

改进 kmeans 算法在学生消费画像中的应用

凌玉龙¹, 张 晓¹, 李 霞², 张 勇¹

(1. 西北工业大学 大数据存储与管理工信部重点实验室, 陕西 西安 710129;

2. 西北工业大学 学生资助服务中心, 陕西 西安 710129)

摘 要: 学生校园消费数据中隐含了大量的低价值信息, 论文从学生消费画像和精准资助两个角度对校园消费数据进行挖掘研究。从数据集本身的特点和 kmeans 算法的缺陷两个角度出发, 论文对 kmeans 算法进行了两点改进: 采用马氏距离代替欧氏距离以适用于具体的校园消费数据应用场景; 为了解决随机选择初始聚类中心的方法受离群样本点的影响, 采用在高密度样本集合中选择相距最远的 k 个样本作为初始聚类中心的改进方法。在西安某高校 3 个月的学生消费数据集上的实验结果表明: 论文设计的学生群体分类模型不仅能有效区分不同行为特征的学生, 而且能很好地刻画学生的消费画像; 通过对比分析聚类标记的贫困生名单和线下认定的贫困生名单, 证明了改进 kmeans 算法在精准资助中的应用价值。

关键词: 改进 kmeans 算法; 马氏距离; 初始聚类中心集合; 学生消费画像; 精准资助

中图分类号: TP311.13

文献标识码: A

文章编号: 1673-629X(2021)10-0122-06

doi: 10.3969/j.issn.1673-629X.2021.10.021

Application of Improved kmeans Algorithm in Student Consumption Portrait

LING Yu-long¹, ZHANG Xiao¹, LI Xia², ZHANG Yong¹

(1. Ministry of Communications Key Laboratory of Big Data Storage and Management,

Northwestern Polytechnical University, Xi'an 710129, China;

2. Student Aid Service Center, Northwestern Polytechnical University, Xi'an 710129, China)

Abstract: There is a large amount of high-value information hidden in the student campus consumption data. We explore the campus consumption data from the two perspectives of student consumption portrait and precision funding. From the perspective of the characteristics of the data set itself and the defect of the kmeans algorithm, two improvements have been made to the kmeans algorithm. On the one hand, Mahalanobis distance is used instead of Euclidean distance to be suitable for specific campus consumption data application scenarios. On the other hand, in order to solve the problem that the method of randomly selecting the initial cluster center set will be affected by outlier sample points, an improved method of selecting the k samples that are the farthest apart from the high-density sample set as the initial cluster center set is adopted. The experiment on three months of student consumption data set in a university in Xi'an shows that the designed student group classification model can not only effectively distinguish students with different behavior characteristics, but also can well display the student consumption portrait. The comparative analysis of the list of poor students marked by the improved kmeans algorithm and the list of poor students identified offline proves the application value of the improved kmeans algorithm in precision funding.

Key words: improved kmeans algorithm; Mahalanobis distance; initial clustering center set; student consumption portrait; precision funding

0 引言

学生群体肩负着祖国的未来, 在社会中扮演着重要的角色, 因此对学生的行为进行分析具有重大意义。数据挖掘作为一种从海量数据中获取潜在知识的技术, 已经在各个领域取得广泛应用。采用数据挖掘技

术挖掘学生消费数据中潜在的知识可以充分发挥现有消费数据的作用, 为学校管理人员的决策提供数据支持。

作为最经典的数据挖掘算法之一, kmeans 算法思想简单, 易于实现, 有着广泛的应用。随着社会的发

收稿日期: 2020-12-02

修回日期: 2021-04-08

基金项目: 国家重点研发计划(2018YFB1004401); 北京市自然科学基金-海淀原始创新联合基金(L192027)

作者简介: 凌玉龙(1995-), 男, 硕士研究生, CCF 会员(93531G), 研究方向为数据挖掘、大数据存储与管理; 张 晓, 博士, 副教授, CCF 高级会员(16288S), 研究方向为大数据存储与管理、云计算与云存储等。

展,出现了一系列使用传统 kmeans 算法难以解决的新问题和新场景,学术界针对传统 kmeans 算法不断进行改进以适应这些场景。例如,谢修娟^[1]为了从微博数据中发现热点舆情,提出一种基于密度的初始聚类中心选择算法,改进算法在微博数据集上拥有更高的准确性和稳定性。马汉达^[2]针对传统 kmeans 在 Web 日志挖掘中性能不高的缺点,提出了基于粒子群算法的改进 kmeans 算法,并在 Hadoop 上实现了并行化,实验证明改进算法不仅提高了聚类准确率,而且提升了运行效率。Lutz^[3]为了解决 kmeans 算法在 GPU 上效能较低的问题,针对 kmeans 算法每次迭代计算质心的过程提出了一种新的算法来更新质心,改进后的 kmeans 算法提高了 20% 的吞吐量。

如何利用校园消费数据分析学生群体行为,实现贫困生精确资助,提高学生学习生活质量,保障学生安全,已经成为高校急需解决的问题。随着数据挖掘技术的发展,很多研究人员尝试将数据挖掘技术与学生行为分析相结合^[4-13],其中黄刚^[14]和姜楠^[15]的研究具有一定的代表性。这两位学者在学生校园消费数据集上采用 kmeans 算法对学生进行聚类,分析学生的消费习惯和群体特征,并进行了画像说明,为高校学生管理工作提供依据。但是他们的研究仅是将 kmeans 算法引入到学生行为分析领域,没有考虑学生消费数据集本身的特点和 kmeans 算法随机选择初始聚类中心的缺点。

文中根据学生消费数据集的特点和 kmeans 算法随机选择初始聚类中心的不足,提出一种基于马氏距离和密度的改进 kmeans 算法,并将其应用于西安某高校的校园一卡通消费数据集上,对学生群体行为进行分析并构建消费画像,从大数据的角度探究了学生行为,同时推动了贫困生精准资助领域的发展。

1 理论介绍

本节分析了学生消费数据集的特点和 kmeans 算法选择随机初始聚类中心的不足,针对 kmeans 算法提出两点改进以更好地适用于学生群体聚类场景。

1.1 欧氏距离和马氏距离

经典的 kmeans 算法采用欧氏距离计算样本之间的距离,欧氏距离单纯考虑数值上的距离,忽略了数据属性之间的依赖性,将数据各个属性之间的差别同等看待。但是大部分实际场景中的数据属性之间并不是独立的,各个属性起到的作用也各不相同。

马氏距离是欧氏距离的一种修正,其修正了欧氏距离中各个属性尺度不一致且相关的问题。马氏距离认为属性之间是存在联系的,所以在计算公式中引入了协方差。对于一个多变量向量 $\mathbf{x} = (x_1, x_2, \dots, x_p)^T$,

设其均值为 $\boldsymbol{\mu} = (\mu_1, \mu_2, \dots, \mu_p)^T$, 协方差矩阵为 $\boldsymbol{\Sigma}$, 则其马氏距离定义为:

$$D_M(\mathbf{x}) = \sqrt{(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})} \quad (1)$$

其中, T 表示矩阵的转置。如果数据之间独立同分布,那么对应的协方差矩阵就变成了单位阵,在这种情况下马氏距离就变成了欧氏距离。

马氏距离由于其设计思想会将某一微小变量的作用放大,这在某些应用场景中可能会导致结果的过拟合。但是不同于其他应用场景下的数据,学生群体是一个具有高度相似性的群体,由于课业的安排和学校的规章制度,大部分学生的生活作息规律极其相似,学生的就餐时间、就餐地点、消费情况相对固定且有规律,不同学生的行为相差较小。由于学生消费数据的特点和马氏距离的特点,该文认为采用马氏距离代替欧氏距离的 kmeans 算法更适合对学生消费数据进行聚类分析。

1.2 初始聚类中心的选择

kmeans 算法是从样本集合中随机选择 k 个样本作为初始聚类中心,这种初始化聚类中心的思想没有考虑到数据的分布情况和离群样本点的影响,很可能产生较差的聚类结果。虽然随机选择初始聚类中心的 kmeans 算法的聚类效果可能不好,但是无论初始聚类中心怎么选择,经过一系列迭代后得到的最终聚类中心的周边的点都是高密度点,即这些聚类中心周边的点都比较密集,不会存在离群样本点。

文中选择初始聚类中心的改进思想为:在高密度样本集上应用最小最大原则得到 k 个样本作为初始聚类中心。具体步骤如下:

- (1) 运行一次 kmeans 算法,得到 k 个聚类中心;
- (2) 选择与 k 个聚类中心距离最近的一定比例(记为 percent)的样本作为高密度样本集;
- (3) 从高密度样本集中随机选择一个样本作为第一个初始聚类中心;
- (4) 从剩下的高密度样本集中选择距离已有的初始聚类中心最远的样本作为第 2 个初始聚类中心;
- (5) 重复执行步骤(4)直到初始聚类中心中包含 k 个样本。

上述步骤中 percent 的值过大会引入噪声样本点,过小又会使高密度样本集中样本数据过少。经过综合考虑,文中设置 percent 为 20%。这个参数并不是固定不变的,可以根据数据集的具体情况进行调整。

在高密度数据集上应用最小最大原则得到的 k 个初始聚类中心,既考虑到了数据的分布情况,又可以避免离群样本点的干扰。文中基于上述两点改进思想实现了改进的 kmeans 算法(记做 Improve-Kmeans 算

法),并将其应用于学生消费数据集,以更好地描述学生群体的共性与特性。

2 实验分析

2.1 实验环境与数据

文中以西安某高校 2019 年 03 月—2019 年 06 月

表 1 一卡通消费数据集(部分)

cardId	name	xh	cost	balance	time	merchantId	POSId
41405	学生 A	2017001641	1	160	2019/3/21 18:12	100149	69
38264	学生 B	2017000794	5	84	2019/3/21 18:12	100210	38
42630	学生 C	2018002436	1	48	2019/3/21 18:12	100020	116

原始的一卡通消费数据集中不仅包含所有学生的一卡通消费记录,还包含部分教职工及学校管理人员的数据,这些数据属于噪声数据,分析这些噪声数据会对挖掘结果产生一定程度的影响。因此,文中的首要工作就是从原始数据集中去除噪声数据(在校学生之外的所有其他人的消费数据)。

一卡通消费数据集中只记录着每一次消费的记

录,对单一记录进行分析无法得出有用的结论,只有根据具体需求构建合适的特征后才能应用于聚类算法中,使数据挖掘更有针对性,从而提高算法性能。

为了全面刻画学生的特点,文中从多个角度分层提取了学生消费特征。表 2 展示了这些特征的基本信息。

为了全面刻画学生的特点,文中从多个角度分层提取了学生消费特征。表 2 展示了这些特征的基本信息。

表 2 学生消费特征

维度	特征	描述
消费水平	午餐平均消费	学生所有午餐的平均消费金额,记为 wcpjxf
	晚餐平均消费	学生所有晚餐的平均消费金额,记为 vcpjxf
消费习惯	食堂消费天数	某天在食堂有一次刷卡就记为一次食堂消费天数,记为 stxfts
	早起次数比例	学生第一次刷卡时间在 6:00—8:00 之间记为早起,早起次数比例就是早起次数占在校天数的比例,记为 zqcsbl
	午晚餐差额比例	学生的午餐次数与晚餐次数的差值占在校天数的比例,记为 wvcebl
在校情况	周内在校天数	周一到周五的在校天数,用于衡量周内在校情况,记为 znzxts
	周末在校天数	周六周天的在校天数,用于衡量周末在校情况,记为 zmzxts

2.2 实验方法

为了确定 Improve-Kmeans 算法中 k 的取值,通过对处理后的学生消费数据集进行 $k = 1$ 到 8 的聚类实验,得到 k 在不同取值下聚类结果的误差平方和(sum of the squared errors, SSE),具体步骤如下:

- (1) 设 $k = 1$, 运行 Improve-Kmeans 算法;
- (2) 记录 $k = 1$ 下得到的各个聚类中心及样本所属的类别;
- (3) 按照公式(2)计算组内方差 SSE, 并记录;
- (4) 设 $k = 2$ 到 8, 重复执行步骤(1)~步骤(3)。

$$SSE = \sum_{i=1}^k \sum_{p \in c_i} |p - m_i| \quad (2)$$

式中, c_i 是聚类结果中的第 i 个类, p 是 c_i 中的样本点, m_i 是 c_i 的聚类中心(即 c_i 中所有样本的均值)。

实验结果如图 1 所示。

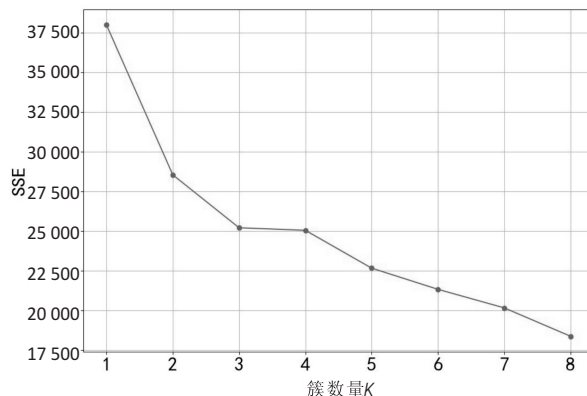


图 1 不同 k 值下聚类结果的 SSE

从图 1 中可以看到,随着 k 值的增大, SSE 逐渐减小,且在 $k = 3$ 时 SSE 的减小幅度的开始减缓,这表明最佳聚类数为 3。

为了减少实验误差,文中运行 Improve-Kmeans

算法 16 次,每次聚类得到的 SSE 和轮廓系数见表 3。从表中可以看出,在第 11 次实验时,SSE 值最小且轮

廓系数最大,因此文中选择该次实验结果得到的学生群体画像进行分析。

表 3 16 次实验的 SSE 和轮廓系数

实验次数	Improve-Kmeans		实验次数	Improve-Kmeans	
	SSE	轮廓系数		SSE	轮廓系数
1	28 334. 579 93	0. 098 640 141	9	26 369. 762 63	0. 280 156 597
2	27 217. 807 91	0. 215 325 502	10	27 300. 394 38	0. 218 668 415
3	33 621. 461 69	0. 047 189 422	11	25 184. 391 06	0. 321 848 592
4	33 621. 461 69	0. 047 189 422	12	25 246. 914 71	0. 320 329 405
5	31 177. 642 47	0. 086 892 379	13	26 379. 618 58	0. 279 631 177
6	28 327. 792 09	0. 099 922 154	14	31 183. 230 48	0. 086 135 27
7	28 409. 386 47	0. 100 302 985	15	28 415. 083 53	0. 100 228 499
8	28 319. 335 6	0. 101 313 385	16	28 334. 579 93	0. 098 640 141

文中设计了两个实验:

(1)使用 Improve-Kmeans 算法对学生刻画群体消费画像,并对画像进行解释分析,应用画像的结果为学校管理人员提供决策支持;

(2)将 Improve-Kmeans 算法得到的贫困生数据与学校线下认定的贫困生数据进行对比分析,以辅助高校精准资助活动。

2.3 学生群体消费画像

文中从三个角度:(1)3 个类的聚类中心点;(2)3 类学生对应的原始数据的平均值;(3)3 类学生对应的原始数据的分布情况,分析学生群体的消费特征。

表 4 列出了第 I、II、III 类学生群体的聚类中心点,图 2 描述了 3 类学生群体对应的原始数据的平均值,图 3 描述了 3 类学生群体对应的原始数据的分布情况。

表 4 学生群体聚类中心

聚类中心	午餐 平均消费	晚餐 平均消费	食堂 消费天数	早起 次数比例	午晚餐 差额比例	周内 在校天数	周末 在校天数	所属类包含 的样本数量
I	-0. 044 3	-0. 010 4	0. 311 4	-0. 232 3	-0. 108 5	0. 357 7	0. 233 8	4 133
II	0. 321 2	0. 197 6	-1. 703 8	-0. 269 0	0. 549 6	-1. 880 3	-1. 298 6	879
III	-0. 239 3	-0. 315 1	0. 507 6	2. 883 6	-0. 083 1	0. 420 8	0. 422 3	415

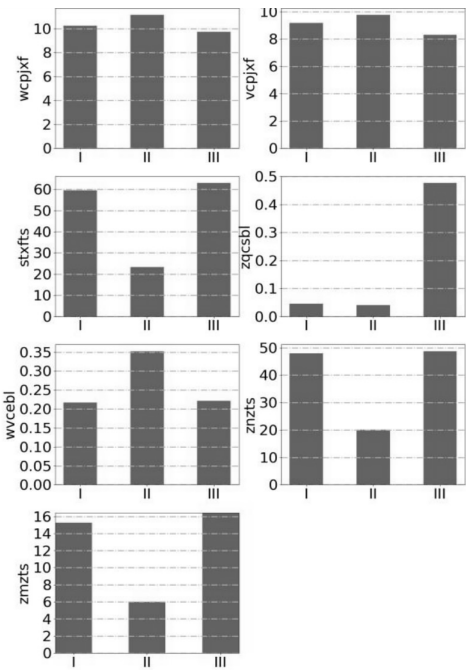


图 2 3 类学生的原始消费数据的平均值

从图 2 中可以看出,对于第 I 类群体,其午餐平均消费、晚餐平均消费明显比第 III 类群体低,同时比第

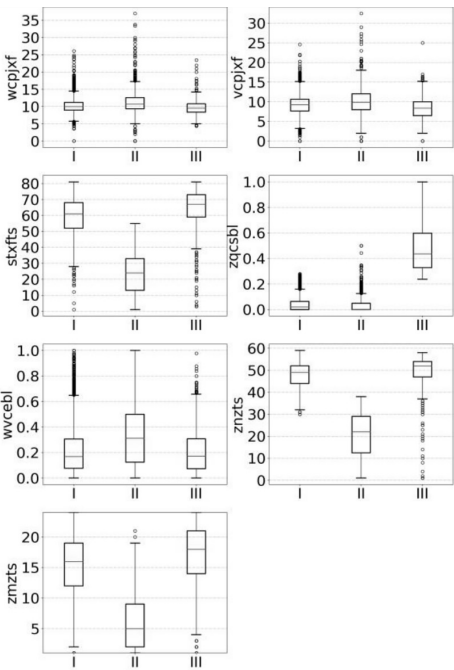


图 3 3 类学生的原始消费数据的分布

II 类群体高,说明其消费水平在三类群体中居中。其食堂消费天数接近 III 类群体且明显高于第 II 类群

体,说明这类学生频繁在食堂就餐。其午晚餐差额比例明显低于第 II 类群体,说明这类学生的饮食比较规律。周内在校天数和周末在校天数接近第 III 类群体且明显高于第 II 类群体,说明这类学生经常在校。早起次数比例接近第 II 类群体且明显低于第 II 类群体,说明这类学生同样很少早起,属于懒癌患者。综上分析,第 I 类学生符合大部分正常学生的消费情况。

对于第 II 类群体,其午餐平均消费、晚餐平均消费明显比第 I 和 III 类群体高,说明其消费水平是三类群体中最高的。其食堂消费天数最低且明显低于其他两类群体,说明这类学生很少在食堂吃饭。其午晚餐差额比例最高且明显高于其他两类群体,说明这类学生经常性的只吃单餐,饮食不规律,很可能是经常点外卖。周内在校天数和周末在校天数最低且明显低于其他两类学生,说明这类学生喜欢经常离校。早起次数比例最低,且明显低于第 III 类群体,说明这类学生很少早起。综上分析,第 II 类学生符合小富群体的行为特点。

对于第 III 类群体,其午餐平均消费、晚餐平均消费最低且明显比第 I 和 II 类群体低,说明其消费水平是三类群体中最底的。其食堂消费天数最高且明显高于其他两类群体,说明这类学生是最频繁在食堂就餐的学生。其午晚餐差额比例接近第 I 类群体且明显低于第 II 类群体,说明这类学生的饮食比较规律,午晚餐消费次数基本上相同。周内在校天数和周末在校天数最高且明显高于第 II 类群体,说明这类学生是最常在校的学生。早起次数比例最高且明显高于其他两类群体,说明这类学生拥有早起的好习惯。综上分析,第 III 类学生符合贫困生群体的行为特点。

表 4 中的“所属类包含的样本数量”列的结果也能验证学生群体分类结果。第 III 类贫困生群体共 415 人,占总人数的 7.64%,符合该校研究生中的贫困生数量和比例,第 I 类普通学生群体共 4 133 人,占总人数的 76.16%,基本上也符合现实情况。

图 3 是 3 类学生群体在消费特征上对应数据的箱

型图,横坐标代表学生群体,纵坐标代表各类群体在各特征上原始数据的分布情况,文中以第 III 类群体为例分析此类群体的消费特点。从图中可以看出第 III 类群体在特征:午餐消费金额、晚餐消费金额、食堂消费天数、午晚餐差额比例、周内在校天数、周末在校天数上的四分位距明显比第 II 类群体小,说明第 III 类群体在这些特征上数据的分布比较集中,波动范围小,消费习惯比较规律。

在特征:午餐消费金额、晚餐消费金额、午晚餐差额比例上的最大值、最小值、中位数、上下限比第 I 和第 II 类群体小,说明这类群体的消费水平较低。在特征:食堂消费天数,周内在校天数,周末在校天数的最大值、最小值、中位数、上下限明显比第 I 和第 II 类群体大,说明这类群体是最经常在校内就餐、很少离校。在早起次数比例上的各特征明显高于其他两类群体,说明这些学生喜欢早起,有着良好的习惯。综上,第 III 类群体可以认定是有着良好生活和消费习惯、基本上不离校且消费水平较低的贫困生。各个群体消费数据波动程度的分析结果和上面各个群体聚类中心的分析结果相同。

2.4 聚类标记的贫困生分析

为了验证 Improve-Kmeans 算法标记贫困生的效果,文中获取了学校线下认定的 2017 级和 2018 级的贫困生名单,共 349 人。聚类标记与线下认定的贫困生名单重合率为 47%,分析原因可能有两方面:(1) Improve-Kmeans 算法还需要进一步完善以更好地适应高校贫困生认定的应用环境;(2) 线下贫困认定名单具有很大的不确定性,老师、学生很多情况下是通过申请表、平时的认知(甚至并不认识)来进行贫困认定,可能存在误判的情况。因此文中以午餐平均消费水平和食堂就餐天数两个特征为例,研究两种方法中不重合的学生的消费情况:分别统计仅在聚类标记名单中出现的贫困生和仅在线下认定名单中出现的贫困生的午餐平均消费水平和食堂就餐天数,并绘制对应的概率密度曲线,如图 4 所示。

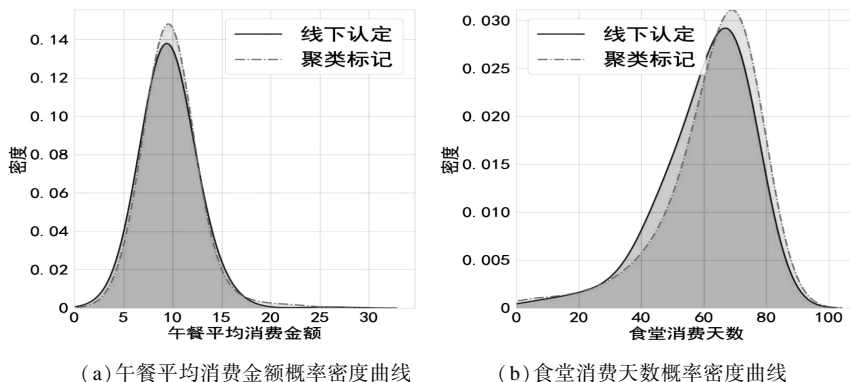


图 4 消费水平的概率密度曲线

从图4(a)中可以看出,聚类标记贫困生的密度曲线比线下认定贫困生的密度曲线更加集中,这意味着聚类标记的贫困生比学生认定的贫困生的午餐平均消费金额波动更小,更加稳定。

从图4(b)中可以看出,聚类标记贫困生的密度曲线所处的位置明显比线下认定贫困生的密度曲线所处的位置整体偏右,这意味着聚类标记的贫困生更偏向于在食堂就餐。

从午餐平均消费和食堂消费天数两个指标上可以看出,相比没有加入数据挖掘算法的线下贫困认定而言,基于客观消费数据聚类挖掘贫困生的方法更加适用。

聚类标记贫困生方法和线下认定贫困生方法的总结如下:

(1)聚类标记贫困生的目标是找出消费水平低的贫困生,不考虑任何人为因素,单纯从客观的学生消费数据出发,挖掘各个群体的学生的消费水平,找出消费水平较低的贫困生。但是没有考虑到学生家庭条件、健康情况、家庭人口情况和是否低保户等信息;

(2)线下贫困生认定的目标是找出家庭经济情况困难的贫困生,以学生家庭收入情况、健康情况、家庭人口情况和是否低保户等信息为标准,按照流程进行贫困生认定。但是没有考虑学生的消费数据,难以发现没有申请贫困认定的隐藏贫困生和申请了贫困认定的伪贫困生。

具体的贫困生认定工作可以结合这两种方式的优点:对于聚类标记认定的贫困生(或只考虑客观消费数据的挖掘算法挖掘出的贫困生)可以发放专项的贫困生助学基金;对于线下流程化认定的贫困生可以按照国家的要求发放贫困生补贴。

文中算法不仅可以用来辅助贫困生的认定,还为以后更深入地利用数据挖掘相关技术研究高校精准资助活动提供了支持,值得进一步研究。

3 结束语

为了全面了解学生的行为特点,文中从学生群体的角度出发,利用学生校园消费数据研究不同学生群体行为特征的相似性与差异性。采用适合校园消费数据场景的 Improve-Kmeans 聚类算法对研究生的消费数据集进行聚类,分析不同学生群体的消费特征,进行画像说明。同时对比分析了聚类得到的贫困生的消费数据和线下认定的贫困生的消费数据,为贫困生认定工作提出了改进意见,为高校的精准资助工作提供数据支持,为学生的校园学习生活保驾护航。

参考文献:

- [1] 谢修娟,李香菊,莫凌飞.基于改进 K-means 算法的微博舆情分析研究[J].计算机工程与科学,2018,40(1):155-158.
- [2] 马汉达,郝晓宇,马仁庆.基于 Hadoop 的并行 PSO-kmeans 算法实现 Web 日志挖掘[J].计算机科学,2015,42(z1):470-473.
- [3] LUTZ C, BRESS S, RABL T, et al. Efficient k-Means on GPUs[C]//14th international workshop on data management on new hardware. Houston, TX, United States: ACM, 2018:157-169.
- [4] LI H, ZHANG X, LI X, et al. An effective method to improve the accuracy of precise funding in campus[C]//Proceedings of the 3rd international conference on computer science and application engineering. Sanya, China: ACM, 2019: 1-6.
- [5] 李博,李霞,张晓,等. MD-KNN 算法在高校精准资助中的应用[J].计算机技术与发展,2020,30(7):91-95.
- [6] 吕品,于文兵,汪鑫,等.基于机器学习的学生成绩预测及教学启示[J].计算机技术与发展,2019,29(4):200-203.
- [7] 徐建军,张国华.基于 Apriori 数据挖掘算法的应用与实践[J].计算机技术与发展,2020,30(4):206-210.
- [8] 任占广,尚福华.基于行为分析的在线课程成绩预测模型[J].计算机技术与发展,2019,29(11):139-143.
- [9] NATEK S, ZWILLING M. Student data mining solution - knowledge management system related to higher education institutions[J]. Expert Systems with Applications, 2014, 41(14):6400-6407.
- [10] FARHAN M, JABBAR S, ASLAM M, et al. A real-time data mining approach for interaction analytics assessment: IoT based student interaction framework[J]. International Journal of Parallel Programming, 2018, 46(5):886-903.
- [11] MENGASH H A. Using data mining techniques to predict student performance to support decision making in university admission systems[J]. IEEE Access, 2020, 8:55462-55470.
- [12] 马培军,王甜甜,苏小红,等.基于程序理解的编程题自动评分方法[J].计算机研究与发展,2009,46(7):1136-1142.
- [13] 陈曦,梅广,张金金,等.融合知识图谱和协同过滤的学生成绩预测方法[J].计算机应用,2020,40(2):595-601.
- [14] 黄刚,刘蓉,刘合富,等.基于校园一卡通数据的人群画像分析[J].计算机与数字工程,2018,46(9):1881-1886.
- [15] 姜楠,许维胜.基于校园一卡通数据的学生消费及学习行为分析[J].微型电脑应用,2015,31(2):35-38.