2024 年第十一届中国可视化与可视分析大会

数据可视化竞赛 赛道 1

（ChinaVis Data Challenge 2024 - mini challenge 1）

答　卷

参赛队名称：上海交通大学-白骐硕-赛道 1

团队成员：　白骐硕，上海交通大学，qishuo-bai@sjtu.edu.cn，队长

吴治远，上海交通大学，wzy605399@sjtu.edu.cn

刘一诺，上海交通大学，lyn15806636972@sjtu.edu.cn

杨雨彤，上海交通大学，flora20@sjtu.edu.cn

曹俊翔，上海交通大学，cjxqaq@sjtu.edu.cn

董笑菊，上海交通大学，xjdong@sjtu.edu.cn，指导老师

团队成员是否与报名表一致（是或否）：是

是否学生队（是或否）：　是

使用的分析工具或开发工具（如果使用了自己研发的软件或工具请具体说明）：D3，Python

例：D3，Excel，MySQL，Qt，CVASTer（天津大学 xxx 中心开发的数据可视分析工具）

共计耗费时间（人天）：　40 人天

本次比赛结束后，我们是否可以在网络上公布该答卷与视频（是或否）：是

1、分析学习者答题行为日志记录，从答题分数、答题状态等多维度属性量化评估知识点掌握程度，并识别其知识体系中存在的薄弱环节。

1.1 系统概览

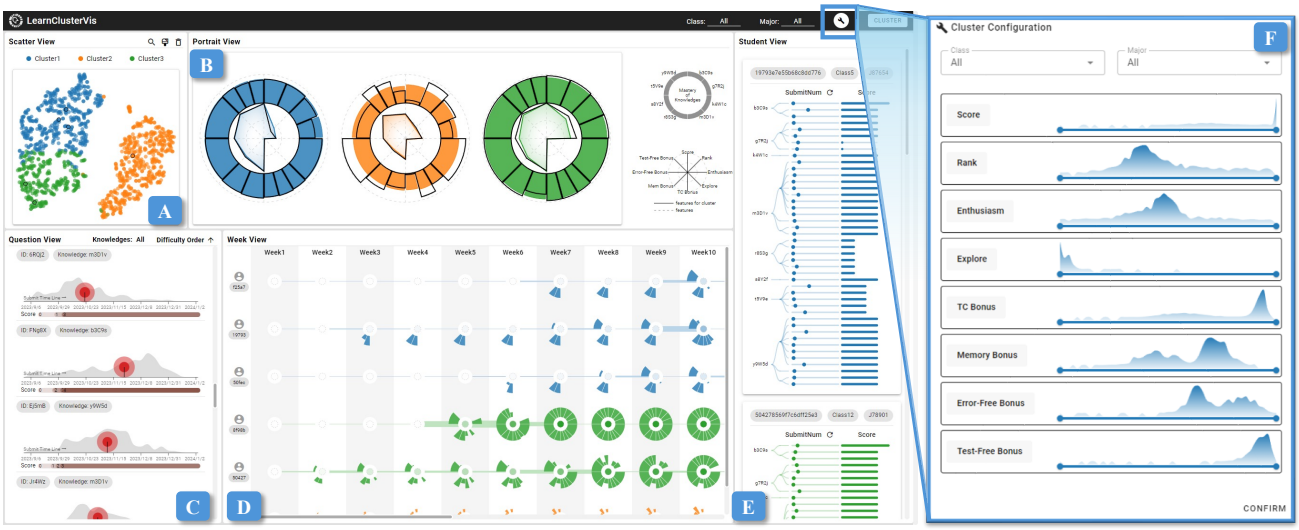

图 1：LearnClusterVis 可视分析系统概览图

为了完成「析数启智」时序多变量教育数据可视分析挑战赛，我们基于赛事提供数据设计了可视分析系统 LearnClusterVis，旨在利用可视化技术帮助教育机构全面了解和优化教学效果。该系统通过多视图展示学习者的学习行为和知识掌握情况，包括：

- A.聚类视图（Scatter View）：展示学习者的聚类情况，便于识别不同群体的学习模式。
- B.画像视图（Portait View）：通过外圈圆环图详细展示群体或个体对不同知识点及子知识点的掌握情况；通过内圈雷达图展示学习者在不同特征上的表现，从而帮助用户深入分析学习行为模式。
- C.题目视图（Question View）：展示各个题目的答题情况，包括每题的提交时间峰值、题目难度、分数分布等信息。
- D.周日志视图（Week View）：按周展示学习者对不同知识点和子知识点的掌握程度，便于跟踪学习者在不同时间段的表现。
- E.学生视图（Student View）：详细展示每个学生的详细答题信息，包括提交次数和最高得分。
- F.聚类设置面板（Cluster Configuration）：用户可以通过该面板选择聚类数据范围和特征，支持对于特征进行选择和范围筛选。

1.2 知识点掌握程度特征选取

为了量化评估知识点掌握情况，我们首先在原始数据的基础上进行了初步分析和处理，最终选取了 6 个特征用来反映学习者对于某道题的掌握情况，然后根据题目的分值为每道题的掌握情况赋予权重，最后通过加权平均来计算学习者对于某个知识点或者子知识点的掌握程度。选取的特征及其解释如**表 1** 所示（所有特征的值均在 0 到 1 之间）：

表 1： 知识点掌握程度

特征字段	含义	解释
score_bonus	答题得分加成	题目最终得分越高，该值越高
tc_bonus	时间复杂度加成	题目时间复杂度越低，该值越高（如题目回答错误，则该值为 0）
mem_bonus	空间复杂度加成	题目空间复杂度越低，该值越高（如题目回答错误，则该值为 0）
_error_type_penalty	错误类型扣减	题目错误类型越少，该值越高
_test_num_penalty	尝试次数扣减	题目尝试次数越少，该值越高
rank_bonus	排名加成	题目排名越高，该值越高（OJ 网站通常会设立用户排名，以用户的提交答案通过数多少或某个题目执行时间快慢为排名依据）

1.3 群体知识点掌握程度分析

由于数据的高维度特性，我们希望通过聚类的方式来进行初始分析。用户可以首先点击右上角的设置按钮打开 Cluster Configuration，选择聚类所使用的维度，并筛选数据范围（**图 1-F**）。例如，通过选择 1.2 中的 6 个特征，将数据范围设置为所有班级和所有专业，我们可以得到**图 2** 所示的聚类和群体画像结果。我们可以通过聚类图发现，根据知识点掌握程度，所有学习者被大致分为了三类。蓝色（第一类）群体数量最多，其次是绿色（第三类），最后是橙色（第二类）。通过观察右侧的画像图的外圈，我们可以发现，第一类学习者群体的各个知识点掌握程度较为平均，同时都保持一个较高水平，因此我们可以认为这一群体的知识点掌握程度属于“较高且均衡”。对于第二类学习者群体来说，部分知识点掌握水平较为薄弱，例如 b3C9s、g7R2j、s8Y2f，但也有个别知识点掌握程度中等，因此这一类群体的知识点掌握程度为“较差且不均”。对第三类学习者群体来说，8 个知识点掌握程度也相对均衡，但整体掌握水平相较于第一类来说仍有进步空间，因此这一群体的知识点掌握程度为“中等且均衡”。

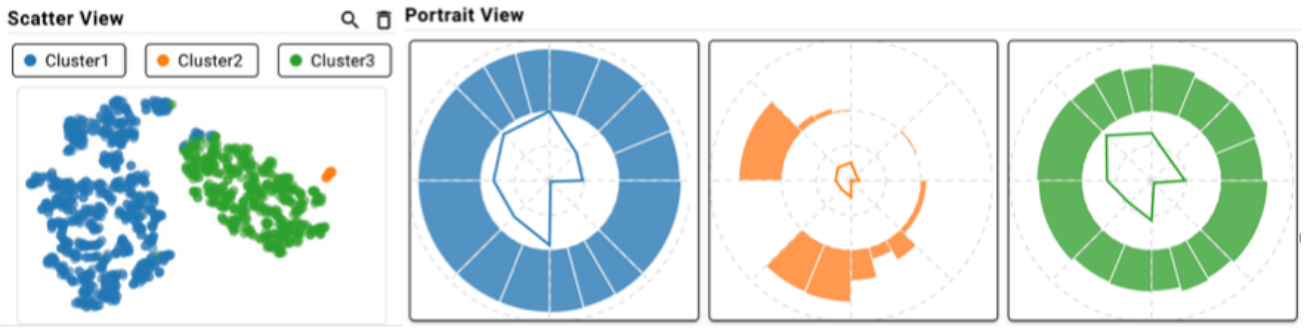

图 2： 群体知识点掌握程度聚类分析

1.4 个体知识点掌握程度分析

我们还可以进一步对个体的知识点掌握程度进行分析。例如我们可以在聚类视图中，点选单个或者框选多个点，以对比探究该点的知识掌握程度与群体平均知识掌握程度的区别。在**图 3** 中，我们可以在聚类视图中看到上方有一个绿色的点靠近蓝色群体，点击该点，我们可以在右侧画像视图中看到该点的知识点掌握程度图（黑色轮廓线标出）。我们可以看到该学习者对大部分知识点的掌握程度较好，高于绿色群体平均值，因此该点具有与蓝色群体相近的特征，这也解释了为什么在聚类图中该点靠近蓝色群体。同时我们发现，该学习者在 b3C9s 和 k4W1c 的知识掌握上较为薄弱，其掌握程度低于绿色群体平均水平。

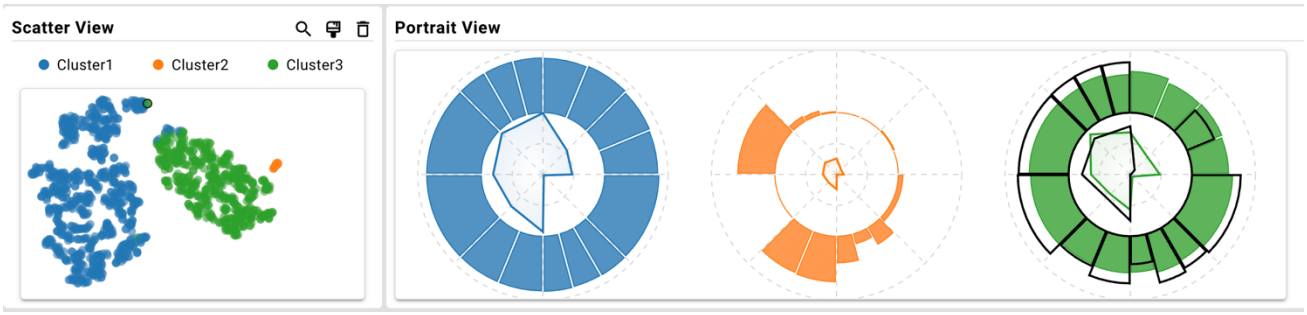

图 3: 个体知识点掌握程度聚类分析

由于数据具有时序性，因此我们还设计了周日志视图（**图 1-D**），以此来探究学习者对知识的掌握程度随着时间的变化。如**图 4** 所示，展示了上述学习者每周的知识点掌握程度变化。外圈与画像视图含义相同，展示了对于各个子知识点的掌握程度，中间的圆通过颜色深浅来映射答题分数的变化，颜色越深，说明目前所得分数越高。同时，我们通过横线对相邻两周的环形图进行连接，线越宽代表学习者在这两周之间的知识掌握程度进步越大。在**图 4** 中，我们发现该学习者在第 1 到 13 周没有进行题目作答，从第 14 周开始，进行了 r8S3g 知识点相关的习题作答。再次作答时间为第 19 周，在该周该学习者进行了大部分问题的作答，因此我们可以观察到第 19 周到第 20 周之间的流动横线宽度较大，表示了该学习者在本周内知识点掌握的进步较大。

图 4: 个体知识点掌握程度周变化分析

2、结合学习者的特征挖掘个性化学习行为模式，从多角度设计并展示学习者画像，如答题高峰时段、偏好题型、正确答题率等。

2.1 学习者画像设计

除了在 1.2 中叙述的用来量化知识点掌握程度的特征，我们还补充了“探索加成”与“热情加成”两个特征，用来更完整和全面地反映学习者的学习习惯和模式。补充特征的具体描述如**表 2** 所示。我们选择

用雷达图的形式对学习者画像进行可视化，学习者的学习行为模式将通过八边形进行呈现，每个顶点所对应的维度如**图 5** 所示。

表 2: 学习模式分析补充特征

特征字段	含义	解释
explore_bonus	探索加成	题目回答正确后，仍然尝试探索，探索次数越多该值越高
enthusiasm_bonus	热情加成	题目发布后，全部提交次数的时间平均值越早，说明该生完成题目的热情越高，则该值越高

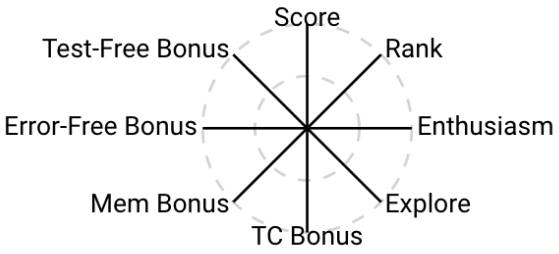

图 5: 学习者画像雷达图设计

2.2 学习者画像分析

以 Class14 为例，我们选择所有 8 个特征进行聚类。通过观察**图 6-A**，我们发现该班级的学习者被分为了三类。第一类学习者群体（蓝色）与第三类学习者群体（绿色）的知识点掌握程度相近，均为“较高且均衡”，但两个群体的学习模式却有差异。具体来讲，第一类学习者群体在“排名”以及“热情”两个维度上表现不如第三类学习者群体。在聚类视图中，任意点击多个蓝色和绿色学习者，我们可以在周日志视图中看到这两位学习者每周的知识点掌握程度变化情况（**图 7**）。我们发现，由于蓝色群体在学习者画像中的“热情加成”相较与绿色群体较低，因此在周日志视图中开始答题的时间也较晚，绿色群体开始于前三周，而蓝色群体开始于第八周以后。此外，通过观察**图 6-A** 我们还可以发现，三个群体的学习者在“探索”维度上的表现都不尽如人意，这说明在题目取得满分后，大部分学习者不会再进行提交和探究。从教学角度来看，这也反映了应当加强引导学有余力的学习者进行不同方法的进一步尝试和探索。

总的来说，蓝色群体的分数和时空复杂度等答题表现较好，但答题热情不够高涨，倾向于拖延到后期完成答题，也因此在排名上不尽如人意，我们可以将这类群体归为“高效拖延者（Efficient Procrastinators）”（**图 6-B**）。橙色群体在答题表现上来看表现较弱，包括分数、排名、时间空间复杂度等均有较大提高空间，但其答题热情比蓝色群体更高，尝试次数也相对另外两个群体更多，提交所涉及的错误类型也较多，我们可以将这类群体归为“积极探索者（Diligent Explorers）”。第三类绿色群体在答题表现上来看表现最佳，拥有较高的分数、排名以及相对理想的时间和空间复杂度，答题热情也较高，通常情况下倾向于较早开始进行题目探究，这类群体可以被称作“卓越学习者（Outstanding Learners）”。但三类群体的共同缺点是在提交正确答案后不会进行复习或尝试新方法继续提交，缺乏深入学习和探索的动力。

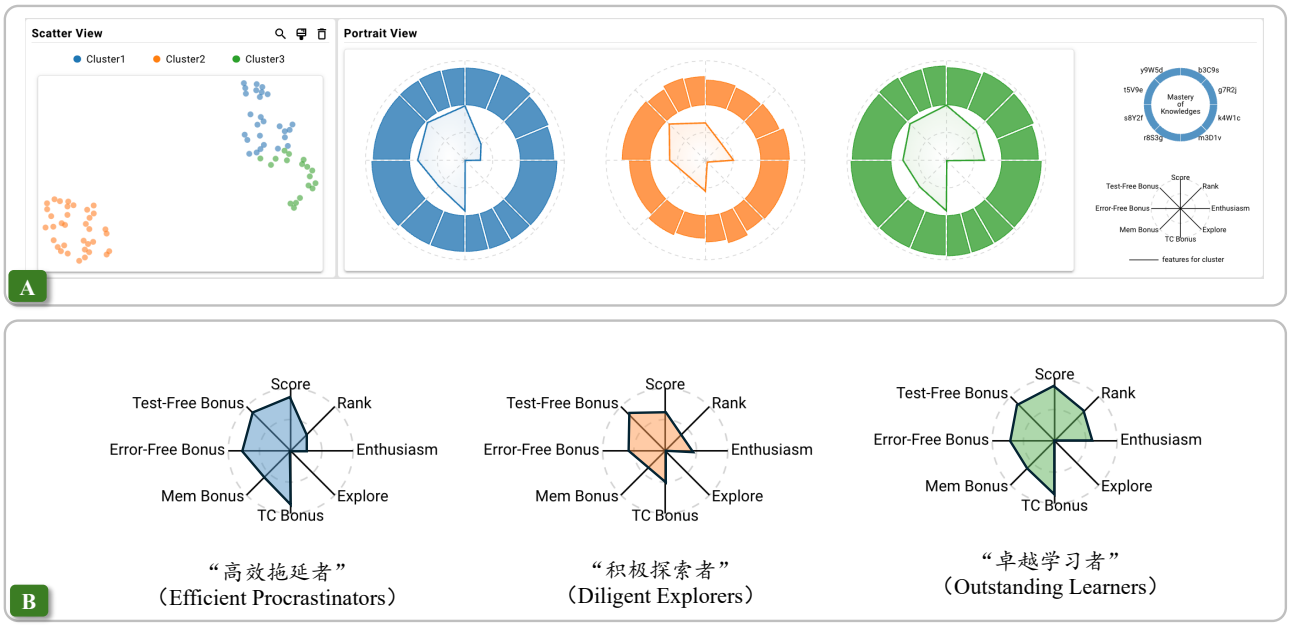

图 6: A. Class14 学习者画像分析；**B.** 三类学习者群体

图 7: 两类学习者群体答题周日志视图对比

3、不同的学习模式直接影响到学习者对知识的吸收、整合及应用能力，高效的学习模式能够促进知识的深度理解和长期记忆。请对学习模式与知识掌握程度之间的潜在关系进行建模，利用图表的形式呈现结果并简要分析。

3.1 三类学习模式与知识掌握程度关系

我们在聚类设置面板中选择了全部 8 个特征进行聚类，班级和专业均设置为全选，聚类结果如**图 8** 所示。我们可以看到学习者们被分为了较为清晰的三类群体。通过观察学习者画像（画像视图中的雷达图部分）我们可以发现学习模式与 2.2 中分析的三个类别相吻合，从左到右分别是“高效拖延者”、“积极探索者”、以及“卓越学习者”。通过观察外圈的知识掌握程度，我们可以发现，“卓越学习者”通常能够全面、均衡牢固地掌握知识点，我们分析这可能是由于他们优秀的学习习惯和相对较强的时间管理能力。“高效拖延者”对知识的掌握程度次之，但也能够相对全面且扎实掌握知识，在效果上略逊于“卓越学习者”，这可能是因为他们拥有较强的学习能力，但是在学习习惯和时间管理上不如“卓越学习者”。“积

极探索者”对知识的掌握程度相对较差，且有不均衡的情况。尽管他们在答题时表现出较高的热情和积极性，尝试多次提交答卷解决问题，但由于基础知识不够扎实，导致答题表现不佳。

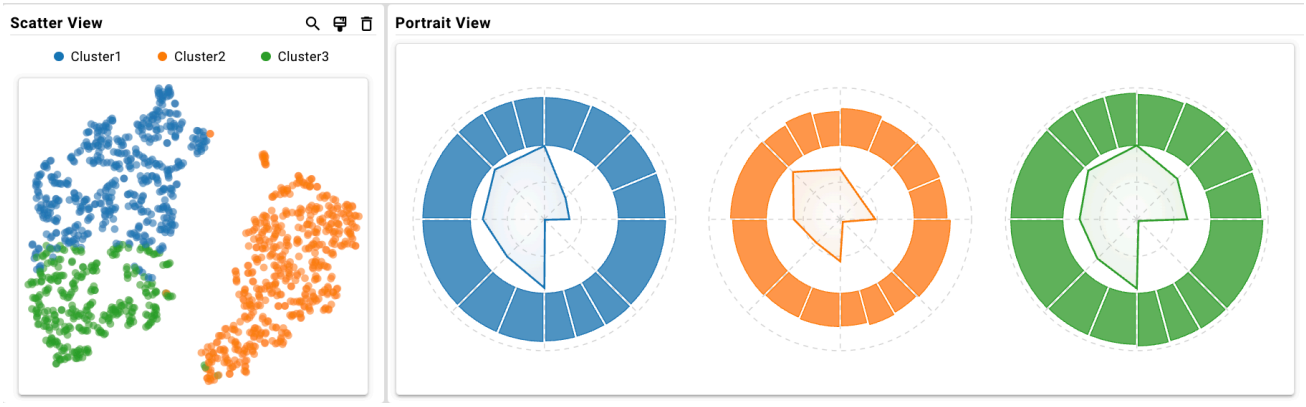

图 8： 三类学习模式不同知识掌握程度对比

3.2 聚类图离群点/边缘点探索

我们可以使用套索工具在聚类图中框选想要探究的学习者。例如在**图 9** 中，我们框选了绿色群体“高效学习者”最下方的边缘点，通过观察画像图我们发现这一簇群体有在“热情加成”和“排名”两个维度高于“高效学习者”的平均值。观察外圈知识点掌握程度我们发现，这部分群体在知识点掌握上也略微优于平均值。通过观察周日志视图（**图 10**），我们发现这部分群体在前三周就开始进行了答题，在第七周时已经全面掌握各个知识点。这再一次证明了**较高的答题热情和更合理的时间安排有利于促进知识的掌握和吸收。**

图 9： “高效学习者”边缘点探索

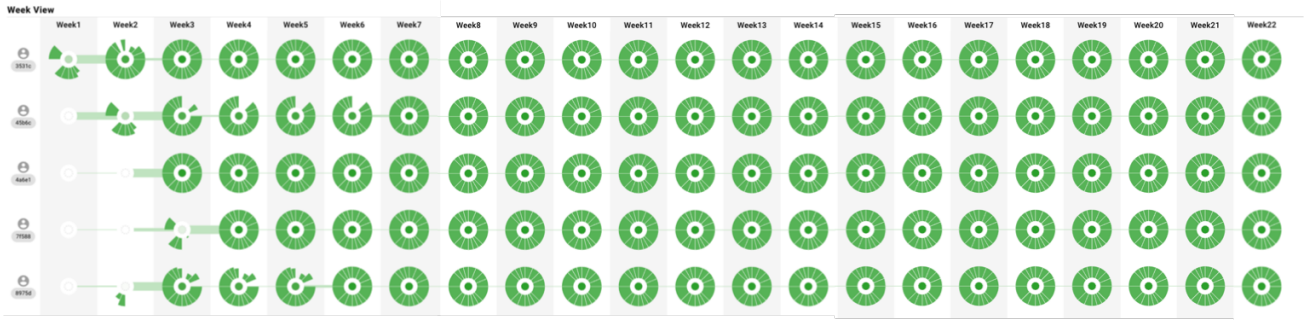

图 10： “高效学习者”周日志视图

我们用同样的方法分别探索了“高效拖延者”的边缘点以及“积极探索者”的离群点，可视化结果如**图 11** 所示。观察左图，我们可以发现这一小簇群体在学习习惯上更加“拖延”，答题热情的值更低，也因此这一簇群体在知识掌握程度上低于“高效拖延者”的平均水平，尤其是对于知识点 b3C9s 和 g7R2j 的掌握受到“拖延”的影响更大。观察右图，我们发现这一簇离群点属于“积极探索者”中的“非积极”群体，他们对于各个知识点的掌握程度较差，但对于这类群体，知识点 t5V9e 和 r8S3g 相对更容易掌握。综上所述，**少部分知识点相对简单，不需要较高学习热情就能基本掌握，但大部分知识点需要较高的学习热情和学习习惯以及较强的时间管理能力去促进其掌握、整合和吸收。**

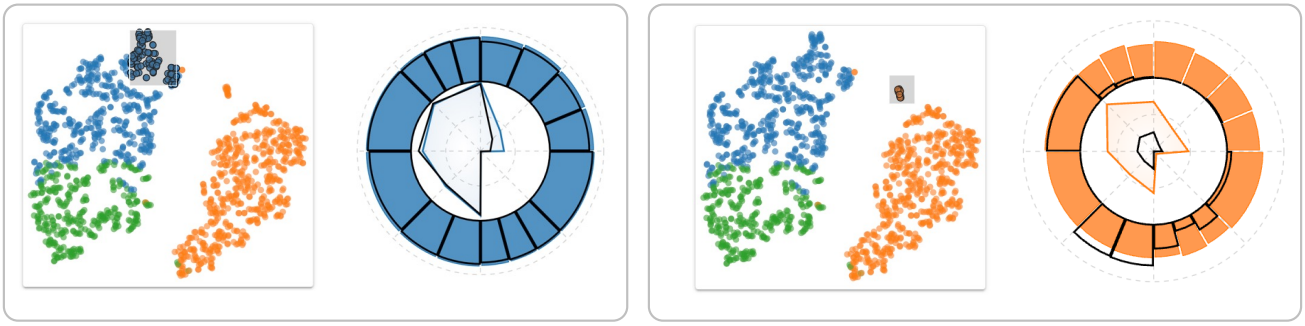

图 11: “高效拖延者”与“积极探索者”离群点探索

4、合理的题目难度应当与学习者的知识掌握程度相匹配，当学习者知识掌握水平很高但答题正确率较低时，意味着题目难度超出了其能力范围。请试着利用可视分析方法找出这些不合理的题目。

4.1 题目表现设计

题目难易程度的讨论应当相对于同知识点下的其余题目对比得出，因此在题目视图（**图 1-C**）中，我们对同一知识点下的题目表现进行可视化，对比展示题目间差距。

一个难易程度适当的题目，其平均得分应当与提交次数相关，学习者在该题上的得分应当随着答题次数的增大而提高，而一个难度偏大的题目，学习者的平均提交次数可能会较高于正常题目，且平均得分较低于正常题目。我们将一道题目的平均提交次数映射为圆环外圈的半径，平均得分情况映射为内部圆圈半径，通过两者之差定义该题的困难程度等级，并将其困难程度高低映射为视图中竖线的高低。另外，该题的提交时间分布通过折线区域图进行展示，竖线的横向位置表示该题的平均提交时间，而题目的得分情况则通过下方的堆叠柱状图进行展示。

该视图与画像视图进行联动，通过画像视图中所选择的知识点，展示该知识点下的题目表现情况，且支持按照题目的困难程度进行排序。

4.2 题目表现分析

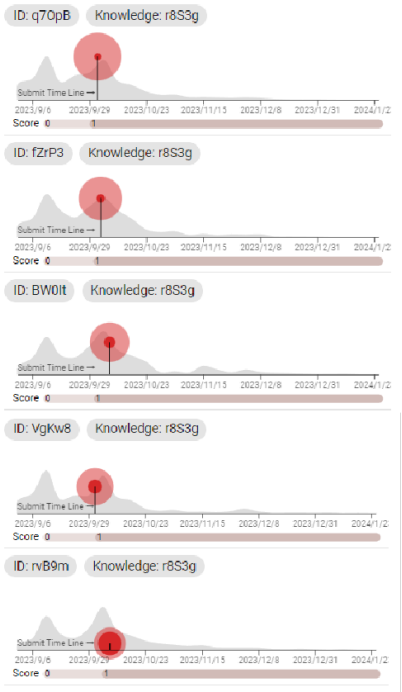

图 12: 学期初期题目视图

选择 r8S3g 知识点，该知识点下题目表现视图如**图 12** 所示，可以发现该知识点下的五道题目中，只有题目 rvB9M 在较低的提交次数中达到了较高的平均得分，其余四道题目则表现出了较高的提交次数和较低的平均得分。观察题目的提交时间分布可以发现，该知识点下的题目提交记录均位于 2023 年 9 月至 10 月，推测学生在学期初由于对编程题作答熟悉程度低，因此答题尝试次数较多，从而拉低了该知识点下题目的平均得分情况。

图 13: 较难题目筛选

选择 t5V9e 知识点，如**图 13** 发现该知识点下的五道题目中，题目#3oPy 出现了最高的难度等级 Hard，其平均提交次数达到 7.32 次而平均得分只有 0.41 分（满分 2 分），题目#3MwA 平均提交次数达到 8.85 次，难度等级为 Medium，其平均得分 0.75 分（满分 2 分），该知识点下的其余 3 道题目则表现出了适当的平均提交次数和平均得分。

图 14: 学生视图

结合画像视图进行分析可以发现学生整体对知识点 t5V9e 的掌握程度处于较高水平，另外，在学生视图（**图 14**）中可以发现，部分同学对知识点 t5V9e 下题目表现整体较好，却在题目#3oPy 上表现较差，学生#1kvfw 在对该题尝试了一定次数之后仍然无法达到满分，学生#40rkj 对该题做出时间复杂度较大的提交后仍然无法获得有效分数。因此我们可以认为在该知识点下，题目#3oPy 的难度较高，超出了同学们的能力范围。

使用题目视图对全体题目按照难度进行排序，可以发现的其余难题包括# q7OpB、#fZrP3、#BW0It、#n2BTx、#xqlJk，我们为这五道题目赋予了 Hard 的难度标签。

5、结合上述分析结果，请你为题目设计者和课程管理人员提供一些宝贵的建议，以优化题库内容设置和改善教学质量，并简要说明理由。

优化题目设置：针对那些学生在题目上花费过多时间却没有取得相应分数的情况，建议题目设计者进行评估并修改这些题目。可能的改进包括简化题目、提供更明确的指导、增加实际应用场景等，以提高学生对这些题目的理解和解答效率。

强调时间管理：针对临近答题截止时间才开始提交的学生群体，建议课程管理人员在教学中强调时间管理的重要性。可以提供时间管理技巧和策略，鼓励学生合理规划答题时间，避免拖延并提高答题效率。

鼓励多次提交：对于那些提交作答次数较少的学生，建议课程管理人员鼓励并促使学生多次提交答题。多次提交可以提供学生更多的机会来改正错误和提高答题质量，同时也可以帮助他们更好地理解题目要求和知识点。

分析早期提交学生的优势：对于那些提交时间较早且得分情况较好的学生，可以对他们的答题行为进行分析并挖掘其成功的因素。这可能包括他们的学习方法、时间管理技巧、优秀的理解能力等。将这些成功因素分享给其他学生，以帮助他们改善学习和答题效果。

处理偏科现象：针对那些存在偏科现象的学生，建议题目设计者和课程管理人员对相关的知识点进行进一步分析和调整。可以增加与偏科知识点相关的练习题目数量，提供更多的辅导材料和支持，以帮助学生加强对弱势知识点的理解和掌握。