文章编号:1671-8879(2018)05-0205-08

基于快速峰值聚类的高速公路异常事件识别方法

赵怀鑫[1,2],张英杰[1],邓然然[1],丁明航[1],孙朝云[1],李 伟[1]
(1. 长安大学 信息工程学院,陕西 西安 710064; 2. 陕西省交通运输厅,陕西 西安 710075)

摘 要:为准确全面感知高速公路交通运行状况,根据高速公路海量收费数据,提出一种高速公路通行异常事件识别的数据挖掘方法。首先,选取贵州省 2017 年 1 月的高速公路收费数据,筛选指定的进站、出站数据并去除多余字段,利用车辆进入和驶出收费站时间计算其在该路段的通行时长。然后,使用快速峰值聚类算法对通行时长和车辆总重进行聚类分析,计算数据间欧式距离,将此距离矩阵作为算法输入,计算各数据点的局部密度 ρ 及与密度更高点的距离 δ 两项指标;这两项指标均以较高的点为聚类中心,进而对非中心点进行分类及优化,输出聚类结果;聚类结果中除被分为若干类的正常数据外,还存在一些数据点明显异于大部分正常数据的噪声点,即异常数据,对这些异常数据进行具体分析。接着,采用孤立点检测法对筛选出的数据进行清洗处理,提取异常数据,检测出通行时间过长、过短及车辆总重过高、过低等异常事件。最后,将孤立点检测法得到的异常数据与快速峰值聚类算法的异常数据进行对比。研究结果表明:快速峰值聚类识别异常事件的准确率高于孤立点检测法约 20%,验证了提出算法的有效性和准确性;提出的算法能有效准确识别收费数据中隐藏的公路拥堵、长时间停留、疑似逃费和网络设备故障等异常事件,进而为高速公路运营服务和管理决策提供数据支持。

关键词:交通信息与控制工程;智能交通;异常事件分析;快速峰值聚类;孤立点检测;高速公路收费数据;数据挖掘
中图分类号:U491;TP301.6 **文献标志码**:A

Expressway anomaly event recognition method based on clustering by fast search and find of density peaks

ZHAO Huai-xin[1,2], ZHANG Ying-jie[1], DENG Ran-ran[1], DING Ming-hang[1],
SUN Zhao-yun[1], LI Wei[1]
(1. School of Information Engineering, Chang'an University, Xi'an 710064, Shaanxi, China;
2. Department of Transport of Shaanxi Province, Xi'an 710075, Shaanxi, China)

Abstract: To sense the expressway traffic operation-status more accurately and comprehensively, a data mining method for identifying abnormal traffic events on an expressway using mass data collection was proposed. First, fee data from January 2017 were selected from the massive data available for the Guizhou Expressway toll. The data on the specific entrance and exit stations were selected, and some redundant fields were deleted, with those data only related to this study being retained. The time for driving into the entrance station and driving out of the exit station

收稿日期:2018-06-20
基金项目:陕西省自然科学基础研究计划项目(2017JQ5014);陕西省交通运输科研项目(15-45r)
作者简介:赵怀鑫(1975-),男,安徽凤阳人,陕西省交通运输厅教授级高级工程师,长安大学工学博士研究生,E-mail:zhaohxin@vip.sina.com。
通讯作者:孙朝云(1962-),女,安徽太和人,教授,博士研究生导师,E-mail:zhaoyunsun@126.com。

was used to calculate the vehicle staying time between the two toll stations. The selected data were analyzed based on the driving time and axle weight using a fast peak clustering algorithm. The distance between each data point was calculated, and the distance matrix was used as the input of the algorithm. The local density of each data point and the distance between the points with higher density were calculated. In addition, the cluster centers were selected based on the principle that the two indicators were higher. The non-central points were classified and optimized, and the clustering result was then outputted. The normal data of clustering results were divided into several categories, and there exists some noise whose data points were significantly different from most of the normal data. A specific analysis was conducted for these abnormal data. An outlier detection algorithm was then used to process the original data, the cleaned abnormal data were extracted, and abnormal events such as excessive transit time, a short transit time, and a high load were detected. Finally, the anomalies in the data obtained using the isolated point detection method were compared with the anomalies in the data of the fast peak clustering algorithm. The results show that the accuracy of fast peak clustering used to identify anomalous events is higher than that of the isolated point detection method by nearly 20%, which verifies the validity and accuracy of the proposed algorithm. The method proposed in this paper can effectively and accurately identify hidden traffic jams such as road congestion, long stays, exit charges, and network equipment failure in the charging data, and provide theoretical support for operational services and management decisions for practical applications of an expressway. 3 tabs, 6 figs, 25 refs.

Key words: traffic information and control engineering; intelligent transportation; anomaly event analysis; clustering by fast search and find of density peak; outlier detection; expressway fee data; data mining

0 引 言

随着中国高速公路网逐步健全和经济的快速发展,高速公路车流量也随之剧增,交通拥堵、交通安全、服务水平降低等频繁发生[1]。高速公路通行费缴纳产生了海量原始收费数据,深入挖掘其信息帮助管理决策成为当前亟待解决的问题[2]。高速公路收费数据的挖掘应用对感知高速公路真实运行状态,提升高速公路运营管理和决策水平具有重要意义。挖掘评估的可靠性取决于数据本身质量,由于种种原因,在原始收费数据中存在着数据缺失、数据错误和数值重大偏差等情况。这些异常数据往往都对应着不同的异常事件,此类事件能够标识出高速公路实际运行状态[3],若忽略这些异常事件的影响,就不能及时发现交通拥堵等问题,更不能为管理部门提供正确的决策支持及科学引导公众出行[4-5]。

高速公路通行异常事件包括交通事故和交通事件两大类[6-7]。交通事故是指车辆撞车、撞人、撞公路设施、翻车等造成人身伤害及车辆或设施损坏的交通异常状况,发生交通事故的路段,轻则局部交通拥堵,重则造成双向交通彻底堵塞或中断;交通事件是指车辆故障、长时停车[8]、车辆逆行、交通瓶颈(由于车速陡减、入口匝道、车辆拥挤等原因)、移动瓶颈(大型低速车辆驶入等原因)、系统故障、设备故障、通行费偷逃等异常情况。

高速公路通行异常事件检测,一般是通过人工巡查或道路监控系统事件检测器来识别。通常的交通异常事件检测方法主要有统计分析、分类、PCA(主成分分析)降维等。吴晓佩提出的基于多 SVM(支持向量机)分类器融合的高速公路异常事件检测方法,可以有效提高高速公路异常事件检测的准确性和可靠性[9]。隋靓等提出一种基于车辆运动轨迹的异常事件挖掘算法,这种异常事件检测模型能够有效检测逆行、违法停车等异常车辆信息[10]。巨永锋等通过对数据融合基本理论的总结与剖析,提出用于交通异常事件检测的数据融合系统模型[11]。应用该数据融合技术的交通事件检测系统能提高系统的有效性,得到最佳协同作用结果。赵勇等提出一种基于 RFID(射频识别)技术和控制器局域网的预警系统模型和具体架构,分析了系统的工作原理

及关键技术，解决了高速公路交通异常事件检测的实时性和可靠性问题[12]。孙静怡等针对高速公路意外事件所导致的异常状态，提出了基于支持向量机的高速公路异常状态的检测方案。应用 VISSIM 仿真软件建立了昆玉（昆明—玉溪）高速公路单向三车道基本路段的车辆抛锚事故模型，对不同条件下的异常事件进行仿真，并分析异常状态对高速公路断面通行能力的影响[13]。国外高速公路异常事件的识别与检测已成为机器视觉和模式识别领域的一个突出目标，并在过去的十年得到显著发展。Sadek 等提出基于流动梯度直方图和统计逻辑回归分析的新框架，对交通事故进行实时识别和检测，实际视频序列的试验结果证明了该框架的效率和适用性[14]。Sheu 提出一种新方法用于高速公路异常事件的实时检测和表征，该技术能够实时检测高速公路事故，并根据时间变化的车道变换和阻塞车道的长度以及事故持续时间等来表征事故[15]。Jin 等提出一种利用建设性概率神经网络（CPNN）进行高速公路事件检测的新技术，CPNN 采用了一种集群技术和自动化培训流程，与传统的概率神经网络相比，采用的网络修剪方法使模型尺寸减小了 11 倍；试验结果表明，CPNN 适用于不断变化的站点交通环境的事件检测问题[16]。近年来，随着数据融合、运动目标轨迹分析等技术的成熟，也出现了一些新的异常事件检测方法。但由于这些方法对数据分布要求过高，并且使用场景及范围有限，导致在实际应用中异常事件的检测效果不理想。当前数据挖掘技术[17]已日趋成熟，多种聚类算法[18-19]在医学、军事、建筑等领域都有较为广泛应用，在综合交通和高速公路收费数据分析方面也有一定应用[20-21]。本文高速公路异常数据分析的聚类算法，可更准确地检测并分析高速公路异常事件。通过对交通数据类型和维度的分析，选择使用一种基于密度的快速峰值聚类方法[22]，对车辆通行时长及车辆总重属性进行聚类，将聚类结果可视化输出，可明显看出噪声点，这些噪声点即为异常数据。然后利用孤立点检测算法[23-24]对此数据清洗，将清洗出来的孤立点数据与聚类算法所得异常数据对比分析，结果表明聚类算法较孤立点算法在异常事件挖掘方面更有效。

本文提出一种基于快速峰值聚类的高速公路数据处理算法，对高速公路收费数据中异常值分析挖掘，建立识别通行异常事件的算法模型，识别车辆超速、长时停留、车辆超载、设备损坏和网路故障等问题，进而识别出高速公路交通异常事件，为高速公路运营管理提供数据支持。本文主要给出数据聚类及可视化、数据清洗及结果分析。其中数据聚类使用的快速峰值聚类算法包括距离矩阵计算、聚类中心选择、其余各点分配及优化；数据清洗使用的孤立点检测算法是对原始收费数据进行清洗，并对处理结果进行对比分析，证实其准确性及有效性，从而有效评估高速公路实际通行状况，为运营管理提供决策支持。

1　收费数据筛选

选取贵州高速公路 2017 年 1 月收费数据字段，说明如表 1 所示，原始数据有百万余条。

表 1　贵州高速公路收费数据字段说明

Tab. 1　Description of fee data in Guizhou

数据字段	对应字段说明	数据示例
ID	系统编号	57 261 916
CardNo	卡号	52 011 328 220 200 196 796
ICCardNo	用户 IC 卡号	52 011 328 220 200 196 796
LastBalance	消费后余额	2 887.00
OutTime	出站时间	2017/1/31 00:00:00
OutLoad	出站荷载	35
OutStationName	出站名称	3
InTime	进站时间	2017/1/30 23:42:48
InLoad	进站荷载	35
InStationName	进站名称	2
VehiclePlate	车牌	贵 GP4820
VerificationCode	验证码	
CreateTime	创建时间	2017/1/31 14:58:47
SettlementTime	清分时间	
ShiftTime	更新账户余额单位时间	2017/1/31 14:58:47
State	状态	0
HandleTime	连续性校验时间	2017/1/30 00:00:00
CardType	卡类型	22
OutBusiNo	外部系统交易流水号	
ExitType	流水类型	
BitchNo	批次号	
VehicleType	车型	1
VehiclePlateColor	车牌颜色	0
TransferType	消费类型	0
TransferState	消费状态	0
TransferMoney	消费金额	9.50
TransferTime	消费时间	

在 2016 年 12 月的收费数据中筛选出进站编号为 569，出站编号为 570 的所有数据，仅保留与本文相关的部分字段（进出站点、进出站时间、车型、车辆总重等），用出站时间减去进站时间，得到车辆在这一路段的通行时长，将其加入原始数据中，最终得到

2 799 条数据。经量纲一化处理后的数据见图 1。

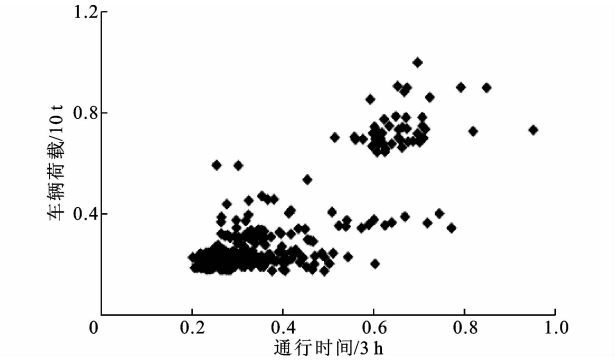

图 1 收费数据筛选结果

Fig. 1 Fee data processed results

2 收费数据异常事件分析

2.1 快速峰值聚类算法挖掘异常事件

快速峰值聚类算法选择聚类中心的原则是其具有较高局部密度且与高密度点的距离较大。

2.1.1 距离矩阵计算

快速峰值算法的输入是原始数据经计算得出的距离矩阵。直接利用原始数据计算距离会因不同属性间量纲的差异影响真实结果。为了消除属性间的量纲影响,在此使用 min-max 标准化方法对相关属性进行标准化处理,然后计算其欧氏距离,并将此距离按 2 个数据点序号及其对应的距离输出作为列数为 3 的矩阵,此距离矩阵即为快速峰值算法输入。

2.1.2 聚类中心选择

本文算法定义聚类中心需满足:局部密度 ρ 值大,与比其密度高的点距离 δ 值也较大。

其中,对于数据点 i,其局部密度 ρ_i 定义为

$$\rho_i = \sum_j \chi(d_{ij} - d_c) \tag{1}$$

式中:d_{ij} 为 i 点与 j 点之间的距离;d_c 为截止距离,由用户自己设定;$\chi(\cdot)$ 为二值函数,当 $d_{ij} - d_c < 0$,$\chi(\cdot)=1$,反之 $\chi(\cdot)=0$。

一般而言,d_c 选取原则为:将距离 δ_i 按递增方式进行排序并编号,找出所有距离个数的 1%~2% 所对应的序号,将 d_c 设置为该序号对应的距离值。本文根据实际设置 d_c 为距离个数的 2% 所对应序号的距离值。由式(1)可知,ρ_i 值基本等于与点 i 之间的距离小于 d_c 的点数。

点 i 与高密度点之间的距离 δ_i 定义为

$$\delta_i = \min_{j:\rho_j > \rho_i} (d_{ij}) \tag{2}$$

对于密度最高的点,取

$$\delta_i = \max_{j:\rho_j > \rho_i} (d_{ij}) \tag{3}$$

计算这 2 个量后,将所有点按 ρ 和 δ 作为 2 个维度进行可视化输出,输出的图形即为决策图。

以图 2 样本数据为例说明该算法聚类中心选择过程。图中各点以密度递减进行排序,1 号点密度最大,28 号点密度最小,其分布见图 2。将 28 个数据逐个按式(1)~式(3)分别计算出 ρ、δ 值,将每个数据点按计算所得的值映射在 ρ - δ 空间里并进行可视化输出,所得决策图见图 3。

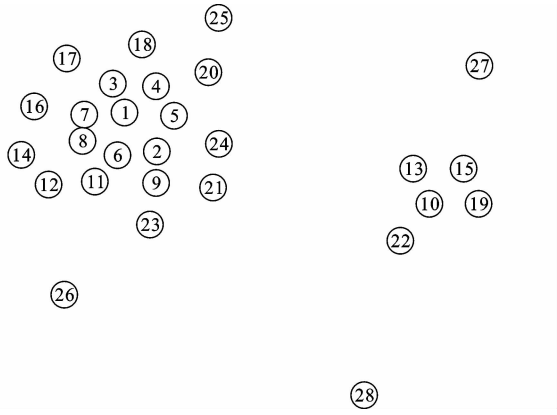

图 2 数据分布示意

Fig. 2 Sketch of data distributions

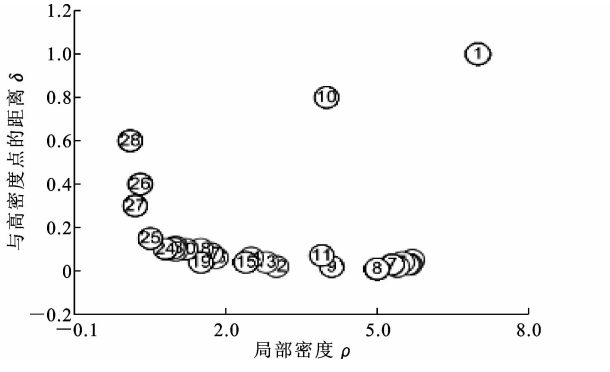

图 3 $\rho\delta$ 决策图

Fig. 3 $\rho\delta$ decision graph

由图 3 可知:1 号、10 号点有较高的 ρ、δ 值,因此这 2 个点被选作聚类中心,而且这是 2 个不同的类簇;26~28 号点有一个相对较高的 ρ 值和低 δ 值,但它们是孤立点,可看作是由单个点组成的集群,也就是异常值;9 号、10 号点的 ρ 值类似,但其 δ 值是截然不同的,而且 9 号点属于类 1,并且距离其很近同一类的点是其他几个 ρ 值相对较高的点,而 10 号点是另一个类的中心,因此,ρ 值和 δ 值都较高的点是聚类中心。本文局部密度是指某点截断距离内点的个数,与高密度点之间的距离表示相对距离的相似性度量。

2.1.3 非聚类中心点的分配及优化

在找到聚类中心后,每个剩余点将逐个被分给

与其距离最近的高密度点所在的类，且此操作以单步执行，直到把所有的点全部分配到对应的类为止。

一般聚类方法的优化都是以使目标函数在每次的迭代中达到最优为原则而实现的。在这些基于函数优化的方法中，函数的收敛值一般也被作为一种度量方式。本文算法中优化方法为：先为每个类找到一个边界区域，边界区域的定义是分配给某一类簇点距离与另一类簇点的距离小于截止距离 d_c；接着在每个类簇的边界区域内找出密度最高的点并标记其密度为 ρ_b，遍历类簇内各点，密度大于 ρ_b 的点为类簇内的点，反之则为噪声点。

2.1.4　收费数据聚类

对筛选后的贵州高速公路收费数据，计算各数据点之间距离，得到距离矩阵作为输入，计算各数据点的 ρ、δ 值。在 ρ-δ 空间里的决策图见图 4，所选聚类中心的 ρ、δ 值都较大。

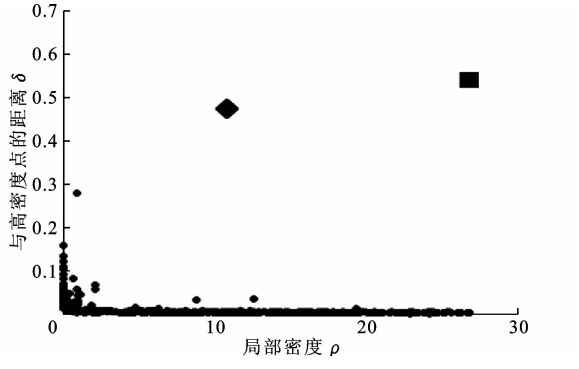

图 4　收费数据决策图

Fig. 4　Fee data decision graph

由图 4 可知，菱形点与方形点分别是由决策图选出的 2 个聚类中心，圆形点是待分配的非聚类中心点。然后根据输入的距离矩阵，将收费数据还原映射到二维空间，即可得到用该算法对收费数据进行聚类的可视化结果。

2.2　孤立点检测算法挖掘异常事件

孤立点是指那些较其他数据在格式、数值等细节都存在较大差异的数据。孤立点检测法能够有效地找出异常数据并将其剔除，通过孤立点检测法得到的异常数据同样能有效跟踪到异常事件。

对于任意点 X_i，全局孤立点很有可能存在于到其距离最大的 n 个点中。算法实现过程为：

(1)原始数据集归一化。采用 min-max 标准化处理原始数据，使各项数据转换为[0,1]区间内的值，有

$$x^* = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{4}$$

式中：x^* 为标准化后的数据值；x 为原数据值；$x_{\max}$ 为数据最大值；$x_{\min}$ 为数据最小值。

(2)计算任意点 X_i 与其他点 Y_j 的欧氏距离。X_i 与 Ω_{ik} 选自 m 维空间内的任意两点，其欧氏距离计算式为

$$D(X_i - Y_i) = \sqrt{(X_{i1} - Y_{j1})^2 + (X_{i2} - Y_{j2})^2 + \cdots + (X_{im} - Y_{jm})^2} \tag{5}$$

式中：$X_{i1}, X_{i2}, \cdots, X_{im}$ 分别为点 X_i 的第 $1,2,\cdots,m$ 维数据值；$Y_{j1}, Y_{j2}, \cdots, Y_{jm}$ 分别为点 Y_j 的第 $1,2,\cdots,m$ 维数据值。

(3)将 X_i 到其他数据点的欧氏距离按照由小到大排序，找出离 X_i 最近的 k 个数据点，加到 X_i 的 k-邻域 Ω_{ik} 中，并找出 X_i 的 k 距离 $D(X_i)$(k 邻域中各点与 X_i 的欧氏距离最大值)，即

$$D(X_i) = \max(D(X_i, Y_j)) \tag{6}$$

同样地，找出到 X_i 距离最大的 n 个数据点，加到 X_i 的 n-最远域 Ω_{ik} 中，并给 n-最远域 Ω_{ik} 中各点的得票数加 1，即

$$T(P_i) = T(P_i) + 1 \tag{7}$$

式中：$T(P_i)$ 为得票数，其阈值用 $T_{\max}$ 表示。

(4)定义 $T(P_i)$ 值大于 $T_{\max}$ 的点是全局孤立点，通过对不同阈值下的试验结果分析可得到，通过设定恰当的阈值，可更好去除全局孤立点。

对选取的数据进行孤立点检测，挖掘出的代表性异常数据样本如表 2 所示。

表 2　异常收费数据样本[25]

Tab. 2　Abnormal fee data[25]

进站时间	出站时间	通行时长/min	车辆总重/t
2017/1/5 13:28:38	2017/1/6 11:42:31	106	26
2017/1/5 17:30:35	2017/1/5 17:30:35	0	0
2017/1/9 15:25:08	2017/1/9 15:25:08	0	53
2017/1/10 07:20:14	2017/1/10 07:00:03	20	210
2017/1/10 22:10:07	2017/1/10 21:45:52	25	0
2017/1/13 20:13:08	2017/1/13 20:13:08	0	61
2017/1/15 12:10:09	2017/1/15 12:32:20	22	153
⋮	⋮	⋮	⋮

由表 2 可知，第 1 条数据记录的通行时间显然过大，将该条数据用于平均通行时间的计算显然会产生较大误差；第 2 条数据的通行时长及车辆总重为 0，显然不合理；第 3 条、第 6 条数据的进出站时间相同，即在该路段的通行时长为 0，显然也是异常的；由于本文在对收费数据进行聚类处理时选择了通行时长和车辆总重 2 个属性，所以车辆荷载异常

也会对聚类结果产生较大影响;第 2 条与第 5 条数据中的车辆总重为 0,而第 4 条和第 7 条数据中的车辆总重过大,所以对应的这些数据也为异常[25]。

2.3 快速峰值聚类算法与孤立点检测算法的异常事件挖掘结果对比

将快速峰值聚类算法与孤立点检测算法在挖掘异常事件的准确性进行对比分析,以验证所使用的快速峰值算法的准确性。通过将 2 种算法所挖掘到的异常收费数据进行可视化输出便可直观得出结论。快速峰值聚类算法对收费数据处理结果见图 5。

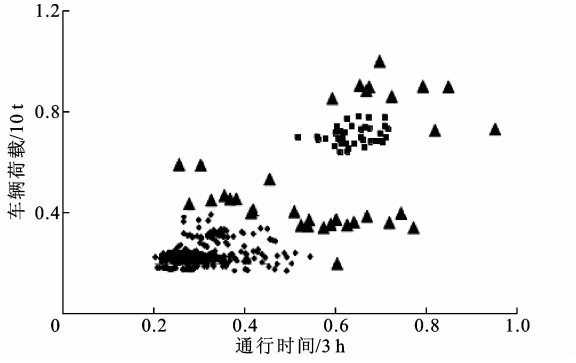

图 5 收费数据聚类结果

Fig. 5 Clustering results of fee data

图 5 中的菱形数据点与方形数据点表示聚类的核心点,三角形数据点为噪声点,即异常数据点,对异常数据进行跟踪分析,便可直接快速发现异常事件。

使用孤立点检测法得到的异常数据见图 6。

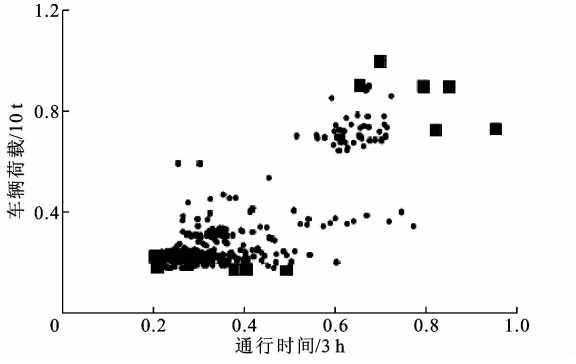

图 6 收费数据孤立点检测结果

Fig. 6 Fee data outlier detected results

图 6 中圆形数据点表示正常数据,方形数据点表示检测出的孤立点,即异常数据。

因孤立点检测算法只能逐维对数据进行检测,对于收费数据,孤立点检测法仅能发现每维数据中的极值,即异常数据中通行时间和车辆荷载均过高的数据点,并错误地将一些通行时间与车辆荷载较低的正常数据判断为异常数据。

对比图 5、图 6 可知,孤立点检测法所挖掘出的异常数据比快速峰值聚类算法数量少 53%。快速峰值聚类算法所挖掘出的部分异常数据主要分布在类簇 1 的上方,而孤立点检测算法在相对应位置所挖掘出的异常数据分布在数据相对集中的类簇 1 的下方,并且挖掘出的异常数据个数远小于快速峰值聚类算法。因此,使用快速峰值聚类算法挖掘出的异常数据更完整、准确和合理,更能真实反映高速公路运行状况。

2.4 快速峰值聚类算法与孤立点检测算法的异常事件识别准确率对比

针对指定进出站口选取的 2 799 条样本收费数据,对 2 种算法挖掘异常数据的效果进行对比分析,可看出,大部分车辆的通行时间都在 1~2 h 之间,存在两类常见的异常数据:通行时间过长和通行时间过短,异常数据全部对应不同的异常事件。根据联网中心值班记录和人工确认,共记录了 72 条各种原因引起的堵车数据,以及 31 条超速、系统故障、逃费嫌疑等事件数据。利用快速峰聚类值算法和孤立点检测算法分别找出的两类数据:快速峰聚类值算法检测出的异常事件中超过 60%的通行时间过长与堵车、事故等异常事件对应,约 50%的通行时间过短与超速、系统故障等异常事件对应;快速峰聚类值算法挖掘异常事件的准确率比孤立点检测算法增加 20%左右。这不仅验证了该算法的有效性和准确性,也表明该算法能快速准确识别收费数据中隐藏的道路拥堵、系统故障和逃费嫌疑等异常事件,进而为高速公路运营管控和维护决策提供科学依据和数据支持。

3 结果与分析

采用快速峰值聚类算法挖掘出的异常数据,可直接定位发生异常事件的车辆、站点和车道。例如在 2017 年 1 月 9~10 日时段内,同一入口或出口出现大量通行时长异常数据,说明有可能是该站收费系统软件、通信网络或车道计算机时钟出了问题,需及时检查维护。如表 3 所示,同一车辆多条数据的通行时间均明显低于正常值,说明该车辆极有可能存在超速或逃费行为,亦或是软件、网络故障等原因,需进行专项核查。

快速峰值聚类识别异常数据主要有以下几类。

(1)通行时间过长,大部分车辆的通行时间为

表 3　某车辆通行时间过短异常数据

Tab. 3　Abnormal data of car with too short time

车牌信息	通行时间/h	车辆总重/kg
* *31 417	0.53 611 111	1 500
	0.75 555 556	1 400
	0.74 444 444	1 400
	0.80 533 333	1 500
	0.65 000 000	1 500
	0.69 000 000	1 500
	0.72 000 000	1 400
	0.66 000 000	1 500
	⋮	⋮

1～2 h，而异常数据的通行事件大多超过 5 h，在 2 个距离较近收费站之间的通行时间过长，可能是事故、停车、时钟不同步、记录错误和疑似逃费等原因。

(2)通行时间过短，由两站间的距离及该路段的最高行驶速度可计算出通行时间最小值，低于该值的数据即为异常数据，可能是车辆超速、网络故障、时钟不同步、记录错误和疑似逃费等原因。

(3)车辆总重过大，主要是货车存在此类问题，可能是车辆超载、称重设备故障、记录错误或疑似逃费等原因。

(4)车辆总重过低，主要是货车存在此类问题，可能是称重设备故障、记录错误或疑似逃费等原因。

4　结　语

(1)本文从海量收费数据入手，考虑异常数据隐藏的异常事件对高速公路运营系统的影响，提出一种识别高速公路异常事件的数据挖掘聚类算法，利用快速峰值聚类算法对车辆在两站间的通行时长和车辆总重 2 个属性进行聚类分析，找出异常数据。将其与孤立点检测算法的清洗结果进行对比，表明快速峰值聚类算法在异常事件识别时更加准确有效。

(2)通过对异常数据挖掘分析，可跟踪车辆的异常行为及其通行信息，为进一步分析异常事件缩小了排查范围，有利于为高速公路运营管控和维护提供可靠数据支撑。本文对通行时长和车辆总重 2 个指标进行了分析，可针对性开展系统故障排查、收费稽查、整治通行费偷逃等，以及对个别站点或车道的称重设备、通讯网络和系统软件等进行重点跟踪，为系统维护提出合理建议。

(3)本文对发生异常事件的车辆、站点、车道等进行了精准定位，对发生的异常事件成因进行了分类，后续可进一步补充数据源进行精准识别和对算法进行改进，从而实现对异常事件原因精准定位及进行其他相应应用。

参考文献：

References：

[1] 陈艳艳，高爱霞，刘小明，等. 道路交通运行状态可靠性评价方法综述及展望[J]. 公路，2003(10)：127-131.
CHEN Yan-yan，GAO Ai-xia，LIU Xiao-ming，et al. Summary and prospect of reliability evaluation methods for road traffic operation status[J]. Expressway，2003(10)：127-131.

[2] ZHOU Rong-gui，ZHONG Lian-de，ZHAO Na-le，et al. The development and practice of China highway capacity research[J]. Transportation Research Procedia，2016，15：14-25.

[3] 沈　强. 基于高速公路收费数据的路网运行状态评价[J]. 公路交通科技，2012，29(8)：118-126.
SHEN Qiang. Evaluation of road network operation status based on highway toll data[J]. Journal of Highway and Transportation Research and Development，2012，29(8)：118-126.

[4] 渐　猛，张俊友. 基于模糊综合评价的道路交通状态判别方法研究[J]. 山东理工大学学报：自然科学版，2013，27(2)：19-22.
JIAN Meng，ZHANG Jun-you. Research on road traffic state discrimination based on fuzzy comprehensive evaluation[J]. Journal of Shandong University of Technology：Natural Science Edition，2013，27(2)：19-22.

[5] KUMARBARAI S. Data mining applications in transportation engineering[J]. Transport，2003，18(5)：216-223.

[6] 李金龙，孙晚华. 高速公路交通事故成因分析及对策研究[J]. 中国安全科学学报，2005，15(1)：62-65.
LI Jin-long，SUN Wan-hua. Analysis of causes of expressway traffic accidents and countermeasures[J]. Chinese Journal of Safety Science，2005，15(1)：62-65.

[7] WENG Jian-cheng，LIU Li-li，DU Bo. ETC data based traffic information mining techniques[J]. Journal of Transportation Systems Engineering and Information Technology，2010，10(2)：57-63.

[8] 潘若禹. 基于数据融合的高速公路交通异常事件检测的研究[D]. 西安：长安大学，2006.
PAN Ruo-yu. Research on highway traffic abnormal events detection based on data fusion[D]. Xi'an：Chang'an University，2006.

[9] 吴晓佩. 基于多 SVM 分类器融合的高速公路异常事

件检测方法[J]. 现代交通技术,2014,11(4):63-67.
WU Xiao-pei. Detection method of expressway abnormal events based on Multi-SVM classifier fusion[J]. Modern Transportation Technology, 2014, 11(4): 63-67.

[10] 隋 靓,党建武. 基于运动目标轨迹的高速公路异常事件检测算法研究[J]. 计算机应用与软件,2018,35(1):246-252.
SUI Liang, DANG Jian-wu. Study on detection algorithm of expressway abnormal events based on moving target trajectory[J]. Computer Applications and Software,2018,35(1):246-252.

[11] 巨永锋,潘若禹,李 磊. 数据融合技术在高速公路异常事件检测中的应用[J]. 现代电子技术,2005(17):70-72.
JU Yong-feng, PAN Ruo-yu, LI Lei. Application of data fusion technology in highway abnormal event detection[J]. Modern Electronic Technique, 2005(17): 70-72.

[12] 赵 勇,刘建华,赵小强. 高速公路交通异常事件预警系统研究[J]. 西安邮电学院学报,2010,15(3):122-124.
ZHAO Yong, LIU Jian-hua, ZHAO Xiao-qiang. Study on early warning system of expressway traffic anomaly[J]. Journal of Xi'an Institute of Posts and Telecommunications,2010,15(3):122-124.

[13] 孙静怡,牟若瑾,苏晓波. 基于支持向量机的高速公路异常状态检测[J]. 重庆交通大学学报:自然科学版,2018,37(9):1-7.
SUN Jing-yi, MOU Ruo-jin, SU Xiao-bo. Detection of abnormal state of expressway based on support vector machine[J]. Journal of Chongqing Jiaotong University: Natural Science Edition, 2018, 37(9): 1-7.

[14] SADEK S, ALHAMADI A, MICHAELIS B, et al. A statistical framework for real-time traffic accident recognition[J]. Journal of Signal & Information Processing, 2010, 1: 77-81.

[15] SHEU J B. A sequential detection approach to real-time freeway incident detection and characterization [J]. European Journal of Operational Research, 2004, 157(2): 471-485.

[16] JIN X, SRINIVASAN D, CHEU R L. Classification of freeway traffic patterns for incident detection using constructive probabilistic neural networks[J]. IEEE Transactions on Neural Networks, 2001, 12(5): 1173-1187.

[17] MARWAN H, THOMAS S. Clustering big data streams: Recent challenges and contributions[J]. It-information Technology, 2016, 58(4): 206-213.

[18] Dmitri Viattchenin. Constructing stable clustering structure for uncertain data set[J]. Acta Electrotechnica et Informatica, 2011, 11(3): 42-50.

[19] 李 艳,张 庆,田苏慧敏. 改进的数据挖掘模糊聚类算法研究与分析[J]. 宁夏师范学院学报,2018,39(1):36-47.
LI Yan, ZHANG Qing, TIAN Su-huimin. Research and analysis of improved data mining fuzzy clustering algorithm[J]. Journal of Ningxia Normal University, 2018, 39(1): 36-47.

[20] 胡 波. 数据挖掘在高速公路联网营运管理及决策上的应用探讨[J]. 中国交通信息产业,2004(12):86-88.
HU Bo. Application of data mining in expressway network operation management and decision-making [J]. China Traffic Information Industry, 2004(12): 86-88.

[21] 何亚龙. 基于模糊聚类分析的高速公路状态识别研究[J]. 山东工业技术,2018(6):208-225.
HE Ya-long. Research on expressway sate recognition based on fuzzy clustering analysis[J]. Shandong Industrial Technology, 2018(6): 208-225.

[22] ALEX R, ALESSANDRO L. Clustering by fast search and find of density peaks[J]. Science, 2014, 344(6191): 1492-1496.

[23] 孙 云,李舟军,陈火旺. 孤立点检测算法及其在数据流挖掘中的可用性[J]. 计算机科学,2007,34(10):200-203.
SUN Yun, LI Zhou-jun, CHEN Huo-wang. Outlier detection algorithm and its availability in data stream mining[J]. Computer Science, 2007, 34(10): 200-203.

[24] 古 平,刘海波,罗志恒. 一种基于多重聚类的离群点检测算法[J]. 计算机应用研究,2013,30(3):751-756.
GU Ping, LIU Hai-bo, LUO Zhi-heng. An outlier detection algorithm based on multiple clusters[J]. Applied Computer Research, 2013, 30(3): 751-756.

[25] 赵怀鑫,邓然然,张英杰,等. 一种用于高速公路通行情况分析的收费数据挖掘方法[J]. 中国公路学报,2018,31(8):155-164.
ZHAO Huai-xin, DENG Ran-ran, ZHANG Ying-jie, et al. Method of mining fee data for expressway traffic analysis[J]. China Journal of Highway and Transport, 2018, 31(8): 155-164.