大学生一卡通数据分析系统的设计与实现

软件学院软件工程专业 14140Y01班张丹

指导教师：章永来

1. 前言

随着数字化时代的发展，全国各个高校建立了属于自己的校园一卡通系统。每位学生每天会使用一卡通在校园内进行各类刷卡，进而产生大量的刷卡记录。但是目前各个高校使用的校园一卡通系统只是针对师生的各种业务需求实现了各种应用，缺乏的是对刷卡记录进行有效的分析，该课题是在此背景下对大学生一卡通数据进行了研究与实现，为学校决策提供充分的科学依据，使管理更规范、更有针对性，同时提高学校各个机构调整和改善的合理性。

1. 课题工作

2.1设计工作

采用B/S模式、微信小程序完成校园一卡通数据分析系统，使用户能够利用此系统清晰的了解该系统的设计思想和一卡通数据的分析情况，为个人生活和工作带来便利。系统提供如下功能：用户能够通过WEB查看系统；各个角色的用户可以在登录界面进行登录、可以从时间维度、消费类型与水平维度、POS机维度、性别维度、年级专业班级维度等对数据可视化、修改个人信息；系统管理员可对用户做增删改查操作；该系统实现者对一卡通数据预处理,将一卡通数据提交给Hadoop集群实现存储和分析,并通过不同维度对数据可视化展示。

2.2 开发技术

本课题利用HDFS技术实现一卡通数据的存储、Hive和MySQL技术实现一卡通数据的统计和分析、Html5展示系统概述、移动端微信一卡通小程序模拟用户行为并与WEB端相结合、K-means聚类算法分析一卡通数据进而得出学生消费水平、等距离离散化一卡通数据得出男女消费习惯、Matplotlib和Echarts进行可视化。

1. 系统实现

3.1 K-means聚类

3.1.1 K-means聚类结果可视化

K-means算法是聚类算法中比较简单、容易使用的算法，适用于处理大量的数据并且算法高效,主要思想是在划分多个中心点，不断取距离中心点最近值的算法。

轮廓系数法确定最佳k值，结合了凝聚度和分离度，轮廓系数取值为[-1,1]，其值越大越好。当值为负时，表明凝聚度<分离度，样本被分配到错误的簇中，聚类结果不可接受。对于接近0的结果，则表明聚类结果有重叠的情况。

手肘法：随着聚类数k的增大，样本划分会更加精细，每个簇的聚合程度会逐渐提高，那么误差平方和SSE自然会逐渐变小。当k到达真实聚类数时，再增加k所得到的聚合程度回报会迅速变小，SSE的下降幅度会骤减，随着k值的继续增大而趋于平缓。

本系统中根据学生一学期的消费金额和消费次数对学生进行聚类。利用Python通过轮廓系数和手肘部法分别计算出最优k值，根据k值对上述两个属性聚类，可以直观地看出聚类效果，最后可以查询学生所属的类别。对于学院管理员来说，可以合理地审核学生的家庭情况。对于餐饮消费位于中上等水平的同学，在每年国家助学金进行补助时可以较后考虑，而优先考虑餐饮费用处于中低消费水平的同学。通过计算所绘最优聚类效果如图1所示。

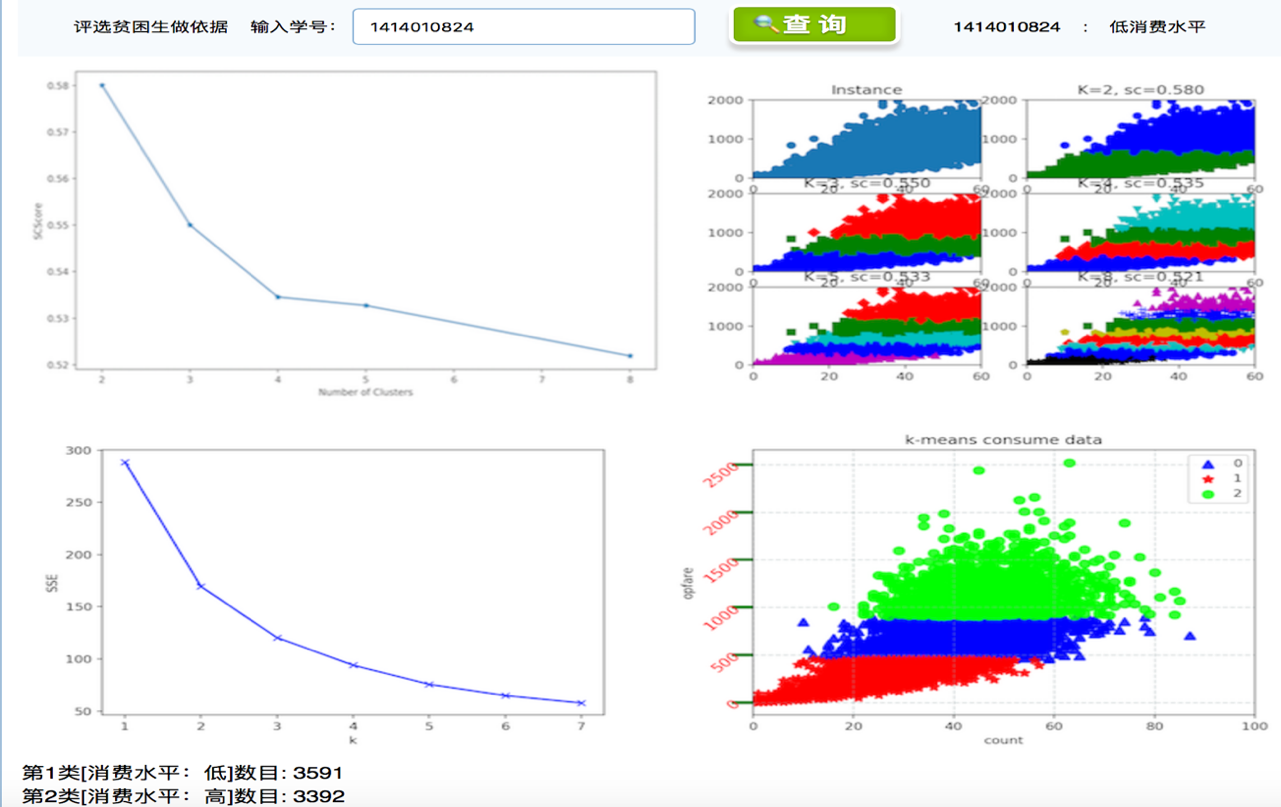


图1 聚类结果展示图

图1展示了K与SSE的关系图、K与轮廓系数的关系图，最终将最优K值设置为3。进而对原始数据按学生、总消费和总消费次数进行聚类分析，可以分析得出两个消费水平，同时可以查询每位学生具体所处的消费水平。

3.1.2 K-means聚类结果分析

上述结果得出，轮廓系数最大的k值是2，表示最佳k为2。但是从k和SSE的手肘图可以看出，当k取2时，SSE还非常大，这是一个不太合理的聚类数，我们退而求其次，考虑轮廓系数第二大的k值3，这时SSE处于一个较低的水平，因此最佳k应该取3而不是2。但是，k=2时轮廓系数最大，聚类效果应该非常好，那为什么SSE会这么大呢？可能在于轮廓系数考虑了分离度b，即样本与最近簇中所有样本的平均距离。从定义上看，轮廓系数大，不一定是凝聚度a（样本与同簇的其他样本的平均距离）小，而可能是b和a都很大的情况下b相对a大得多，这样a是有可能取得比较大的。a大，样本与同簇的其他样本的平均距离就大，簇的紧凑程度就弱，那么簇内样本离质心的距离也大，从而导致SSE较大。因此，虽然轮廓系数引入了分离度b而限制了聚类划分的程度，但是同样会引来最优结果的SSE比较大。

3.1.3 K-means聚类结果结论

从上述两例可以看出，轮廓系数法确定出的最优k值不一定是最优的，有时候还需要根据SSE去辅助选取。如果没有特殊情况的话，我建议首先考虑用手肘法。

3.2 等距离离散化

3.2.1 离散化结果可视化

离散化在数据分析中是将连续型数据进行分段。等距离离散化正是分段的一种原则，此方法是将连续型数据分为距离相等的“段”，即将连续变量均匀划分成n等份并且每一份的间距相等，进而分析多个段之间的关系。

在本系统中，对消费时间进行等距离离散化，进而分析消费者性别与消费时间的关系。

消费记录数量与消费时间点关系如图2所示。消费整体情况如图3所示。

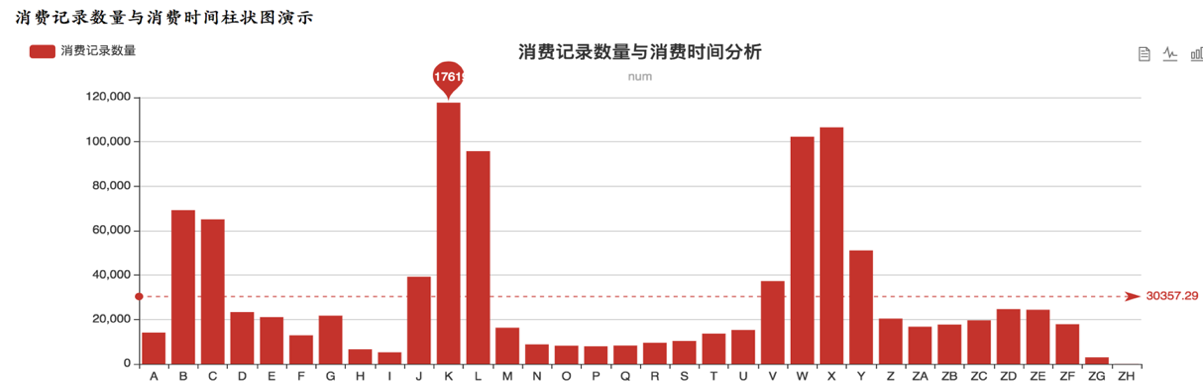


图2消费记录数量与消费时间分析情况可视化折线图

图2中的数据是14级同学一个学期消费的数据，使用Hive对原始数据按时间点进行统计分析，利用折线图可以分析得出每个时间点的消费记录数量。

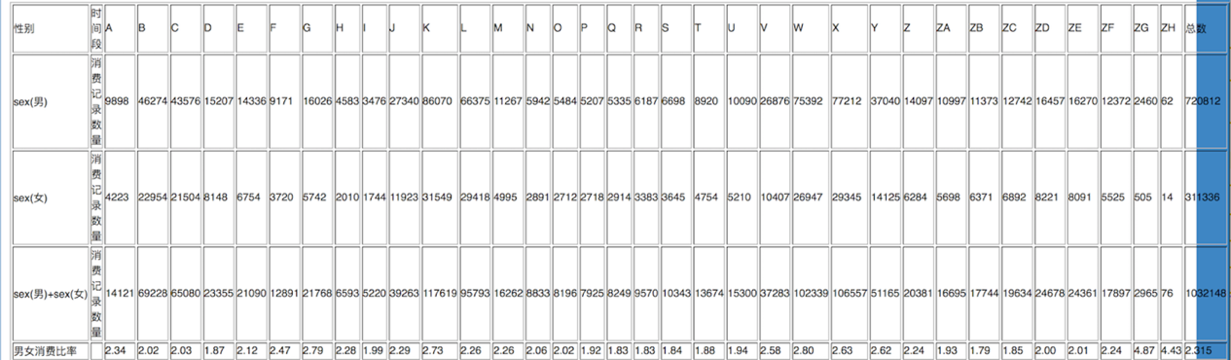


图3 消费整体情况展示图

图3展示了男女分别在每个时间段的消费记录数、消费比例。男性消费数量约为女性消费数量的 2.315 倍。

3.2.2 离散化结果分析

为寻找消费者性别与消费时间的规律，从图3.2交叉表中找出男女消费比显著偏离 2.315:1 的记录。设计函数 F=T(T>2.315)，其中 T=Ti(sex1)/Ti(sex2) 即第 i 个时间段的男女消费记录比。可得到如下结论：

1．对于某个时段 Ti，sex(男)+sex(女) 的值越小说明在此时间段产生的消费记录越少，即使产生较大的 F 值，相应的说服力也不大，为减少此类数据的影响，在计数大于平均数的时段内，选取算子 F 较大值才是有意义的。

2．由于每个时段消费的平均值约为 30357，因此在消费总数大于 30357 的时间段内选取 F 较大的值，相应时段分别为:K、V、W、X、Y。分别代表在 11:30--12:00,17:00-19:00期间，性别消费差异更明显，在这个期间男生相对于其他时段的消费比例更高一些。

3．在消费总数大于 30357 的时间段内选取 F 较小的值，相应时段分别为:B、C、J、L，分别代表在 7:00--8:00,11:30-12:30期间，性别消费差异更明显，在这个期间男生相对于其他时段的消费比例更高一些。

3.2.3 离散化结果结论

整体来看，男生晚上消费更多，应考虑女生减肥原因；女生早晨消费更多，应考虑男生早起问题。对于早晚饭菜可以针对男女生喜好改进，提高收益。

1. 结论

本课题完成了从前端设计到后台开发、后台用户的管理、数据采集、存储、处理、可视化到微信一卡通小程序，具备一定的应用性。对学校高层决策者来说，对数据存储与管理分析，对多年采集的高达几十GB的庞大数据进行存储与管理决策提供依据；对于后勤管理员来说，对消费数据分析，使各经营单位尽量做到供求偏差与挤压现象减少；对消费时间分析，查看性别与消费时间点关系，调整窗口饭菜质量以提高收益；关注POS机收益情况，调整窗口卫生情况及饭菜质量；对于学院管理员来说，对消费数据分析，审核学生的家庭情况。对于餐饮消费位于中上等水平的同学，在国家助学金进行补助时可以较后考虑，而优先考虑平时餐饮费用处于中低消费水平的同学。对学生来说，根据一卡通消费情况的分析可以调整自己的消费时间点，避免就餐和淋浴拥挤；学生能及时展示当前的一卡通消费和余额情况,结合当前的一卡通状况改善消费状况。