文章编号：2095-2163(2021)04-0039-06 中图分类号：U491.12 文献标志码：A

城市轨道交通客流时段 OD 对挖掘及快慢车停站方案确定算法研究

杨恺鹤，丁小兵，刘志钢，陈家萍，万 苏
（上海工程技术大学 城市轨道交通学院，上海 201620）

摘 要：本文首先引入大数据理论，构建具备补充和修正功能的时段聚集机理数据挖掘算法，随后建立 $Time\_cluster_k^{m,n}$ 聚类算法挖掘客流峰值时段划分，从行车调度指挥角度研究客流时段聚集规律；其次，根据客流聚集的时段规律挖掘，以 AFC 客流数据为支撑挖掘基于最小支持度 minsupport 的客流 OD（Origin- Destination），以优化轨道交通市郊线路的行车组织方案；最后，通过算例分析验证了 $Time\_cluster_k^{m,n}$ 模型优越于目前经验法确定快慢车开行比例。

关键词：数据挖掘算法；时段划分；行车组织；快慢车开行比例

Research on Mining Algorithm of Time Aggregation and OD Distribution of Urban Rail Transit Passenger Flow

YANG Kaihe, DING Xiaobing, LIU Zhigang, CHEN Jiaping, WAN Su
(School of Urban Rail Transportation, Shanghai University of Engineering Science, Shanghai 201620, China)

【Abstract】 Firstly, the big data theory is introduced to construct the mining model of temporal aggregation mechanism with supplement and correction function, then the clustering algorithm $Time\_cluster_k^{m,n}$ is used to mine the peak time interval of passenger flow, and the passenger flow time aggregation rule is studied from the angle of traffic dispatching command. Secondly, according to the law of mining traffic aggregation, passenger flow calculation can be determined by the time of train lines in the suburbs of vehicle speed ratio, to match the time period of the uneven distribution of passenger flow. Finally, an example is used to prove the superiority of model in determining train ratios with the experience method.

【Key words】 Data mining algorithm; Time division; Traffic organization; The proportion of express and local train

0 引 言

中国轨道交通建设和运营高速发展，在运营管理方面积累了经验，也带来了一定的问题：城市区域内交通通行拥堵状况加重，城市的部分功能开始转移向郊区，市郊区间的时段客流明显增加，市郊线路、地铁客流的出行规律产生了明显改变，且预测难度较大，增加了车站客流组织的难度，若车站站台的滞留客流过大，还可能引发严重的安全问题。本文通过大数据理论分析，结合先进的数据挖掘技术研究轨道交通客流在车站的聚集规律，可以及时获取不同规模车站，不同手段的客流出行特征，有助于运营企业及时调整行车计划，为日常行车调度指挥提供决策依据。

在国内外研究地铁客流出行分布所构建的模型中，以重力模型和增长系数法用得较多，奠定了客流研究的基础。1940 年，Stouffer 最早提出了介入机会模型；1955 年，Casey 提出了重力模型，后来该模型在最大熵原理和最大似然原理方面均得 NT 解释；1965 年，Fumess 提出了著名的增长系数法。随后，美国在交通规划中开始使用线性回归模型来预测出行发生和吸引量；1960 年末，英国专家首先提出一种交叉分析法的客流研究算法，基于此 Gordon W.S 提出了不以家庭为出行单位的研究模型，研究了客流的分布于聚集模型[1]；Richard C.M.Yam 建立专门模型来研究专有客流的出行特征[2]；J.L.Bowman 等以出行者的活动安排为切入点建立模型，来预测出行的产生[3]。这些国外学者在客流量预测和分布的模型上作了较为深入的研究，奠定了理论基础。国内近几年在该方面作了较多值得借鉴作用的研

作者简介：杨恺鹤（1993-），女，硕士研究生，主要研究方向：轨道交通行车组织优化；丁小兵（1982-），男，博士，讲师，主要研究方向：轨道交通运营组织优化；刘志钢（1974-），男，博士，教授，主要研究方向：轨道交通安全与优化；陈家萍（1996-），女，硕士研究生，主要研究方向：轨道交通运营安全及应急处置；万 苏（1997-），女，硕士研究生，主要研究方向：轨道交通延误传播机理研究。
通讯作者：丁小兵 Email：dxbsuda@163.com

收稿日期：2020-11-04

究:杨晓光构建了研究客流分布滞后的预测模型,并兼顾了其他关键因素的影响,所建模型更加完善,计算结果也更加精确[4];姚恩建指出非集计模型在理论上的不足,提出了一种集计数据的客流 OD 分布模型,并进行了实例验证[5]。在 OD 分布及路径选择的联合模型上,由于缺少关于目的地选择的调查结果,在参数标定上采用外推方式,致使两阶段分开,影响了模型的预测精度。随后,赵鹏提出了基于状态空间方法的短时客流 OD 估计模型,预测结果的精确度得到较为理想的提高[6];陈小鸿以客流 OD 矩阵为出发点,建立了以广义最小二乘理论为基础的 OD 矩阵估计模型,并采用递增拉格朗日算法给出了相应的求解过程,但反推结果的精度不太理想[7]。

上述研究成果在海量客流数据预处理方面均未深入研究,对数据字段值缺失等未形成补充机制;对地铁客流积累的数据的应用尚不够充分,还有很多方面可以深入挖掘。若能对客流数据进行精细化预处理和缺失补充,探索乘客的时段聚集实时状态,为轨道交通市郊线路行车组织方案优化提供数据支持。本文拟从运营企业获取的 AFC 进出站客流数据为支撑,研究基于乘客出行交通卡卡号及出行时段的信息跟踪,并结合相关时段参数数据,计算乘客的走行时间,从而构建模型挖掘乘客的聚集机理,以聚集机理为支撑获取乘客的出行时段特征及时段分布,以此指挥轨道交通的行车组织,列车运行计划的编制及临时调整等,亦可为客流预测及客流应急疏导模型的补充等提供理论及方法参考。

1 城市轨道交通客流数据预处理算法

轨道交通运营企业在日常运营中积累了大量客流数据,该客流数据包含了刷卡卡号、进出站时间、进出站车站编号等重要数据。而 Oracle 数据库容量较大,且早期在建设数据库时,字段、存储过程、数据规则、范式等设计合理性欠佳等原因,数据库中的客流数据字段及其存储值有冗余甚至错误,不仅浪费存储空间,还降低了数据挖掘与分析的效率和稳定性。所以对客流数据的异常识别很有必要,本文研究的异常数据包括:按字段填充缺失的数据、冗余的数据、错误数据等。

采集的原始数据通常存在冗余、字段值不完整、数据内容不规范等问题,不能直接用于数据挖掘,需要设计数据预处理算法。但由于数据量大,预处理的效率较低,若算法设计不完善,则会出现冗余计算,甚至错误计算,导致处理性能和计算结果的准确度降低。本文拟从数据清洗算法角度进行优化,从而提高清洗运算效率,主原始数据预处理设计过程如图 1 所示。

在轨道交通日常运营维护等积累的数据中,有的是手工填写、有的是根据系统前台运行采集并经过转换自动填写而成,数据不完整,字段值缺失等是极有可能存在的。故本文提出一种改进的朴素贝叶斯分类算法,用于对重要数据字段缺失的填充,进而数据的有效性能得到保证,基本流程如图 2 所示。

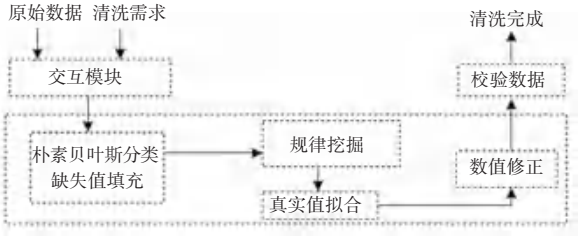

图 1 原始数据预处理算法设计

Fig. 1 The flow chart of data cleaning and pretreatment

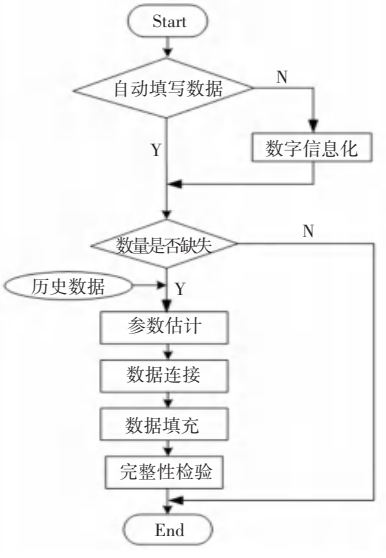

图 2 缺失值填充流程图

Fig. 2 The flow chart of filling missing value

在对缺失值的参数估计方面,通过式(1)来计算各缺失字段值的可能概率,其中 $P(X)$ 的取值在各个可能取值状态下都可视为常数。

$$P(C_i \mid X) = \frac{P(X \mid C_i) \times P(C_i)}{P(X)}. \qquad (1)$$

在该状态下计算出 $P(X \mid C_i) \times P(C_i)$,即可完成填充缺失值。若缺失值的先验概率不可提前预知的情况下,可将其预先设为等概率发生,所以 $P(X \mid C_i)$ 可由式(2)变形求得:

$$P(X \mid C_i) = P(X_1 \mid C_i) \times P(X_2 \mid C_i) \cdots P(X_n \mid C_i). \qquad (2)$$

通过分析和推导，缺失值的参数估计算法可以转化为求所缺失字段的取值概率 $P(X \mid C_i)$。当研究数据的样本达到标准时，其概率近似等价于发生的频率，可用数据字段完整且数据值符合规则的记录取值出现频率来估计缺失值的发生概率 $P(X_n \mid C_i)$。

式(2)在计算含有缺失值的记录值时，因数据库范式的依赖性关系，主码的取值很大程度上决定了各待填充值的概率，由于 MapReduce 算法在 Map 阶段和 Reduce 阶段将处理时间切割成小碎片时，其实际一次仅能处理一条实体记录，基于此必须将依赖属性取值和其条件概率取值关联起来，以期求得的值更加准确。

数据的填充模块主要通过 MapReduce 算法实施，通过对比数据连接模块运算结果与原始输入数据的偏差值进行连接模块运算的输入，Map 阶段的计算和连接模块一致；而在 Reduce 阶段，则利用式(2)计算出每个 C_i 对应的条件概率，选择其中 $P(C_i \mid X)$ 概率最大的 C_i 作缺失值填充。

2　客流时段聚集机理挖掘模型 $Time\_cluster_k^{m,n}$ 的构建

当数据库中的数据实体量或字段值发生变化后，原来的模糊频繁属性集则可能发生变化而不再是频繁属性集。针对模糊频繁属性集集合、负边界的定义及计算方法，计算出模糊属性集在原始数据库中的最小模糊支持度，本文提出了一种动态模糊频繁属性集的负边界的计算方法，并将该方法用于轨道交通时段客流量聚集度的挖掘。

2.1　轨道交通客流时段聚集规律挖掘

将轨道交通运营的时段进行模糊离散化，令时间序列 $S=(x_1,x_2,x_3\cdots x_n)$，取宽度为 w 的时间窗作用于 S 形成的子序列 $S_i=(x_1,x_2\cdots x_{i+w-1})$，单步滑移形成一系列宽度为 w 的子序列 $x_1,x_2,x_3,\cdots,x_{n-w+1}$，式(3)：

$$W(s,w)=\{S_i \mid i=1,2,\cdots n-w+1\}. \quad (3)$$

将 $W(s,w)$ 看作 w 维欧氏空间中的 $n-w+1$ 个点，随机地分到 k 类中，计算每类中心作为每类的代表，计算集合 $W(s,w)$ 的元素 $S_i(i=1,2,\cdots n-w+1)$，第 j 类代表的隶属度属性函数 $u_j(S_i)$ 为式(4)：

$$u_j(S_i)=\frac{(\frac{1}{\|s_i-x_j\|^2})^{\frac{1}{b-1}}}{\sum_{c=1}^{k}(\frac{1}{\|s_i-x_c\|})^{\frac{1}{b-1}}} j=1,2\cdots,k;b>1. \quad (4)$$

其中：$b>1$，b 为常数，用于控制聚类的模糊度，$\|s_i-x_j\|^2$ 为第 i 个点到第 j 类点距离绝对值的平方。

如果支持度满足 A 条件，则在时间 T 范围内 B 发生，即 $A \xrightarrow{T} B$；$A,B \in \{x_1,x_2,\cdots x_k\}$，需要确定 A 发生的频率数为式(5)：

$$F(A)=\sum_{i=1}^{n-w+1} u_A(S_i). \quad (5)$$

其中：$u_A(S_i)$ 为 S_i 点属于第 A 个代表属性的隶属度。

关联规则 $A \xrightarrow{T} B$ 的支持度 c 为式(6)：

$$c(A \xrightarrow{T} B)=\frac{F(A,B,T)}{F(A)}. \quad (6)$$

其中：$F(A,B,T)$ 表示 A 发生的条件下，T 时间内 B 事件发生的概率(频率数)，主要计算过程如 $A \xrightarrow{T} B$ 的 $C-means$ 方法，式(7)所示：

$$C(B_T:A)=p(A)\times\{p(B_T \mid A)log\frac{p(B_T \mid A)}{p(B_T)}+[1-p(B_T \mid A)]log\frac{1-p(B_T \mid A)}{1-p(B_T)}\}. \quad (7)$$

其中：$p(A)$ 表示满足 A 规则情况的发生概率；$p(B_T \mid A)$ 表示 A 事件发生的情况下，在时间 T 内 B 事件发生的概率；式(7)右侧表示从先验概率 $p(B_T)$ 到后验概率 $p(B_T \mid A)$ 的信息传递及获取过程。

2.2　基于关联规则支持度的时段客流 OD 对挖掘

经过数据清洗所得轨道交通车站 AFC 客流数据已经较为精准，符合数据挖掘的基本要求，拟构建 $Time\_cluster_k^{m,n}$ 算法挖掘轨道交通车站时段客流的聚集关联规则，时段客流量的聚集机理及以交通卡卡号为主码的客流 OD 对的最小支持度是挖掘的主要对象，本文构建的 $Time\_cluster_k^{m,n}$ 算法如下：

Step 1　首先计算 $u_A(S_i)$，根据轨道交通客流时段聚集支持度进而计算 $c(A \xrightarrow{T} B)$，并设好 $minsupport=c(A \xrightarrow{T} B)$；

Step 2　将清洗和预处理后的数据作为客流聚集关联规则挖掘初始数据，根据预先设定的条件对 AFC 客流数据库进行初始化，扫描事务数据库 T_{ID}，从事务库中遍历搜索出所有的项集长度为 $k=l$ 的项集，形成候选 1 项集 C_1，将 C_1 代入式(5)、(7)，计算每项的支持度，依次与最小支持度比较，支持度大于 minsupport 的则形成频繁 1 项集 L_1；

Step 3 以频繁 k 项集为基础产生候选($k+1$)项集 C_{k+1},如果 C_{k+1} 为空集,则进入下一步,否则继续循环;

Step 4 遍历客流数据库所有记录,计算各候选项集的基于时段聚集的客流支持度;

Step 5 将候选项中支持度计算值小于 minsupport 的移出,形成($k+1$)项频繁项集 L_{k+1};

Step 6 $k=k+1$,转入 Step 3;

Step 7 由此得出分时段客流聚集量的频繁项集关联规则;

Step 8 输出以 2 h 时粒度的分时段客流聚集量及 OD 对结果,以指导轨道交通运营计划的编制与调整。

3 算例分析

3.1 客流数据清洗与缺失补充

本文客流数据来源于上海轨道交通 9 号线 AFC 闸机,客流时段覆盖 6:00-22:00,数据库表数据字段覆盖交通卡卡号、进站时间、进站编号、出站编号、出站时间、是否使用了 P+R 停车场。所构建的算法的参量需要对原始数据字段编程计算得到,如乘客的总在途时间等,但数据库存储的 AFC 客流数据,会出现乘客因逃票、漏刷、卡丢失等原因,导致数据缺失、无效等而无法使用,部分原始数据见表 1。

由原始数据可知,部分数据出现字段缺失,尤其是必填字段,运用数据清洗及填充方法,处理原始数据,以方便数据挖掘。

为填充原始数据的丢失值,将式(2)代入式(1)得式(8):

$$P(Ci \mid X)=\frac{P(X \mid C_i)=P(X_1 \mid C_i)\times P(X_2 \mid C_i)\cdots P(X_n \mid C_i)\times P(C_i)}{P(X)}. \tag{8}$$

以 AFC 数据的卡号字段值为依据,运用 SQL 语句检索统计每位乘客的出行次数及 OD 百分比,代入式(8),并通过参数估计、连接、填充等模块,补充原始数据,得表 2。

表 1 上海轨道交通 9 号线 AFC 客流数据(原始)

Tab. 1 The AFC passenger flow data of Shanghai Metro line 9(Original)

卡号	进站时间	进站站名	出站时间	出站站名	P+R
U12452172127	09:12	九亭	09:43	徐家汇	
U07452172874	05:58	松江大学城	06:49	肇嘉浜路	是
U31842254712	06:58	松江大学城	07:28	宜山路	是
U12452172127	08:31	九亭	09:03	徐家汇	
U07452172874	06:21	松江大学城	07:23	肇嘉浜路	是
U31842254712	06:24	松江大学城	07:28	宜山路	是
U07452172874	16:46	肇嘉浜路	17:51	松江大学城	是
U07452172874	17:02	徐家汇			
……	……	……	……	……	……

表 2 上海轨道交通 9 号线 AFC 客流数据表(补充修正后)

Tab. 2 The AFC passenger flow data of Shanghai Metro line 9(supplementary and correction)

卡号	进站时间	进站站名	出站时间	出站站名	P+R	总时间/min
U12452172127	09:12	九亭	09:43	徐家汇	否	31
U07452172874	05:58	松江大学城	06:49	肇嘉浜路	是	51
U31842254712	06:58	松江大学城	07:28	宜山路	否	30
U12452172127	08:31	九亭	09:03	徐家汇	否	32
U07452172874	06:21	松江大学城	07:10	肇嘉浜路	是	41
U31842254712	06:24	松江大学城	07:28	宜山路	是	30
U07452172874	16:46	肇嘉浜路	17:32	松江大学城	是	46
U07452172874	17:02	徐家汇	17:48	松江大学城	是	30
……	……	……	……	……	……	……

3.2　客流时段聚集量挖掘

根据时间序列离散化设计：将客流时段按照 2 h 粒度划分，并单步滑移形成一系列宽度为 2 的时间序列 $S=(x_1,x_2,x_3\cdots x_n)$。计算 $W(s,w)=\{S_i \mid i=1,2,\cdots n-w+1\}$，以及隶属度函数，客流时段覆盖 6:00-22:00，以 2 h 粒度划分，进而得到每个 2 h 粒度的车站客流聚集隶属度，见表 3。

用挖掘算法可计算出上海轨道交通 9 号线部分车站客流聚集量：世纪大道、陆家浜路、马当路、肇嘉浜路、徐家汇、宜山路、桂林路、七宝、九亭、佘山、松江大学城的 2 h 粒度时段客流量，各站分别对应 St_1-St_{11}，计算结果见表 4。

表 3　上海轨道交通 9 号线松江大学城站 2 h 粒度客流特征值

Tab. 3　Characteristic values of 2 h particle size of da xue cheng station

时段序列	类目		
	隶属度 $u_j(S_i)$	支持度计算 $C(A-B)$	客流量
6:00-8:00	0.85	0.90	47 804
8:01-10:00	0.74	0.86	31 012
10:01-12:00	0.91	0.84	11 430
12:01-14:00	0.90	0.80	9 010
14:01-16:00	0.89	0.84	13 014
16:01-18:00	0.82	0.81	30 254
18:01-20:00	0.69	0.83	30 147
20:01-22:00	0.92	0.84	29 874

表 4　车站分时段客流聚集量

Tab. 4　The aggregation passengers' flow based on distribution time

时段	车站										
	St1	St2	St3	St4	St5	St6	St7	St8	St9	St10	St11
6:01-8:00	7 031	6 031	9 214	15 274	22 215	14721	16 014	8 721	10 721	30 264	28 214
8:01-10:00	12 031	13 031	8 610	11 201	20 203	6 321	11 782	6 321	9 843	19 721	18 734
10:01-12:00	10 574	7 574	7 141	9 845	6 201	5014	9 487	7 034	7 123	10 351	16 321
12:01-14:00	9 414	3 414	6 573	9 216	5 845	4782	5 014	6 791	8 787	9 014	10 024
14:01-16:00	9 018	10 034	5 362	8 142	9 216	3587	6 571	8 482	5 591	7 782	9 789
16:01-18:00	8 124	8 124	5 217	9 345	13 201	8721	20 362	6 217	10 258	9 587	8 137
18:01-20:00	10 250	22 230	6 057	9 714	9 231	6321	18 217	9 354	9 239	8 214	10 354
20:01-22:00	9 217	13 217	5 325	8 521	7 227	5014	9 057	8 242	8 427	9 241	9 792

线路分时段数据变化趋势如图 3 所示，可较为直观的显示沿线车站在 2 h 粒度区间的客流量分布情况，该客流分布情况可直接指导运营产生实践。

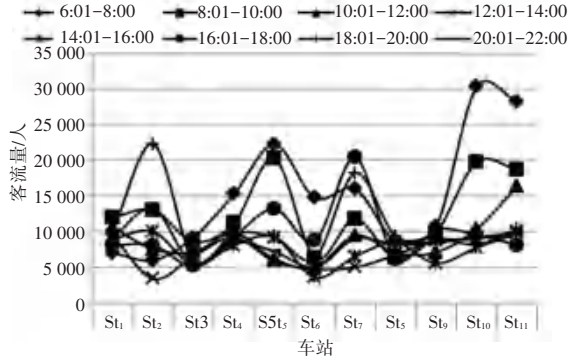

图 3　车站分时段客流聚集量

Fig. 3　The aggregation passengers' flow based on distribution time

表 4 的挖掘结果可为乘客出行高峰期，轨道交通车站制定行车调整方案提供数据支撑，使得列车运能与客流相匹配，快速疏散乘客至路网中。亦可为轨道交通市郊线路快慢车开行比例的确定提供数据支持，从而节省能耗和乘客的出行时间。

3.3　基于关联规则的客流 OD 对挖掘

计算每张交通卡号的 OD 对的 $c(A \xrightarrow{T} B)$，并与 $minsupport = c(A \xrightarrow{T} B)$ 对比，高于该值的确定为 OD 对，挖掘结果 见表 5。该挖掘结果可为快慢车模式下，快车停站方案的确定提供精准的决策支持。

表 5　乘客出行 OD 对规律挖掘（上行方向）

Tab. 5　The mining of passengers' travel OD law (upward direction)

卡号	O	D	P+R
U12452172127	九亭	徐家汇	否
U07452172874	松江大学城	肇嘉浜路	是
U31842254712	松江大学城	宜山路	是
U01552172876	松江新城	七宝	是
U07452172874	松江新城	宜山路	是
U12452376572	松江体育中心	世纪大道	否
U06571172546	洞泾	世纪大道	否
U318745684324	佘山	徐家汇	否
U01555476158	七宝	世纪大道	否
U15478475786	九亭	世纪大道	否
……	……	……	……

由表 5 可知，若能结合乘客的出行时段来挖掘 OD 对，则可为快慢车开行方案的优化提供更加精细化的决策支持。

4　结束语

特大城市的轨道交通市郊线路客流具有比较明

（下转第 48 页）