文章编号:1673-0291(2026)03-0063-10 **DOI**:10.11860/j.issn.1673-0291.20250139

基于聚类算法的高铁旅客市场细分方法研究

范家乐[1],景 云[1],徐 彦[2]
(1.北京交通大学 交通运输学院,北京 100044;2.中国国家铁路集团有限公司 客运中心,北京 100844)

摘 要:针对传统聚类算法进行高速铁路旅客市场细分结果不稳定,难以有效提取各类别旅客特征差异等问题,提出基于Halton序列的K-均值-自适应学习粒子群(K-Means-Adaptive Learning Particle Swarm Optimization,KM-ALPSO)算法.该算法基于京沪高铁客票数据研究高铁旅客市场细分.首先,将旅客年龄、提前购票时间、出发时间段作为特征变量,利用近邻传播(Affinity Propagation,AP)算法识别数据集中的样本代表点;然后,基于Halton序列生成初始粒子群,使用KM-ALPSO算法对样本代表点进行聚类,并与经典的K-均值算法、K-均值-粒子群(K-Means-Particle Swarm Optimization,KM-PSO)算法等进行对比分析,选择轮廓系数(Silhouette Coefficient,SC)、Davies-Bouldin(DB)、Calinski-Harabasz(CH)指标评价算法的聚类效果,确定最终聚类数目;最后,分析不同类别旅客群体的特征差异,应用频繁模式增长(Frequent Pattern-Growth,FP-Growth)算法提取强关联规则.研究结果表明:使用AP算法预处理可使基于Halton序列的KM-ALPSO算法的运行时间缩短至原来的23.26%,同时保持与未预处理时相近的评价指标值;利用Halton序列初始化粒子群位置,可以提升KM-ALPSO算法全局搜索能力;基于Halton序列的KM-ALPSO算法在客票数据分析中的聚类评价指标SC为0.332,DB为0.933,CH为708.5,表明聚类数目为5时性能最佳,且优于其他算法;5类旅客在年龄、提前购票时间和出发时间段上差异显著,表现出不同的出行计划性和出发时间偏好.

关键词:铁路运输;市场细分;聚类算法;高铁旅客;强关联规则
中图分类号:U293.1 **文献标志码**:A

Research on market segmentation method for high-speed rail passengers based on clustering algorithm

FAN Jiale[1], *JING Yun*[1], *XU Yan*[2]
(1.School of Traffic and Transportation, Beijing Jiaotong University, Beijing 100044, China;
2.Passenger Transport Center, China National Railway Group Co., Ltd., Beijing 100844, China)

Abstract: To address the instability and difficulty in effectively extracting characteristic differences among passenger groups when applying traditional clustering algorithms to high-speed rail passenger market segmentation, this study proposes a K-Means-Adaptive Learning Particle Swarm Optimization (KM-ALPSO) algorithm based on the Halton sequence. This algorithm is applied to segment the high-speed rail passenger market using ticket data from the Beijing-Shanghai high-speed railway.

收稿日期:2025-09-06;**修回日期**:2025-12-26
基金项目:国家自然科学基金(52372300)
Foundation item: National Natural Science Foundation of China (52372300)
第一作者:范家乐(2003—),男,河南信阳人,博士生.研究方向为铁路运输组织理论与技术.**email**:298250836@qq.com.
通信作者:景云(1981—),男,陕西宝鸡人,教授,博士,博士生导师.**email**:yjing@bjtu.edu.cn.

引用格式:范家乐,景云,徐彦.基于聚类算法的高铁旅客市场细分方法研究[J].北京交通大学学报,2026,50(3):63-72.
FAN Jiale, JING Yun, XU Yan. Research on market segmentation method for high-speed rail passengers based on clustering algorithm[J]. Journal of Beijing Jiaotong University, 2026, 50(3): 63-72. (in Chinese)

First, considering passenger age, advance booking time, and departure period as feature variables, the Affinity Propagation (AP) algorithm is adopted to identify representative sample points within the dataset. Second, an initial particle swarm is generated based on the Halton sequence, and the KM-ALPSO algorithm is used to cluster the representative sample points. A comparative analysis is then conducted against the classic K-Means algorithm and the K-Means-Particle Swarm Optimization (KM-PSO) algorithm. The Silhouette Coefficient (SC), Davies-Bouldin (DB) index, and Calinski-Harabasz (CH) index are selected to evaluate the clustering performance and determine the optimal number of clusters. Finally, the characteristic differences among various passenger groups are analyzed, and the Frequent Pattern-Growth (FP-Growth) algorithm is employed to extract strong association rules. The results indicate that preprocessing with the AP algorithm reduces the runtime of the Halton sequence-based KM-ALPSO algorithm to 23.26% of its original duration while maintaining evaluation metrics comparable to those obtained without preprocessing. Furthermore, initializing particle swarm positions with the Halton sequence enhances the global search capability of the KM-ALPSO algorithm. In the analysis of ticket data, the Halton sequence-based KM-ALPSO algorithm achieves an SC of 0.332, a DB of 0.933, and a CH of 708.5, demonstrating that a cluster number of five yields optimal performance and outperforms the baseline algorithms. The five identified passenger groups exhibit significant differences in age, advance booking time, and departure period, revealing distinct travel planning tendencies and departure time preferences.

Keywords: railway transportation; market segmentation; clustering algorithm; high-speed rail passengers; strong association rules

随着我国高速铁路的不断发展,旅客出行需求呈现出多样化的特点,客运市场的竞争日趋激烈.为满足多样化的旅客出行需求、增强高铁市场竞争力,有必要引入收益管理理论,最大化铁路部门的总收益.收益管理理论以市场为导向,对市场进行细分,通过分析各子市场的群体特征,制订定价策略,构建存量控制模型[1].市场细分是收益管理理论的重要环节,为后续定价策略和存量控制奠定基础.

国内外学者通常使用聚类算法研究市场细分,传统聚类算法按照算法思想的不同通常分为原型聚类、密度聚类和层次聚类.原型聚类按聚类结构刻画方式的不同可分为基于原型向量的聚类算法(如K-means算法)和基于概率模型的聚类算法.文献[2]使用经典的K-means算法对Sacrament地区的共享单车市场进行细分研究,从用户对共享单车服务系统的感知、用户使用共享单车的出行目的和用户的通勤模式3个角度进行聚类分析,挖掘用户出行习惯,为提高共享单车使用率提供政策建议.文献[3]为分析机场旅客出行行为,基于K-means算法对旅客特征数据集进行聚类分析,并结合轮廓系数法和主成分分析法优化旅客聚类结果,为构建机场旅客出行链提供保障.文献[4]将量化处理后的态度变量作为特征变量,使用K-means算法进行聚类分析,细分公交出行信息使用市场.文献[5]为研究低收入通勤者公交出行选择差异,考虑出行者的心理潜变量需求作为特征变量,利用K-means算法细分公交出行市场,剖析产生选择差异的主观态度方面的原因.文献[6]对传统的高斯混合聚类(Gaussian Mixture Model, GMM)进行改进,提出双层高斯混合聚类(Two-Level Gaussian Mixture Model, TL-GMM),用于优化共享单车系统.文献[7]基于证据理论、GMM和期望最大化(Expectation Maximization, EM)算法,在证据c均值(Evidential C-Means, ECM)算法[8]的基础上,提出了证据高斯混合聚类(Evidential GMM, EGMM)算法,结果表明其性能优于传统的GMM算法.文献[9]对传统的基于密度的带噪声空间聚类(Density-Based Spatial Clustering of Applications with Noise, DBSCAN)算法进行改进,引入基于时间序列的动态密度聚类算法,对不同的旅客群体行为特征进行对比分析,为铁路客运营销策略研究提供支持.然而,以上研究均具有局限性,GMM算法及其改进算法要求聚类特征变量服从或近似服从高斯混合分布;DBSCAN算法及其改进算法无法直接计算类均值,各类别特征差异难以精准解释;经典的K-means算法应用广泛,但基于贪心策略进行迭代优化,极易受到初始聚类中心的影响,并且聚类结果不稳定.

针对K-means算法的缺陷,有关学者基于群体

智能的思想,使用启发式算法中的群智能算法(遗传算法、粒子群算法等)改进K-means算法,以得到更加有效的聚类结果.文献[10]将聚类作为一个子问题进行研究,使用非支配排序遗传算法Ⅱ(Non-dominated Sorting Genetic Algorithm Ⅱ,NSGA-Ⅱ)改进K-means算法.文献[11]使用改进粒子群算法(Particle Swarm Optimization,PSO)优化K-means算法的目标函数,提出KM-ALPSO算法,为PSO算法设计参数动态更新的自适应学习(Adaptive Learning,AL)策略,最终根据收集到的3 230份有效调查问卷对中国葡萄消费市场进行细分,为相关企业制定营销策略奠定基础.KM-ALPSO算法的核心在于PSO算法,文献[12]回顾PSO算法的发展历程,详细说明原始PSO算法的基本原理,并简单评述PSO算法的改进方向.文献[13]提出基于Halton序列的PSO算法,使得初始各粒子位置随机且均匀,利于提升算法的全局搜索能力.然而,KM-ALPSO算法因计算流程复杂,导致其效率低下,运算时间长,需要结合其他算法进行改进.近邻传播算法(Affinity Propagation,AP)可以为其改进提供思路.AP算法由文献[14]首次提出,该算法不需要事先确定聚类数目,通过设置参数p值,算法可基于样本点之间的相似度和竞争关系自动识别若干样本代表点.然而,该算法的局限性为最终的样本代表点不确定,若聚类数目过多,实际应用性较差.文献[15]基于AP算法对乘坐普速列车和高速铁路动车组列车出行的旅客进行聚类,所得聚类效果可以接受,一定程度上说明了该算法在铁路客运市场细分领域中的适用性.以上研究在分析聚类结果时,仅仅依赖样本描述性统计,无法深入分析各聚类特征变量取值对最终聚类结果的影响程度.关联规则分析可作为解决该问题的一种方法.关联规则分析通常使用Apriori算法和FP-Growth算法,文献[16]使用Apriori算法对贵广高铁某年11月份1周的客票数据进行关联规则挖掘,分析旅客提前购票时间、旅行时间、发车时间等特征之间的关联,为研究旅客出行行为和客流时空特征提供依据.文献[17]利用FP-Growth算法研究道路交通领域违法行为与交通事故之间的关联性,并说明了FP-Growth算法在处理大规模数据集时的优越性.以上研究的方法整合度不足,仅应用单一方法分析研究对象.聚类与关联规则分析具有天然互补性,聚类可实现群体细分,而关联规则能深入解释子群体的内在规律.

因此,本文利用客票数据研究市场细分,客票数据结构是研究的基础,通过对客票数据的特征变量进行聚类对比试验,提取旅客年龄、提前购票时间、出发时间段作为最终聚类特征,聚类特征变量提取没有统一标准,与待分析数据集的结构密切相关.旅客年龄可直接从客票数据中提取;提前购票时间以天为单位,根据发车时间与购票时间之差得到;出发时间段根据实际数据集取值情况,考虑范围为7:00至23:00,每1 h为一个区间,分别记为1,2,…,16.上述3个特征变量均为离散变量,但是各变量的取值范围大.因此首先利用AP算法识别客票数据中的样本代表点,使用基于Halton序列的KM-ALPSO算法对样本代表点进行聚类分析,考虑SC、DB、CH作为评价指标说明算法的有效性,之后针对旅客聚类结果进行特征分析,最后使用FP-Growth算法挖掘强关联规则.该聚类算法具备一定的普适性,可为高铁客运市场细分提供研究思路,为铁路领域开展收益管理实践奠定基础.

1　高铁旅客市场细分研究框架

收益管理理论以最大化总收入为目标,通常由市场细分、客流预测、定价策略和存量控制等模块组成.市场细分作为应用收益管理理论的先决步骤,对后续分析起到至关重要的作用.市场细分研究框架如图1所示,包含预处理、聚类和结果分析在内的3个处理步骤.

步骤1:预处理.客票数据是铁路部门的基础数据,可用于旅客行为分析.但是,其数据量大的特点会导致聚类算法的运行效率低、运行时间长.因此可应用AP算法识别大规模数据集中的样本代表点,即潜在的聚类中心,通过数据点之间的交互信息,生成样本代表点数据集.

步骤2:聚类.针对样本代表点,分别利用基于Halton序列的KM-ALPSO算法、KM-ALPSO算法、KM-PSO算法和K-means算法进行聚类分析,得到不同聚类数目下的运行结果,结合SC、DB、CH评价指标,判断各算法的聚类效果,并确定最终聚类数目.

步骤3:结果分析.根据旅客聚类结果,对每一子市场旅客群体分析其行为特征,总结群体出行偏好,为铁路部门设计客运产品奠定基础.同时,将旅客行为特征作为关联规则前项,旅客类别作为关联规则后项,挖掘强关联规则.

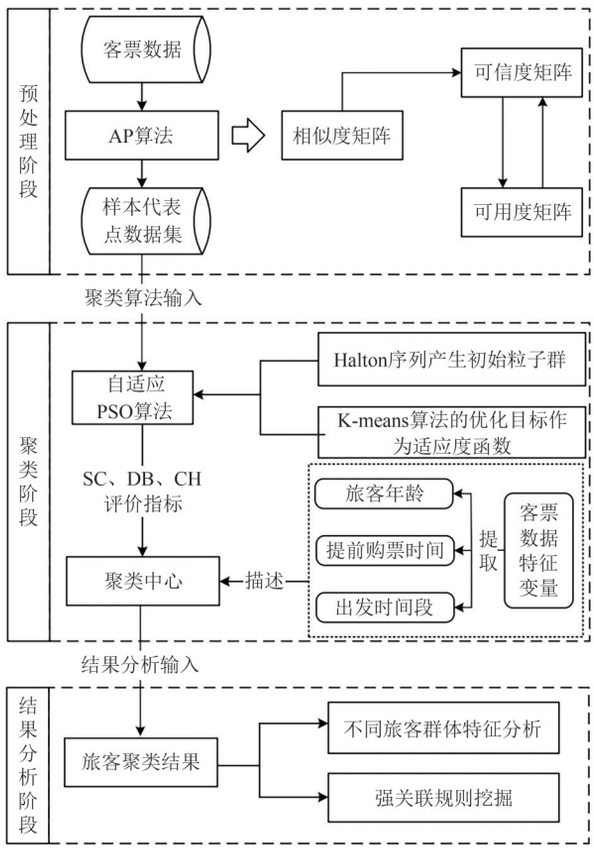


图1 市场细分研究框架

Fig.1 Research framework of market segmentation

2 AP-KM-ALPSO聚类算法

2.1 AP算法

AP算法相较于传统聚类算法，最大的优势在于不需要提前设置聚类数目.通过引入样本相似度矩阵、可信度矩阵、可用度矩阵，AP算法可在众多样本点中自动识别样本代表点，能够有效降低数据集规模，提升聚类效率.在旅客市场细分任务中，1个样本点表示1位旅客的特征信息，使用维度数量为聚类特征数的向量表示.算法包含3个处理步骤.

步骤1：初始化.对于任意两个样本点x_i和x_k，i和k表示索引，计算两者之间的相似度，进而构成相似度矩阵元素为

$$s(i,k)=\begin{cases}-\left\| x_i-x_k \right\|_2^2 & i\neq k\\ p & i=k\end{cases} \tag{1}$$

式中：p为相似度矩阵对角线上的元素，衡量样本点选择自身作为代表点的程度，p值越大，聚类数目越多.

样本的可信度矩阵和可用度矩阵初始设置为零值矩阵，两者的具体内涵参见文献[14].

步骤2：迭代更新.可信度矩阵中的元素$r(i,k)$更新式为

$$r(i,k)=s(i,k)-\max_{k'\neq k}\left\{a(i,k')+s(i,k')\right\} \tag{2}$$

可用度矩阵中的元素$a(i,k)$更新式为

$$a(i,k)=\begin{cases}\min\left\{0,r(k,k)+\sum\limits_{i'\neq\{i,k\}}\max\left\{0,r(i',k)\right\}\right\} & i\neq k\\ \sum\limits_{i'\neq k}\max\left\{0,r(i',k)\right\} & i=k\end{cases} \tag{3}$$

步骤3：输出聚类结果.达到最大迭代次数后，对任意样本点x_i，求可信度矩阵与可用度矩阵之和为

$$l=\arg\max_{k}\left\{r(i,k)+a(i,k)\right\} \tag{4}$$

得到样本点x_i对应的样本代表点x_l.

2.2 基于Halton序列的KM-ALPSO算法

KM-ALPSO算法以K-means算法的目标函数作为ALPSO算法的适应度函数，进行迭代优化.其中，K-means算法是一种经典的聚类算法，优化目标为

$$\min Z=\sum_{j=1}^{J}\sum_{x\in C_j}\left\| x-x_j \right\|_2^2 \tag{5}$$

$$x_j=\frac{1}{\left| C_j \right|}\sum_{x\in C_j}x \tag{6}$$

式中：J表示聚类数目；C_j代表第j个类的样本点集合.

该问题需要优化各类的样本点集合C_j和类中心x_j，是一个NP-hard问题.因此K-means算法采用了贪心策略，通过迭代以求得近似解.经典的K-means算法易受到初始聚类中心的影响，并且聚类结果不稳定.改进算法，如K-means++，虽然在一定程度上缓解了K-means算法过度依赖初始聚类中心的缺点，但算法全局搜索能力依旧没有改善.基于Halton序列的KM-ALPSO算法可作为解决算法全局搜索能力差的一种方法.

Halton序列用于初始化粒子群，使得粒子群均匀且随机，可以增强算法全局搜索能力.Halton序列依据基数逆函数生成.基数逆函数为

$$\phi_p(e)\triangleq\frac{b_0}{p}+\frac{b_1}{p^2}+\cdots+\frac{b_m}{p^{m+1}} \tag{7}$$

式中：p是质数；$e=b_0+b_1p+\cdots+b_mp^m$，满足$0\leqslant b_v<p$，$0\leqslant v\leqslant m$且v为整数.

在s维空间中，初始化g个粒子的Halton序列$\boldsymbol{Y}_g$为

$$\boldsymbol{Y}_g=\left(\phi_{p_1}(g),\phi_{p_2}(g),\cdots,\phi_{p_s}(g)\right) \tag{8}$$

式中:$p_1, p_2, \cdots, p_s$均为质数.

Halton序列生成初始解后,运行KM-ALPSO算法,包含6个处理步骤.

步骤1:粒子编码和适应度函数定义.假设聚类数目为J,聚类特征数为n.基于Halton序列初始化粒子群位置向量为$(X_{11}, X_{12}, \cdots, X_{1n}), (X_{21}, X_{22}, \cdots, X_{2n}), \cdots, (X_{J1}, X_{J2}, \cdots, X_{Jn})$.随机初始化粒子群速度向量为$(V_{11}, V_{12}, \cdots, V_{1n}), (V_{21}, V_{22}, \cdots, V_{2n}), \cdots, (V_{J1}, V_{J2}, \cdots, V_{Jn})$.选取适应度函数为

$$f(x_1, x_2, \cdots, x_J) = \sum_{j=1}^{J} \sum_{x \in C_j} \left\| x - x_j \right\|_2^2 \tag{9}$$

步骤2:更新最优向量.计算粒子群中各粒子的适应度,适应度值越小,则认为粒子位置向量越好,据此更新粒子群的个体最优向量和群体最优向量.

步骤3:计算惯性权重w^t.设t为当前迭代次数,T_{rec}为提前设定的参数,反映惯性权重更新式的变化.当$t \leqslant T_{\text{rec}}$时,粒子群的迭代次数较低,应注重全局搜索能力,$w^t$更新式为

$$w^t = \begin{cases} \left(\dfrac{2(w_{\min} - w_{\max}) t^2}{T_{\max}^2} + w_{\max} \right) \exp(-\alpha^t t^2) + \beta^t \\ \qquad \gamma^t \geqslant \tau \\ w_{\max} - \dfrac{(w_{\max} - w_{\min})}{T_{\max}} t \\ \qquad \gamma^t < \tau \end{cases} \tag{10}$$

当$t > T_{\text{rec}}$时,粒子群的迭代次数较高,应注重局部搜索能力,w^t更新式为

$$w^t = \begin{cases} \left(\dfrac{2(w_{\max} - w_{\min}) t^2}{T_{\max}^2} - \dfrac{4(w_{\max} - w_{\min}) t}{T_{\max}} + \right. \\ \left. 2w_{\max} - w_{\min} \right) \exp(-\alpha^t t^2) + \beta^t \quad \gamma^t \geqslant \tau \\ w_{\max} - \dfrac{(w_{\max} - w_{\min})}{T_{\max}} t \quad \gamma^t < \tau \end{cases} \tag{11}$$

式中:$w_{\min}$、$w_{\max}$表示惯性权重的下界和上界,一般取值0.4和0.9;$T_{\max}$表示最大迭代次数;阈值τ一般取值0.001;α^t、β^t和γ^t分别表示粒子群进化程度、聚合程度和提前收敛因子,表达式分别为

$$\alpha^t = \frac{f_{\text{gbest}}^t}{f_{\text{gbest}}^{t-1}} \tag{12}$$

$$\beta^t = \frac{f_{\text{gbest}}^t}{f_{\text{avg}}^t} \tag{13}$$

$$\gamma^t = \left| f_{\text{gbest}}^t - f_{\text{avg}}^t \right| \tag{14}$$

式中:f_{gbest}^t表示第t代粒子群最优适应度值;f_{avg}^t表示第t代粒子群平均适应度值.

步骤4:计算个体学习因子c_1^t和种群学习因子c_2^t分别为

$$c_1^t = \frac{w^t + 3}{2} \tag{15}$$

$$c_2^t = \frac{-w^t + 3}{2} \tag{16}$$

式(15)和式(16)的推导原理涉及序列的平稳性条件[18],详细过程参见文献[11].

步骤5:更新粒子的位置和速度分别为

$$X_{hd}^t = X_{hd}^{t-1} + V_{hd}^t \tag{17}$$

$$V_{hd}^t = w^{t-1} V_{hd}^{t-1} + c_1^{t-1} r_1 \left(\text{pbest}_{hd}^{t-1} - X_{hd}^{t-1} \right) + c_2^{t-1} r_2 \left(\text{gbest}_{hd}^{t-1} - X_{hd}^{t-1} \right) \tag{18}$$

式中:$h = 1, 2, \cdots, J$; $d = 1, 2, \cdots, n$; r_1和r_2分别为0~1的随机数;pbest_{hd}^{t-1}和gbest_{hd}^{t-1}分别表示$t-1$代的个体最优向量和群体最优向量在第h个聚类中心、第d个聚类特征上的取值.

步骤6:迭代停止条件.判断粒子迭代过程是否达到设置的最大迭代次数,是则算法结束,否则返回步骤2.

基于Halton序列的KM-ALPSO算法流程图如图2所示.

2.3　聚类算法评价指标

在没有任何外界参考模型的条件下,一般使用内部评价指标衡量聚类算法的优劣,原则为类内相似度越高,类间相似度越低,聚类结果越好.通常使用轮廓系数(Silhouette Coefficient, SC)、Davies-Bouldin(DB)、Calinski-Harabasz(CH)作为聚类算法的评价指标[11].SC指标越大,DB指标越小,CH指标越大,聚类效果越好.

3　关联规则分析

3.1　关联规则

关联规则分析作为一种较为实用的数据挖掘技术,旨在从数据集中提取数据属性频繁出现的先验关联关系,基本概念[19]包含3个方面.

方面1:项与项集、事务.设I_i为一个项,项作为数据的基本组成单位,构成项集,比如$I = \{I_1, I_2, \cdots, I_z\}$是具有$z$个项目的集合,简称$z$-项集;事务用$D$表示,每一个样本可以理解为一个事务,事务的集合为事务集,事务集中的事务数记为$|D|$.

方面2:支持度、置信度、提升度.设X和Y为项集,满足$X \cap Y = \varnothing$,则关联规则可用蕴含式来表达:$X \Rightarrow Y$,其中X为关联规则前项,Y为关联规则后项.记$\text{num}(X \cup Y)$表示X和Y同时出现在一个

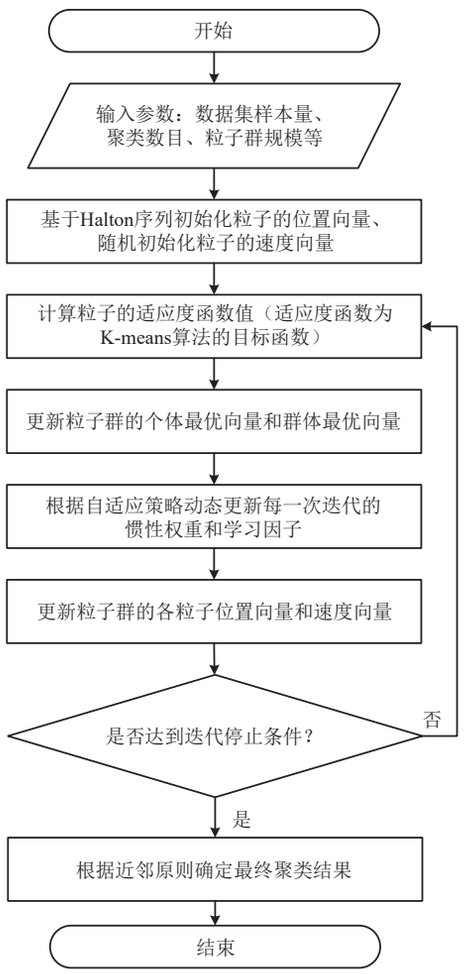


图 2 基于 Halton 序列的 KM-ALPSO 算法流程图

Fig. 2 Flowchart of the KM-ALPSO algorithm based on Halton sequences

事务中的次数，则关联规则 $X \Rightarrow Y$ 的支持度、置信度和提升度分别为

$$\text{Support}(X \Rightarrow Y) = \frac{\text{num}(X \cup Y)}{|D|} \tag{19}$$

$$\text{Confidence}(X \Rightarrow Y) = \frac{\text{Support}(X \cup Y)}{\text{Support}(X)} \tag{20}$$

$$\text{Lift}(X \Rightarrow Y) = \frac{\text{Confidence}(X \Rightarrow Y)}{\text{Support}(Y)} \tag{21}$$

方面 3：最小支持度、最小置信度、频繁项集、强关联规则．最小支持度和最小置信度均需根据实际情况提前预设，若某项集的支持度大于或者等于最小支持度，则称该项集为频繁项集；若某关联规则的支持度和置信度均大于或者等于最小支持度和最小置信度，且提升度大于 1，则该关联规则称为强关联规则．通常情况下，提升度越大，强关联规则越有效[2]．

3.2 FP-Growth 算法

关联规则分析常用 Apriori 算法和 FP-Growth 算法．FP-Growth 算法通过构造频繁模式树（FP-Tree）来提升搜索效率，在处理大规模数据集时，较 Apriori 算法更高效[20]．FP-Growth 算法包含 6 个处理步骤．

步骤 1：遍历数据集，计算每个项目的支持度，根据预设的最小支持度，删除小于最小支持度的项目，建立剩余项目的 FP-Tree 和项头链表．

步骤 2：依据项头链表自下而上访问 FP-Tree 生成条件模式基．

步骤 3：根据条件模式基构建条件 FP-Tree．

步骤 4：判断 FP-Tree 是否为空或仅含单路径，如果否，则返回步骤 2．

步骤 5：通过条件 FP-Tree 生成频繁项集．

步骤 6：对频繁项集进行划分，生成所有非空真子集，计算由真子集构成的关联规则的置信度和提升度，根据预设的最小置信度和强关联规则定义，输出强关联规则．

4 实例分析

4.1 聚类算法运行结果

高铁客票数据是铁路部门的基础数据，包含的原始信息有车次、乘车日期、上车站、发车时间、到站、席别、票价、购票时间、购票方式、旅客年龄、旅客性别[21]．根据上述信息可以进行特征提取与特征构造，最终考虑旅客年龄、提前购票时间、出发时间段作为最终聚类特征，旅客年龄属于旅客个人特征变量，提前购票时间和出发时间段属于旅客出行行为特征变量．本文使用的数据集为京沪高铁某年 7 月份的 10 000 条数据，将数据集聚类特征进行归一化处理，便可进行聚类分析．本文所有算法测试均在配置为 AMD Ryzen 7 5800H with Radeon Graphics CPU @3.20 GHz，内存为 16.0 GB 的计算机上执行．首先使用 AP 算法挖掘初始数据集中的先验信息，以达到降低数据量，提高算法运行效率的目的．通过设置 p 值为 -0.1[22]，AP 算法共识别 1 331 个样本代表点，如图 3 所示，实现了样本数据量的降低．

通过设置参数 $T_{\max} = 1\,200$，$T_{\text{rec}} = T_{\max}/2$，粒子群规模为 50，KM-PSO 算法、KM-ALPSO 算法和基于 Halton 序列的 KM-ALPSO 算法的迭代均收敛，基于 Halton 序列的 KM-ALPSO 算法能够以较快速度达到收敛状态，且基于 Halton 序列的 KM-ALPSO 算法和 KM-ALPSO 算法的收敛稳定性较 KM-PSO 算法强．相同参数设置下，若不使用 AP 算法预处理，基于 Halton 序列的 KM-ALPSO 算法迭代 1 200 次

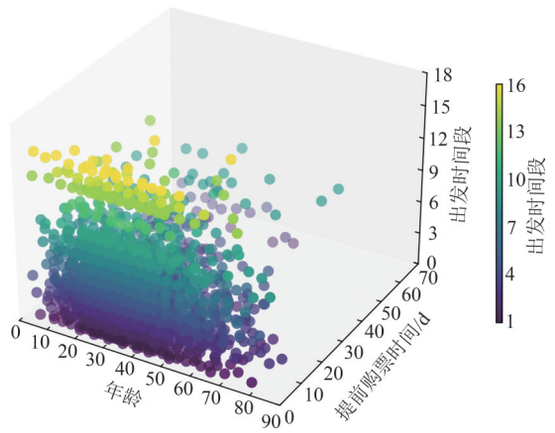


图 3　AP 算法识别的样本代表点

Fig.3　Exemplary samples detected by the AP algorithm

需 18 169.46 s,而使用 AP 算法的情况下运行时间减少至 4 227.01 s,为不作预处理时的 23.26%,且两者的评价指标值差距较小.

算法适应度函数值低,仅能代表类内相似度高,无法衡量类间相似度的大小,算法迭代收敛仅能说明算法的有效性,无法全面地评价算法的聚类效果.因此需要 SC 指标、DB 指标、CH 指标综合考虑类内相似度和类间相似度的大小,评选出聚类效果最佳的算法,并确定最终的最优聚类数目.不同聚类数目下各算法的评价指标值如表 1 所示.

由表 1 可知,最终的聚类数目应当确定为 5.当聚类数目为 5 时,通过对比基于 Halton 序列的 KM-ALPSO 算法、KM-ALPSO 算法、KM-PSO 算法、经典的 K-means 算法的 SC、DB、CH 值,表明基于 Halton 序列的 KM-ALPSO 算法的各项指标均优于其他算法,由此说明该算法在综合考虑类内相似度和类间相似度的条件下,聚类效果较好.

4.2　不同类别旅客群体特征分析

通过样本描述性统计,5 个类别旅客群体的年龄、提前购票时间和出发时间段对比分析如图 4 所示.由图 4 可知:5 个类别的旅客群体在年龄、提前购票时间、出发时间段这 3 个方面存在较大差异.第 1 类旅客年龄分布离散程度最高,提前较长时间开始购票,出行的计划性非常强,出发时间段偏向于 9:00 至 11:00 和 15:00 至 16:00;第 2 类旅客年龄分布较为集中,26~40 岁的旅客偏多,当天购票和提前 1 天购票的旅客占比接近 65%,出行计划性较弱,出发时间段集中于 12:00 至 16:00,偏好中午、下午出行;第 3 类旅客年龄分布呈现近似双峰形态,23~39 岁的旅客偏多,大多提前 1~2 天购票,出发时间段集中于 9:00 至 10:00,其次是 7:00 至 8:00,偏好早间出行;第 4 类旅客年龄分布离散程度较高,多为

表 1　不同聚类数目下各算法的评价指标值

Tab.1　Evaluation metric values of different algorithms under various cluster numbers

聚类方法	聚类数目	评价指标		
		SC 指标	DB 指标	CH 指标
K-means	2	0.342	1.216	632.3
	3	0.295	1.165	607.9
	4	0.294	1.172	536.6
	5	0.268	1.111	620.4
	6	0.291	1.051	659.6
KM-PSO	2	0.342	1.211	632.1
	3	0.296	1.156	607.1
	4	0.333	0.955	713.4
	5	0.324	0.938	693.7
	6	0.261	1.029	641.2
KM-ALPSO	2	0.342	1.217	632.3
	3	0.295	1.157	608.7
	4	0.322	0.977	696.8
	5	0.328	0.954	696.6
	6	0.310	0.968	630.5
基于 Halton 的 KM-ALPSO	2	0.333	1.175	605.0
	3	0.356	1.058	611.5
	4	0.327	0.956	638.9
	5	0.332*	0.933*	708.5*
	6	0.311	0.955	672.6

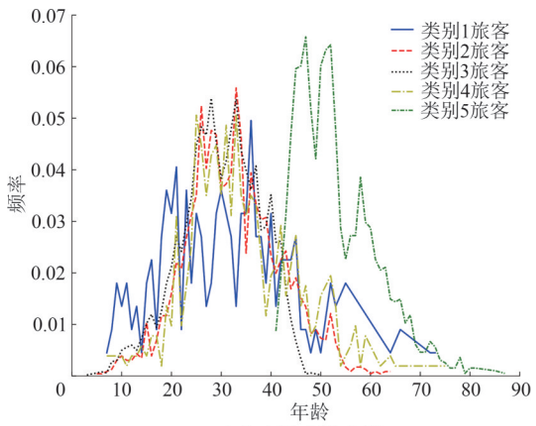
注:*表示最终的选择结果.

年轻旅客,当天购票的人数占比超过 50%,出行计划性最弱,出发时间段集中于 20:00 至 21:00,倾向于晚间出发;第 5 类旅客年龄分布也呈现近似双峰形态,年龄全部在 40 岁以上,为年长旅客群体,大多提前 1~2 天购票,出发时间段集中于 9:00 至 12:00,偏好早间、午间出行.

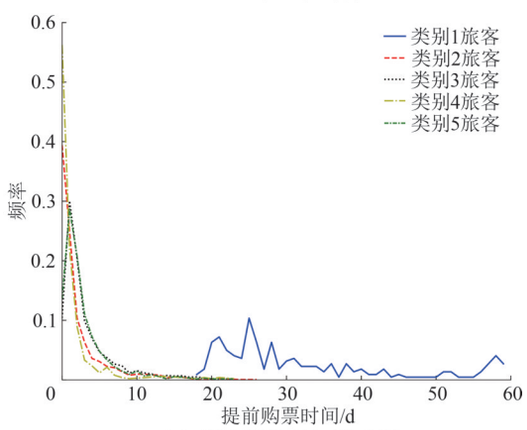
旅客年龄不同,对客运服务质量的感知程度不同,通常年长旅客对出行舒适性的要求更高;旅客对提前购票时间和出发时间段的偏好差异性,可以分别为预售期内的动态定价和一天内的分时定价提供依据.铁路部门可基于旅客市场细分结果,根据不同类别旅客群体的出行异质性,分别设计相对应的客运产品,在满足旅客异质性需求的基础上,提高自身的总收益.

4.3　强关联规则挖掘

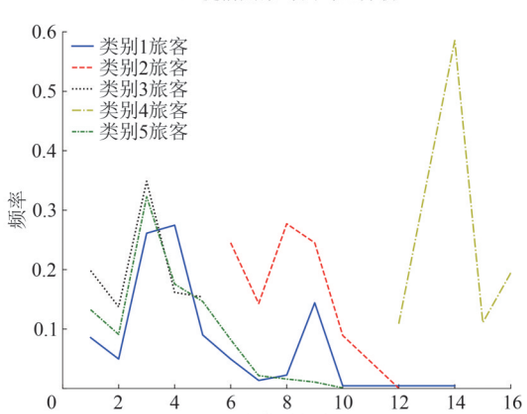
为了方便铁路部门判断旅客类别,对客票数据集中的 3 个特征和旅客类别之间的强关联性进行分析.首先对特征进行离散化处理,使用五数(最小值、第一四分位数、中位数、第三四分位数、最大值)进行统计分析,3 个特征的五数总结如表 2 所示.关联规则的前项设置为 3 个特征的分级水平值,后项设置为旅客所属类别.根据表 2 中 3 个特征的最小



（a）年龄对比分析



（b）提前购票时间对比分析



（c）出发时间段对比分析

图4 5个类别旅客群体的特征对比分析

Fig.4 Comparative analysis of characteristics among five passenger groups

值、第一四分位数、中位数、第三四分位数、最大值，可对特征进行离散化分级处理.

在jupyter notebook上使用Python语言运行FP-Growth算法，设置最小支持度为0.15，最小置信度为0.3，共得到24个频繁项集，其中频繁1项集为15个，频繁2项集为9个.根据关联规则前项和后项的设置，通过判断提升度是否大于1，共得到符合要求的强关联规则9个，强关联规则如表3所示.

根据表3中的强关联规则，铁路部门能够实现仅依据旅客的少量信息就推断出旅客所属类别，并根据支持度、置信度和提升度来评价推断结果的有效性和准确性.

表2 3个特征的五数总结

Tab.2 Five-number summary of the three features

五数	年龄	提前购票时间/d	出发时间段
最小值	3	0	1
第一四分位数	27	1	3
中位数	33	2	4
第三四分位数	43	4	6
最大值	87	59	16

表3 强关联规则

Tab.3 Strong association rules

强关联规则	支持度	置信度	提升度
“年龄在33～42岁”⇒“旅客为第3类”	0.180	0.639	1.276
“提前2～3天购票”⇒“旅客为第3类”	0.152	0.587	1.172
“出发时间段为12:00以后”⇒“旅客为第2类”	0.231	0.737	3.193
“提前1天购票”⇒“旅客为第3类”	0.151	0.547	1.092
“年龄在26岁及以下”⇒“旅客为第3类”	0.164	0.660	1.318
“出发时间段为7:00至9:00”⇒“旅客为第3类”	0.168	0.783	1.564
“出发时间段为9:00至10:00”⇒“旅客为第3类”	0.175	0.718	1.434
“出发时间段为10:00至12:00”⇒“旅客为第3类”	0.158	0.691	1.379
“年龄在43岁及以上”⇒“旅客为第5类”	0.189	0.727	3.737

5 结论

1）基于自适应启发式方法改进的KM-ALPSO算法进行高铁旅客市场细分时，收敛速度快，收敛稳定性强，可以解决K-means算法极易受到初始聚类中心影响和聚类结果不稳定的缺点.

2）使用AP算法提升KM-ALPSO算法的运算效率；利用Halton序列初始化粒子群位置，提升KM-ALPSO算法全局搜索能力.在客票数据分析中，AP算法从10 000条数据中识别出了1 331个样本代表点；基于Halton序列的KM-ALPSO算法的聚类评价指标SC为0.332，DB为0.933，CH为708.5，在聚类数目为5时，性能最佳，且优于其他算法.5类旅客在年龄、提前购票时间和出发时间段上差异明显，说明了算法的有效性.

3）引入FP-Growth算法挖掘强关联规则.在客票数据分析中，通过箱线图对聚类特征进行离散化分级处理，揭示3个特征和旅客类别之间的关联性.在设置最小支持度为0.15，最小置信度为0.3的条件下，提取出9条强关联规则，提升度均大于1，结果可靠，能够为铁路部门推断旅客类别提供理论支持.

本文分析的数据集包含的OD数量多（既有京沪高铁本线的客流OD，也有跨线的客流OD），而且

聚类特征变量的选取含有局限性.首先需探讨聚类特征变量的选取原则,将其作为聚类分析的前置步骤.其次,利用基于Halton序列的KM-ALPSO算法,在选取合适的聚类特征变量基础上,对单一OD对进行市场细分,从而为动态定价和席位分配奠定基础,将是进一步研究的主要内容.

参考文献(References):

[1] 单杏花.铁路客运收益管理模型及应用研究[D].北京:中国铁道科学研究院,2012.
SHAN Xinghua. Study on railway revenue management model and its application in passenger transportation[D]. Beijing: China Academy of Railway Sciences, 2012.(in Chinese)

[2] MOHIUDDIN H, FITCH-POLSE D T, HANDY S L. Examining market segmentation to increase bike-share use and enhance equity: The case of the greater Sacramento region[J]. Transport Policy, 2024, 145: 279－290.

[3] 尹溪琛.基于旅客画像和出行链模型的交通出行推荐方法及应用[D].北京:北京交通大学,2021.
YIN Xichen. Travel recommendation method and application based on passenger profile and trip chain model[D]. Beijing: Beijing Jiaotong University, 2021.(in Chinese)

[4] 侯现耀,陈学武.基于态度的公交出行信息使用市场细分[J].吉林大学学报(工学版),2018,48(1):98－104.
HOU Xianyao, CHEN Xuewu. Use of public transit information market segmentation based onattitudinal factors[J]. Journal of Jilin University (Engineering and Technology Edition), 2018, 48(1): 98－104.(in Chinese)

[5] 程龙,陈学武,杨硕,等.基于市场细分的低收入通勤者公交出行改善对策[J].交通运输系统工程与信息,2016,16(3):8－14.
CHENG Long, CHEN Xuewu, YANG Shuo, et al. Public transit improvement measures of low-income commuters based on market segmentation [J]. Journal of Transportation Systems Engineering and Information Technology, 2016, 16(3): 8－14.(in Chinese)

[6] JIA W Z, TAN Y Y, LIU L, et al. Hierarchical prediction based on two-level Gaussian mixture model clustering for bike-sharing system [J]. Knowledge-Based Systems, 2019, 178: 84－97.

[7] JIAO L M, DENŒUX T, LIU Z G, et al. EGMM: An evidential version of the Gaussian mixture model for clustering[J]. Applied Soft Computing, 2022, 129: 109619.

[8] MASSON M H, DENŒUX T. ECM: An evidential version of the fuzzy c-means algorithm[J]. Pattern Recognition, 2008, 41(4): 1384－1397.

[9] 俞诚成,吕红霞.基于大数据铁路旅客画像的客运营销策略研究[J].铁道学报,2020,42(8):23－28.
YU Chengcheng, LÜ Hongxia. Research on passenger transport marketing strategy based on big data railway passenger portrait[J]. Journal of the China Railway Society, 2020, 42(8): 23－28.(in Chinese)

[10] TAO X R, PAN Q K, SANG H Y, et al. A multiobjective optimizer with a K-means cluster algorithm for a distributed flexible flowshop rescheduling problem [J]. Applied Soft Computing, 2024, 166: 112124.

[11] LI Y, CHU X Q, TIAN D, et al. Customer segmentation using K-means clustering and the adaptive particle swarm optimization algorithm[J]. Applied Soft Computing, 2021, 113: 107924.

[12] JAIN M, SAIHJPAL V, SINGH N, et al. An overview of variants and advancements of PSO algorithm [J]. Applied Sciences, 2022, 12(17): 8392.

[13] WEERASINGHE G, CHI H M, CAO Y Z. Particle swarm optimization simulation *via* optimal Halton sequences [J]. Procedia Computer Science, 2016, 80: 772－781.

[14] FREY B J, DUECK D. Clustering by passing messages between data points [J]. Science, 2007, 315(5814): 972－976.

[15] 吕红霞,王文宪,蒲松,等.基于聚类分析的铁路出行旅客类别划分[J].交通运输系统工程与信息,2016,16(1):129－134.
LV Hongxia, WANG Wenxian, PU Song, et al. Classification of railway passengers based on cluster analysis [J]. Journal of Transportation Systems Engineering and Information Technology, 2016, 16(1): 129－134.(in Chinese)

[16] 陈权超.基于大数据的高速铁路客流分析[D].成都:西南交通大学,2020.
CHEN Quanchao. Research on passenger flow analysis of High Speed Railway based on big data[D]. Chengdu: Southwest Jiaotong University, 2020.(in Chinese)

[17] 方腾源,徐峰祥,朱其茂,等.时空约束下的道路交通事故与违法行为关联规则研究[J].交通运输系统工程与信息,2025,25(3):246－254.
FANG Tengyuan, XU Fengxiang, ZHU Qimao, et al. Association rules between urban road traffic accidents and violations considering temporal and spatial constraints[J]. Journal of Transportation Systems Engineering and Information Technology, 2025, 25(3): 246－254.(in Chinese)

[18] SUN J, WU X J, PALADE V, et al. Random drift particle swarm optimization algorithm: Convergence analysis and parameter selection [J]. Machine Learning, 2015, 101(1): 345－376.

[19] 李海杰，苗蕾，聂磊，等. 基于关联规则和主成分分析的高铁旅客购票行为特征研究[J]. 铁道科学与工程学报，2023，20(6)：2013－2025.
LI Haijie，MIAO Lei，NIE Lei，et al. High-speed railway passenger ticketing behavior characteristics based on association rules and principal component analysis[J]. Journal of Railway Science and Engineering，2023，20(6)：2013－2025.(in Chinese)

[20] LIN K C，LIAO I E，CHEN Z S. An improved frequent pattern growth method for mining association rules[J]. Expert Systems with Applications，2011，38(5)：5154－5161.

[21] 乔珂，赵鹏，文佳星. 基于潜在类别模型的高铁旅客市场细分[J]. 交通运输系统工程与信息，2017，17(2)：28－34.
QIAO Ke，ZHAO Peng，WEN Jiaxing. Passenger market segmentation of high-speed railway based on latent class model[J]. Journal of Transportation Systems Engineering and Information Technology，2017，17(2)：28－34.(in Chinese)

[22] 冯晓磊. 近邻传播聚类算法研究[D]. 郑州：解放军信息工程大学，2011.
FENG Xiaolei. Research on affinity propagation clustering algorithm [D]. Zhengzhou：People's Liberation Army Information Engineering University，2011.(in Chinese)