基于一卡通数据优化川大助学金发放

邓钰川 2021141460159

四川大学，计算机科学与技术，成都，610207

dengyuchuan@stu.scu.edu.cn

摘 要 本文介绍了一种基于校园一卡通数据的数据挖掘和分析方法，旨在解决高校中贫困生认定和助学金发放的公正性和准确性问题，同时提高学校后勤服务的质量和效率。该方法包括收集学生一卡通消费信息、数据清洗和预处理、数据挖掘和分析，以及根据挖掘结果进行学生认定和助学金发放等步骤。通过该方法，可以客观了解学生的家庭经济状况，减少人为因素的干扰，提高认定和发放的准确性和公正性，并根据学生的消费习惯和行为特点优化食堂服务和经营管理，提高学生和教职工的满意度。我们认为，该方法具有重要的实践意义和推广价值，可以促进教育公平。

关键词 校园一卡通、数据挖掘、后勤食堂、消费情况分析

Optimizing Sichuan University Scholarship Distribution Based on Card

Yuchuan Deng

Sichuan University, Computer Science and Technology, Chengdu, 610207

dengyuchuan@stu.scu.edu.cn

Abstract This article introduces a data mining and analysis method based on campus all-in-one card data, aiming to solve the fairness and accuracy issues of identifying impoverished students and distributing financial aid in universities, while improving the quality and efficiency of school logistics services. This method includes steps such as collecting student consumption information on one card, data cleaning and preprocessing, data mining and analysis, as well as identifying students and distributing scholarships based on the mining results. Through this method, it is possible to objectively understand students' family economic situation, reduce human interference, improve the accuracy and fairness of identification and distribution, and optimize cafeteria services and management based on students' consumption habits and behavioral characteristics, thereby improving the satisfaction of students and faculty. We believe that this method has important practical significance and promotional value, and can promote educational equity.

Keywords Campus One Card, Data Mining,Canteen, Consumption Analysis

1 引言

在高校中，食堂是学生和教职工每天必须光顾的场所，也是高校后勤服务的重要组成部分。随着学生和教职工数量的增加，食堂的服务质量和效率成为备受关注的问题。本文针对川大后勤食堂，提

出了一种基于校园一卡通数据的数据挖掘和分析方法。由于目前学校认定困难学生和发放助学金主要依赖困难家庭调查表和同学评价，考虑到人为因素，有时难免会出现一些不公平现象，导致一些需要受资助的贫困学生未得到资助，而另一些不需要资助的学生则得到了补助。因此，如何更好地完成精准资助就成为了亟待解决的问题。

通过对学生一卡通消费信息的挖掘，可以及时掌握学生在校消费情况，为贫困生认定和助学金发放提供辅助决策。这种方法可以更客观地了解学生的家庭经济状况，减少人为因素的干扰，提高认定的准确性和公正性。同时，这种方法还可以有效地提高学校后勤服务的质量和效率，根据学生的消费喜好和行为特点，优化食堂服务和经营管理，提高学生和教职工的满意度。因此，本文提出的基于校园一卡通数据的数据挖掘和分析方法具有重要的实践意义和推广价值。

2 相关工作

聚类分析是一种无监督学习方法，用于对未分类的记录集合进行分类。通过一定的规则和算法，将记录集合划分为不同的类别，并通过显式或隐式的方式描述不同类别的特征和属性。与监督学习中需要预先定义类别和标记的训练实例不同，聚类分析可以自动发现数据中的潜在模式和结构，属于观察式学习。聚类分析也被称为概念聚类。聚类分析的输入是一组未分类的记录，而分类分析则需要预先定义好类别。通过分析数据库中的记录数据，聚类分析可以确定每一个记录所属的类别，而聚簇则是被划分为一组有意义的子集记录。聚类分析的目标是最大化类内相似性，最小化类间相似性，使得同一簇中的对象具有高度相似性，而不同簇中的对象具有显著差异。

基于划分的聚类算法通过构造一个迭代过程来优化目标函数，当目标函数优化到最小值或极小值时，可以得到数据集的一些不相交的子集，通常认为此时得到的每个子集就是一个聚类。多数基于划分的聚类算法都是非常高效的，但需要事先给定一个难以确定的聚类数目。其中最著名的两个算法是 k-means 算法[1]和 FCM (Fuzzy C-Means) 算法[2]。此外，还有 k-中心点算法 PAM (Partitioning Around Medoid) 和 CLARA (Clustering Large Applications)[3]、k-模算法（用于聚类分类数据）和 k-原型算法（用于聚类混合数据)[4]，它们也属于这种类型的算法。

基于划分的聚类算法易于实现，聚类速度快，其时间复杂度与数据点数目 n、数据的维度 d 及预先设定的聚类数目 k 成线性关系。例如，k-means 算法和 FCM 算法的时间复杂度为 $T = O(ndkt)$，其中 t 为算法的迭代次数。由于 k-means 算法和 FCM 算法的目标函数优化是个 NP(Non-deterministicPolynomial) 难问题，要搜索到最小值，所花费的时间代价非常高。采用迭代重定位的目标函数优化方法很容易陷入到局部极小值。对于有些数据集，即使优化到目标函数的全局最小值，此时对应的聚类簇也未必与数据集的实际分布结构相吻合。这类算法通常适合那些具有近似 (超) 球体形状、簇半径相近的数据集；而对于极不规则、大小相差很大的数据集，基于划分的聚类算法显得力不从心。因为这类算法具有这样的优点和缺点，所以自 k-means 算法和 FCM 算法发表以后，一直有大量的研究人员从事这方面的理论改进及扩展研究。例如，对于一些局部分布稀疏不均、聚类区域的形状及大小很不规整的数据集，k-means 和 FCM 算法常常不能很好地探测出其聚类分布结构。为克服 k-means 算法和 FCM 算法与初始值有关的两个重大缺陷 (聚类数目 k 的确定、初始中心点集的选择)，许多研究者进行了更为深入的研究。基于划分的聚类算法具有易实现和高速的优点，其时间复杂度与数据点数目 n、数据的维度 d 及预先设定的聚类数目 k 成线性关系。例如，k-means 算法和 FCM 算法的时间复杂度为 $T = O(ndkt)$，其中 t 为算法的迭代次数。但是，由于 k-means 算法和 FCM 算法的目标函数优化属于 NP（Non-deterministicPolynomial，非确定性多项式）难问题，要搜索到最小值需要非常

高的时间代价。而且，采用迭代重定位的目标函数优化方法很容易陷入到局部极小值。在一些数据集中，即使优化到目标函数的全局最小值，对应的聚类簇也未必与数据集的实际分布结构相吻合。这类算法适合于近似（超）球体形状、簇半径相近的数据集，但对于极不规则、大小相差很大的数据集，基于划分的聚类算法显得力不从心。

因为 k-means 算法和 FCM 算法具有这些缺点，自它们被发表以来，一直有大量的研究人员从事理论改进及扩展研究。例如，针对一些局部分布稀疏不均、聚类区域的形状及大小很不规则的数据集，k-means 和 FCM 算法常常不能很好地探测出其聚类分布结构。为了克服 k-means 算法和 FCM 算法的两个重大缺陷（聚类数目 k 的确定和初始中心点集的选择），许多研究者进行了更为深入的研究。

在聚类中心点集的初始化方向上，Arthur 和 Vassilvitskii[5] 提出了复杂的 k-means++ 改进算法。但实际上，该算法的改进效果并不是十分明显。Zalik[6] 提出的算法较好地解决了 k 值和初始聚类中心的选择问题，具有较好的应用参考价值。此外，Cao 等人[7] 提出了一种利用数据点的邻居信息来确定初始聚类中心的方法，也具有一定的参考价值。在确定合适的聚类数目方面，许多研究者在聚类有效性函数的构造方面进行了研究，取得了许多成果，其中包括：基于对象三元组的指标度量，由 Hubert 和 Arabie[8] 提出；基于簇相似性的度量指数，由 Davies 和 Bouldin[9] 提出；Dunn 指数，由 Dunn[10] 提出；Dunn 推广指数，由 Bezdek 和 Pal[11] 提出等。

3 数据收集与预处理

3.1 数据收集

我们的数据集采集自川大校园一卡通系统数据库，包括江安、望江、华西三个校区所有在校学生的一卡通消费信息，时间跨度为过去三年。出于对学生隐私的保护，我们仅采集并保留了数据挖掘所需的必要字段，例如学生的学号、类别（本科生、硕士生、博士生）、所在系、专业、刷卡日期、刷卡时间、刷卡地点、刷卡金额、圈存信息、证件号码、商品名称和使用的 POS 机器代码等。

3.2 数据预处理

在基于校园一卡通的数据挖掘中，数据预处理是非常重要的一步。原始提取的信息可能存在噪声、冗余等。数据预处理的目的是清理、集成、变换和规约原始数据，以便进行后续的数据分析和挖掘工作。

数据清洗 包括去除重复数据、缺失值处理、数据类型转换、数据去噪和数据集成等步骤。在处理校园一卡通数据时，需要只保留学生数据，删除因服兵役、休学、退学、开除等原因引起的缺失数据样本，以及去除一卡通消费中的开水消费和洗浴消费。此外，考虑到特殊日期如节假日等，需要将节假日的售卖数量定为节假日前后两天的平均值，以消除特殊日期对数据分析和挖掘的影响。

数据集成 将不同系统或数据表中的数据根据需要进行集成，保持数据格式和内容的一致性。由于本文用到的数据来源于不同的数据库表，因此需要将不同数据表的数据整合到一起，方便后续的数据分析和挖掘。

数据变换 以提高挖掘的效率和维度，让数据分析更容易实现。在本研究中，对学生学号的前 4 位提取单独处理，使其分为大一到大四四个年级。

数据规约 选择与挖掘目标相关性强的属性，放弃与挖掘目标无相关性或弱关联的属性。在本研究中，放弃了学生的身份证号、电话、家庭住址以及卡号、余额等属性。根据数据变换的结果整理出每位学生的总刷卡次数和总的消费金额，用均值方式得出月均刷卡数和月均消费额，从而大大缩小数据量，提高挖掘的效率。并加入学校校区、学生类别、学号属性

通过以上数据预处理过程，可以使待挖掘的数据合乎规范且精简，为后续的数据分析和挖掘建模提供良好的数据基础。

4 数据挖掘建模

为研究学生在校消费水平，本文使用了模糊聚类算法 FCM[2] 和改进的 FCM 算法 DWFCM[12] 对学生在校消费水平的数据集进行聚类，并对聚类结果进行比较。在 FCM 算法中，对初值的依赖比较强，如果初值选取不合适，可能会导致聚类效果不佳，得到的最终聚类中心与实际有一定差距。而 DWFCM 算法则通过加权密度来更新目标函数，使得密集区域内的样本点权重高于稀疏区域的样本点，并且可以加快聚类中心的迭代速度，从而取得更好的聚类效果，优化算法的性能。

4.1 FCM 算法

假设有 n 个样本点和 c 个聚类中心，每个样本点 $x_i \in \mathbb{R}^d$，每个聚类中心 $v_j \in \mathbb{R}^d$，聚类中心的隶属度为 $u_{ij} \in [0,1]$，表示样本点 x_i 属于聚类中心 v_j 的隶属度。那么 FCM 算法的目标函数可以表示为：

$$J_{FCM} = \sum_{i=1}^{n}\sum_{j=1}^{c} u_{ij}^{m} \left\| x_i - v_j \right\|^2 \tag{1}$$

其中，m 是常数，一般取值大于 1。这个目标函数表示的是样本点与聚类中心之间的欧几里得距离平方，隶属度 u_{ij} 表示样本点 x_i 属于聚类中心 v_j 的程度，m 控制了隶属度的模糊程度，m 越大，隶属度的差异越明显，聚类结果越精确。

根据隶属度矩阵 U，可以计算每个聚类中心的位置 v_j，表示为：

$$v_j = \frac{\sum_{i=1}^{n} u_{ij}^{m} x_i}{\sum_{i=1}^{n} u_{ij}^{m}} \tag{2}$$

根据当前聚类中心和样本点的位置，可以更新隶属度矩阵 U，表示为：

$$u_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{\|x_i - v_j\|}{\|x_i - v_k\|} \right)^{\frac{2}{m-1}}} \tag{3}$$

这里的 m 是前面目标函数中的常数，表示样本点 x_i 属于聚类中心 v_j 的隶属度，隶属度越高表示样本点 x_i 越接近聚类中心 v_j。

FCM 算法的迭代过程将重复进行，直到满足停止条件。常用的停止条件包括目标函数的变化量小于某个阈值或者隶属度矩阵 U 的变化量小于某个阈值。

可以总结为下列详细步骤

(1) 初始化隶属度矩阵 U 和聚类中心 V。

(2) 根据隶属度矩阵 U 计算每个聚类中心的位置 v_j。$v_j = sum_{i=1}^{n}(u_{ij}^{m} * x_i)/sum_{i=1}^{n}(u_{ij}^{m})$

(3) 根据当前聚类中心和样本点的位置，更新隶属度矩阵 U。$u_{ij} = 1/(sum_{k=1}^{c}(||x_i - v_j||/||x_i - v_k||)^{(}2/(m-1)))$

(4) 检查是否满足停止条件。如果满足停止条件，则跳到步骤 6。

(5) 重复步骤 2 和 3，直到满足停止条件为止。

(6) 输出聚类结果。

4.2 DWFCM 算法

假设有 n 个样本点和 c 个聚类中心，每个样本点 $x_i \in \mathbb{R}^d$，每个聚类中心 $v_j \in \mathbb{R}^d$，聚类中心的隶属度为 $u_{ij} \in [0,1]$，表示样本点 x_i 属于聚类中心 v_j 的隶属度。那么 DWFCM 算法的目标函数可以表示为：

$$J_{DWFCM} = \sum_{i=1}^{n}\sum_{j=1}^{c} \rho\left(u_{ij}\right) \left\| x_i - v_j \right\|^2 \tag{4}$$

其中，$\rho(u_{ij})$ 是样本点 x_i 的权重，可以表示为：

$$\rho\left(u_{ij}\right) = \frac{1}{\sum_{k=1}^{n} u_{kj}^{m}} \tag{5}$$

其中，m 是常数，一般取值为 2。这里的权重表示样本点 x_i 在聚类中心 v_j 所处的密度，距离 v_j 越近的样本点会被赋予更高的权重，距离 v_j 较远的样本点会被赋予较低的权重。

根据隶属度矩阵 U，可以计算每个聚类中心的位置 v_j，表示为：

$$v_j = \frac{\sum_{i=1}^{n} \rho\left(u_{ij}\right) u_{ij}^{m} x_i}{\sum_{i=1}^{n} \rho\left(u_{ij}\right) u_{ij}^{m}} \tag{6}$$

根据当前聚类中心和样本点的位置，可以更新隶属度矩阵 U，表示为：

$$u_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{\|x_i - v_j\|}{\|x_i - v_k\|} \right)^{\frac{2}{m-I}}} \tag{7}$$

这里的 m 同样取值为 2，表示样本点 x_i 属于聚类中心 v_j 的隶属度，隶属度越高表示样本点 x_i 越接近聚类中心 v_j。

DWFCM 算法的迭代过程将重复进行，直到满足停止条件。常用的停止条件包括目标函数的变化量小于某个阈值或者隶属度矩阵 U 的变化量小于某个阈值。可以总结为下列详细步骤

(1) 初始化隶属度矩阵 U 和聚类中心 V。

(2) 根据隶属度矩阵 U 计算每个聚类中心的位置 v_j。$v_j = sum_{i=1}^{n}(u_{ij}^{m} * w_i * x_i)/sum_{i=1}^{n}(u_{ij}^{m} * w_i)$

(3) 根据当前聚类中心和样本点的位置，更新隶属度矩阵 U。$u_{ij} = 1/(sum_{k=1}^{c} w_i * (||x_i - v_j||/||x_i - v_k||)^(2/(m-1)))$

(4) 检查是否满足停止条件。如果满足停止条件，则跳到步骤 6。

(5) 重复步骤 2 和 3，直到满足停止条件为止。

(6) 输出聚类结果。

4.3 评估指标

当用聚类算法将数据点分为 k 个簇时，常用的聚类评估指标包括以下几种：

簇内平方和 (Sum of Squared Errors) 簇内平方和是最常用的聚类评估指标之一，它表示所有数据点到其所属簇的质心的距离平方和，即

$$SSE = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2 \tag{8}$$

其中 C_i 表示第 i 个簇，μ_i 表示该簇的质心。

簇间平方和 (Sum of Squares Between) 簇间平方和表示所有簇质心之间的距离平方和，即

$$SSB = \sum_{i=1}^{k} n_i \|\mu_i - \mu\|^2 \tag{9}$$

其中 n_i 表示第 i 个簇中的数据点数量，μ 表示所有数据点的平均值，μ_i 表示第 i 个簇的质心。

Calinski-Harabasz 指数 CH 指数是一种基于簇间平方和和簇内平方和的评估指标，定义为

$$CH = \frac{SSB/(k-1)}{SSE/(n-k)} \tag{10}$$

其中 n 表示数据点的总数。

Davies-Bouldin 指数 Davies-Bouldin 指数是一种基于簇内平均距离和簇间距离的评估指标，定义为

$$DB = \frac{1}{k} \sum_{i=1}^{k} \max_{j \neq i} \left(\frac{\sigma_i + \sigma_j}{\|\mu_i - \mu_j\|} \right) \tag{11}$$

其中 μ_i 表示第 i 个簇的质心，σ_i 表示第 i 个簇中所有数据点到质心的距离的平均值。

Silhouette 系数 Silhouette 系数是一种基于簇内距离和簇间距离的评估指标，定义为

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{12}$$

其中 $a(i)$ 表示数据点 i 到其所在簇中其他数据点的平均距离，$b(i)$ 表示数据点 i 到其他簇中数据点的平均距离。整个数据集的 Silhouette 系数定义为所有数据点的 Silhouette 系数的平均值。

5 结果分析与建议

数据集包括学校校区、学生类别、学号、月均消费金额和月均刷卡次数等属性。在进行聚类之前，需要设置一些参数，包括模糊指数 m、收敛精度 、最大迭代次数 K、分类数 c 等。本文设置模糊指数为 2.0，收敛精度为 0.0001，最大迭代次数为 50，分别取分类数 c 为 3、4、5 进行多次试验。对得到的聚类结果进行比较，发现当分类数取 5 时，能够更好地区分两类极端群体，因此最终选择将数据集分为 5 类。

我们认为当学生的月刷卡数较高且月消费额较低时，属于困难家庭学生的几率比较大。对这些学生在进行困难家庭认定时应着重注意，避免漏选；而对于那些虽然有困难家庭调查表但在校消费较高的同学也应着重调查，避免错选。

改进后的 DWFCM 算法无论是聚类中心还是聚类结果都比 FCM 算法更优，更符合学校困难学生认定现状。如果将 DWFCM 算法与学校现有的贫困生认定方法相结合，可以通过利用算法的聚类结果来确定哪些学生应该被认定为贫困生，从而提高贫困生认定的准确性和覆盖率。这种方法可能会使学校更加有效地识别和支持贫困学生，从而有助于实现精准资助的目标。

6 总结

四川大学是一所重点高校，积极推进数字化校园建设，其中校园一卡通已成为该校数字化校园建设的重要组成部分之一。本文针对校园一卡通食堂消费数据挖掘，提出使用 DWFCM 方法，以为贫困生认定和助学金发放提供客观数据支持。校园一卡通能够帮助学校掌握学生校内的消费及就餐行为，提高后勤食堂管理水平和服务效率，提升师生们就餐体验。通过对校园一卡通用户就餐消费这一特定行为进行数据挖掘与分析，可以为贫困生认定和助学金发放提供客观数据支持。这对于学校提高管理水平、推动数字化校园建设具有重要的现实意义和使用价值。

参 考 文 献

[1] PEIZHUANG W. Pattern recognition with fuzzy objective function algorithms (james c. bezdek)[J]. Siam Review, 1983, 25(3): 442.

[2] MACQUEEN J. Classification and analysis of multivariate observations[C]//5th Berkeley Symp. Math. Statist. Probability. University of California Los Angeles LA USA, 1967: 281-297.

[3] KAUFMAN L, ROUSSEEUW P J. Finding groups in data: an introduction to cluster analysis[M]. John Wiley & Sons, 2009.

[4] HUANG Z. Extensions to the k-means algorithm for clustering large data sets with categorical values[J]. Data mining and knowledge discovery, 1998, 2(3): 283-304.

[5] ARTHUR D, VASSILVITSKII S. k-means++: The advantages of careful seeding[R]. Stanford, 2006.

[6] ŽALIK K R. An efficient k -means clustering algorithm[J]. Pattern Recognition Letters, 2008, 29 (9): 1385-1391.

[7] CAO F, LIANG J, JIANG G. An initialization method for the k-means algorithm using neighborhood model[J]. Computers & Mathematics with Applications, 2009, 58(3): 474-483.

[8] HUBERT L, ARABIE P. Comparing partitions[J]. Journal of classification, 1985, 2: 193-218.

[9] DAVIES D L, BOULDIN D W. A cluster separation measure[J]. IEEE transactions on pattern analysis and machine intelligence, 1979(2): 224-227.

[10] DUNN J C. A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters[M]. Taylor & Francis, 1973.

[11] BEZDEK J C, PAL N R. Some new indexes of cluster validity[J]. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 1998, 28(3): 301-315.

[12] LI P, CHEN Z, HU Y, et al. A weighted fuzzy c-means clustering algorithm for incomplete big sensor data [C]//Wireless Sensor Networks: 11th China Wireless Sensor Network Conference, CWSN 2017, Tianjin, China, October 12-14, 2017, Revised Selected Papers 11. Springer, 2018: 55-63.