

学生消费行为的聚类分析优化研究和应用^①

游香霁¹, 王 业², 杨 抒¹, 王 斌², 赵新苗¹, 姚海伦¹

¹(新疆农业大学 计算机与信息工程学院, 乌鲁木齐 830052)

²(新疆农业大学 现代教育技术中心, 乌鲁木齐 830052)

摘 要: 基于现有的数据挖掘技术, 采用优化初始聚类中心的方法来改进 k-means 聚类算法, 对新疆农业大学学生一卡通消费数据进行研究和分析, 为相关部门提供决策支持. 首先, 根据需求分析, 选取部分学生于 2014-2015 学年在一卡通系统中产生的真实数据作为分析数据, 并进行数据预处理, 同时选择食堂消费次数和金额、超市消费次数和金额、就餐场所为特征属性; 其次, 使用改进的聚类算法进行分析, 并且对比分析了基于三种距离度量方式下的 k-means 聚类算法; 然后, 得出分析结论, 学生的食堂消费行为和超市消费行为; 最后, 探讨了如何根据分析所得结论为学校提供决策支持.

关键词: 优化 k-means 算法; 特征属性; 消费行为; 距离度量方式; 决策支持

Research and Application of Clustering Analysis Optimization for Students' Consumer Behaviors

YOU Xiang-Ru¹, WANG Ye², YANG Shu¹, WANG Bin², ZHAO Xin-Miao¹, YAO Hai-Lun¹

¹(College of Computer & Information Engineering, Xinjiang Agricultural University, Urumqi 830052, China)

²(Modern Education Technology Center, Xinjiang Agricultural University, Urumqi 830052, China)

Abstract: Based on the existing data mining technology, this article adopts the method of optimizing the initial clustering center to improve the k-means clustering algorithm, we can study of Xinjiang agricultural University student id card consumption data for the research and analysis, and provide decision support for the related departments. First of all, according to the demand analysis, we will choose some students for school year 2014-2015 real data in one cartoon system as data analysis, and data preprocessing, at the same time, we will choose the dining room number and amount, the supermarket consumption number and amount, the dining place for experimental characteristic attributes; Secondly, we use the improved clustering algorithm to analyze the data, and comparative analysis based on three kinds of distance measure under the k-means clustering algorithm; Then, the analysis conclusion, the student canteen consumption behavior and supermarket consumption behavior; Finally, the study was based on the conclusions of analysis provides decision support for schools.

Key words: optimization of the k-means clustering algorithm; feature properties; consumer behavior; distance measure; decision support

随着数据挖掘技术的不断发展, 对校园历史数据进行分析研究, 已经成为高校有关管理部门有效开展学生工作的重要参考依据. 本文使用数据挖掘方法对学生一卡通消费数据进行分析对于学校管理学生和做出一些重要决策提供有效的参考依据. 例如, 为每年的助学金发放工作提供参考依据; 另一方面, 通过学生对餐厅的喜好程度, 加强对餐厅的监管力度, 做到

及时发现问题, 及时解决问题.

1 研究现状及意义

随着大数据时代的到来, 目前有很多学者都在研究数据挖掘在校园一卡通消费中的研究与应用. 例如, 2011 年, 裴嫣璐^[1]利用数据挖掘技术对校园一卡通数据进行分析, 了解学生校内活动情况, 为教学和

① 基金项目: 校前期资助课题(XJAU201426); 自治区自然科学基金(2014211B023)

收稿时间: 2016-09-09; 收到修改稿时间: 2016-11-03 [doi: 10.15888/j.cnki.csa.005798]

管理提供辅助依据;2014年,陈锋^[2]从学生消费情况和食堂经营现状对学生食堂消费数据进行分析讨论,提出相应的营销策略.这些研究通过对学生各种消费数据进行数据挖掘,使管理人员可以了解学生的消费习惯和消费水平.一方面,通过分析数据挖掘结果,客观了解学生日常消费行为,发现影响消费行为的因素,分析原因,提出决策.

本文通过对新疆农业大学学生一卡通消费数据进行研究和分析可以在以下几个方面为学校决策提供参考依据:第一,在分析学生思想方面,发现隐藏在学生消费习惯中的隐患,相关辅导人员应及时调查原因,尽早做思想工作,防止隐患扩大化.例如,学生的家庭情况会影响学生的在校行为能力,特别是少数民族聚集的地区,更应该通过对学生消费行为进行客观有效的分析,了解学生真实的生活状态;第二,在学校管理工作方面,上述结果可以对每年的助学金发放工作提供一定参考依据,增强评选结果的客观性和有效性;第三,在餐厅建设发展方面,按照学生的消费喜好,可以客观的反映餐厅的经营状况,对有问题的餐厅,应及时调查原因,一旦发现问题,相关管理部门应及时采取相应的措施,合理应对.例如,由于高校扩招,民族学生数量逐年增加,不少清真餐厅出现供不应求的状况,或者经常发生同一时间餐厅人满为患的景象,相关部门应该具体问题具体分析,增加餐厅数量,合理分配资源,加强监管力度.

2 相关技术

2.1 聚类算法

数据挖掘的重要任务之一就是发现大数据中的积聚现象,并加以定量化描述,即从海量数据中挖掘出隐藏在其中的有用信息.聚类分析是一种探索性的分析,在分类的过程中,算法接受一个未分类的数据训练集,然后数据在非监督的情况下,自动聚集成不同的组,使得组内成员密集程度高,组间成员密集程度低.聚类算法起源于分类学,随着理论研究和计算机技术的不断发展,使其广泛的应用于社会的各个领域.

传统的聚类方法按照数据源的特点不同大致分为五类,包括基于层次的算法、基于划分的算法、基于密度的算法、基于网格的算法和基于模型的算法^[3],其中基于层次和划分的算法是应用最广泛的聚类算法.基于层次的聚类算法一般分为两种,凝聚算法和分裂

算法,其中CURE算法是典型的凝聚层次算法^[4].基于划分的聚类算法指定聚类个数,试图找到最优的划分,其中k-means聚类算法是最普及的一种算法^[5].

1967年,MacQueen^[6]在前人研究成果的基础上提出了k-means聚类算法,算法原理简单易行,时间复杂度接近线性,在大规模数据挖掘中具有可伸缩性和高效性,主要针对数值型数据的分类问题.

2.2 改进 k-means 算法

本实验数据为数值型数据,故选用基于划分的k-means聚类算法来完成实验任务.虽然,k-means聚类算法得到了广泛的应用,但该算法仍然存在不足.针对不足提出了相应解决方案如下:

① 初始聚类中心随机选择.

传统的k-means聚类算法是在样本数据中随机抽取K个对象作为初始簇中心,这样可能会出现初始簇中心过于集中或者不能均匀的分布在样本数据集中,从而导致数据收敛所需的迭代次数增加,甚至陷入局部最优解,影响最后聚类结果的准确性^[7].为了避免这样的结果出现,本实验按照具体情况对初始簇中心的选取做了改进,使簇中心能够均匀的分布在样本数据集中,避免孤立点和噪点,增强算法的鲁棒性^[8],公式如下:

$$m_i' = \min_{j=0}^f + \frac{f \cdot (\max_{j=0}^f - \min_{j=0}^f) \cdot (i+1)}{k+1} \quad (0 \leq i < k) \quad (1)$$

其中 m_i' 是初始聚类中心, f 是参加聚类的特征数, $\max$ 是由每个特征的最大值组成的集合, $\min$ 是由每个特征的最小值组成的集合;

② 聚类分析中,选择不同的距离度量方式,最终聚类结果也有所差异.

传统的k-means聚类算法通常采用的Euclidean Distance来度量对象之间的聚集程度,该距离度量方式对孤立点、噪音点敏感,可能导致低质量的聚类结果,并且一般只能发现球状簇.故本实验选取了三种距离度量方式进行分析比较,根据原始数据不同,采用不同的距离算法,其中包括Euclidean Distance(dE)、CityBlock Distance(dC)和Minkowski Distance(dM).聚类结果平面图如图1所示.

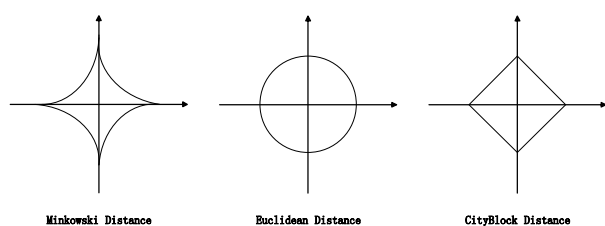


图1 三种距离度量方式聚类结果平面图

由图1可得, Minkowski Distance 的聚类结果为星形簇, Euclidean Distance 的聚类结果为圆型簇, CityBlock Distance 的聚类结果为菱形簇. 同一组数据, 聚类分析所使用方法的不同, 则会得到不同的结果. 故要根据数据源的特征来选取合适的距离度量算法, 寻求最优的算法, 提高聚类结果的准确性和有效性. 具体公式如下, 其中 s_i 为第 i 类的对象, m_i 为第 i 类的聚类中心.

$$d_E = \sqrt{\sum_{i=1}^n (s_i - m_i)^2} \quad (2)$$

$$d_C = \sum_{i=1}^n |s_i - m_i| \quad (3)$$

$$d_M = \left(\sum_{i=1}^n |s_i - m_i|^p \right)^{\frac{1}{p}} \quad (4)$$

3 实验设计

3.1 实验过程

1) 数据抽取和整理

本实验所采用的数据是新疆农业大学一卡通数据库中 2013 级学生在 2014-2015 学年的各类刷卡记录, 其中包括 3447 名学生. 首先获取本次实验所需要的数据, 预处理原始数据, 数据进行合并、整理, 其中包括数据转换、数据筛选、数据集成和数据归约^[9]等操作; 其次, 进行需求分析, 选择符合实验要求的特征属性; 最后, 将预处理后的数据存入数据仓库中, 为后续聚类分析做好准备.

现有 47 个民族共同生活在新疆维吾尔自治区, 各高校都是民汉合校, 各民族之间存在巨大的文化差异, 考虑到这些差异会在不同程度上影响学生的在校行为, 故将预处理后的数据分为少数民族学生消费数据和汉族学生消费数据来进行研究, 其中少数民族学生有 1436 组和汉族学生有 2011 组. 选取食堂消费金额、食堂消费次数、超市消费金额、超市消费次数和就餐场所这 5 个特征属性进行分析和研究. 具体过程如图 2.

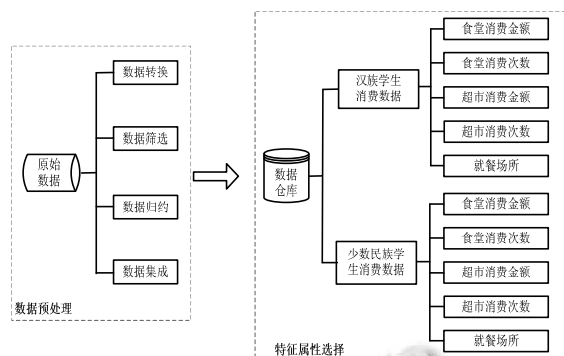


图2 数据抽取整理图

2) 算法应用

本实验将高校学生消费额的样本集合定义为对象集合 S , 其中每一个学生就是一个对象 s . 任意一个簇表示一组 $C_i \subset S (i=1,2,\dots,k)$ 相似对象的集合, 即 $\bigcup_{i=1}^k C_i = S$, 其中 k 表示簇的总数.

其具体流程如下:

步骤 1. 根据公式(1), 从数据集合中选择 $K^{[10]}$ 个对象作为初始聚类中心.

步骤 2. 求每个对象与初始聚类中心点的 Euclidean Distance、CityBlock Distance 和 Minkowski Distance, 将其与距离最近的聚类中心点关联起来, 将与同一个聚类中心相关联的所有点聚成一簇.

步骤 3. 计算簇中属性特征的均值, 组成一个新的均值对象, 将该对象作为下一次的聚类中心 $m_i (i=1,2,\dots,k)$.

步骤 4. 重复上述步骤, 直到得到的聚类中心不再变化, 则迭代停止.

通过对比三种度量方式, 寻求合适的聚类算法, 从而提高聚类结果的有效性. 同时, 采用平方误差总和 P 、迭代次数和迭代时间对三种方式进行评价, 其中 P 的公式如下:

$$P = \sum_{i=1}^k \sum_{j=1}^n d(s_j, m_i) \quad (5)$$

3.2 三种距离度量方式结果的比较

本实验使用 Euclidean Distance、CityBlock Distance 和 Minkowski Distance 三种距离度量方式的聚类算法分别对汉族学生消费数据和少数民族学生消费数据进行聚类分析, 并通过上述三个指标来选择出最符合数据情况的聚类算法, 实现最终聚类结果最优. 汉族学生三种距离度量方式对比情况如表 1 所示,

少数民族学生三种距离度量方式对比情况如表 2 所示.

表 1 三种距离度量方式对比表 (汉族)

距离度量方式	迭代次数	P 值($\times 10^5$)	时间(s)
Euclidean Distance	66	26.29802	1.690112
CityBlock Distance	91	33.94961	2.580224
Minkowski Distance	68	25.42643	3.550208

由表 1 可得, 在三种方式中迭代次数最少的是 Euclidean Distance; P 值最小的是 Minkowski Distance; 用时最小的是 Euclidean Distance. 通过综合分析, Euclidean Distance 在汉族学生消费数据上的聚类效果比较好, 故本实验将选择基于 Euclidean Distance 的优化聚类算法来获取最优的聚类结果.

表 2 三种距离度量方式对比表 (少数民族)

距离度量方式	迭代次数	P 值($\times 10^5$)	时间(s)
Euclidean Distance	20	17.34837	0.870048
CityBlock Distance	31	22.72901	0.630016
Minkowski Distance	20	16.74533	0.770048

由表 2 可得, 迭代次数相对较少的是 Euclidean Distance 和 Minkowski Distance; 平方误差和 P 值最小的是 Minkowski Distance; 算法运行时间最少的是 CityBlock Distance. 因为实验数据规模不是很大, 所以三者时间上的差距不是很明显, 故重点考虑迭代次数和 P 值的情况. 对于聚类结果来说, P 值越小, 类内对象相似度越高, 类间对象相似度就会越低^[1]. 通过综合比较之后, 对于少数民族学生消费数据来说, 本实验将选择基于 Minkowski Distance 的优化聚类算法来获取最优的聚类结果.

4 结果分析

4.1 汉族同学聚类结果分析

通过上述对比分析, 汉族学生消费数据采用基于 Euclidean Distance 的优化聚类算法进行聚类, 按照需求分析, 将聚类个数定为 5 类, 最终结果如表 3 所示.

根据表 3 分析得, 第五类学生的食堂消费金额、消费次数和超市消费次数在所有人中最多, 就餐场所多为消费较低的学院汉、清餐厅和消费适中禾源汉、清餐厅, 学生数量占总人数的 7%; 第二类学生的超市消费金额属于最高消费水平, 这类学生的就餐地点一般为消费较高的学苑食府、学轩西餐厅和学士餐厅, 占总人数的 25%; 通过上表发现, 第三类学生的人数最多, 就餐地点多为消费较低的学院汉、清餐厅和消

费适中禾源汉、清餐厅, 占总人数的 33%.

表 3 聚类结果表(汉族)

特征属性	第一类	第二类	第三类	第四类	第五类
食堂消费金额	1023	1888.67	2728.33	3472.5	4379.67
食堂消费次数	206	389	649.33	739.83	980.16
超市消费金额	29.17	167.33	87.17	41.66	147
超市消费次数	2.83	15.83	9.5	11.5	16.16
就餐场所	30.41	30.62	30.57	30.44	30.23
人数	176	499	679	502	155

根据表 3 分析得, 第五类学生的食堂消费金额、消费次数和超市消费次数在所有人中最多, 就餐场所多为消费较低的学院汉、清餐厅和消费适中禾源汉、清餐厅, 学生数量占总人数的 7%; 第二类学生的超市消费金额属于最高消费水平, 这类学生的就餐地点一般为消费较高的学苑食府、学轩西餐厅和学士餐厅, 占总人数的 25%; 通过上表发现, 第三类学生的人数最多, 就餐地点多为消费较低的学院汉、清餐厅和消费适中禾源汉、清餐厅, 占总人数的 33%.

通过食堂消费对比分析得到, 学生单次消费平均在 4.20~4.96 元. 其中, 第一类学生单次消费平均最高, 占总人数的 8%. 第三类学生单次消费平均最低. 根据超市消费对比分析得到, 学生单次消费平均在 3.62~10.57 元. 其中, 第二类学生单次消费平均最高, 第四类学生单次消费平均最低, 占总人数的 27%. 综合分析可得, 第三类学生的消费习惯为大众消费, 以这类人群消费习惯为标准, 第一类和第二类学生群体为高消费人群, 第四类和第五类学生群体为低消费人群.

4.2 少数民族同学聚类结果分析

对于少数民族学生消费数据, 选择基于 Minkowski Distance 的优化聚类算法, 聚类个数为 5, 聚类结果如表 4 所示.

根据表 4 分析得, 第五类学生的食堂消费金额和消费次数在所有人中最多, 这类学生的就餐地点多为消费较低的学院清餐厅和禾源清餐厅, 占总人数的 5%; 第二类学生的超市消费金额属于最高消费水平, 这类学生的就餐地点一般为消费适中禾源汉、清餐

厅、消费较高的学苑食府和学轩西餐厅,学生数量占总人数的30%;第三类学生的人数最多,就餐地点多为消费较低的学院清餐厅和消费适中的禾源清餐厅,占总人数的31%;第四类学生就餐场所通常为消费较高的学苑食府、学轩西餐厅和学士清餐厅,占总人数的19%。

表4 聚类结果表(少数民族)

特征属性	第一类	第二类	第三类	第四类	第五类
食堂消费金额	726.7	1574.17	2252.33	2991.84	3846.33
食堂消费次数	150	299	649.67	788.33	789.67
超市消费金额	99.5	174.67	171.33	73.17	44.66
超市消费次数	13	23.5	29.33	8.67	3.5
就餐场所	29.86	29.83	29.54	29.89	29.78
人数	216	435	443	270	72

通过食堂消费对比分析得到,学生单次消费平均在3.41~5.26元。其中,第二类学生单次消费平均最高,第三类学生单次消费平均最低。根据超市消费对比分析得到,学生单次消费平均在5.84~12.76元。其中,第五类学生单次消费平均最高,第三类学生单次消费平均最低。综合分析可得,第三类学生的消费习惯为大众消费,以这类人群消费习惯为标准,第一类和第二类学生群体为低消费人群,第四类和第五类学生群体为高消费人群。

综合民、汉学生消费聚类结果分析可得,在食堂消费方面,汉族学生比民族学生的消费金额区间要小,产生的原因可能有两个方面:一方面,学校民、汉餐厅食品定价不同,造成学生消费差异增大;另一方面,少数民族学生家庭条件差距较大,增大少数民族学生消费区间。在超市消费方面,汉族学生和民族学生的消费金额区间近似,但是少数民族学生的超市消费金额普遍较高。

4.3 学校决策

根据上述分析,可以从学生行为、学校管理工作和餐厅建设发展方面为学校相关部门提供决策支持。

助学金发放方面,根据学生在校超市和学校食堂的消费情况,客观的分析他的消费能力,学校使用该分析方法的数据作为学校助学金的发放的重要参考

依据,帮助家庭经济困难的学生顺利完成学业。

思想教育方面,学校通过食堂一卡通刷卡记录发现有个别学生在这一年之内没有在食堂刷卡消费过,通过沟通后发现,因为该学生不能够接受学校食堂的各种条件,几乎不在学校吃饭。针对以上情况,学校针对以上情况做出了以下决策:

(1) 学校对该生的情况做了进一步的了解并走访了该生的父母,告知了具体的情况。

(2) 学校通过心理服务中心对该生进行一对一的心理辅导,积极改善该生的思想状态。

(3) 为了防止再次出现这样的情况,学校在明德广场电子大屏幕上滚动播放食堂的宣传片消除学生心中的各种隐患。

餐厅建设方面,学校通过该方法分析出的结果观察到如下情况:第一,有个别接受学校补贴的食堂价格却比其他食堂的价格贵;第二,有个别餐厅就餐的人数很少。经过学校的调查后发现就餐人数少的餐厅环境卫生、饭菜的质量和服务态度都不是很好,因此针对以上情况学校做出以下决策:

(1) 完善学校食堂承包合同,规定食堂未经过学校允许不能私自涨价,饭菜价格需要合情合理。

(2) 制定更加完善的学校食堂考核制度,定时定点安排人员检查食堂的卫生、饭菜质量和服务态度。每个学期末时,通过问卷调查的形式向学生了解食堂的情况和一些建议。

5 结论

本文探讨了基于聚类算法,学校如何根据学生消费行为进行决策,为学校工作提供了新的依据视角。实验对汉族学生和少数民族学生消费数据进行分别研究,选择基于最优距离度量方式的优化聚类算法,挖掘数据中蕴藏的知识。通过对学生消费数据进行分析研究,对学生的消费水平和消费习惯有了初步了解,最终结果将为相关管理部门提供决策支持,提高工作效率,实现资源优化配置。

参考文献

- 1 裴嫣璐.数据挖掘在学生校内活动情况分析中的应用.上海第二工业大学学报,2011,(1):60-64.
- 2 陈锋.基于校园一卡通系统的高校用户就餐消费行为分析与数据挖掘.中国教育信息化,2014,(9):47-49.

- 3 石飞飞.智慧校园数据挖掘及其应用[硕士学位论文].北京:北京邮电大学,2014.
- 4 李莹.聚类结果评价方法和聚类知识提取技术的研究[硕士学位论文].南京:南京航空航天大学,2008.
- 5 郝春梅,吴波.基于量子蚁群改进的 K-means 算法.计算机测量与控制,2013,(4):1011-1013.
- 6 MacQueen J. Some methods for classification and analysis of multivariate observations. Proc. of the 5th Berkeley Symposium on Mathematical Statistics and Probability. Berkeley. University of California Press. 1967. 281-297.
- 7 崔丹丹.K-Means 聚类算法的研究与改进[硕士学位论文].合肥:安徽大学,2012.
- 8 吴晓蓉.K-均值聚类算法初始中心选取相关问题的研究[硕士学位论文].长沙:湖南大学,2008.
- 9 徐云.多通道数据采集系统数据预处理方法研究[硕士学位论文].杭州:浙江大学,2014.
- 10 曾秋凤.数据挖掘在校园一卡通数据库中的应用研究[硕士学位论文].南昌:江西师范大学,2008.
- 11 杨怡,王江晴,朱宗晓.基于仿射传播聚类的自适应手写字符识别.计算机应用,2015,35(3):807-810.