浙江财经大学

中国大学生计算机设计大赛

项目名称 基于学生画像的数据挖掘与分析可视化系统

项目成员:

姓名学号: 尹玉龙 200110900432
学院专业: 浙江财经大学-信智学院-软件工程

姓名学号: 赵坚 210110900640
学院专业: 浙江财经大学-信智学院-软件工程

姓名学号: 倪承枫 200110900418
学院专业: 浙江财经大学-信智学院-软件工程

指导教师: 张翔

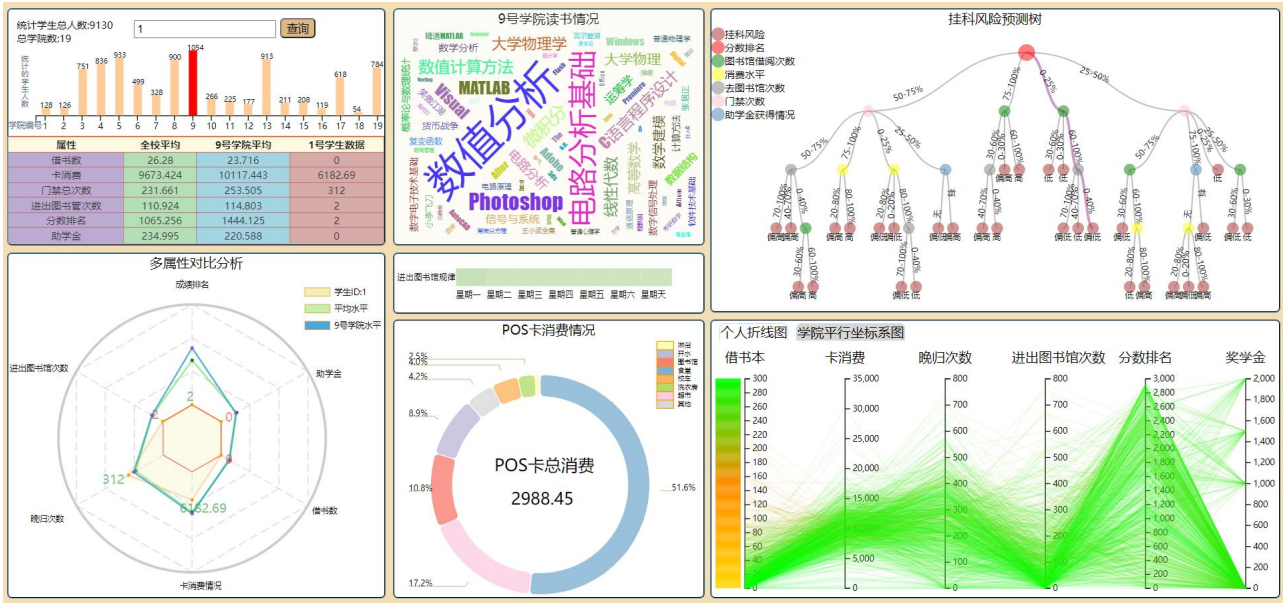
摘　要

学生画像是基于学生在校数据呈现的学生在校特征。学生本人和教务处能通过学生画像能了解学生自身情况，这能让学生自身注意学习和生活方面的问题，也便于教务处管理学生的生活作息和学习。

我们的作品通过可视化的技术直观地呈现学生在校期间的各项数据，这包括用饼图、柱状图、雷达图、树图等呈现学生在校特征信息。此外，我们还利用决策树模型对学生挂科风险进行预测，根据预测结果来警示学生做好学习规划，及时调整学习状态。

目 录

一、作品简介



作品主题

本作品基于学生每天产生的一卡通实时数据，通过数据整合、分析，利用大数据挖掘与分析技术、数学建模理论，可以挖掘出每个学生的学习、生活状态，预测出学生的挂科危险。基于学生在校产生的数据进行学生画像分析，从而解决学生管理难，学生信息掌握不准确和不及时等问题，及时定位和纠正学生发展过程中存在的问题。

主要功能

本作品搭建的学生画像可视化系统总共涉及 10 种可视化图表，分别是：查询框、柱形图、文本框、折线图、饼图、雷达图、词云图、文本框、平行坐标系图和渐变图，以此来从不同方面分析学生数据，实现了例如学院的图书阅读情况，校园卡消费情况，挂科的风险预测等等的易理解的可视化，其中各自的具体功能在**系统搭建**一栏中有详细介绍。

目标用户

本系统主要用于分析学生在校的生活、学习等多属性状态，面向的用户为高校学工，教师以及辅导员等从事学校生活管理的工作者与组织。

作品意义

①设计出多种可视化图表相结合的数据可视化系统，直观地展示学生信息；

②利用决策数模型对学生是否有挂科的风险进行预测；

③设计出多图形的交互，从多方位分析学生在校期间的学习、生活等状态

二、数据处理

数据的处理我们主要分为四个部分，分别是：删除冗杂数据、整合学院总体数据、 整合学生个人数据和决策树模型构建及其数据分类。

主要操作如下：

①删除冗余数据：借助 python 的字典，我们对原有的 txt 文件里的数据进行了初 步的数据清洗，快速地删除掉了原有数据里面的冗余数据。

② 整合学院总体数据：通过 python 的 os、json 以及 csv 库，我们分别整合了各学院整体的图书借阅信息、总人数信息和平均消费水平信息等信息，用于后期在各学院书 籍阅读词云图、各学院人数柱状图和学生平均学习及生活情况表中展示出各学院学生的 整体学习和生活状态，同时作为判断学生是否有挂科风险或出现“特殊情况”的依据。

③ 整合学生个人数据：我们借助 python 将学生个人的借阅图书及进出图书馆的数据以季度为单位进行统计整理，并将学生个人包含图书借阅和消费情况等学习及生活情 况的数据进行整合，将其以字典的形式储存在 json 格式的文件中。该数据将用于后期学生个人学习及生活情况的可视化展示，以及挂科危险和出现“特殊情况”的预测。

④ 决策树模型构建及数据分类：我们从上步得到的学生个人数据中随机抽取了一百位学生的学习及生活情况数据，进行打标签分类处理后作为训练集构建了决策树模型。之后将全部学生的个人数据进行处理后导入决策树模型中尽心预测分类处理，得到含有挂科危险及出

现“特殊情况”风险标签程度标签的数据，用于决策树的绘制及学生有挂科危险及出现“特殊情况”的风险程度的预测展示。

三、系统搭建

3.1 算法模型

决策树算法是一种基于数据的归纳推理分类方法，它可以通过对决策表形式的无序样本进行推理学习，建立用于分类判断的决策树，从中提取出一些隐含的规则。由于学生挂科风险预测的决策表涉及的属性分支大于等于二，而 CART 算法只能产生二叉树，C4.5 算法相对 ID3 算法有改进，因此我们选择了 C4.5 算法作为来生成我们的挂科风险预测的决策树。

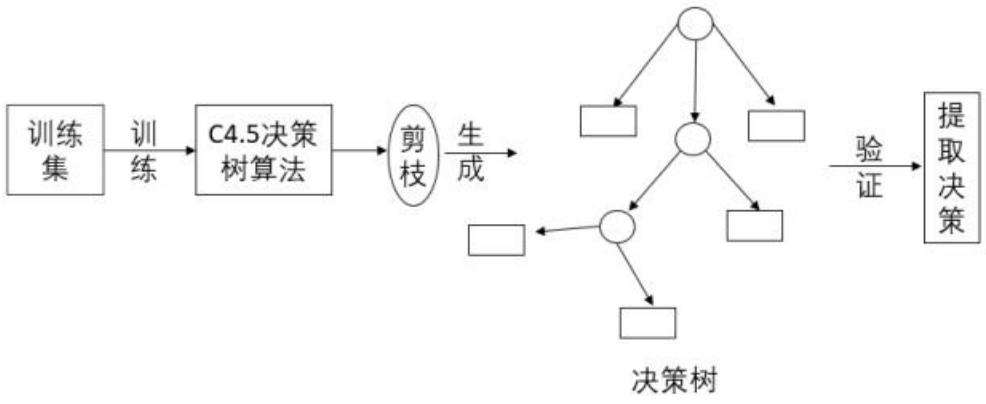


图 1

如图 1 所示，决策树的建立需要经过 C4.5 的决策树算法以及剪枝两步。

C4.5 算法

为利用 C4.5 决策树算法，我们定义了一个 7 元组的属性集< 成绩排名，图书借阅数，进出图书馆次数，晚归次数，卡消费情况，助学金获得情况，挂科风险>以及与之对应按照划

分标准创建的训练集决策表，如图 2 所示。C4.5 的算法将以不同的属性集中每一个的属性进行划分得到一个让预测结果的信息增益率最大的一个属性作为根节点，并以这个属性的划分条件作为不同的分支，以这种相同的方式递归找寻到的节点作为对应分支的节点。以此类推递归地使信息熵逐渐降低，直至子样本或子样本集中的所有样本均属于同一预测结果时停止递归，并将其子样本集中属于挂科风险属性的最大概率的划分标准作为叶子节点。最后得到图一中的决策树，树上的叶子节点都属于挂科风险。上述的信息增益率公式如下:

$$-\sum_{i=1}^{m} P_i \log_2\left(\frac{P_i}{N}\right)$$

剪枝：为了减少的决策树的过拟合性的产生，我们在 C4.5 的算法基础上增加了两种剪枝体条件。①.在递归找寻节点时若未找到可以使信息熵降低的划分方式，则停止对该分支的划分，并以子样本集中属于挂科风险属性的最大概率的划分标准作为叶子节点。②.当子集中的属于挂科风险属性的划分标准的纯度大于 95%时将该划分标准作为叶子节点。

挂科风险	成绩排名	图书借阅数	进出图书馆次数	晚归次数	卡消费情况	助学金获得情况
高	70-100%	30-60%	25-50%	0-40%	20-80%	有
偏高	75-100%	0-30%	0-25%	0-40%	20-80%	无
低	0-25%	0-30%	0-25%	40-70%	20-80%	无
……	……	……	……	……	……	……

图 2. 训练集决策表

初始决策表的建立 我们将给定的数据进行分析，可知学生的挂科风险可能与成绩排名、进出图书馆次数、图书借阅数、晚归次数、卡消费情况、助学金获得情况有关。并根据其可能存在的相关性分别给定如下的划分标准：

成绩排名：0-25% 、25-50%、50-75%、75-100%，排名百分比越大排名越靠后。

图书借阅数：0-30%、30-60%、60-100%，百分比越大借书数越少。

进出图书馆次数：0-25% 、25-50%、50-75%、75-100%，百分比越大次数越少。

晚归次数：0-40%、40-70%、70-100%，百分比越大次数越少。

卡消费情况：0-20%、20-80%、80-100%，百分比越大消费金额越少。

助学金获得情况：有、无

同时预测的挂科风险分为：低、偏低、偏高、高这四个等级。在上述的规则下我们将真实的数据处理为图二.训练集决策表类似的 json 格式并选取其中的部分，给其中的挂科风险属性打上对应的标签，最终到的数据作为生成决策树的训练集 json 格式）。

提取决策 对于给定的学生 ID，可以找到其对应属性所在的划分区间。如图 3，根据训练得到的决策树，我们可以按照学生的属性信息以及对应的划分区间依次在决策书中找到对应的路径，该路径是唯一的且，其通往的叶子节点即预测得到的该学生的挂科风险。

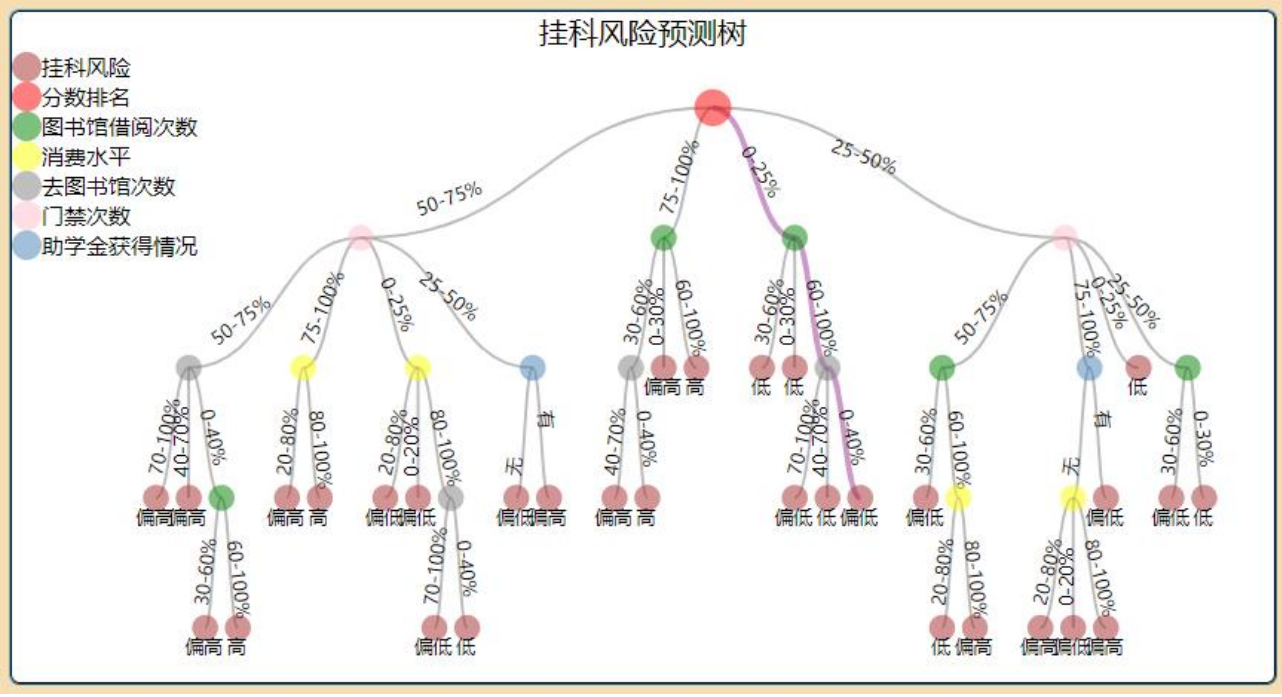
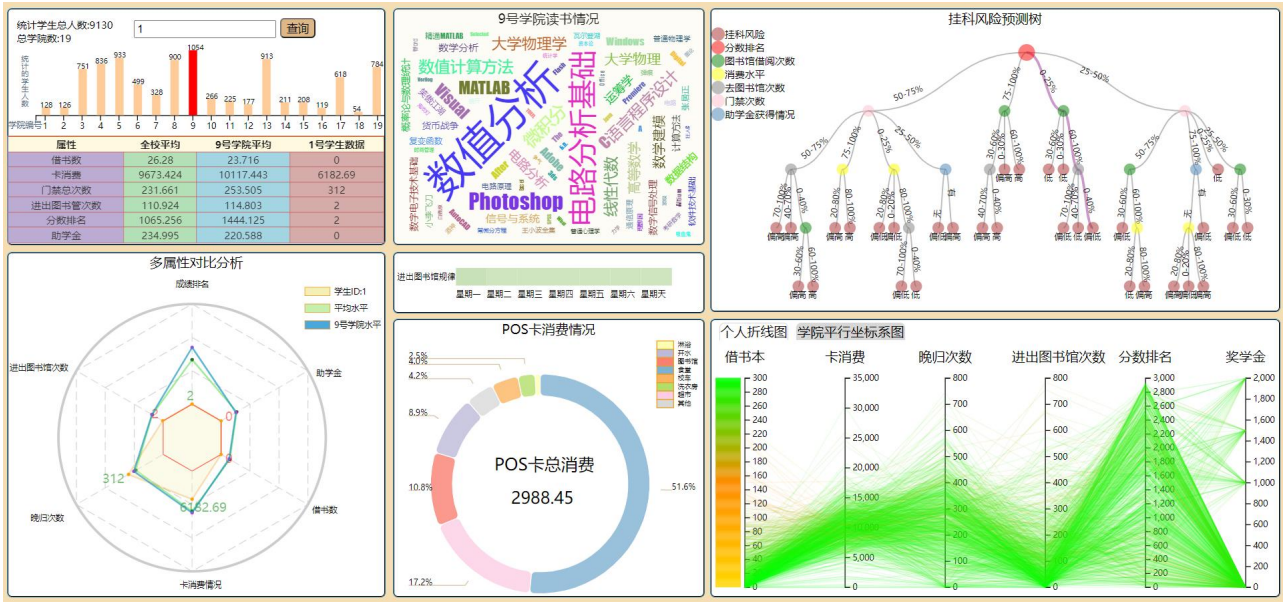


图 3

3.2 可视化系统界面



3.2.1 可视化图表与其功能

查询框：查询框是通过学生的 ID 来进行学生信息的查询的，是本系统最主要的输入端口以及实现信息交互的大门。左侧有显示学生和学院的总数量，方便了解大体内容。而且在输入其他文字或空白 ID 时，会弹出提示。 如图 4 所示：

图 4:查询框

柱形图：柱状图是用于直观地展示各学院的学生数量，横坐标为学院的编号，纵坐标为各院学生的数量，并在上方显示具体数据。当鼠标点击选中时变为红色并与其他窗口交互，交互的具体功能在交互栏目有详解。如图 5 所示：

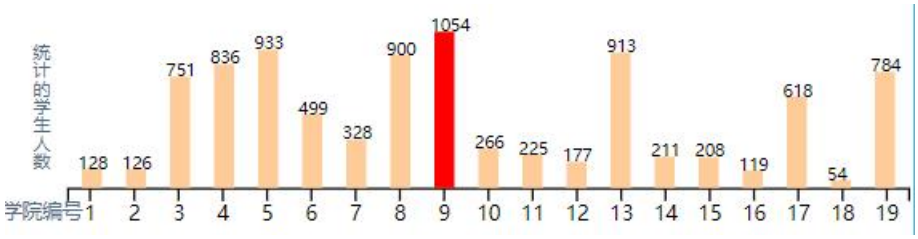


图 5：柱形图

文本框：文本框可以用来展示文字信息， 对比不同类的数据。本系统中，我们以此来展示不同属性的校平均， 学院平均与学生之间二点对比， 分析个人与整体的关系， 如图 6:

属性	全校平均	9号学院平均	1号学生数据
借书数	26.28	23.716	0
卡消费	9673.424	10117.443	6182.69
门禁总次数	231.661	253.505	312
进出图书管次数	110.924	114.803	2
分数排名	1065.256	1444.125	2
助学金	234.995	220.588	0

图 6：文本框

折线图：折线图可以很好地展示出事物变化的趋势，本文中，我们用折线图来展示学生进出图书馆的次数随着不同学年季度的变化和学生借书数量的变化趋势。横坐标代表不同学年季度，纵坐标代表数量，紫色的线条表示图书馆进入次数变化，橘色的线条表示的是借书本书的变化，如图 7。如果该学生去图书馆次数太少，则会如下图 8:

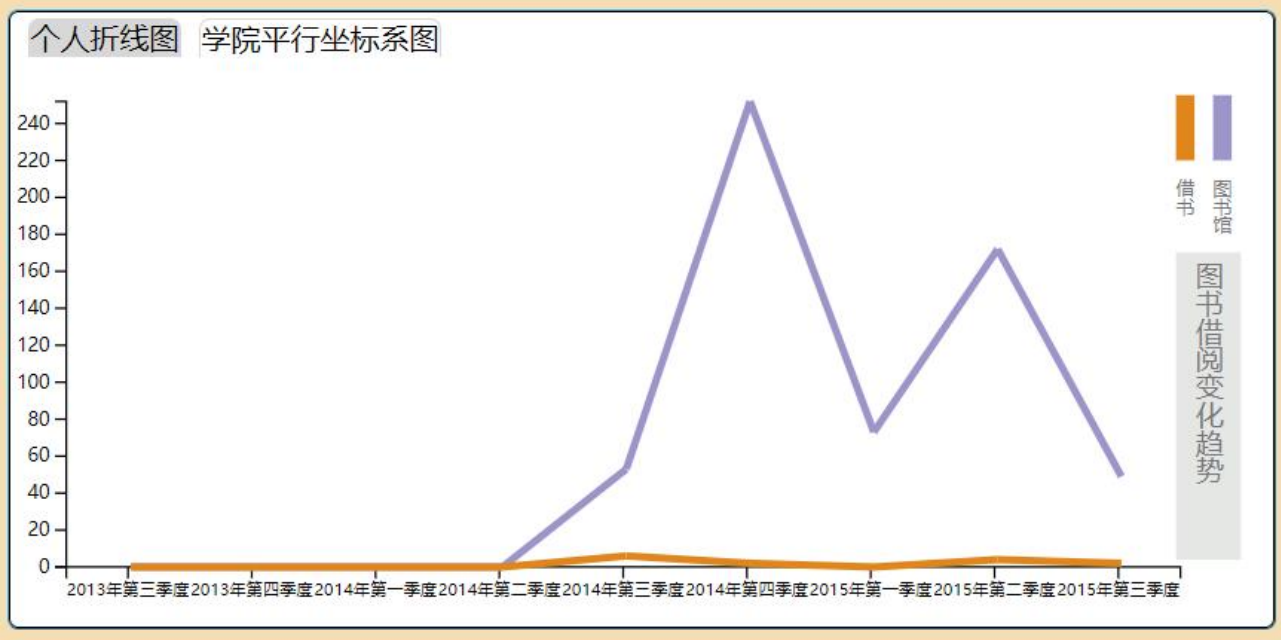


图 7

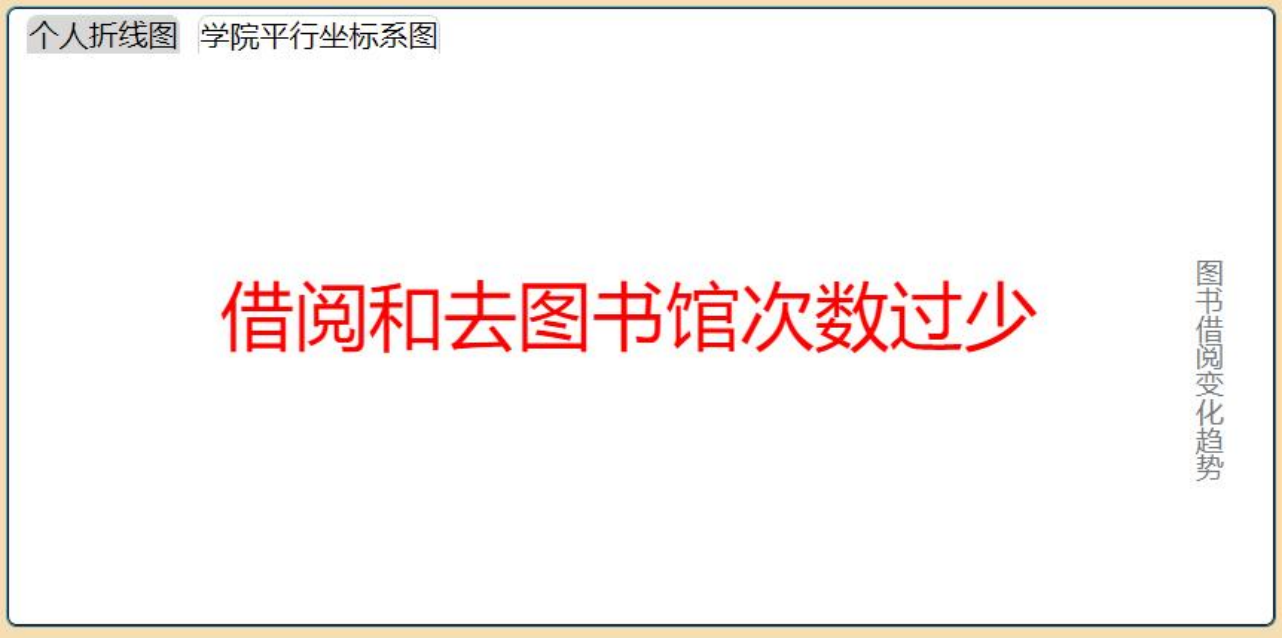


图 8

雷达图：雷达图可以同时展现出多维属性信息，以便于从多个方面分析数据。本系统中，我们从成绩排名、助学金、图书借阅、POS 卡消费、晚归次数、进出图书馆次数六个维度来分析该学生的水平，并将其与校平均水平、学院水平相对比，以此来更为完善，全面的对比学生的生活，学习情况，方便个体的问题分析。图中黄色区域代表学生数据，绿色区域代表校平均水平，蓝色区域代表 学生所在学院的水平，如图 9:

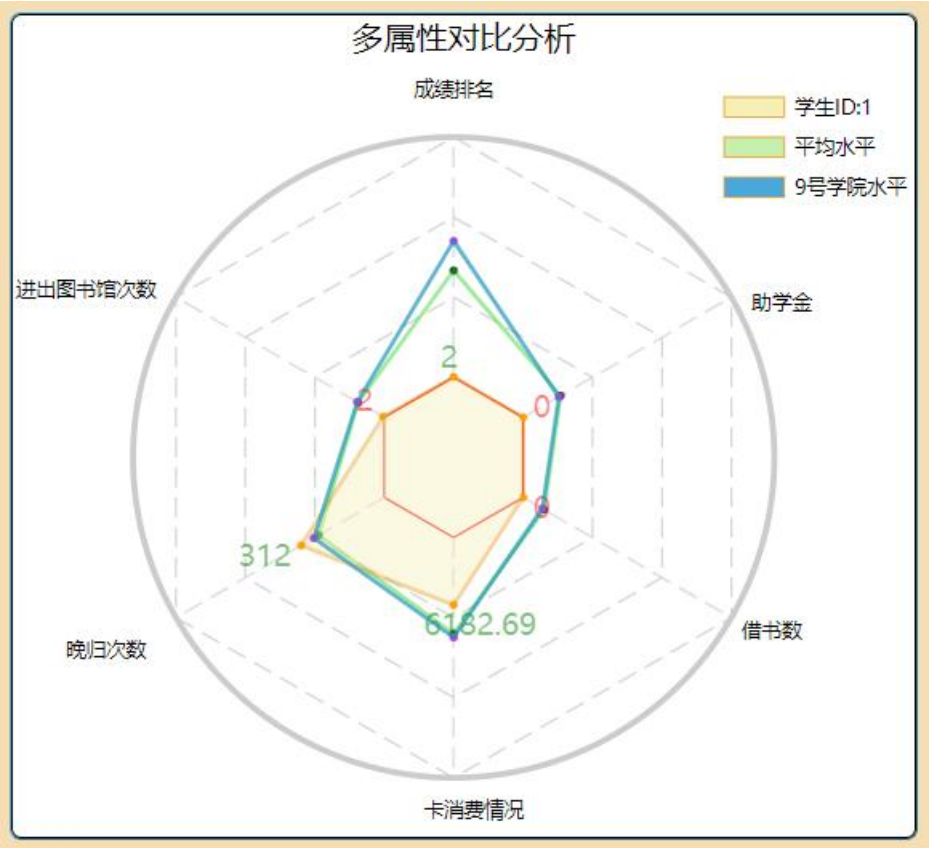


图 9:雷达图

饼图：饼图可以用来具体清晰显示部分占全体的百分比。我们用饼图来展示学生的消费占比，在旁标出消费所占的数据。以此来分析学生在校额生活消费情况，如图 10。

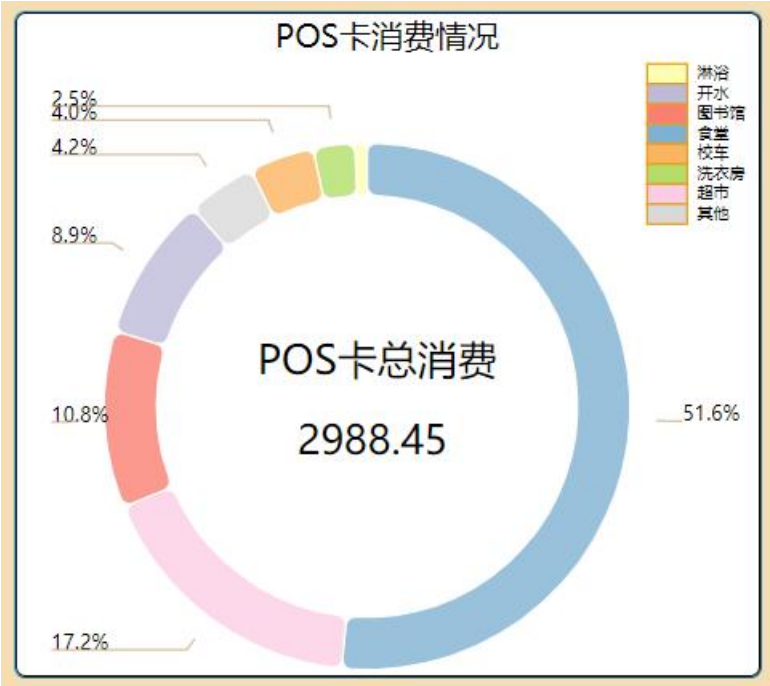


图 10：饼图

词云：词云可以根据不同词语出现的频率来对控制改词可视化出来的效果，出现的频率越大，显示出来的词语字体越大，反之出现的词语频率越小，显示出来的词语字体越小，甚至不显示。本文中，我们对按学院进行分类，来展示不同学院学生阅读书本的兴趣，以此分析不同学院学生额差异和学生对书本的兴趣倾向，如图 10。

图 10:词云

平行坐标系图：平行坐标系图通常用来展示高维空间上数据的变化。本系统中，我们借平行坐标图来展现不同学院学生的水平趋势，每一条线代表不同的学生，从不同的维度：借书本数、消费数据、晚归次数、进出图书馆次数、分数排名、助学金来分析出学院学生的画像，线条初始颜色由借书本书来确定。通过整体趋势使分析更加具有全面性，如图 11：

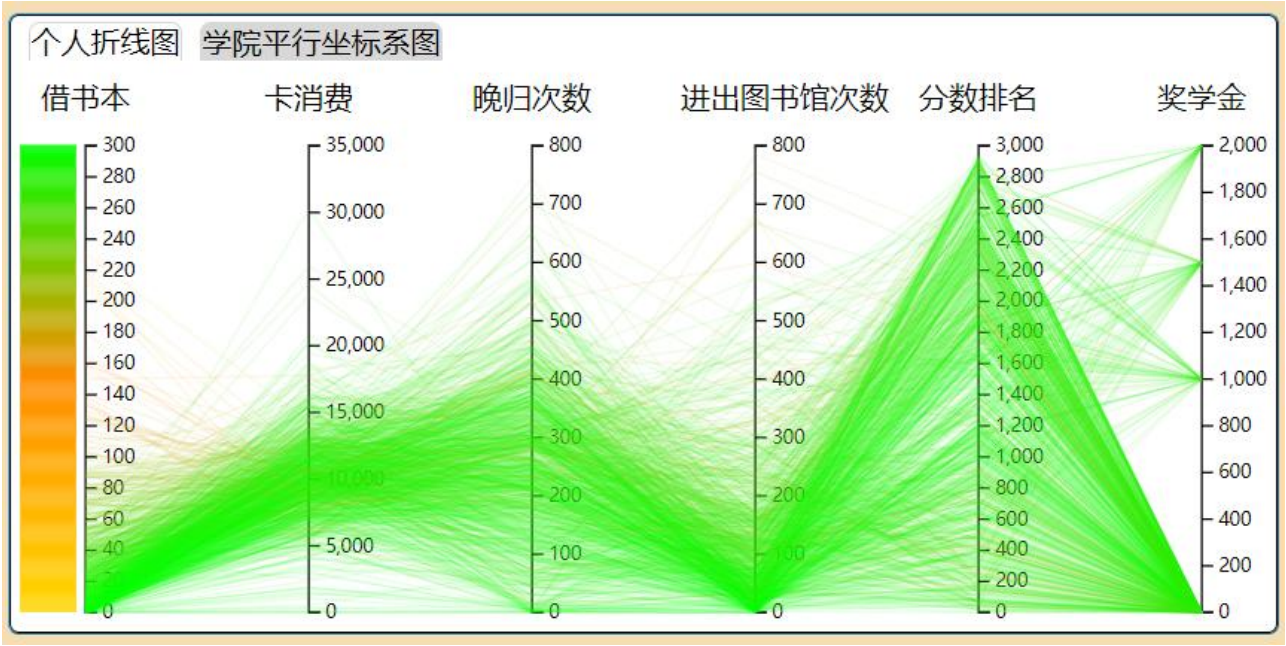


图 11:平行坐标系图

渐变图：渐变图根据颜色的深浅来表示事件出现的频率。本系统中，我们对学生两年来所有星期的进图书馆次数使用渐变图进行展示，结合其他图表分析出学生的学习状态，颜色越深表示去图书馆的频率越高，以此来分析出此学生的常用学习时间。如图 12：

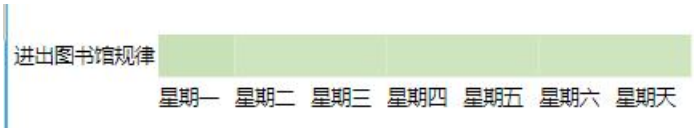


图 12：渐变图

3.2.2 交互

交互是可视化系统最为重要的一环，可以说灵活交互的系统带给将成倍提升用户的使用体验。这里我们分为两方面：**整体性交互**和**局部交互**来说明本系统所含有的交互功能。

整体性交互：主要通过查询框来查询学生信息， 进而使整个系统页面交互变化成该

学生的学生画像系统。

现在让我们随便输入一位学生的学号，就比如 id：1 的学生吧，系统展示如图 13:

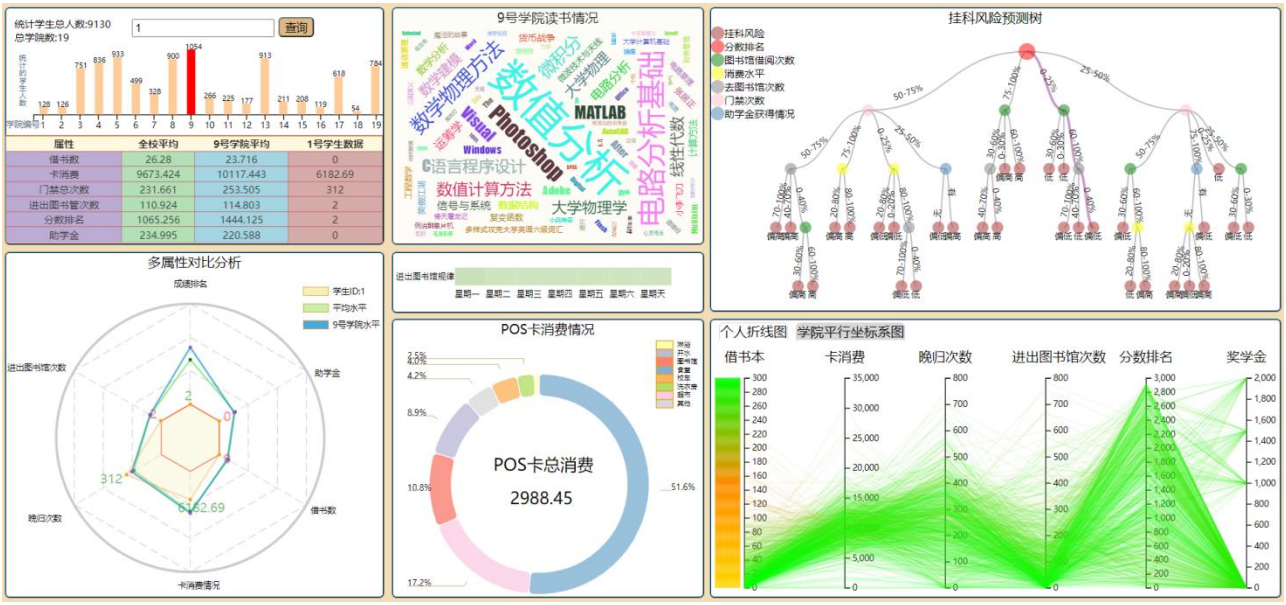


图 13:学号为 1 的学生页面显示

不难发现，柱形图所对应的学生学院编号变成了该学生所在的 9 号学，表格也变为该学生以及其所在学院的数据，词云相应地显示出 9 号学院学生的阅读书本的情况，雷达图的各颜色区域的面积分布相应地改变了，饼图的百分比也显示成该学生的具体消费百分比，而决策树模型将预测出该学生的挂科风险。 渐变图颜色的深浅也发生了变化。

局部交互：局部交互在本系统中是指当鼠标悬浮在不同可视化图表之上或者点击选择部分图表时所相应显示出的交互。还是 1 号学生。

当鼠标悬浮在雷达图的黄色区域（学生水平区域）时，会显示如图 14 效果:

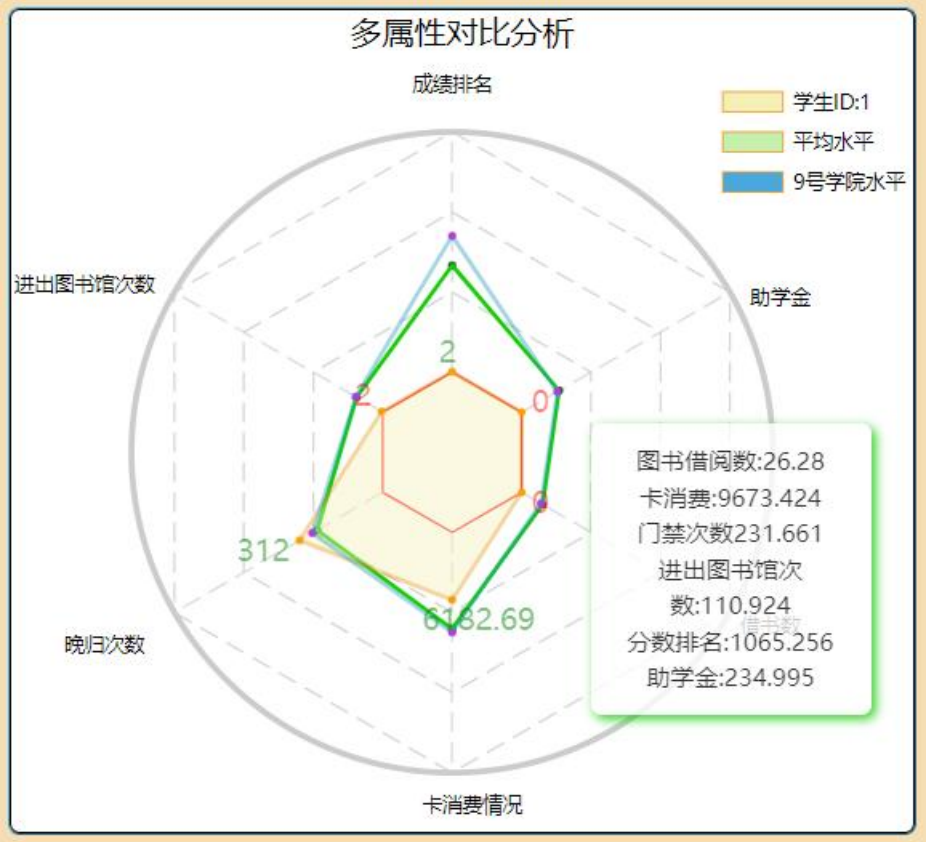


图 14：雷达图的交互

柱状图中红色代表该学生所在学院。当点击其他学院时，系统所有关于学院的显示数据和可视化图片都会变为点击学院的数据。如图 15 所示：

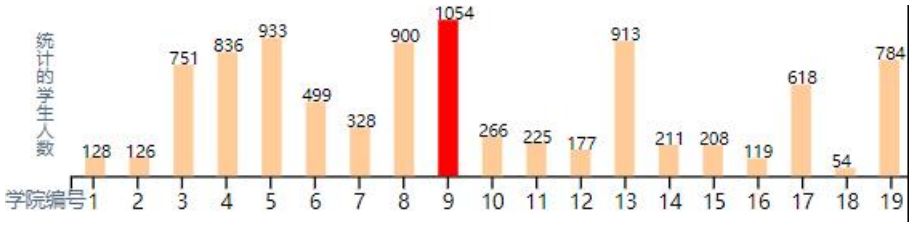


图 15：柱形图交互

当鼠标悬浮在词 云图上所对应的书本名称时，会显示出该书本借出的具体次数，如图 16 所示：

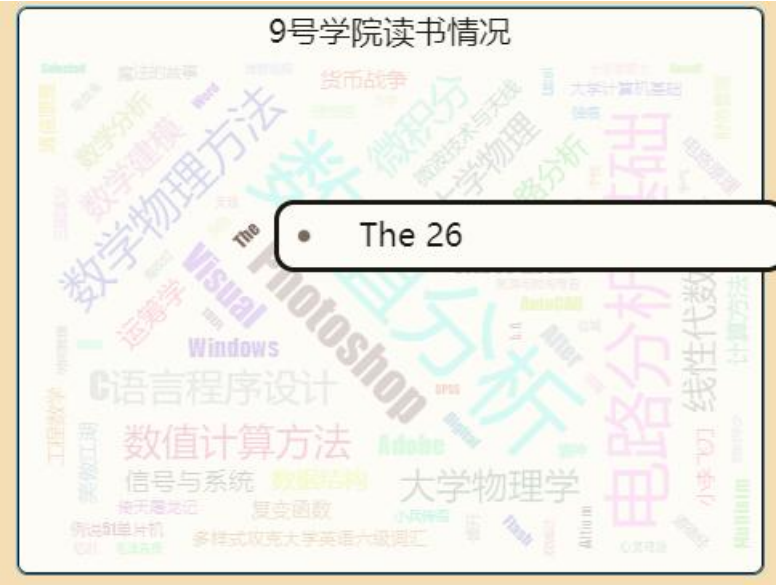


图 16:词云图交互

不难看出，此书被借阅了 26 次。

而当鼠标悬浮在平行坐标系中不同的线时，会在左侧的颜色条上显示出其相应的数量大小，如图 17:

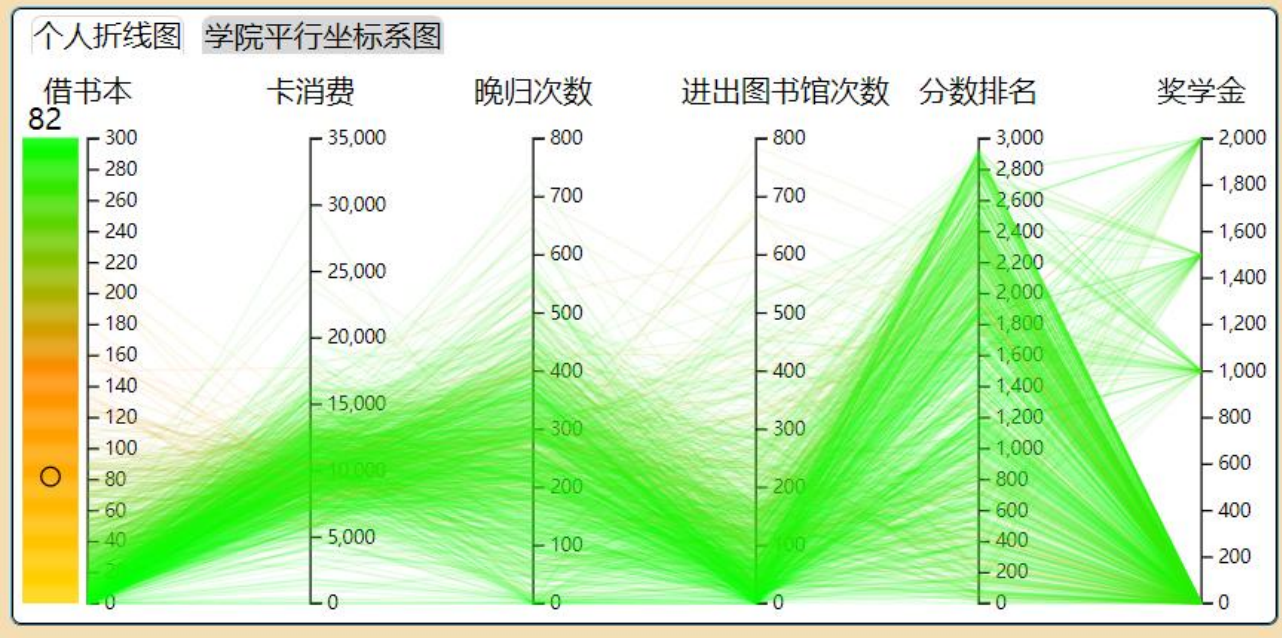


图 17:平行坐标系交互

通过整体交互与局部交互相结合的方法，可以增加系统分析的灵活性，使数据得到

最大化的利用。

四、案例分析

对于学生画像可视化系统的构建，我们主要从两个方面来进行案例的分析，分别是：学生群体画像分析与学生个体画像分析。

学生个体画像分析我们又分成：学习方面和生活方面和挂科风险预测三点。

让我们选用 id 为 13509 的学生来作为分析的对象，来演示并分析该学生和其所在学院的生活学习状态，系统表现如下图 18：

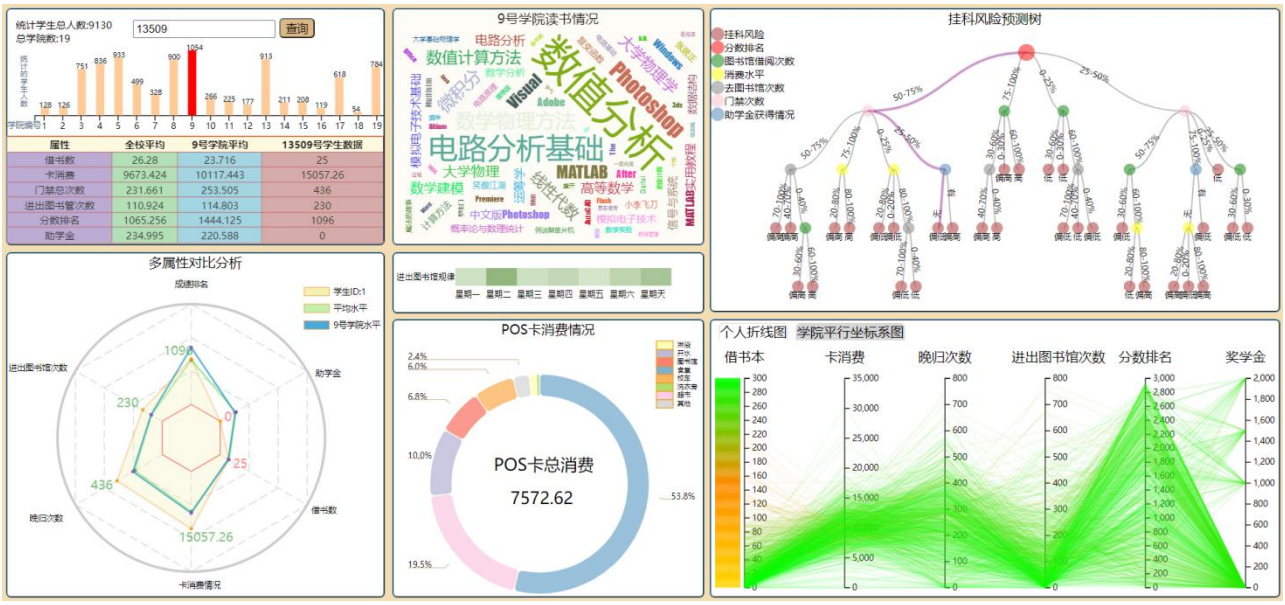


图 18：id 为 13509 学生画像可视化系统展示

学生群体画像分析：（id 为 13509）

为了分析学生群体画像， 我们主要通过柱形图、文本框、词云图和学院平行坐标系图来进行分析， 为了方便观察， 我们将系统展示的图片拆分后， 整合在一起如下图 19 所示：

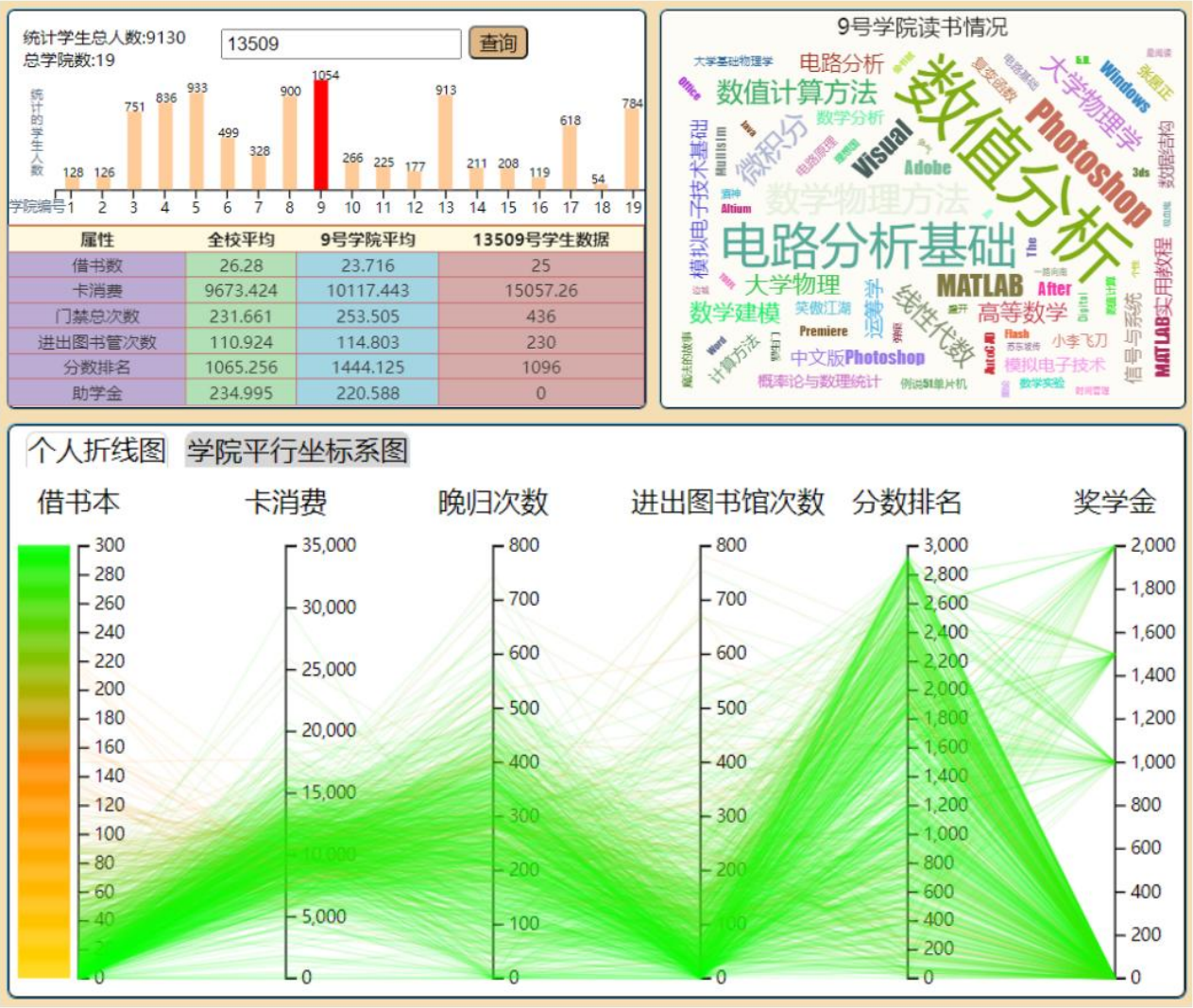


图 19：id 为 13509 学生所在学院群像分析

由柱状图我们可以看出该学生所在的九号学院人数在学校中占比较高，说明九号学院的学生所在的专业可能是该学校的重要专业，但在观察表格时我们发现该学院分数排名较低于全校水平，由此可以分析出这可能是学生数量太大使得难以集中管理导致的成绩偏低，学校管理者应该考虑加大辅导员的人数或是严格管理。

而通过九号学院图书阅读情况词云可以看出，《数值分析》和《电路分析基础》分别有 249 次和 171 次借阅量（图 21，22）。同时《大学物理》、《数学物理方法》等与物理相关的书借阅量较大，由此可以分析出该专业大概率为物理方向的理科专业，男生相对也就比较多，学院方面应从男生的角度分析其存在的问题。

通过平行坐标系图看出九号学院的学生借书阅读量集中于 0-50 本，通过分数排名的轴的分布看出学生在各个阶段都有分布，且在后面的线条数量略显密集分析出，即使为王牌专

业，但是可能学习难度较大、学生大部分水平靠后，也是这个学院平均水平低于学校的原因之一。

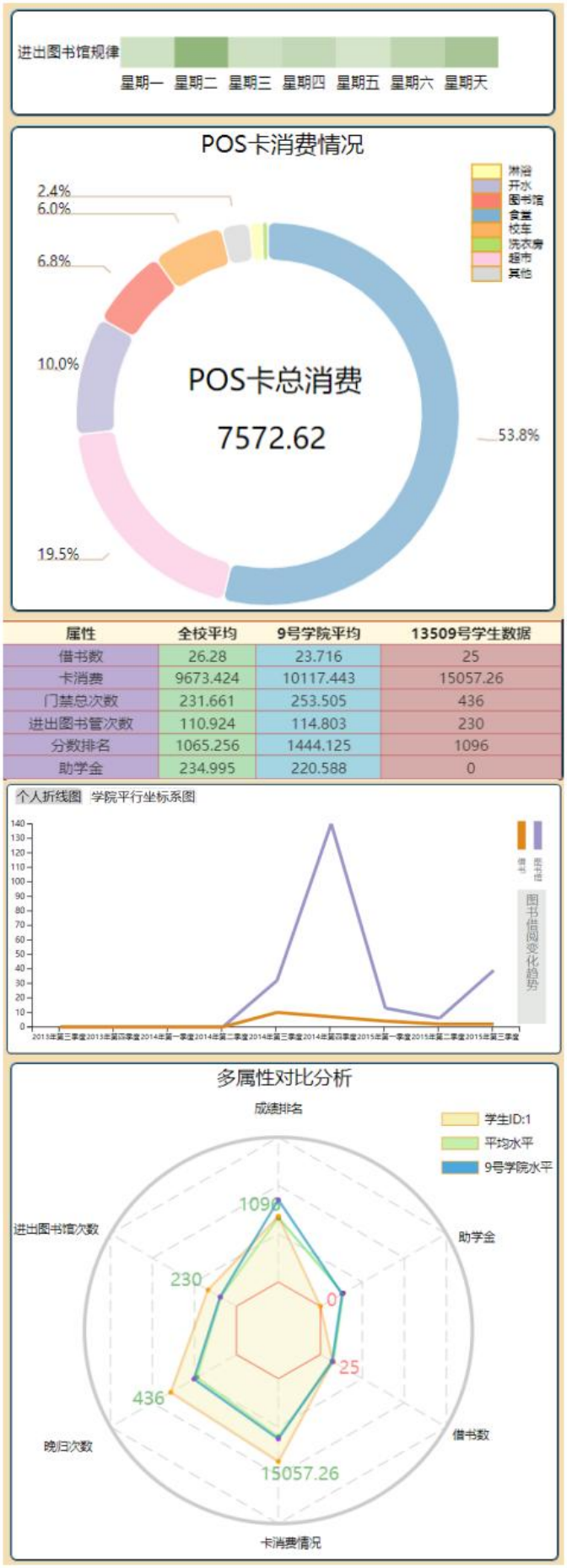
由此分析，该学院虽是学校重点学院但是管理人员不足，学生平时有在学习但是可能因为专业难度成绩并不理想。建议加大管理人员的投入，合理安排教课计划从而强化教学效果。

图 20:《数据分析》阅读量

图 21：《电路分析基础》阅读量

学生个体画像分析：

学生个体画像主要是通过文本框、雷达图、饼图、渐变图、折线图和决策树模型来进行分析。同样为了方便观察，我们系统展示的图片拆分并整合如下图 23 所示：



属性	全校平均	9号学院平均	13509号学生数据
借书数	26.28	23.716	25
卡消费	9673.424	10117.443	15057.26
门禁总次数	231.661	253.505	436
进出图书管次数	110.924	114.803	230
分数排名	1065.256	1444.125	1096
助学金	234.995	220.588	0

图 23：分析学生个人画像可视化图表-->

（一）学习方面：

通过雷达图和文本框可以看出该学生进出图书馆次数上高于学校和学院平均水平，并且晚归次数超出学院和学校水平，推测出该学生习惯在图书馆学习，并学习到很晚，但是通过成绩排名对比发现该学生成绩并不突出，推测可能没有掌握正确的学习方法，渐变图显示，该学生在周二和周日进入图书馆学习的次数较多，推测出该学生在周二和 周日的课比较少，自由管理的时间较多，学校可在周二和周日对其定制个性化学习方法

（二）生活方面：

通过饼图展现的 POS 卡消费情况来看，该学生消费水平高于学院和学校水平，并 且多用于食堂消费（53.8%）和校园超市消费（19.5%），表明该学生生活水平较高且主 要在学校活动， 不常出去聚餐， 结合上述学习方面的分析得出该学生是一个内向， 不爱与人交流的学生。

（三）挂科风险预测

通过决策树模型可以看出， 如图 24：根据决策树的高亮节点以及高亮路径我们可以 获取到该学生的部分属性的排名情况以及使用决策树预测得到的该学生的挂科风险， 如 图 24。根据高亮出来的部分表明该学生的分数排名在 50-75%区间内、去图书馆的次数 在全校水平的 25-50%之间、他未获得助学金。路径的叶子节点为偏低，即预测的该学 生的挂科风险偏低（红色圆圈圈出）。

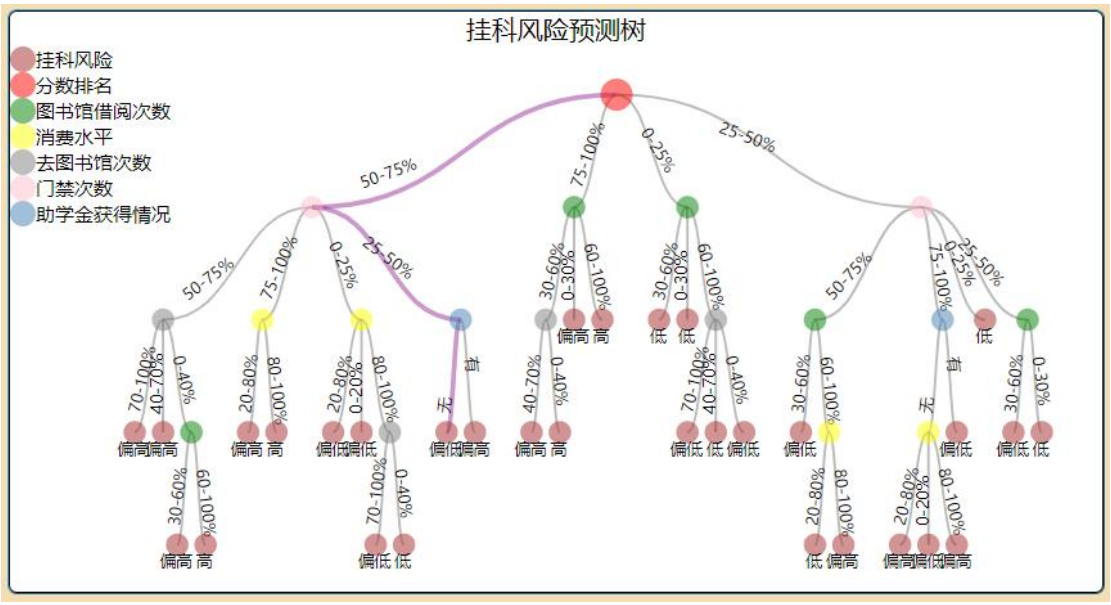


图 24:学生在决策树模型下的挂科风险

综上分析出，该学生在大学不爱社交，或有明确的学习目标但是并未掌握良好的学习方法，长此以往可能导致孤僻的性格。学校应对其疏导，鼓励学生参加课余活动。