基于手机信令数据的快递人员辨识方法

方珊珊[1]，陈艳艳[1]，刘小明[1]，魏攀一[2]，赖见辉[1]
(1. 北京工业大学北京市交通工程重点实验室，北京 100124; 2. 交通运输部公路科学研究院，北京 100088)

摘 要： 提出一种基于朴素贝叶斯分类法(naive Bayesian classifier，NBC)的城市快递人员辨识方法. 首先，通过相关问卷调查，研究快递派送人员的手机信令发生规则. 然后，依据北京市移动用户手机通信信令数据，利用问卷调查数据和手机信令数据2种数据源中同时包含的通信数据属性，建立通信数据与调查数据中类别变量(快递人员/非快递人员)之间的贝叶斯概率关系，以此为基础构建NBC模型并对其进行训练. 最后，使用未参与训练的样本数据测试标定后模型的准确性，测试结果显示快递人员的预测成功率达到88.3%. 结果表明：该方法具有较高的精度，可以满足实际应用需求.

关键词： 城市配送；快递人员识别；朴素贝叶斯分类法；手机信令数据
中图分类号： U 461 **文献标志码：** A **文章编号：** 0254－0037(2017)03－0413－09
doi： 10.11936/bjutxb2016070035

Identification of City Couriers Based on Mobile Phone Data

FANG Shanshan[1], CHEN Yanyan[1], LIU Xiaoming[1], WEI Panyi[2], LAI Jianhui[1]
(1. Beijing Key Laboratory of Traffic Engineering, Beijing University of Technology, Beijing 100124, China;
2. Research Institute of Highway Ministry of Transport, Beijing 100088, China)

Abstract: An identification method of urban express based on the naive Bayesian classifier (NBC) was proposed in this paper. Firstly, the rules of express delivery personnel phone signaling was researched. And then, based on the mobile phone communication signaling data in Beijing, the Bayesian probabilistic relations were established between the interviewer's category variable (couriers/non-couriers) and the bus travel information which were contained in both questionnaire and mobile phone signaling data. On the basis of this, the Bayesian model was constructed and its training was carried out. Finally, the accuracy of the calibrated model was tested by using the sample data which was not involved in training, and the test showed that the success rate of the courier prediction reached 88.3%. It is shown that the method has high accuracy, which can meet the demand of practical application.

Key words: logistics engineering; courier identification; naive Bayesian classifier (NBC); mobile phone signaling data

信息化浪潮使得电商物流迅猛发展. 2014年，我国快递业务量首次突破130亿，跃居世界第一，实现全年业务收入2 045.4亿元，人均快递量达到10.2件，人均快递支出达到149.5元[1-2]. 快递单量的飞速增长给城市快递网点的合理布局、快递人员的合理配备及快递行业发展战略调整都提出了新的要求[3]. 快递运输车辆的快速激增对城市交通秩序也带来了一定的冲击. 研究城市内部快递单量时空分布特征，有助于快递企业合理规划网点布局、配备快递人员、及时调整企业发展策略，能为交通部门实施相关交通管制政策以减少快递车辆对交通秩序的影响提供一定的数据支撑[4].

收稿日期： 2016-07-13
基金项目： 北京市自然科学基金资助项目(8131001)；湖北省交通运输厅科技项目(2014721311)
作者简介： 方珊珊(1983—)，女，博士研究生，主要从事城市物流方面的研究，E-mail: sabrina0820@126.com

手机作为快递人员通知客户取件的移动终端，已经成为快递人员最基本的工作装备，每天大量的话单数据是快递人员与非快递人员区别的显著特征. 而快递人员的话单数据辨识和分析是研究城市快递单量时空分布特征的前提. 目前，国内外已经有大量与基于移动通信信令数据的人群识别相关的研究[5-7]，主要针对人口出行分布调查、实时人流量统计、出行方式识别、职住空间分析等方面[8-10]，而针对快递人员辨识方法的研究尚无.

本文以北京市移动通信信令数据为依托，结合相关问卷调查数据，提出一种基于朴素贝叶斯分类法(naive bayesian classifier，NBC)的快递人员辨识方法，为城市物流配送人员活动规律研究、城市内部快递单量分布特征研究等提供了一定的数据分析基础.

1 移动信令数据介绍

1.1 数据描述及预处理

移动信令数据主要包含了 4 个信息：1) 匿名用户编号 MSID. 相当于手机号码，但取代手机号码保护了个人隐私，该字段是识别移动用户的唯一编码. 2) 时间戳. 记录信令发生的时间. 3) 基站小区编号. 信令事件发生时所在的基站小区. 4) 信令事件类型. 记录了用户信令属性，如发短信、接短信、打电话、接电话、正常位置更新、小区切换等. 手机用户在出行过程中，基站小区的切换，以及其他的信令动作，都会被系统采集到. 这些手机信令记录了用户的出行轨迹点. 通过这些信令特征和轨迹点可以对用户的社会属性和出行行为进行分析.

本文所用的数据为 2016 年 3 月份北京市移动用户所产生的全部信令数据. 为保证数据分析的高效性和后期模型训练的准确性，对样本数据做如下预处理：

1) 完整自然周的通信信令数据提取. 为对模型计算结果进行对比分析，将研究的总体样本划分为 4 个自然周，并添加“是否为工作日”的识别字段.

2) 按照城市用地性质划分标准，根据信令数据中基站编号，识别事件发生位置区，为信令数据添加用地性质字段.

3) 无效伪码数据的清洗. 在信令数据的产生过程中，会出现数据库系统信息接收错误的情况，造成该条记录中手机伪码被默认为“全 0”，需要将这部分无效卡号进行筛选和清洗.

1.2 快递人员移动信令特征

快递人员在包裹配送过程中完全依靠移动终端与客户取得联系，因此无论是打电话或是发短信，其信令产生的频率与非快递人员均有一定的差别. 同时由于快递人员工作时间主要集中在白天，且同一快递员的服务点位一般相对固定，因此，快递人员与非快递人员相比其信令高频产生的分布时段及分布地点差异性较大.

按照上述分析，本文取“用户全天打电话平均次数”“全天高频打电话的时段分布”“全天发短信平均数量”“全天高频发短信的时段分布”“全天高密度打电话或发短信的点位个数”5 个属性作为区分快递人员与非快递人员的属性特征值. 并通过对北京市 5 种不同类型用地性质(生活住宅、文化教育、体育娱乐、行政办公、商业办公)配送点位 7 家快递公司(邮政 EMS、顺丰、申通、中通、圆通、韵达、宅急送)快递人员的调查，统计其以上 5 个属性特征值一周 7 天的分布概况，结果显示如图 1 ~ 图 5 所示.

短信是快递人员与客户取得联系通知客户取件的首选联系方式，一般情况下快件到达时，快递人员至少给取件人发送 1 条短信. 当超出通知时间取件人依然没有取件时，则快递人员会发出第 2 条短信或者打电话通知取件人.

从图 1 可以看出，北京市约有 50% 的快递人员全天短信发送量在 300 条以上，98% 以上的快递人员全天短信发送量在 50 条以上. 这一特征与非快递人员(除电话推销等特殊人群外)区别较为明显，可以作为区分快递人员与非快递人员的属性条件之一.

图 2 显示了快递人员全天话务量的分布，由图可以看出，约有 50% 的快递人员全天平均话务量在 35 单以上，约 97% 的快递人员全天话务量在 20 单以上. 这一结果与短信发送量差距较大，仅为快递人员全天短信发出量的 10% 左右. 造成这种现象的原因为打电话一般为快递人员通知货主取件的辅助手段，当短信通知 1 次以上，货主依然没在通知时间内取走快件，快递人员才会选择电话通知. 但是平均全天 20 单的话务量，对于大多数城镇居民来说已超出正常生活中的日均话单量. 因此，日均话务量可以作为区分快递人员与非快递人员的属性条件之一.

图 3 显示了 200 名快递人员全天所有短信发出的时间分布，可看出分布结果具有明显的高峰时段和低峰时段. 其中上午 9:00—10:00 平均短信发送量约 200 条，占全天短信发送量的 47%；下午 16:00—17:00 平均短信发送量约 90 条，占全天短信发送量的 21%. 2 个高峰时段短信发送总量约占全天

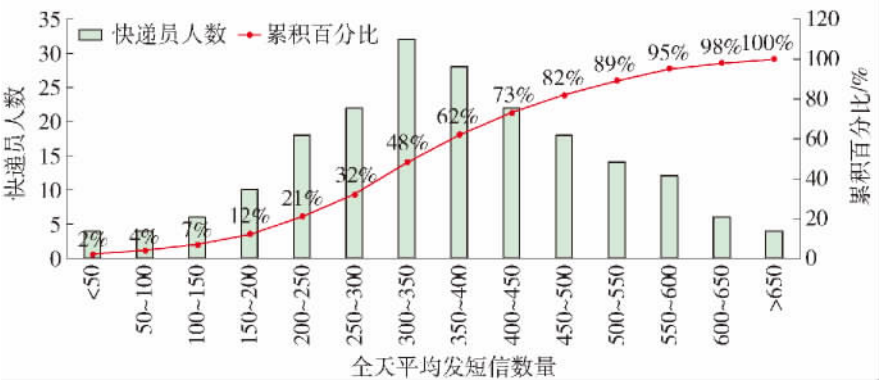

图 1　北京市快递人员全天平均发短信数量

Fig. 1　Short-message amount of couriers in Beijing

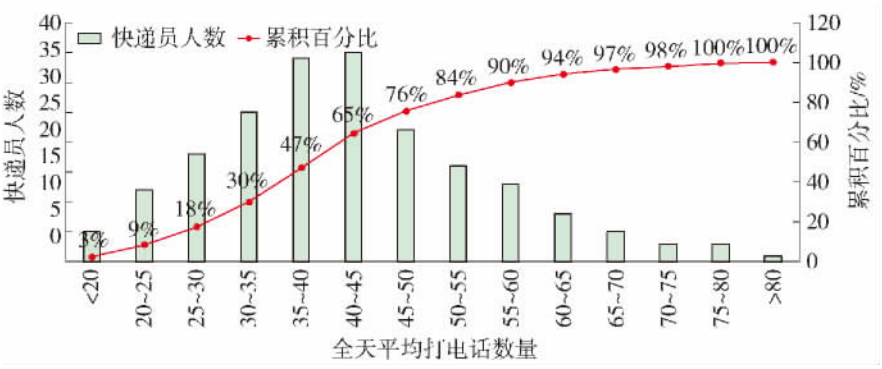

图 2　北京市快递人员全天平均打电话数量

Fig. 2　Call amount of couriers in Beijing

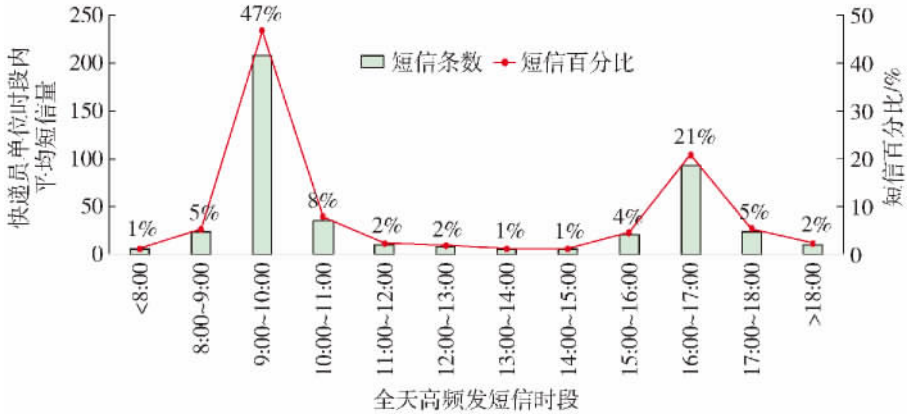

图 3　北京市快递人员全天高频发短信时段分布

Fig. 3　Time distribution of couriers' high frequency texting in Beijing

短信发送总量的 68%. 产生这种现象的原因为快递公司的到货一般分为上下午 2 个批次，早上 9:00 之前一般进行货物分拣和末端配送，9:00 之后快递员才能到达配送小区发货点位，9:00—10:00 为通知货主取件的高峰时段. 同时在上午发送短信的对象中包含前一日通知之后没有及时取件的客户. 而下午的到货时间一般为 14:00—15:00，短信通知取件的高峰时段为 16:00—17:00. 因此，这种明显的短信发送时间峰值特征可以作为区分快递人员与非快递人员的属性条件之一.

图 4 显示了 200 名快递人员打电话的时间分布，与发出短信的时间分布特征类似，具有明显高峰时段和低峰时段，分别为上午 10:00—11:00 和下午 16:00—17:00，这 2 个时间段内的快递员拨打电话的总

量占全天话务总量的 48%. 但是与发短信的时间分布特征不同,拨打电话的上午峰值时段比发送短信的峰值时段滞后 1 h. 拨打电话的下午峰值要高于上午峰值. 这是由于短信一般是快递员通知货主取件的首选联系方式,在上午 9:00—10:00 时段内快递员会优先以短信方式通知当天到达快件的货主取货,当天快件货主通知完后再以短信或电话形式通知往日积压快件货主. 因此,在上午时段内以电话通知的快件多为往日积压没有及时取走的快件. 而为了尽量避免快件的隔夜配送,当上午通知的货主没有按时取走快件时,多数快递员会直接在下午以电话通知货主取件. 因此,高频拨打电话的时段分布可以作为区分快递员与非快递员的属性之一.

图 5 显示了快递员全天高频打电话或发短信所处位置区个数的分布状况. 即以快递员定点配送快件的区域作为信令高频发出的聚焦点,能表示快递员全天主要活跃区域的个数. 由图可以看出,全天约有 80% 的快递人员活跃区域大于或等于 3 个. 而白领、上班族或电话推销人员其全天发出信令的聚焦区域多为 1 个,因此,全天高频发出信令的聚焦点个数可以作为区分快递人员与非快递人员的属性特征之一.

图 4 北京市快递人员全天高频打电话时段分布

Fig. 4 Time distribution of couriers' high frequency calling in Beijing

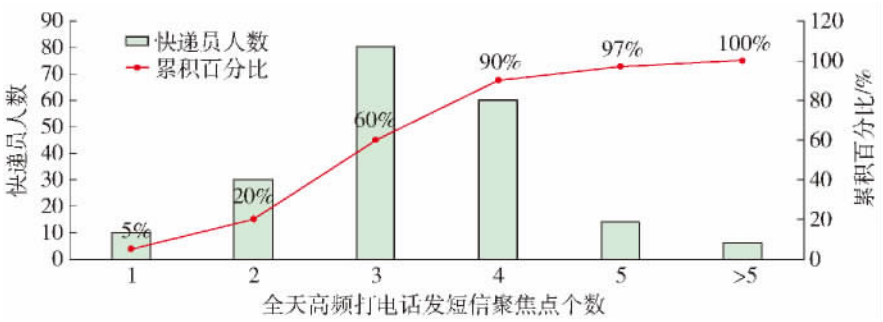

图 5 北京市快递员全天高频发出信令聚焦点个数

Fig. 5 Service points number of couriers' high frequency mobile communication in Beijing

根据上述分析,快递人员全天所发出的移动信令在话务数量、短信数量、高频打电话时段分布、高频发短信时段分布、高频发出信令聚焦点个数等方面均与非快递人员具有较大的差异. 但由于各属性指标均存在不同程度的边缘模糊区域,如果单凭一个属性指标对快递人员进行辨识,会产生较大误差. 为了更加精确地从移动通信信令中辨识出快递人员,本文利用上述 5 个属性值提出一种基于朴素贝叶斯分类法(NBC) 的快递人员辨识方法.

2 移动信令数据中快递人员辨识方法

2.1 朴素贝叶斯分类法(NBC)

数据分类是大数据挖掘中一项非常重要的基础工作,高质量的分类是做好数据统计和分析的前提. 分类是描述重要数据类的模型,这种模型称为分类器[11]. 分类器的构造方法有很多,常见的有决策树法、遗传算法、贝叶斯分类法、神经网络法、粗糙集理论法等.

贝叶斯分类法用于预测类隶属关系的概率，包括朴素贝叶斯分类法和贝叶斯网分类法，属于统计学分类方法. 分类算法的比较研究发现，朴素贝叶斯分类法能进行自我监督和推理，克服了基于规则的系统所具有的许多概念和计算上的困难，分类的高准确率和高速度，在大数据分析中得到了广泛的应用.

如图 6 所示，假设 $U=\{X,C\}$ 是随机变量有限集，其中 n 维属性变量 $X=\{X_1,\cdots,X_n\}$，$C=\{C_1,\cdots,C_n\}$ 代表变量的类别，样本 $x_i=\{x_1,\cdots,x_n\}$ 属于 c_i 的概率由贝叶斯公式可表示为

$$P(C=c_j\mid X=x_i)=\frac{P(c_j)\,P(x_1,\cdots,x_n\mid c_j)}{P(x_1,\cdots,x_n)}=\alpha P(c_j)\,P(x_1,\cdots,x_n\mid c_j)\tag{1}$$

式中：α 为正则化因子，$\alpha=\frac{1}{P(x_1,\cdots,x_n)}$；$P(c_j)$ 为 c_j 的先验概率；$P(x_1,\cdots,x_n\mid c_j)$ 为 c_j 关于 x_i 的似然.

由概率的链式法则，式(1)可以表示为

$$P(c_j\mid x_1,\cdots,x_n)=\alpha P(c_j)\prod_{i=1}^{n}P(x_i\mid x_1,\cdots,x_{i-1},\cdots,x_n,c_j)\tag{2}$$

因为手机信令数据是具有多重属性的数据集，为了降低 $P(x/c_j)$ 的计算量可以做“类条件独立”的朴素假设，即每个属性 X_i 只与类变量 C 相关，因此式(2)中的 $P(x_i\mid x_1,\cdots,x_{i-1},\cdots,x_n,c_j)$ 可以简化为 $P(x_i\mid c_j)$. 即

$$P(c_j\mid x_1,\cdots,x_n)=\alpha P(c_j)\prod_{i=1}^{n}P(x_i\mid c_j)\tag{3}$$

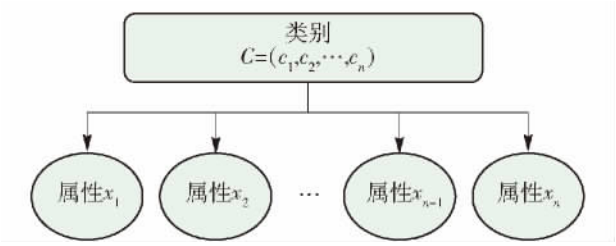

图 6　朴素贝叶斯分类器结构图

Fig. 6　Structure of Naïve Bayesian classifier (NBC)

对分类器进行分类训练时，朴素贝叶斯模型先按照类标签把训练样本集分成几个训练子样本集 $D_j(j\geqslant 1)$，并利用训练样本对每个类的先验概率进行预估，然后在每一个由 C_j 标定的子集样本中，对类条件属性的概率进行估测. 当对一个未知样本 X 的类别进行预测时，可对每个类别 c_j 计算相应的 $P(X=x\mid C=c_j)\,P(C=c_j)$，当且仅当对于 $j\geqslant 1,j\neq i$ 时，

$$P(C=c_i\mid X)>P(C=c_j\mid X)\tag{4}$$

朴素贝叶斯分类法预测 X 属于 C_i，$P(C=c_i\mid X)$ 最大的类 C_i 称为最大后验假设. 对样本 X 的类别进行判别的过程中，需要将类别已知的样本总体均分为 2 份，根据样本的属性特征数据，先对其中一份样本的概率分布值训练估计，然后利用另外一份未参与训练的样本对模型预测的准确性进行测试，并根据测试结果对模型进行修正. 最后利用式(4)即可计算未知集合 X 最大后验概率的从属类别.

2.2　基于问卷调查数据与移动信令数据的 NBC 模型构建

本文采用调查问卷与移动信令数据相结合的方式，将 NBC 模型应用于快递人员的辨识研究中. 利用问卷和移动信令数据中同时包含的基本属性，例如全天发送短信数量、短信发送时段分布、高频发送信令数据的聚焦点个数等，建立基本属性集合 X 与问卷中独有的属性 C(快递人员/非快递人员)之间的贝叶斯概率关系，并通过计算得到的概率值赋予移动信令数据中缺失的“快递人员”C 属性，见图 7.

本研究拟在调查问卷中对受访者是否为快递人员，以及过去一周(含周末)的手机通信频率等相关信息进行采集，并通过 5 个属性变量构建贝叶斯概率模型中的 X 属性集合($X=\{x_1,x_2,x_3,x_4,x_5\}$)：其中：$x_1$ 为平均每天发送短信数量；x_2 为每天高频发送短信时段(每小时发送 20 条以上)；x_3 为平均每天拨打电话数量；x_4 为每天高频拨打电话时段(每小时拨打 10 个以上)；x_5 为平均每天高频发短信或打电话所在位置区域个数. 以上信息均以过去一周为单位进行统计. 是否为快递人员用分类变量 C 表示($C=\{c_1,c_2\}$)，其中：c_1 为该受访者是快递人员；c_2 为该受访者非快递人员. 在手机信令数据中，对总体样本每个手机号伪码一周内平均每天发送短信数量、每天高频发送短信时段、平均每天拨打电话数量、每天高频拨打电话时段以及平均每天高频发短信或打电话所在位置区域个数进行统计，分别对应问卷调查数据中 X 集合中的 5 个属性变量.

2.3　基于问卷调查数据的 NBC 模型训练与测试

本文于 2016 年 3 月份在北京市选取生活住宅、文化教育、体育娱乐、行政办公、商业办公等 5 个不同用地性质的代表区域(包含城市生活中居民日常所接触的领域)，对 1 000 名受访者进行了调查(每种类型区域调查 200 名受访者，其中快递人员 100 名，非快递人员 100 名)，调查内容包括：个人年龄、社会经济属性、工作类型、过去一周内平均每天发送短信数量、全天 8:00—18:00 单位小时发送短信数

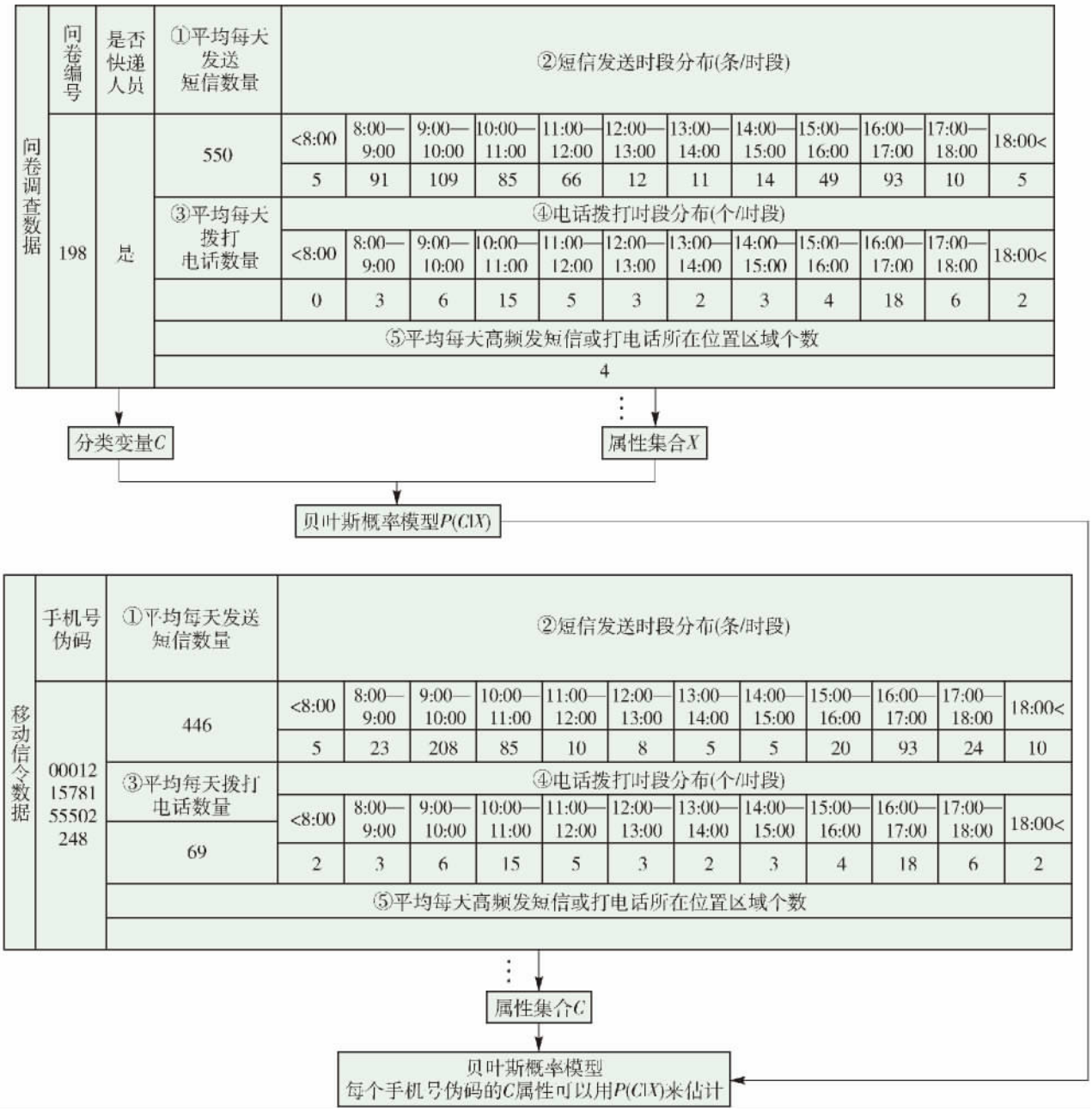

图 7 基于 NBC 模型的快递人员辨识图

Fig. 7 Identification of couriers in mobile phone signaling data based on NBC model

量分布、平均每天拨打电话数量、全天 8:00—18:00 单位小时拨打电话数量分布、平均每天高频发短信或打电话所在位置区域个数等，共计回收有效问卷 980 份(其中快递人员 494 份，约占样本总量的 50.4%)，样本有效份数满足在 95% 置信水平，最大允许绝对误差为 4% 下的样本量需求，可靠性统计量指标值大于 0.8，调查结果信度较高. 由于不同用地类型间快递员移动信令数据分布特征差异较大，因此，按照数据预处理中为信令记录所添加的用地性质字段(与调查所选区域对应，含生活住宅、文化教育、体育娱乐、行政办公、商业办公 5 类)，本文建立了区分用地性质的 NBC 模型.

以商业办公区域为例，将商业办公区回收的 196 份有效调查问卷分为 2 份(每份 98 个样本)，并保证每份样本中快递人员的比例与总体样本相同. 选取其中 1 份样本，根据 2.2 节中对 X 属性集合，以及类别变量 C 的定义结果，获取两者之间的贝叶斯概率关系 $P(X_n|C)$ $(n=1,2,3,4,5)$，如图 8 所示.

如图 8 所示，商业办公区快递人员与非快递人员在 X 属性集合上的概率分布差异较大. 在平均每

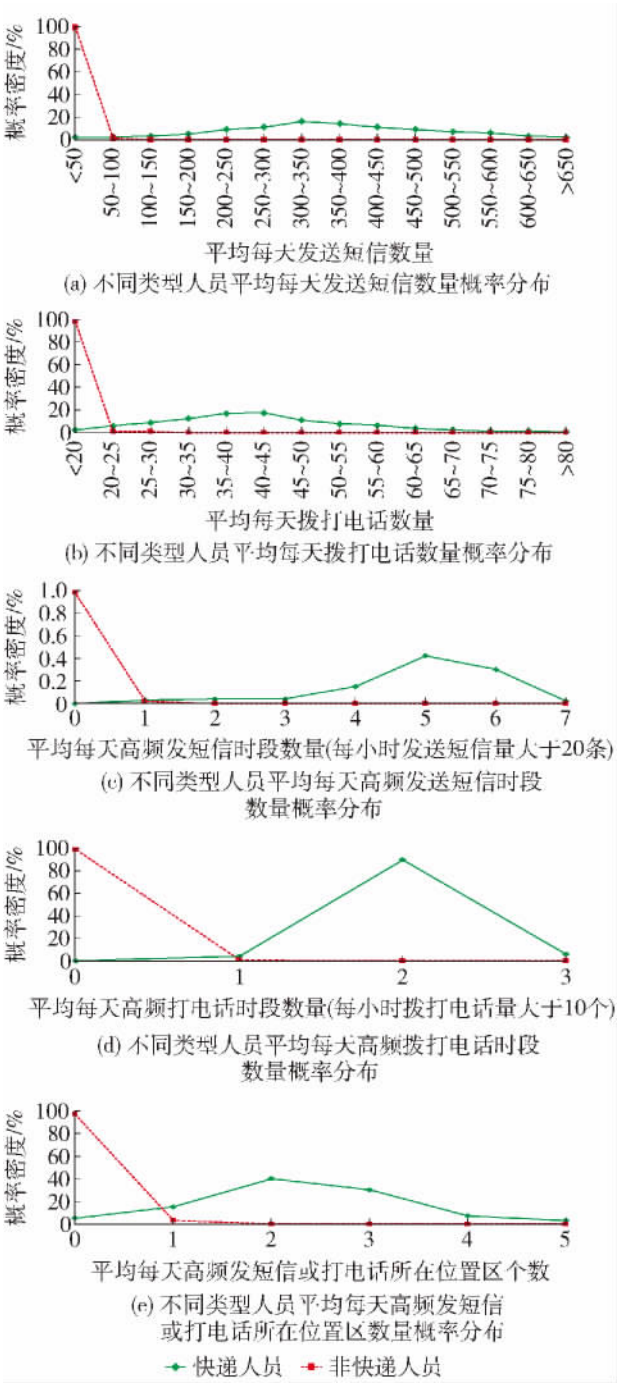

图 8　基于问卷调查数据的 $P(X_n|C)$ 贝叶斯概率分布(商业办公区)

Fig. 8　Bayesian probabilistic distribution of $P(X_n|C)$ based on survey data (commercial office district)

天发送短信和拨打电话数量方面，约 99% 的非快递人员全天发送短信数量都小于 50 条，拨打电话数量少于 20 个，约 98% 的快递人员发送短信数量大于 50 条，拨打电话数量大于 20 个；在高频发送短信或拨打电话时段数量方面，约 99% 的非快递人员全天没有高频拨打电话或发送短信时段，约 93% 的快递人员全天高频发送短信的时段数量在 3 个以上，约 96% 的快递人员全天高频拨打电话的时段数量在 2 个以上；在高频发短信或打电话所在位置区域个数方面，约 3% 的非快递人员全天在 1 个区域出现高频打电话或发短信现象，其他非快递人员则无高频信令发出的位置区，约 80% 的快递人员全天有 2 个以上的高频信令发出位置区. 上述结果表明：选取的 X 属性集合能很好地区分 2 类人员，适用于本次研究中 NBC 模型的标定.

将标定后的 NBC 模型用于第 2 份样本中人员类别的判断，例如某一受访者其属性集合 $X=\{x_1=135, x_2=34, x_3=2, x_4=2, x_5=3\}$，通过式(3)计算得知其属于快递人员的概率为 $P(C_1|X)=96.4\%$，则预测该受访者属于快递人员，最终将判别预测结果与实际分类情况进行对比，得出该 NBC 模型的预测成功率，如表 1 ~ 5 所示.

表 1　生活住宅区 NBC 模型测试结果

Table 1　Test results of the NBC model in living residential district

参数	快递人员(49 份)	非快递人员(47 份)
成功预测数	42	46
错误预测数	7	1
预测成功率	85.71%	97.87%

表 2　文化教育区 NBC 模型测试结果

Table 2　Test results of the NBC model in cultural education district

参数	快递人员(50 份)	非快递人员(50 份)
成功预测数	45	49
错误预测数	5	1
预测成功率	90.00%	98.00%

表 3　体育娱乐区 NBC 模型测试结果

Table 3　Test results of the NBC model in sports entertaining district

参数	快递人员(49 份)	非快递人员(47 份)
成功预测数	41	45
错误预测数	8	2
预测成功率	83.67%	95.74%

表 4 行政办公区 NBC 模型测试结果

Table 4 Test results of the NBC model in administrative office district

参数	快递人员(50 份)	非快递人员(50 份)
成功预测数	46	43
错误预测数	4	7
预测成功率	92.00%	86.00%

表 5 商业办公区 NBC 模型测试结果

Table 5 Test results of the NBC model in commercial office district

参数	快递人员(49 份)	非快递人员(49 份)
成功预测数	44	47
错误预测数	5	2
预测成功率	89.8%	95.9%

对以上 5 种类型用地性质的预测结果进行统计分析,得出 NBC 模型的平均预测成功率,见表 6.

表 6 NBC 模型总测试结果

Table 6 Test results of the NBC model overall

参数	快递人员(247 份)	非快递人员(243 份)
成功预测数	218	230
错误预测数	29	13
预测成功率	88.3%	94.7%

由汇总结果可以看出,标定后的 NBC 模型对于第 2 份样本人员类别的预测成功率超过 85%,其中对快递人员的预测成功率达到了 88.3%,可以接受该模型的预测结果,模型通过测试.

3 移动信令数据中快递人员的辨识结果

选用北京市 2016 年 3 月 4 个自然周的移动信令数据进行分析,并根据不同运营商号码在北京城区快递员群体中所占比例进行适当扩样,表 7 给出了快递人员的辨识结果(已扩样). 从表中可以看出,4 个自然周中通过 NBC 模型辨识出来的快递人数之间存在一定的波动,但总体波动不大,基本保持在 3% 以内. 而通过与问卷受访快递人员的深入交谈,基本上可以将造成辨识结果波动的原因归纳为以下几个方面:

1) 快递包裹能否及时取走,受天气状况和节假日影响较大. 通常情况下雨天快件领取拖延时间较长,客户一般会等待天气转晴之后再取包裹,因此快递人员可能会打电话或发短信与顾客沟通. 若遇到周末或者节假日,顾客有可能外出游玩,而无法及时领取寄往单位或住宅的快递包裹,此时快递人员会多次发短信或者打电话催促顾客或者与顾客沟通. 造成了数据库中额外增加大量的信令记录,导致模型辨识准确度降低.

2) 受快递员收发快递量影响,当快递员服务区域发出快递量较少,而收取快递量较多时(如高校区域),此时快递员全天发出信令数量较多且时间分布较为集中,模型较易进行判断; 当快递员服务区域发出快递量较多,而收取快递量较少时(如中关村等商业区域),此时快递员需分散较多精力进行寄件处理,快递员全天发出信令数据较少且时间分布较为分散,此时模型对快递人员的识别较为困难,识别准确度降低.

3) 部分快递人员在发送短信的时候利用特定手机软件进行发送,因此在移动信令数据库中不能记录该信令发出的手机号伪码,从而导致模型辨识错误.

表 8 中给出了 NBC 模型的最终判断结果,以及相关指标的统计结果. 最终,共有约 2 万个移动终端用户被判定为快递人员,鉴于标定后 NBC 模型自身有 10% 左右的误差,实际快递人员的数量应介于 1.8 万 ~2.2 万,平均每个快递人员每天发送的短信数量为 269 条,拨打电话的数量为 32 个,平均每天有 2.2 个高频发送短信的时段,有 2.1 个高频拨打电话的时段,平均每天高频发送短信或拨打电话所在位置区数量为 2.5 个.

表 7 NBC 模型在各自然周移动信令数据中的辨识结果

Table 7 Identification results of the NBC model in each week

数据周期	手机伪码数量	平均每天发送短信数量	平均每天拨打电话数量	平均每天高频发送短信的时段数量	平均每天高频拨打电话的时段数量	平均每天高频发短信或打电话所在位置区数量
第 1 周	23 157	275	38	2.1	2.0	3.3
第 2 周	24 564	267	36	2.3	2.2	3.5
第 3 周	21 987	274	33	2.3	2.3	2.3
第 4 周	21 655	281	42	2.0	2.3	2.1

表 8　NBC 模型的最终判断结果

Table 8　Identification results of the NBC model

识别人数	平均每天发送短信数量	平均每天拨打电话数量	平均每天高频发送短信的时段数量	平均每天高频拨打电话的时段数量	平均每天高频发短信或打电话所在位置区数量
20 152	269	32	2.2	2.1	2.5

4　结论

1）本文以北京市移动用户通信信令数据为依托，结合相关问卷调查，研究快递人员手机信令的发生规则，提出一种基于朴素贝叶斯分类法的城市快递人员辨识方法. 本研究利用问卷与信令数据库中同时包含的信令属性信息，建立其与调查问卷中独有的类别变量（快递人员/非快递人员）之间的贝叶斯概率关系，并以此对 NBC 模型进行构建、训练与测试.

2）利用通过测试的 NBC 模型对移动通信信令数据中城市快递人员进行识别. 统计结果显示，北京市快递人员数量在 1.8 万～2.2 万，平均每个快递人员每天发送的短信数量为 269 条，拨打电话的数量为 32 个. 发送短信的主要集中时段为上午 9:00—10:00 和下午 16:00—17:00，拨打电话的主要集中时段为上午 10:00—11:00 和下午 16:00—17:00. 针对同一快件快递人员重复通知的情况，根据信令伪码将 NBC 模型识别出的信令数据进行数据清洗，保留一件快递一条信令的处理结果. 最终结果显示，2016 年 3 月北京市平均每天快件收取量约 161.5 万件，快件配送的高峰小时为上午 9:15—10:15 和下午 15:50—16:50.

参考文献：

[1] 王宝义. 中国快递业发展的区域差异及动态演化[J]. 中国流通经济，2016，30(2)：36-44.
WANG B Y. An analysis on regional disparities and dynamic evolution of express industry development in China [J]. China Business and Market，2016，30(2)：36-44. (in Chinese)

[2] 金玉清. 中国快递业现状及发展对策研究[J]. 商业流通，2013，25(5)：51-53.
JIN Y Q. The present situation and development countermeasure research of express industry in China [J]. The Business Circulate，2013，25(5)：51-53. (in Chinese)

[3] 张雪芹，窦德强，付燕荣，等. 从统计数据看我国快递业的发展现状[J]. 交通与运输，2015(1)：215-217.
ZHANG X Q，DOU D Q，FU Y R. et al. Observing the present state of express industry in China based on the statistical data analysis [J]. Traffic & Transportation，2015(1)：215-217. (in Chinese)

[4] 匡旭娟，荣朝和. 快递企业与专业运输企业合作战略的稳定性分析[J]. 交通运输系统工程与信息，2008，8(5)：21-25.
KUANG X J，RONG C H. Analysis on the stability of cooperative strategy between express company and specialized transportation enterprise [J]. Journal of Transportation Systems Engineering and Information Technology，2008，8(5)：21-25. (in Chinese)

[5] 何兆成，陈展球，范秋明，等. 基于手机基站数据的混合地图匹配算法研究[J]. 交通运输系统工程与信息，2014，14(3)：34-42.
HE Z C，CHEN Z Q，FAN Q M，et al. Hybrid map matching algorithm based on mobile base station data [J]. Journal of Transportation Systems Engineering and Information Technology，2014，14(3)：34-42. (in Chinese)

[6] CAYFORD R，JOHNSON T. Operational parameters affecting the use of anonymous cell phone tracking for generating traffic information [C/CD] // Institute of Transportation Studies for the 82th TRB Annual Meeting. Washington D.C.：Transporation Reseach Board，2003.

[7] YGNACE J L，REMY J G，BOSSEBOEUF J L，et al. Travel time estimates on Rhone corridor network using cellular phones as probes：phase 1 technology assessment and preliminary results [R]. [S.l.]：INRETS，2000.

[8] FISHBEIN M，AJZEN I. Predicting and changing behavior：the reasoned action approach [M]. New York：Taylor & Francis，2011.

[9] BAMBERG S，HUNECKE M，BLÖBAUM A. Social context，personal norms and the use of public transportation：two field studies [J]. Journal of Environmental Psychology，2007，27(3)：190-203.

[10] THOGERSEN J. Norms for environmentally responsible behaviour：an extended taxonomy [J]. Journal of Environmental Psychology，2006，26(4)：247-261.

[11] HAN J W，KAMBER M，PEI J. Data mining concepts and techniques [M]. 3 ed. Beijing：China Machine Press，2012.

（责任编辑　张　蕾）

特约组稿专家简介

丁治明，教授、博士生导师，国务院政府特殊津贴获得者，北京工业大学信息学部副主任、计算机学院院长。中国计算机学会（CCF）大数据专委会委员、CCF数据库专委会委员、CCF电子政务与办公自动化专委会委员、中国指挥与控制学会数据处理与集成专委会委员、中国健康大数据产业技术创新战略联盟(CHBDA)常务理事。曾工作于德国时空数据管理领域的国际顶级专家Ralf Hartmut Güting教授团队；长期在中国科学院系统（计算技术研究所、软件研究所）学习与工作。担任IEEE智能交通系统学会(ITSS)社会交通技术委员会主席、ACM SIGSPATIAL China副主席、担任国际顶级刊物《IEEE Transactions on Intelligent Transportation Systems》《IEEE Intelligent Transportation Systems Magazine》的编委。是中国物联网研究发展中心总体专家组成员、中国科学院“感知中国”先导专项计划总体组成员、工业和信息化部软件与集成电路促进中心（CSIP）专家组成员。担任国家“863”计划领域评审专家、国家自然科学基金评审专家、国家科技奖评审专家。主要研究方向为数据库与知识库系统，时空数据库及移动对象数据库，新型计算环境（云计算、物联网、移动计算、智慧城市等）下的大数据存储、查询与分析技术等。负责和承担了国家或省部级科研项目26项，其中主持的国家自然科学基金重点项目“面向非常规应急管理的物联网技术与系统”结题验收获得“优秀”。近年来，在数据库及大数据处理领域发表论文130余篇、出版学术著作4部（其中英文专著1部）。以第一发明人获授权国家发明专利5项、软件著作权8项。曾获北京市科技进步二等奖、国家优秀科技信息成果奖、中国科学院朱李月华优秀教师奖、中国科学院软件研究所优秀研究生导师奖、北京工业大学优秀教育工作者等荣誉。

陈艳艳，教授、博士生导师，北京工业大学城市交通学院院长，中国公路学会城市交通分会副理事长、交通运输部智能公交行业重点实验室主任、交通运输部城市综合交通协同运行及超级计算协同创新中心常务副主任、交通工程北京市重点实验室及北京城市交通顺畅保障工程技术中心常务副主任、全国智能运输系统标准化技术委员会委员、北京市规划学会常务理事、美国交通研究委员会（TRB）发展中国家分会理事。在交通大数据、城市交通规划与管理、智能交通与系统仿真等领域取得了一系列创新性成果。入选新世纪百千万人才工程及北京市科技新星计划。出版著作6部，发表学术论文100余篇，授权国家发明专利10项。获国家教育部自然科学奖、中国公路学会奖、住房建设部华夏奖、北京市科技进步奖共11项。

贾克斌，教授、博士生导师，国务院政府特殊津贴获得者。现任北京工业大学“211工程”办公室主任兼研究生院副院长。兼任中国自动化学会智能建筑与楼宇自动化专业委员会主任委员、中国生物物理学会分子影像专业委员会常务委员、中国电子教育学会研究生教育分会常务理事、国际期刊《Journal of Information Hiding and Multimedia Signal Processing》副主编。2001年在日本早稻田大学国际信息与通信研究中心做客座研究员，2006年、2008年和2009年分别到英国伯明翰大学、香港理工大学及美国纽约州立大学布法罗大学（UB）进行短期合作研究。2004年以来，先后主持国家“973”项目子课题、国家自然科学基金重点项目、国家科技支撑项目（子课题）、国家重大科技专项（专题）各1项，主持国家基金面上项目4项、北京市自然科学基金项目4项（其中重点项目3项），主持国家教育部、北京市科学技术委员会、国家军工等重点项目共计20余项。在SCI、EI及国内核心期刊发表学术论文250余篇，出版学术专著2部，编著国际会议论文集1册。作为第一申请人申请国家发明专利25项（其中19项已获授权）、申请国际发明专利3项。主要研究方向为图像/视频处理技术、生物智能化信息处理技术、基于Internet的多媒体信息系统。