文章编号:1671-1637(2015)05-0118-09

基于交通轨迹数据挖掘的道路限速信息识别方法

廖律超[1,2],蒋新华[1,2],林铭榛[3],邹复民[2]

(1. 中南大学 信息科学与工程学院,湖南 长沙 410075; 2. 福建工程学院 福建省汽车电子与电驱动技术重点实验室,福建 福州 350108; 3. 中南大学 软件学院,湖南 长沙 410075)

摘 要:分析了道路限速信息的时空变化性,提出一种基于轨迹数据挖掘技术的道路限速信息自动识别方法。为了实现海量交通轨迹数据的快速处理,研究了快速地图匹配与数据清洗等预处理算法,分析了交通轨迹数据的速度分布特性与最高车速限制指标。基于路段行车速度的统计特性,构建了道路特征向量模型,以快速提取海量轨迹数据的潜在特征信息。提出了多投票 K 近邻分类算法对数据特性进行训练与学习,以实现对道路限速信息的快速识别。以福州市交通路网及其浮动车轨迹数据构建试验样本集进行训练、学习与交叉验证试验。试验结果表明:在训练过程中,当样本数量达到 1 200 时,方法的识别准确率最高达到 93%,在仅有 150 个小训练样本下,方法的识别准确率也达到 75%;方法具有近线性的处理性能,处理 1.0×10^6 条道路的限速信息仅用时 46 ms。

关键词:道路限速;轨迹数据挖掘;浮动车数据;交通流;地图匹配;K 近邻算法

中图分类号:U491 **文献标志码**:A

Recognition method of road speed limit information based on data mining of traffic trajectory

LIAO Lu-chao[1,2], JIANG Xin-hua[1,2], LIN Ming-zhen[3], ZOU Fu-min[2]

(1. School of Information Science and Engineering, Central South University, Changsha 410075, Hunan, China; 2. Fujian Key Laboratory for Automotive Electronics and Electric Drive, Fujian University of Technology, Fuzhou 350108, Fujian, China; 3. School of Software, Central South University, Changsha 410075, Hunan, China)

Abstract: The spatiotemporal variability of speed limit information was analyzed, and an automatic recognition method of road speed limit information was proposed based on the mining technique of trajectory data. To fast process the massive traffic trajectory data, the pretreatment algorithms such as rapid map matching and data cleaning were researched. The speed distribution features of traffic trajectory data and the maximum speed limit index were analyzed. Based on the speed features at road section, a road feature vector model was constructed to rapid extract the latent characteristics information from the massive trajectory data was achieved. In order to implement a rapid recognition of speed limit information, a classification algorithm based on multi-voting K-nearest neighbor (MV-KNN) algorithm was proposed for the training and learning process of data feature. The training, learning and cross-validation experiments were completed by using the sample sets constructed by actual floating car trajectory data and traffic network in Fuzhou City. Experimental result indicates that the highest system recognition

收稿日期:2015-04-16

基金项目:国家自然科学基金项目(61304199);福建省中青年教师科技项目(JA14209);福建省自然科学基金项目(2012J06015,2013J01214);福建省科技重大专项专题项目(2013HZ0002-1);福建省科技计划项目(2012I0002,2014H0008)

作者简介:廖律超(1980-),男,福建长汀人,福建工程学院高级工程师,中南大学工学博士研究生,从事交通数据挖掘与处理技术研究。

导师简介:蒋新华(1956-),男,湖南长沙人,中南大学教授。

accuracy of proposed method is up to 93% by using 1 200 samples in the training process, and the system recognition accuracy is 75% by using only 150 samples. The near-linear processing performance of proposed method is revealed, and the system operating time is only 46 ms in processing 1 000 000 samples of road speed limit information. 1 tab, 12 figs, 29 refs.

Key words: road speed limit; trajectory data mining; floating car data; traffic flow; map matching; *K*-nearest neighbor algorithm

Author resumes: LIAO Lu-chao(1980-), male, doctoral student, +86-591-22863333, lcliao@csu.edu.cn; JIANG Xin-hua(1956-), male, professor, +86-591-22863333, xhj@csu.edu.cn.

0　引　言

道路限速作为交通管理的一项重要安全措施，在交通安全管理和城市交通规划中占有重要的地位[1-5]。交通信息服务部门主要通过人工实地采集道路限速信息，但道路限速并不是一成不变的，交通管理部门会根据道路的交通流量、周边区域的开发情况、交通饱和度等条件的变化，适时地进行不同时段与路段限速的调整[6-9]。正因为数据采集困难且具有动态变化性，使得现有的 GIS 地图数据中，只有少量的道路具有相应的限速信息。以高德地图 2013 年数据为例，在福州市交通路网的 48 355 个路段中，仅有 1 973 个路段明确标注了限速信息，而且这些数据往往未及时更新，难以满足实际应用需求。通过自动采集并识别现有道路(甚至是新增道路)的限速信息，可以为公众出行提供精准的道路限速提醒服务，有效辅助驾驶人安全驾驶，有利于避免潜在的交通安全事故[2]，因此，道路限速信息的自动识别技术已成为国内外智能交通领域的研究热点与难点之一[10]。

目前，国内外对道路限速信息识别的研究主要是基于图像识别技术对道路上设置的限速标志牌进行识别提取[11-16]。Greenhalgh 等提取了道路限速标志牌图像的 HOG(Histogram of Oriented Gradient)特征，基于 SVM 方法进行训练学习，实现了限速信息的自动识别，系统的识别准确度最高达到 89.2%[13]；Zaklouta 等提出了一种道路限速信息的实时识别方法，核心原理为基于 SVM 方法对道路限速标志牌图像进行 HOG 特征识别，并加入了自适应阈值机制，系统识别准确率为 90%，单个标志牌识别最快为 357 ms[15]；Liu 等提出了一种基于稀疏编码的道路限速信息识别方法，系统识别率最高达到 97.83%，但对于道路限速变化(如取消限速)的识别率仅为 85.33%[16]。尽管现有方法已具有较好的识别准确率，但需要勘测人员到每条道路上进行实地信息采集，且由于其通过图像处理手段采集道路限速标志牌信息，容易受道路复杂性与天气、光照环境等因素的影响，尤其在雨雾天气与黑夜很难进行有效采集。在轨迹数据挖掘方面，Pascale 等将意大利 A4 高速公路(都灵-威尼斯)划分为 166 个路段，并基于连续 63 个工作日的浮动车数据(每路段每小时的车辆数约为 10 veh)进行道路行车速度特性分析，研究表明在不同路段具有不同的典型行车速度模式[17]，但并没有进一步深入研究其形成相应速度模式的道路限速等内在约束信息。

面对国内城市化进程与交通道路建设的高速发展，实地采集的道路限速信息往往严重滞后，而且由于道路周边环境等因素的变化，交通道路限速标识会进行相应的调整，现有的采集方法需要全路网排查更新，不仅更新成本高，而且更新周期长，难以有效满足实际应用需求[18]。

实际上，尽管每个驾驶人的驾驶习惯迥异，但在实际驾驶中，驾驶人还是会在一定程度上根据实际道路的限速信息进行相应的调整[19]，基于此，本文提出了一种根据海量浮动车轨迹数据进行道路限速信息识别的方法。分析了交通轨迹数据速度信息的统计特性，提出了一种路段地图匹配方法，实现了海量交通轨迹数据与道路的快速匹配，并构建了交通道路速度特征向量模型，进而提出了一种多投票 *K* 近邻(MV-KNN)算法对交通道路的速度特征进行有监督分类训练，以实现未知道路的限速信息分类识别。以福州市交通路网及其浮动车轨迹数据构建试验样本集进行训练、学习与交叉验证试验，试验结果表明识别方法能够很好地解决现有道路与新增道路限速信息识别问题，具有良好的处理效率和动态更新能力。

1　交通轨迹速度特性

为了深入挖掘浮动车轨迹数据的速度信息与道路限速信息之间的潜在关联特性，首先对道路浮动车速度的统计特性进行分析。速度信息主要有 2 种

获取方式:区间平均速度与瞬时速度。区间平均速度的获取需要计算相邻 2 次采样数据之间的路程,然而由于浮动车数据是一种稀疏数据[20],以 30 s 的采样间隔计算,若平均速度为 60 km·h^{-1},则在相邻采样间隔之间车辆可能行驶了 500 m,难以在同一路段内进行区间平均速度估计。另外区间平均速度的计算往往需要通过插值法估计浮动车的实际行驶道路[21],计算过程复杂,计算量大,难以满足大规模浮动车数据的快速处理要求[22]。为此,本文采用车辆瞬时速度进行道路行车速度统计特性分析。

交通轨迹数据的速度时域特性反映了在不同限速条件下各道路交通流的时变演化规律。以福州市晋安南路为例,该路段限速为 40 km·h^{-1},从 0:00 开始,以每个小时为 1 个时间段,将全天划分为 24 个时间段,分别编号为 0、1、…、23,对每个时间段的速度信息进行盒图分析,结果见图 1。从图 1 可知,受路况变化等因素的影响,在不同时间段,路段平均瞬时速度并不相同,其车辆速度最大值与最小值、上下四分位的速度分布以及每个时间段的数据量都不尽相同。可见,完全依赖传统的统计分析方法难以直接有效地获得准确的限速,若将数据按天进行分析,而不考虑不同时间段的特征差异,则容易受个别时间段道路拥堵等因素的影响,因此,需要充分融合道路在不同时间段的行车速度特征信息,并进行知识训练与学习,以分析道路速度演化表征下的限速信息等内在约束机理。

为进一步研究不同时间段内的速度特性,引入速度均值、中值与样本方差、样本标准差、车速极差等频域空间的评价指标进行速度特性分析。由于车速极差受车速随机性影响较大,同时考虑到国内外研究常将交通车辆的频数-速度分布曲线的 85%位置定义为重要的限速参考线,并作为最高车速限制指标[23],故根据 85%位置线引入车辆的速度离散度指标,以分析数据内部的离散特性,并进行离群点检测以屏蔽路段噪声数据的影响。

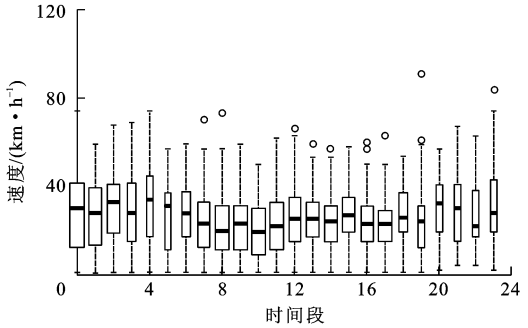

图 1 速度信息盒图分析

Fig. 1 Box analysis of speed information

定义 1(车速离散度指标):路段浮动车数据频数-速度分布曲线图中 85%位置的速度为 v_1,15%位置的速度为 v_2,车速离散度指标 Δv 为

$$\Delta v = v_1 - v_2$$

车速离散度指标描绘了路段中浮动车速度的离散特性,该值范围越大,说明速度信息的离散程度越高,反之则说明速度信息分布越集中。车速离散度指标也是路段频数-速度分布曲线中主要速度分布的区间范围,包含了道路上绝大多数的浮动车数据。

在路段的浮动车轨迹数据中,由于 GPS 数据错误或道路拥堵等原因,出现速度信息远偏离其均值的轨迹数据,如停车数据、异常低速和异常高速的数据等,这些数据都是不利于道路限速信息挖掘的噪声数据。以图 1 的盒图分析为例,空心圆点为数据中的离群噪声数据。为此,在限速信息挖掘过程中,需要对这些噪声数据进行清洗处理。

2 系统关键算法设计

2.1 轨迹数据预处理

2.1.1 浮动车数据地图匹配

要利用浮动车数据获取路段的速度信息,首先必须通过地图匹配算法将浮动车数据匹配到正确的路段上[24],然而由于浮动车系统具有数据量大、实时性要求高和稀疏采样等特点[25],本文基于搜索引擎倒排快速索引机理[26],提出了一种基于路网倒排索引方式的地图匹配算法,以提升系统处理精度与效率,并支持现代城市路网密集且结构复杂的应用场景需求。

定义 2(路网倒排索引):路网倒排索引是一种特殊的网格索引方法,也称为路网反向索引,通过预处理计算道路与网格之间的映射关系,根据空间网格序号快速获取网格对应的候选路段列表。路网倒排索引的创建主要包括 2 个步骤。

Step 1:根据精度要求,创建空间信息网格,并根据网格中心点坐标创建网格关键字,关键字创建机制主要根据网格精度对坐标经纬度值分别进行截尾并转化为字符串,产生的两个字符串的连接作为网格关键字。

Step 2:以网格关键字为索引构建路网倒排表,计算交通道路与各个网格之间的空间关系,并以相应网格关键字为索引值保存到路网倒排表中,具体流程见图 2。

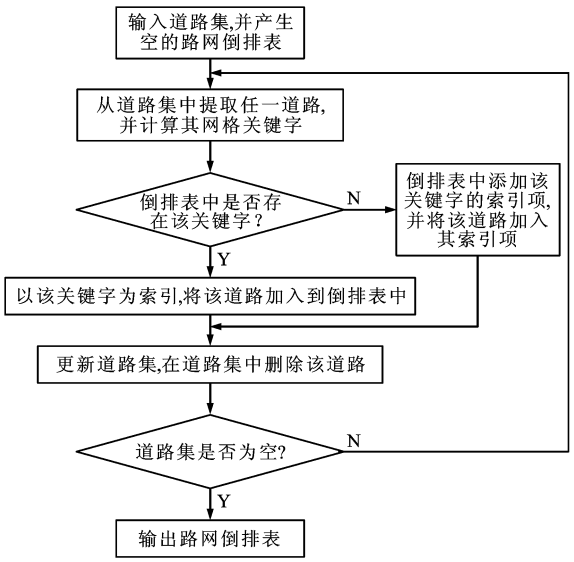

图 2　路网倒排表创建

Fig. 2　Establishment of road network inverted list

通过路网倒排索引，可以快速检索浮动车数据点对应的候选匹配路段。但当候选路段有多个时，无法直接确定具体路段，此时可以通过将浮动车数据点根据索引到的路段进行投影匹配，并结合投影距离和车辆行驶方向与路段夹角偏差等因素选取最佳匹配路段 ξ 作为浮动车的行驶匹配路段。即对于任一位置点 p，若其对应网格关键字在路网倒排表中有多条候选道路，则其最佳匹配路段 ξ 为

$$\xi = \arg\left\{\min_{r\in R}\left[\frac{D(p,r)}{\kappa_1}+\frac{\eta\alpha}{\kappa_2}\right]\right\} \tag{1}$$

式中：$D(p,r)$ 为位置点 p 到道路 r 的垂直投影距离；R 为位置点 p 在路网倒排表中索引到的侯选路段集合；α 为位置点 p 与道路 r 的夹角；κ_1、κ_2 分别为投影距离和夹角的最大阈值；η 为方向权重因子。

2.1.2　数据清洗

在进行速度特征建模前，应通过构造有效的损失评估函数，以评估数据的离群信息，并清洗轨迹数据中的异常数据，避免异常数据的存在干扰模型的建立，也使构建的模型更加逼近真实模型。

根据中心极限定理，在采集路段的大量浮动车数据的条件下，路段速度数据集 X 将趋于正态分布，即其频数-速度曲线将呈现高斯分布规律，也可分别利用 K-S 检验和 S-W 检验来验证这个假设，即 $X\sim N(\mu,\sigma^2)$，其中，μ 为路段样本平均瞬时速度，决定了曲线的位置，体现了不同限速情况下路段速度分布的集中趋势，σ 为路段样本数据的速度标准差，主要描述了路段样本总体速度数据分布的离散程度，其值越大，表示速度数据分布越分散，相反则分布越集中。

为此，通过对浮动车轨迹数据中的速度进行分析，构建了噪声数据清洗阈值模型，以检测海量数据中离群点信息并予滤除。模型的基本思想是根据距离数据统计分布的中心点阈值 v_T 来进行判决，并将在这个阈值之外的数据剔除，进而在海量的轨迹数据中快速滤除噪声数据，表示为

$$v_T = \mu \pm 2.62\sigma \tag{2}$$

噪声数据清洗阈值模型根据数据的统计分布特性，以数据均值为中心设置阈值过滤频带，具有实现简单且快速高效等特点，从而实现在海量数据中快速滤除路段噪声数据。

2.2　速度特征向量建模

对交通轨迹数据的速度特性统计分析表明，在不同时间段，道路行车速度具有不同的特性。为此，速度特征向量的建立需要充分考虑不同时间段的影响，尤其平均速度较大的时间段，该时间段行车速度受道路拥堵的影响较小，而主要受道路限速约束。通过频域空间的评价指标对这些速度特征进行描述。结合主要时间段的特征数据及相关评价指标，构建速度特征向量模型，将道路的速度特征转化为一个 13 维度的特征向量 $\boldsymbol{\Lambda}$，表示为

$$\boldsymbol{\Lambda} = (\lambda_1,\lambda_2,\cdots,\lambda_{12},\lambda_{13})^{\mathrm{T}} \tag{3}$$

式中：属性 $\lambda_1\sim\lambda_6$ 为数据样本中数值最大的 6 个时间段的平均瞬时速度，即将路段全天数据按小时划分为 24 个时间段，计算每个时间段的平均瞬时速度，并由大到小进行排序，取前 6 个值，作为属性 $\lambda_1\sim\lambda_6$，对于在此属性中缺失的数据，可以用属性中最大值填补，该属性系列的设置，主要避免了在某些时间段里由于特殊路况或交通拥堵等造成的瞬时速度相对偏低所带来的扰动；属性 λ_7、λ_8 分别为路段总体平均瞬时速度 μ 和标准差 σ，对于每一个道路样本的交通轨迹数据，这 2 个属性代表了与其近似拟合的高斯分布曲线；属性 λ_9、λ_{10} 分别为路段频数-速度分布曲线上的 15%位置车速 v_2 和 85%位置车速 v_1，这 2 个属性代表了路段车速较为集中的一个范围，在交通管理上通常将 v_1 作为限速的参考值；属性 λ_{11} 为路段频数-速度分布曲线上 95%位置车速 v_3，即路段车辆速度落在超出该速度之外的概率为 5%，该属性可对限速相近（例如 40、50 $\mathrm{km\cdot h^{-1}}$）的道路分类产生影响；属性 λ_{12} 为车速离散度指标 Δv，在一定程度上反映实际交通流速度数据的变化范围和离散幅度；属性 λ_{13} 为路段速度众数，即频率最高的瞬时速度，代表车辆速度统计规律的一般水平。

以上属性从数据统计分析角度描述了路段的主要速度特征，将所有特征值构造成向量形式，便得到了路段的行车速度特征向量。

2.3 基于 MV-KNN 的道路限速信息训练识别

K 近邻方法(K-Nearest Neighbor, KNN)为一种基于实例的惰性学习方法[27-28]，具有简单高效、精度高等特点，且不要求建立显示规则，对异常数据不敏感，能对超多边形的复杂决策空间建模且实现方便，并支持增量学习(如对新增路段处理)等优点，但样本库中存在的噪声样本容易降低分类准确率[29]。

为此，提出一种多投票 K 近邻道路限速信息识别算法，其核心步骤主要包括基于单日数据的道路限速信息分类识别以及基于多日数据分类识别结果集的最佳限速识别，即对同一道路不同日期的速度特征向量同时进行处理，并通过投票方式获取最佳的分类结果，进而实现更为准确可靠的道路限速信息识别。具体算法步骤如下。

Step 1：构造训练样本集与测试集。

由已知限速信息的道路原始数据抽象出的 13 维路段速度特征向量组成道路训练样本，设第 i 个训练样本为 s_i，由 n 个训练样本构成训练样本集为 S。

每个训练样本都有其已知的道路限速信息，设训练样本 s_i 对应的道路限速分类标记值为 l_i，则构成与训练样本集对应的训练样本分类标记集 L。

构建训练样本集的目的是通过给定若干已知限速信息的路段速度特征，为其他道路的限速信息估计提供参考依据。训练样本集 S 的构造有内在约束机制，其对应的训练样本分类标记集 L 必须覆盖所需要识别道路的各种限速值。

提取某一待识别道路 d 天的行车轨迹数据，按路段速度特征向量方法，每天构建一个特征向量作为测试样本，设第 m 天轨迹数据构成的测试样本为 q_m，所有测试样本构成整个测试集 Q，测试集的测试样本越多，越有利于提高系统识别的准确率。

Step 2：道路限速信息分类识别。

每个训练样本可理解为 13 维特征空间中的一个点，测试样本与各个训练样本的相似度就是它们在特征空间中的欧氏距离，即对于任意 2 个特征向量样本 a、b，其欧氏距离 $E(a,b)$ 表示为

$$E(a,b)=\sqrt{\sum_{t=1}^{13}(a_t-b_t)^2} \tag{4}$$

式中：a_t、b_t 分别为路段样本 a、b 第 t 维的属性。

根据 K 近邻的思想，给定测试样本 q_m，其 K 近邻集就是在特征空间中与 q_m 最相近的 K 个训练样本。其中，K 为需设定的参数，一般情况下，增大 K 可减小噪声的影响，但也可能会降低准确率，因此，需要对 K 进行试验调优，通常采用交叉验证的方式进行 K 调优。

基于训练样本集可实现任一测试样本的限速分类标记值估计，其核心思想是统计分析测试样本在特征空间中的近邻集分类情况，当其近邻集中的训练样本大部分为某一限速分类标记时，则该测试样本也以较大概率服从该分类标记。进一步考虑到近邻集中各个训练样本对测试样本的影响程度各不相同，距离越近则其影响越大。为此，对每个训练样本 s_i 赋予相应的反距离权重因子 ω_i，即若测试样本为 q_m，则训练样本 s_i 的反距离权重因子为

$$\omega_i=E(s_i,q_m)^{-1}$$

根据上述分类识别思想，任一测试样本 q_m，根据 K 近邻算法提取其特征空间的 K 个近邻样本，以构成其近邻集 P，进而分析近邻集 P 中各个训练样本的对应分类标记值分布情况，并结合其反距离权重因子，提取最大概率的限速分类标记值作为 q_m 的限速值 c_m，则有

$$c_m=\arg\left\{\max_{l_i\in L}\left[\sum_{j=1}^{K}\omega_j\vartheta(h_j=l_i)\right]\right\} \tag{5}$$

式中：$\vartheta(\cdot)$为示性函数，当其参数为真时为 1，否则为 0；h_j 为近邻集 P 中第 j 个近邻样本的道路限速分类标记值；ω_j 为第 j 个近邻样本的反距离权重因子。

对该道路基于测试样本 q_m 估计的道路限速 c_m，设其匹配度为 $M(c_m)$，则有

$$M(c_m)=\max_{l_i\in L}\left[\sum_{j=1}^{K}\omega_j\vartheta(h_j=l_i)\right] \tag{6}$$

通过测试集中的任一测试样本 q_m，即可得到该道路限速信息的初步估计值 c_m。为了进一步提高识别的准确率，对测试集 Q 中的 d 个测试样本分别进行道路限速估计，可得到 d 个限速分类标记值及其各自对应的匹配度，以寻找与该道路最匹配的道路限速信息 $C(Q)$

$$C(Q)=\arg\left\{\max_{l_i\in L}\left[\sum_{m=1}^{d}M(c_m)\vartheta(c_m=l_i)\right]\right\} \tag{7}$$

针对各类限速条件下道路情况，构造车速特征向量形成路段训练样本集，并通过 MV-KNN 分类处理算法，实现未知道路限速信息的快速识别。通过 MV-KNN 算法的多投票机制，可提高系统识别的准确率与鲁棒性，同时，由于在识别过程中该算法只与极少量的相邻样本相关，有望解决小样本学习问题，且可避免各类别样本间的不平衡问题。

3　试验结果分析

3.1　数据准备与预处理

3.1.1　试验数据集

试验数据集为 2013 年 12 月的福建省福州市浮动车轨迹数据，数据集覆盖地理空间的经度范围为[119.113°，119.684°]，纬度范围为[25.904°，26.155°]，区域面积约为 1 430 km^2，覆盖了福州市主城区与周边区域。数据集包括出租车、两客一危、重型货车与半挂牵引车等 10 类车辆，车辆数约为 3.0×10^4 veh，每天数据量约为 2.1×10^7 条，整个数据集共包含约 6.0×10^9 条浮动车位置信息，交通轨迹总里程约为 1.056×10^8 km。

试验道路地图数据为福州市路网，见图 3，来源于高德地图提供的测试数据。地图共包括 48 355 个路段，其中已知限速信息的路段为1 973 个。

图 3　福州市区路网

Fig. 3　Road network of Fuzhou City

3.1.2　地图匹配处理

对交通轨迹数据进行地图匹配将为数据的挖掘处理提供重要的车路关联数据支持。根据路网倒排索引方法，首先将试验区域全路网划分为 6.06×10^5 个子路段，共创建了 108 631 个倒排索引号，每个倒排索引号对应若干个子路段，通过距离与方向等因素，将整个试验数据集的浮动车数据快速匹配到各个具体路段。图 4 为地图路段匹配后的轨迹数据点(浦上大桥)。

进一步的试验表明，创建全路网索引并进行路网划分的时间为 13 s，而匹配一个 GPS 数据到相应路段的时间少于 0.1 ms，实现了浮动车轨迹数据的快速地图匹配。

3.1.3　数据清洗处理

通过地图匹配预处理，抽取任一路段的轨迹数据进行特性分析，以橘园洲特大桥的浮动车数据为

图 4　地图匹配后的轨迹数据点

Fig. 4　Trajectory data points after map matching

例，其频数-速度分布曲线见图 5，其数据分布基本符合高斯正态分布特性。进一步通过正态检验分析其高斯分布性质，结果见图 6，尽管总体数据近似服从正态分布，但仍有噪声数据需要清洗处理，尤其低速数据较大幅度地偏离拟合直线。

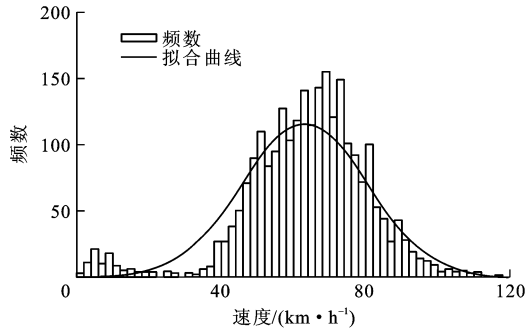

图 5　原始数据频数-速度分布

Fig. 5　Frequency-speed distribution of original data

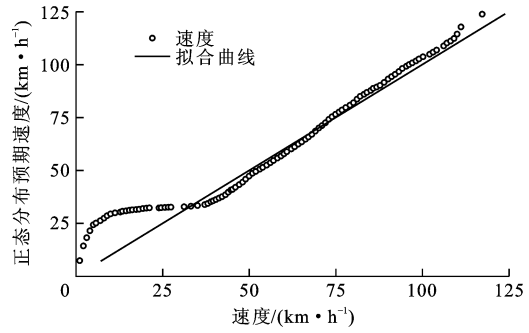

图 6　原始数据正态检验结果

Fig. 6　Normal test result of original data

进一步通过噪声数据清洗阈值模型对数据进行清洗处理，得到频数-速度关系见图 7，符合高斯正态分布，其正态检验结果见图 8，符合直线，表明已经清洗了数据中的大部分噪声数据。通过数据的降噪清洗，有利于进一步挖掘分析交通轨迹数据的潜在特性。

3.2　特征向量构建

道路的行车速度变化往往呈现以天为周期的演

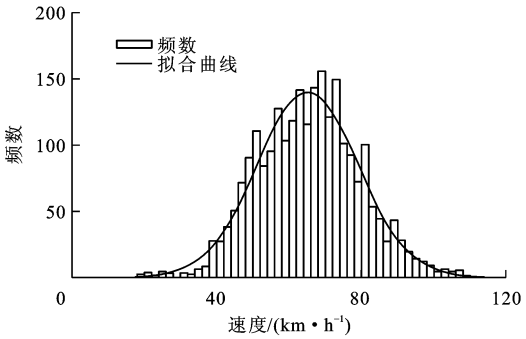

图 7 降噪后的频数-速度分布

Fig. 7 Frequency-speed distribution after noise reduction

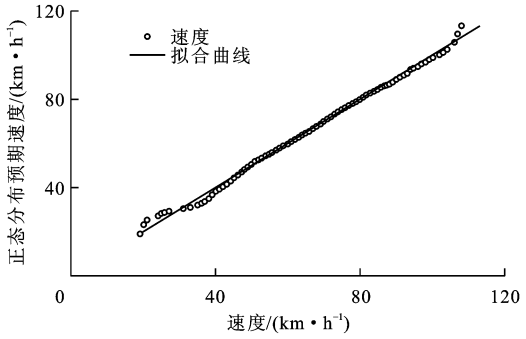

图 8 降噪后的正态检验结果

Fig. 8 Normal test result after noise reduction

化特性，而同一天内的不同时段往往具有不同的速度分布模式。通过速度特征向量模型，将整合不同时段的速度分布与离散程度等数据特征，有利于挖掘道路限速等影响行车速度演化的内在约束信息。

通过速度特征向量模型的 13 个统计特征，构建了道路速度特征向量集，见表 1，每个向量共包含 13 个维度的属性及其样本分类标记，其中样本分类属性 l 为路段限速，整个数据集的分类值包含了《公路工程技术标准》(JTG B01—2003)和《城市道路设计规范》(CJJ 37—90)中规定主要的道路限速 40、50、60、80、100 km·h^{-1}。

3.3 系统识别准确率试验

影响系统识别准确率的主要因素有样本量大小及 K 近邻算法中的 K 值选择，为此，通过调整不同的参数进行准确率试验测试。

3.3.1 试验 1

试验 1 测试不同样本量及 K 值选择对系统识别准确率的影响。首先将测试数据集分为不同规模的样本子集 1～3，分别包含各类限速信息的道路样本数 1 300、580、130 个。在 K 分别为 1、2、…、6 情况下，对样本集1～3分别进行模型识别，并通过测试样本进行系统识别准确率测试。

表 1 测试路段特征向量

Tab. 1 Feature vectors of test road sections km·h^{-1}

路段名称	日期	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7	λ_8	λ_9	λ_{10}	λ_{11}	λ_{12}	λ_{13}	l
洋中路	2013-12-26	38	37	37	36	34	33	23	12	11	37	49	26	24	40
晋安南路	2013-12-10	32	32	29	28	28	28	26	11	13	39	46	26	30	40
	2013-12-16	33	30	30	29	29	28	25	11	12	37	44	25	31	40
杨桥中路	2013-12-17	41	40	39	39	38	38	29	16	10	48	58	38	30	50
	2013-12-18	45	41	41	40	37	37	31	14	14	48	57	34	35	50
浦上大桥	2013-12-16	75	65	62	61	60	59	56	18	35	76	85	41	69	60
	2013-12-12	59	54	54	52	51	51	48	17	33	70	83	37	69	60
橘园洲特大桥	2013-12-25	68	66	66	66	65	64	62	15	46	78	86	32	72	60
南二环路	2013-12-25	72	69	69	69	66	65	61	21	38	83	94	45	80	80
	2013-12-26	70	68	67	65	65	64	61	19	41	81	91	40	81	80
福州机场高速	2013-12-29	66	64	62	62	61	60	58	19	39	79	89	40	72	100
	2013-12-30	65	64	64	63	62	62	58	18	39	79	87	40	74	100

系统识别准确率测试结果见图 9，当选择相同 K 值时，样本量越大，系统识别准确率越高，其中，样本集 1 的系统识别准确率达到 94%；对于相同样本量情况下，K 值越小，系统识别准确率越高。

3.3.2 试验 2

试验 2 测试不同样本量对系统识别准确率的影响。为了进一步测试不同样本量对系统识别准确率的影响，设定 K 为 1，并采用 150、300、450、600、750、900、1 050、1 200 个样本进行系统识别准确率测试，结果见图 10。测试结果表明：随着样本量的增大，系统识别准确率快速提高，当样本量达到 1 200 个时，系统识别准确率达到 93%；系统总体上具有较高的识别准确率，即便在较小样本量(样本量为 150 个)的情况下，仍具有 75%的系统识别准确率。

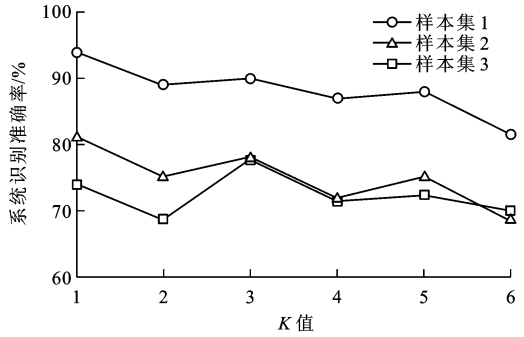

图 9　不同 K 值的系统识别准确率

Fig. 9　System recognition accuracies of different K values

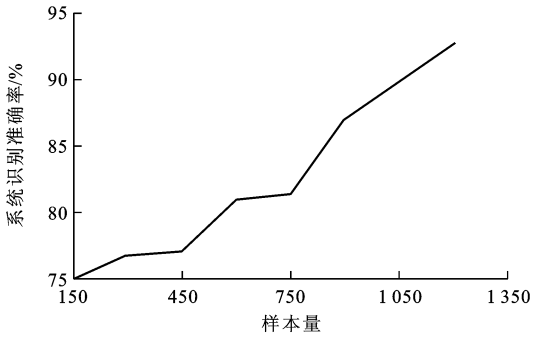

图 10　不同数据量的系统识别准确率

Fig. 10　System recognition accuracies of different data quantities

3.3.3　系统处理性能试验

理论上，MV-KNN 算法作为一种基于惰性学习的道路限速信息识别算法，其空间复杂度为 $O(n)$，其中 n 为路段样本量大小，而当 K 值为 1 时，其时间复杂度也为 $O(n)$，系统的存储与计算开销随着 n 的变化呈线性增长，且方法不需要提前训练，节省了系统额外的性能开销。为了测试系统的实际处理性能，通过分别设置不同样本量与 K 值，对处理性能进行试验分析。

生成大规模道路特征向量数据进行系统性能测试。固定 K 为 1，设置初始样本量为 10^5 个，并以 10^5 个的量级递增，测试结果见图 11。结果表明，随着数据量的增加，算法运行时间以毫秒级呈近似线性递增。进一步固定数据量为 10^6 个，将其 K 取值递增进行系统性能测试，结果见图 12。结果表明，随着 K 值的增大，算法耗时平稳增加，表明算法具有良好的处理性能，可支持大规模道路限速信息的快速识别处理。

4　结　语

(1)每条道路的行车速度在不同时间段具有不同的统计特性，而交通轨迹数据的速度时域特性正反映了在不同限速条件下道路交通流的时间变化规律，并可用于推测道路限速等内在约束信息。

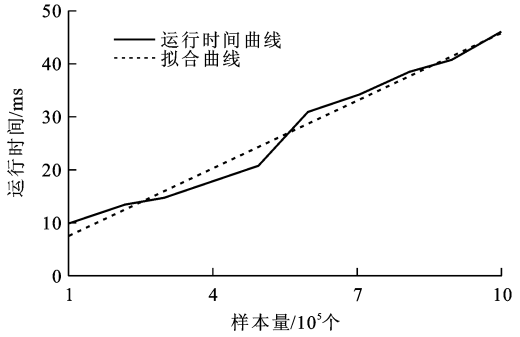

图 11　不同样本量系统运行时间

Fig. 11　System operating times of different data quantities

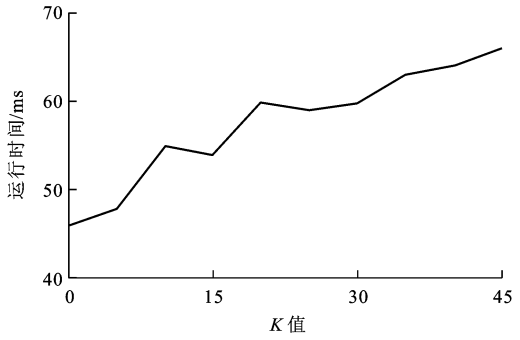

图 12　不同 K 值的系统运行时间

Fig. 12　System operating time of different k values

(2)构造包含不同时间段速度分布与离散程度等信息的特征向量模型，提出多投票 K 近邻算法(MV-KNN)对特性向量进行有监督分类标记可识别道路的限速信息，算法具有良好的系统识别准确率与处理性能。系列试验表明，当训练样本达到 1 200 个时，算法的系统识别准确率最高可达 93%，在仅有 150 个的小训练样本下，算法的系统识别准确率也达 75%。系统具有近线性的处理性能，其处理 10^6 条道路的限速信息仅用时 46 ms，可支持大规模道路的限速信息识别提取与动态更新。

(3)本文将所有车辆统一建模，但考虑到不同车辆类型具有不同的驾驶速度特性，如能进一步根据车辆类型分类，并对不同类别的车辆单独建模，对其特征赋予不同的权重，可有望进一步提高识别准确率，这也是下一步研究的方向。

参 考 文 献：

References：

[1] 姜　康，张梦雅，陈一锴. 山区圆曲线路段半挂汽车列车行驶安全性分析[J]. 交通运输工程学报，2015，15(3)：109-117.
JIANG Kang, ZHANG Meng-ya, CHEN Yi-kai. Driving safety analysis of semi-trailer train at circular curve section in mountain area[J]. Journal of Traffic and Transportation

Engineering, 2015, 15(3): 109-117. (in Chinese)

[2] AARTS L, SCHAGEN I V. Driving speed and the risk of road crashes: a review[J]. Accident Analysis and Prevention, 2006, 38(2): 215-224.

[3] QUDDUS M. Exploring the relationship between average speed, speed variation, and accident rates using spatial statistical models and GIS[J]. Journal of Transportation Safety and Security, 2013, 5(1): 27-45.

[4] BELLA F. Driving simulator for speed research on two-lane rural roads[J]. Accident Analysis and Prevention, 2008, 40(3): 1078-1087.

[5] HOSSEINLOU M H, KHEYRABADI S A, ZOLFAGHARI A. Determining optimal speed limits in traffic networks[J]. IATSS Research, 2015, 39(1): 36-41.

[6] SUN Rui, HU Jian-ming, XIE Xu-dong, et al. Variable speed limit design to relieve traffic congestion based on cooperative vehicle infrastructure system[J]. Procedia-Social and Behavioral Sciences, 2014, 138: 427-438.

[7] HEYDECKER B G, ADDISON J D. Analysis and modelling of traffic flow under variable speed limits[J]. Transportation Research Part C: Emerging Technologies, 2011, 19(2): 206-217.

[8] LI Zhi-bin, LI Ye, LIU Pan, et al. Development of a variable speed limit strategy to reduce secondary collision risks during inclement weathers[J]. Accident Analysis and Prevention, 2014, 72: 134-145.

[9] GRUMERT E, MA Xiao-liang, TAPANI A. Analysis of a cooperative variable speed limit system using microscopic traffic simulation[J]. Transportation Research Part C: Emerging Technologies, 2015, 52: 173-186.

[10] SOUANI C, FAIEDH H, BESBES K. Efficient algorithm for automatic road sign recognition and its hardware implementation[J]. Journal of Real-Time Image Processing, 2014, 9(1): 79-93.

[11] 王 进,孙开伟,李钟浩. 超网络道路限速标志识别[J]. 小型微型计算机系统,2012,33(12):2709-2714.
WANG Jin, SUN Kai-wei, LEE C H. Hypernetworks for road speed limit sign recognition[J]. Journal of Chinese Computer Systems, 2012, 33(12): 2709-2714. (in Chinese)

[12] STALLKAMP J, SCHLIPSING M, SALMEN J, et al. Man vs. computer: benchmarking machine learning algorithms for traffic sign recognition[J]. Neural Networks, 2012, 32(2): 323-332.

[13] GREENHALGH J, MIRMEHDI M. Real-time detection and recognition of road traffic signs[J]. IEEE Transactions on Intelligent Transportation Systems, 2012, 13(4): 1498-1506.

[14] LILLO-CASTELLANO J M, MORA-JIMÉNEZ I, FIGUERA-POZUELO C, et al. Traffic sign segmentation and classification using statistical learning methods[J]. Neurocomputing, 2015, 153: 286-299.

[15] ZAKLOUTA F, STANCIULESCU B. Real-time traffic sign recognition in three stages[J]. Robotics and Autonomous Systems, 2014, 62(1): 16-24.

[16] LIU Hua-ping, LIU Yu-long, SUN Fu-chun. Traffic sign recognition using group sparse coding[J]. Information Sciences, 2014, 266(10): 75-89.

[17] PASCALE A, DEFLORIO F, NICOLI M, et al. Motorway speed pattern identification from floating vehicle data for freight applications[J]. Transportation Research Part C: Emerging Technologies, 2015, 51: 104-119.

[18] 吴佩莉,刘奎恩,郝身刚,等. 基于浮动车数据的快速交通拥堵监控[J]. 计算机研究与发展,2015,51(1):189-198.
WU Pei-li, LIU Kui-en, HAO Shen-gang, et al. Rapid traffic congestion monitoring based on floating car data[J]. Journal of Computer Research and Development, 2015, 51(1): 189-198. (in Chinese)

[19] WALKER G, CALVERT M. Driver behaviour at roadworks[J]. Applied Ergonomics, 2015, 51: 18-29.

[20] RAHMANI M, JENELIUS E, KOUTSOPOULOS H N. Non-parametric estimation of route travel time distributions from low-frequency floating car data[J]. Transportation Research Part C: Emerging Technologies, 2015, 58: 343-362.

[21] RAHMANI M, KOUTSOPOULOS H N. Path inference from sparse floating car data for urban networks[J]. Transportation Research Part C: Emerging Technologies, 2013, 30(5): 41-54.

[22] JIMÉNEZ-MEZA A, ARÁMBURO-LIZÁRRAGA J, FUENTE E. Framework for estimating travel time, distance, speed, and street segment level of service (LOS), based on GPS data[J]. Procedia Technology, 2013, 7(4): 61-70.

[23] ALJANAHI A A M, RHODES A H, METCALFE A V. Speed, speed limits and road traffic accidents under free flow conditions[J]. Accident Analysis and Prevention, 1999, 31(1): 161-168.

[24] CHEN Bi-yu, YUAN Hui, LI Qing-quan, et al. Map-matching algorithm for large-scale low-frequency floating car data[J]. International Journal of Geographical Information Science, 2014, 28(1): 22-38.

[25] 王美玲,程 林. 浮动车地图匹配算法研究[J]. 测绘学报,2012,41(1):133-138.
WANG Mei-ling, CHENG Lin. Study on map-matching algorithm for floating car[J]. Acta Geodaetica et Cartographica Sinica, 2012, 41(1): 133-138. (in Chinese)

[26] BRIN S, PAGE L. Reprint of: the anatomy of a large-scale hypertextual web search engine[J]. Computer Networks, 2012, 56(18): 3825-3833.

[27] BIJALWAN V, KUMAR V, KUMARI P, et al. KNN based machine learning approach for text and document mining[J]. International Journal of Database Theory and Application, 2014, 7(1): 61-70.

[28] JIANG Sheng-yi, PANG Guan-song, WU Mei-ling, et al. An improved *k*-nearest-neighbor algorithm for text categorization[J]. Expert Systems with Applications, 2012, 39(1): 1503-1509.

[29] LIU Hua-wen, ZHANG Shi-chao. Noisy data elimination using mutual *k*-nearest neighbor for classification mining[J]. The Journal of Systems and Software, 2012, 85(5): 1067-1074.