验效果。

曹天生等[11]考虑了学习者的信息素养、学习风格、认知能力和知识基础等因素,使用聚类算法分组。Sanz-Martínez 等[12]通过页面浏览量、作业提交量等活动量化学习者的参与度,再依据参与度使用 K-Means 聚类算法分组。罗凌等[13]考虑了学习者的学习风格、知识水平和学习目标等因素,使用模糊 C 均值算法分组。陈甜甜等[14]考虑了在线学习者的人口因素(如性别、国籍)、课程参与度、行为信息,分组时也使用了模糊 C 均值算法。

然而,现有的研究中还没有同时兼顾学习者的主题意愿、学习时间规律、小组规模这三个重要因素的协作学习分组方法。首先,相同的主题有助于减少组内分歧,且给予学习者对主题的自由选择权可以提高其参与度和积极性[15]。其次,在线学习平台中的课程大多没有固定的时间表,学习者在任意时刻都有可能进入平台,将学习时间规律相同的学习者分配到同一小组中有助于提高组员间的沟通效率。最后,小组规模过大时会提高组员间的沟通成本[16],进而降低学习效率,小组规模过小会使单个学习者承担更多的工作,不利于学习目标的完成。综上所述,对于协作学习分组,主题意愿、学习时间规律、小组规模是三个必要的考虑因素。

为了填补上述分组研究中的空白,该文综合考虑了学习者的主题意愿、学习时间规律、小组规模三个因素,提出了基于聚类的协作学习分组方法。

1 基于聚类的协作学习分组方法

1.1 相关符号与定义

N 个学习者的集合表示为 $S=\{s_i \mid 1 \leqslant i \leqslant N\}$,将集合 S 划分为 M 个小组 $\{G_j \mid 1 \leqslant j \leqslant M\}$,这里 $G_j \cap G_k = \varnothing (j \neq k)$,$\cup_{j=1}^{M} G_j = S$。$L$ 个主题的集合表示为 $P=\{p_l \mid 1 \leqslant l \leqslant L\}$。$s_i$ 的选题意愿表示为向量 $\boldsymbol{w}_i = [w_{i1}, w_{i2}, \cdots, w_{il}, \cdots, w_{iL}]$,这里 $1 \leqslant w_{il} \leqslant L$ 是整数,表示学习者 s_i 将主题 p_l 作为第 w_{il} 个意愿。例如,有 3 个选题,即 $L=3$,假设学习者 s_5 的选题意愿向量为 $\boldsymbol{w}_5 = [2,3,1]$,则表示 s_5 将 p_1 作为第二意愿,将 p_2 作为第三意愿,将 p_3 作为第一意愿。s_i 的学习时间规律表示为 $\boldsymbol{t}_i = [t_{i1}, t_{i2}, \cdots, t_{iq}, \cdots, t_{iQ}]$,这里将学习周期(通常为 7 天)划分为 Q 个相等的时间段,t_{iq} 等于 0 或 1,表示学习者 s_i 的第 q 个时间段是否可以参与协作学习。例如,s_5 的学习周期被划分为 4 个相等的时间段,假设 $\boldsymbol{t}_5 = [1,0,0,1]$ 表示 s_5 的第 1、4 个时间段可参与协作学习,第 2、3 个时间段不可参与协作学习。学习者在选择选题的同时可以声明自己的哪些时间段可参与协作学习。最后,将小组成员数上限表示为 ceil,小组成员数下限表示为 floor。

定义 1(满意度):若学习者所在的小组成员数在上下限之间(包括上下限),且该小组被分配的主题是当前学习者的第一个意愿,则认为当前学习者是满意的。满意的学习者数占学习者总数的比例称为满意度。

定义 2(时间重合度):时间重合度衡量了小组中学习者的学习时间规律的相似程度。时间重合度的计算方式如下:

$$\frac{1}{M}\sum_{j=1}^{M}\sum_{\substack{s_i,s_k \in G_j \\ i \neq k}} \frac{(t_i \bullet t_k)/Q}{C^2_{|G_j|}}$$

式中,“·”是向量的内积运算符,“$C^2_{|G_j|}$”表示小组 G_j 中任取两个学习者的组合数。

定义 3(协作学习分组):给定主题集合 P,学习者集合 S,学习者的主题意愿向量 $\{\boldsymbol{w}_i \mid 1 \leqslant i \leqslant N\}$,学习者的学习时间规律 $\{t_i \mid 1 \leqslant i \leqslant N\}$,以及小组人数上下限 ceil 和 floor。最终,将学习者集合 S 划分为互不相交、组内学习者数在给定上下限之间的协作学习小组 $\{G_j \mid 1 \leqslant j \leqslant M\}$。使得满意度和时间重合度最大化。

1.2 协作学习分组方法

基于聚类的协作学习分组方法(Cooperative Learning Grouping Method Based on Clustering, CLGC)大致包含以下四个步骤:第一步,预分组;第二步,处理预分组集合;第三步;使用聚类算法分组;第四步,处理聚类后的分组结果集合。分组过程如图 1 所示,图中

floor = 3, ceil = 5。

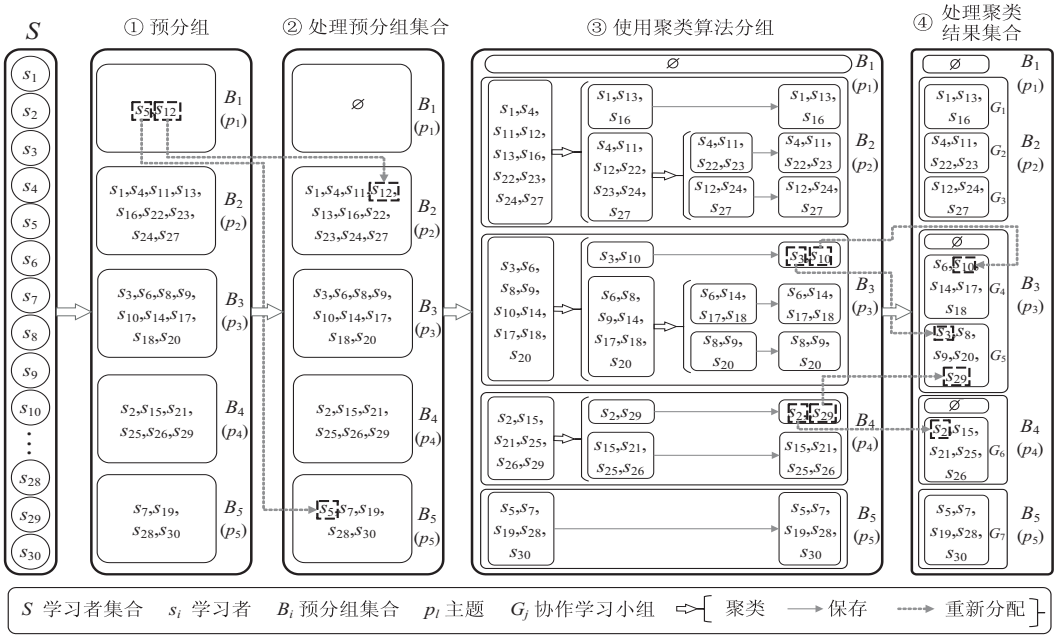

图 1　基于聚类的协作学习分组方法示意图

1.2.1　预分组

依据学习者的主题意愿向量 $\{w_i \mid 1 \leqslant i \leqslant N\}$,将学习者集合 S 划分为 L 个预分组集合 $\{B_l \mid 1 \leqslant l \leqslant L\}$,这里 $B_l \cap B_k = \varnothing (l \neq k)$, $\cup_{l=1}^{L} B_l = S$ 。划分预分组集合时,确保第一意愿相同的学习者在同一个预分组集合中,预分组集合 B_l 对应的主题为 p_l 。

1.2.2　处理预分组集合

依次遍历预分组集合 $\{B_l \mid 1 \leqslant l \leqslant L\}$,若预分组集合中学习者数大于等于 floor ,则进入下一个步骤,使用聚类算法进行分组(详见 1.2.3 节)。若当前预分组集合中学习者数小于小组成员数下限 floor ,则依次将该预分组集合中的学习者重新分配到当前学习者第二意愿对应的预分组集合中,如图 1 步骤②中将 s_5 从 B_1 分配到 B_5 ,如果第二意愿对应的预分组集合中学习者数仍小于 floor ,则将其分配到该学习者第三意愿对应的预分组集合,依此类推,直到将该学习者分配到某个预分组集合中为止。

1.2.3　使用聚类算法分组

依次遍历经过处理后的预分组集合 $\{B_l \mid 1 \leqslant l \leqslant L\}$,若预分组集合中的学习者数大于小组成员数上限 ceil ,则依据学习者的学习时间向量,使用聚类算法将预分组集合中的学习者划分为 2 个子集,若子集中的学习者数仍大于 ceil ,则继续将子集划分为 2 个更小的子集,重复这个过程,直到子集中的学习者数小于等于 ceil ,最后,将满足条件的子集加入到分组结果集合。如图 2 所示,若预分组集合非空且其中的学习者数小于等于 ceil ,则直接将该预分组集合加入到分组结果集合。

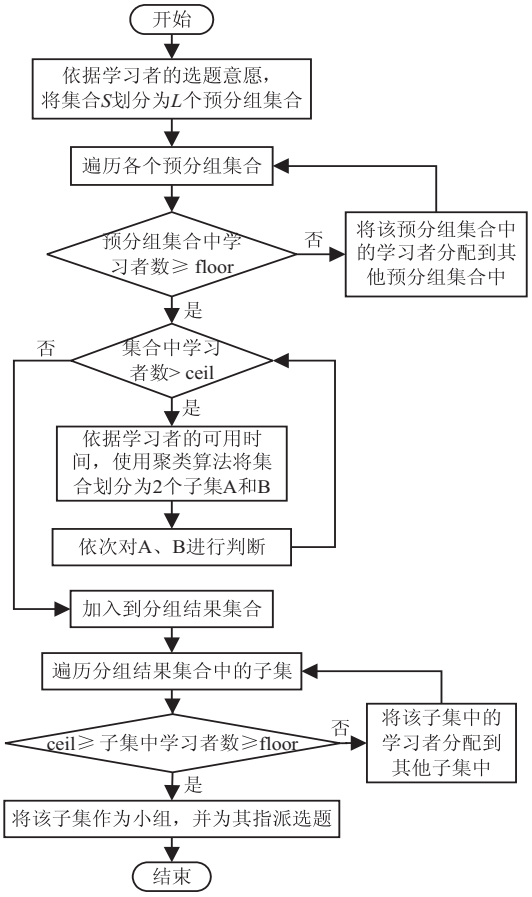

图 2　基于聚类的协作学习分组方法流程

上述过程中可使用的聚类算法大致分为三种:基于划分的聚类算法,如 K-Means[17];层次聚类算法,如 Agglomerative Clustering[18](简称为 AC 算法)、BIRCH[19];以及近邻传播算法 Affinity Propagation[20](简称为 AP 算法)。需要注意的是,使用基于划分的

聚类算法和层次聚类算法时,指定聚类中心数为 2,而 AP 算法不需要指定聚类中心数,可以对所有划分后的子集递归地调用自身再次聚类,所以也可用于分组。其他的聚类算法,如基于密度的聚类算法需要指定聚类半径,针对不同的学习者集合该聚类半径也不相同,因此难以确定合适的聚类半径,而基于网格的聚类和谱聚类算法由于其计算复杂度高,不适用于学习者规模较大的在线学习平台。

1.2.4 处理聚类结果集合

遍历聚类后分组结果集合中的所有子集,若子集中的学习者数小于 floor ,则将该子集中的学习者依次分配到和当前学习者在同一个预分组集合中的其他子集,目标子集须满足: floor ≤ 目标子集中的学习者数 < ceil 。若在同一个预分组集合中没有满足条件的目标子集,则将当前学习者暂存到待分配集合中;若有多个满足条件的目标集合,则优先分配到和该学习者的时间重合度最高的目标子集。

遍历结果集合中的所有子集后,若待分配集合非空,则依次遍历待分配集合中的学习者,先找到该学习者的第二意愿对应的预分组集合中的结果集合,再从这些结果集合中找到满足条件的目标子集,将当前学习者分配到该目标子集中,如图 1 步骤④中将 s_{29} 分配到 B_3;若第二意愿对应的预分组集合中找不到满足条件的目标子集,则依次在后续意愿对应的预分组集合中寻找,直到将学习者分配到某一个目标子集中。

最后,遍历所有经过处理后的结果集合,若当前结果集合满足: floor ≤ 结果集合中的学习者数 ≤ ceil ,则将其作为一个协作学习小组,小组所在的预分组集合对应的主题即为指派给该小组的主题,如图 1 步骤④中指派给小组 G_1 的主题是 p_2。

2 实验及结果分析

2.1 数据集和预处理

学习者的主题意愿是由计算机模拟生成的,学习时间规律来源于公开的真实数据集,由在线学习平台 XuetangX 收集[2]。

2.1.1 主题意愿

在现实场景中,学习者对主题的偏好往往是不均匀的,在某些主题上会比较集中。所以模拟生成时假设学习者的主题意愿服从正态分布 $N(\mu,\sigma^2)$,其中参数 $\mu=\frac{1}{L}\sum_{l=1}^{L} l$, $\sigma^2=\frac{1}{L}\sum_{l}^{L}(l-\mu)^2$。

2.1.2 学习时间规律

XuetangX 平台收集的数据集中包含了从 2015 年 8 月至 2017 年 8 月的 1 213 门课程中共 378 237 名学习者的活动记录[2],活动发生的时间精确到秒。学习者的时间数据需要将其预处理为标准化的 0-1 向量用于后续的聚类。

时间数据的预处理过程如下:首先,将活动记录文件分别按照学习者的学号和课程分割,若某个学习者参与了多门课程,那么将对应多个分割后的文件。然后,随机选取一门课程,并统计该课程中的学习者在平台的累计学习时长,为了保证有充足的数据量化学习时间规律,仅保留了累计学习时长超过 20 小时的学习者。最后,遍历该门课程中被保留的学习者的记录,将其记录编码为 0-1 向量,具体地,若 s_i 的活动发生在 t_i 的第 q 个时间段,则 $t_{iq}=1$,否则 $t_{iq}=0$。

2.2 基于聚类的协作学习分组方法有效性

为了验证提出的协作学习分组方法的有效性,将其与随机分组方法(Random Grouping Method,RGM)以及 Akbar 等[10]提出的 IFST 分组算法进行对比。

随机选取了课程“TsinghuaX_30640014X”中自 2015.10.13-0:0:0 至 2015.10.19-23:59:59 期间有记录的 1 754 名学习者,平均每个学习者每天的学习时长是 138 分钟。将他们在这 7 天内的活动记录预处理为 0-1 向量,将其作为学习者的学习时间规律。编码时,将 Q 设置为 56,即一周被均分为 56 个时间段,每个时间段的长度为 3 小时。主题数 L 设置为 8, floor 设置为 3, ceil 设置为 5。

CLGC 在分组时可使用的聚类算法有多个,这里仅选择其中的三个聚类算法。将基于 K-Means 算法的协作学习分组方法记作 CLGC(KM),将基于 BIRCH 算法的分组方法记作 CLGC(BC),将基于 AP 算法的分组方法记作 CLGC(AP)。上述三种分组方法和 IFST 以及随机分组方法的实验结果见表 1,表中展示的结果是重复 10 次实验后的平均值。

表 1 五种分组方法的实验结果

分组方法	满意度/%	时间重合度/%	运行时长/s
IFST	86.58	1.26	34.62
RGM	42.73	1.23	0.95
CLGC(KM)	100	4.34	11.43
CLGC(BC)	99.94	4.44	7.16
CLGC(AP)	100	2.88	7.01

从表 1 可以看出,相对于 IFST 和 RGM,CLGC 在满意度和时间重合度上都有更好的表现。其中,基于 BIRCH 算法的 CLGC 表现最好,时间重合度约是 IFST 和 RGM 的 3 倍,相当于平均每个学习者每天可参与协作学习的时长为 60 分钟,大约占每天学习时长的一半。在运行时长方面,IFST 的开销最大,CLGC 次之,RGM 的运行时长最短。综合考虑满意度、时间重合度和运行时长,该文提出的 CLGC 优于 IFST 和 RGM,验

证了其有效性。

2.3　基于聚类的协作学习分组应用

将 2.2 节中的 1 754 名学习者作为实验对象, Q 设置为 56, L 设置为 8, floor 设置为 3, ceil 设置为 5。采用与 2.2 节相同的方式预处理后,使用 CLGC(BC)将学习者划分为 361 个协作学习小组。其中,组员数为 3 的小组有 23 个,组员数为 4 的小组有 5 个,组员数为 5 的小组有 333 个。指派给各个小组的主题都是组员的前三个意愿之一,较好地满足了学习者的自由选择权。组员平均每天可参与协作学习的时长是 64 分钟,大约占平均每天总学习时长的一半,保证了学习者有充足的时间协作完成学习目标。

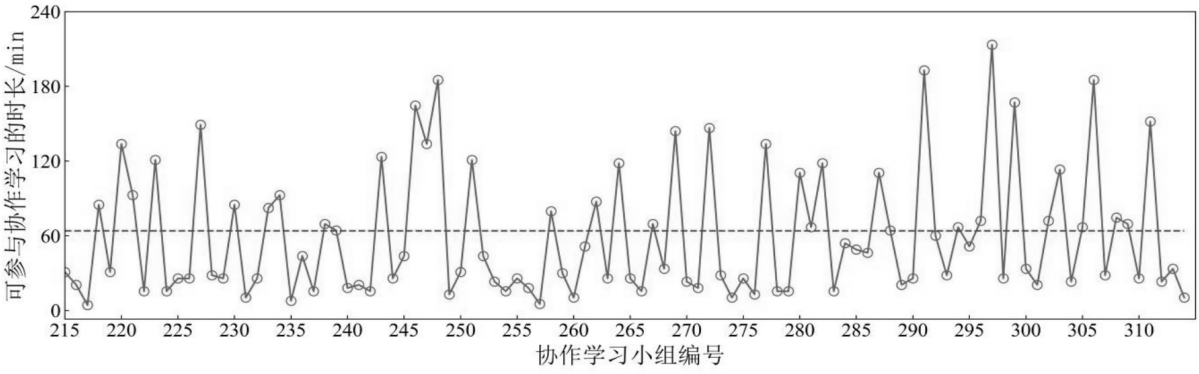

图 3　部分小组的可参与协作学习时长

图 3 显示了其中 100 个小组可参与协作学习的时长,图中虚线表示平均值。从图中可以看出,不同的小组可用于协作学习的时长差异很大,这主要是因为内学习者在选定的这段时间内活跃程度各不相同,该文提出的分组算法倾向于将活跃程度高的学习者划分到同一个小组中,因此这些小组可用于协作学习的时长普遍高于其他小组。

3　结束语

提出了一种基于多种聚类算法的协作学习分组方法,该方法综合考虑了学习者的选题意愿、学习时间规律和小组规模,形成的小组有充分的协作学习时间,指派给小组的主题能够很好地满足学习者的意愿,且各个小组之间成员数均衡。以上三点都为提高协作学习小组的沟通效率提供了保证,不仅有助于学习目标的完成,还可以有效地降低辍学率。

参考文献:

[1] SHEN X. Current situation, problems and countermeasures of MOOC in Chinese universities[C]//Proceedings of 1st international symposium on innovation and education, law and social sciences (IELSS 2019). Shenyang: Atlantis Press, 2019:299-305.

[2] FENG W, TANG J, LIU T. Understanding dropouts in MOOCs[C]//Proceedings of the AAAI conference on artificial intelligence. Palo Alto: AAAI Press, 2019:517-524.

[3] 栾　华,李庆忠. 本体在协作学习中的应用[J]. 计算机集成制造系统,2003,9(z1):28-32.

[4] ANDREJCZUK E, BISTAFFA F, BLUM C, et al. Synergistic team composition: a computational approach to foster diversity in teams[J]. Knowledge-Based Systems, 2019, 182: 104799.1-104799.16.

[5] NAND R, SHARMA A. Meta-heuristic approaches to tackle skill based group allocation of students in project based learning courses[C]//Proceedings of 2019 IEEE congress on evolutionary computation (CEC). Wellington: IEEE, 2019: 1782-1789.

[6] FLORES-PARRA J, CASTAÑÓN-PUGA M, EVANS R D, et al. Towards team formation using belbin role types and a social networks analysis approach[C]//Proceedings of 2018 IEEE technology and engineering management conference (TEMSCON). Evanston: IEEE, 2018:1-6.

[7] 桑治平,何聚厚. 基于改进细菌觅食的协作学习分组算法[J]. 计算机工程,2014,40(10):137-142.

[8] 潘　芳,仲伟俊. E-Learning 协作学习中的分组问题研究[J]. 中国远程教育,2014(1):59-63.

[9] 李浩君,杜兆宏,邱飞岳. 基于混合遗传算法的任务驱动分组优化研究[J]. 计算机科学,2017,44(S1):105-108.

[10] AKBAR S, GEHRINGER E F, HU Z. Poster: improving formation of student teams: a clustering approach[C]//Proceedings of the 40th international conference on software engineering: companion. Gothenburg: IEEE, 2018:147-148.

[11] 曹天生,孔凡士,朱　珂,等. 促进学习者之间交互深度的分组策略研究[J]. 现代教育技术,2020,30(6):55-60.

[12] SANZ-MARTÍNEZ L E, MARTÍNEZ-MONÉS E, DIMITRIADIS A, et al. Creating collaborative groups in a MOOC: a homogeneous engagement grouping approach[J]. Behaviour & Information Technology, 2019, 38(10/12): 1107-1121.

[13] 罗　凌,杨　有,马　燕. 基于模糊 C 均值的在线协作学习混合分组研究[J]. 计算机工程与应用,2017,53(16):68-73.

[14] 陈甜甜,何秀青,葛文双,等. 大规模在线协作学习分组方

(下转第 201 页)