doi: 10.3969/j.issn.0490-6756.2018.05.010

基于贝叶斯与因果岭回归的物联网流量预测模型

陈 翔[1], 唐俊勇[2]
(1. 西安工业大学建筑工程学院, 西安 710021; 2. 西安工业大学计算机科学与工程学院, 西安 710021)

摘 要: 针对物联网流量预测困难的问题,提出了一种基于贝叶斯与因果岭回归的物联网流量预测模型. 该模型首先根据物联网流量传输波动影响链路变化等因果关系,深入刻画物联网流量局部特征,并利用薛定谔方程优化识别模型,同时结合贝叶斯拟合因果关系联合岭回归方法建立预测模型. 最后,通过仿真实验研究了该模型与其他方法之间的性能状况,结果表明该模型在平均队列、阻塞率和延迟率等方面具有较大优势.

关键词: 物联网; 流量; 预测; 贝叶斯; 因果岭回归
中图分类号: TP393 **文献标识码**: A **文章编号**: 0490-6756(2018)05-0965-06

The flow prediction model in internet of things based on Bayesian and causal ridge regression

CHEN Xiang[1], *TANG Jun-Yong*[2]
(1. School of Civil Engineering, Xi'an Technological University, Xi'an 710021, China;
2. School of Computer Science and Engineering, Xi'an Technological University, Xi'an 710021, China)

Abstract: In order to solve the flow prediction problem of Internet of Things, a flow prediction model is proposed based on Bayesian and causal ridge regression. At first, the local characteristic of flow is deeply depicted considering the causal relationship between the fluctuation of the traffic flow and the change of the link; in addition, Schrodinger equation is used to optimize the recognition model. Then, the prediction model is built with Bayesian and causal ridge regression. Finally, the performance of this model and other methods is studied by simulation experiment. The results show that this model has a great advantage in average queue, blocking rate, delay rate and so on.

Keywords: Internet of Things; Flow; Prediction; Bayesian; Causal ridge regression

1 引 言

随着物联网技术的快速发展,物联网在各行各业迅速普及开来,因此对其服务质量的要求也在不断提高[1-5]. 并且伴随着各类数据数量的指数型增长,导致服务质量降低. 同时物联网与互联网具有明显的不同属性特征,导致产生的流量特性也是多种多样,所以利用传统的基于互联网类型的流量预测模型来刻画物联网流量具有一定困难. 因此,迫切需要构建符合物联网流量特征预测方法,以此提高服务质量.

物联网广泛存在于不同通信链路和终端设备中,其网络时延和带宽都有着不同的要求,因此在不同类型的流量聚合之后其特征可能会产生明显变化,加大了对流量预测的难度. 对此,国内外学者进行了大量研究. 文献[6]提出了基于物联网思想

收稿日期: 2017-09-21
基金项目: 陕西省科技厅工业科技攻关项目(2016GY-088)
作者简介: 陈翔(1977-), 男, 四川蒲江人, 讲师, 硕士, 研究方向为计算机软件与理论 . E-mail: chengxiang1977@hainan.net

的物联网空间科学技术框架,使得物联网领域研究进一步加强.文献[7]提出了基于 MMPP(Markov Modulated Possion Process)的物联网流量模型,通过考虑流量的自相似性和长短相关性,对网络流量进行统计和预测.文献[8]在此基础上对物联网通信特点进行了具体比较,对数据包和约束时长进行比较分析,但是由于对数据的排查还局限于少数应用建立的物联网,针对性不强,不能实现对大型流量的预测.文献[9]针对物联网之间的业务流,通过分析业务流的尺度特征,以此实现对物联网流量的控制和预测,但针对突发型的物联网模型,该方法容易出错,所以利用尺度流量建立预测模型实用性不强.此外,采用射频识别技术实现对目标的自动识别并获取相关数据[10-12],建立物联网基本模型,以此达到对物联网流量的智能预测.但是由于受到系统限制,能够应用并且实现流量认证的比例还有待提高.文献[13,14]提出了物联网的参考模型,利用标准建模语言构建了抽象物联网分析用例模型,对物联网概念有了进一步的分析,但该用例模型建立在完整用例之间,对于不具备完整关联之间的用例定义并不准确.但是,目前提出的物联网流量预测模型复杂度较大,虽然在精度上不断得到改进,但预测时间也增加较大,很难有效满足实时需求.

在上述工作的基础上,本文结合贝叶斯[15,16]和因果岭回归[17,19]提出一种新的物联网流量预测模型.因果岭回归法能对多重共线性问题进行统一分析,根据流量输出的因果关系和贝叶斯方法来刻画数据流量传输特征.该模型刻画物联网流量局部特征,并结合贝叶斯拟合因果关系联合岭回归方法建立预测模型.最后通过仿真实验研究了该模型与其他方法之间的性能状况.

2 物联网流量特征

当链路处于高度繁忙状态时,数据传输可能出现"阻塞"或"时间阻塞",为了刻画物联网流量,本文首先引出链路阻塞率进行刻画.假设随机抽取一个物联网中单独的预测个体 M,为了对预测个体进行信息采集,每个信息集成点为 i、j,设 δ 为目标的感知范围,且保证感知活动只有一种基本趋势,并确定聚合周期时间序列 S 和事件驱动时间序列 S',根据聚合周期性流量变化和事件驱动性流量变化,提出不同种类流量的阻塞率 e,如下式.

$$e(M_{i,j}) = \frac{f(S_M) - f(S'_M)}{\sum_{i=1}^{k}(f(S_0) - f(S'_{M_i})) + \sum_{j=1}^{l} f(S) - f(S'_{M_J})} \tag{1}$$

当网络流量大量聚合时,一般会形成缓慢的排队情况,而传输链路 k 的数量和链路长度 l 会影响阻塞率,对于链路数量和链路长度决定的低阻塞情况,使用平均队列时延反应缓冲区阻塞状态.令 $S'_{\max}$ 表示流量时延上限值;$F(x)$ 是流量集合函数;$o(f)$ 是对应 f 的功能复杂度评价值;a 和 b 是常参数.那么其平均队列时延 E 可表示为

$$E = 1 - \gamma\left[0.2\left(\frac{S(x)}{S'_{\max}}\right)^{\alpha} + 0.8\frac{\sum_{f\in F(x)} b^{o(f)}}{b^{5}\,|\,F(x)}\right] \tag{2}$$

同时,针对高阻塞链路,结合流量的自相似性来刻画流量变化模型,网络的骨干网和接入网具有不同的流量密度,流量聚合时,稀疏程度不同的流量会发生抖动,利用抖动率进行刻画.同时,结合流量的自相似性来刻画流量变化模型.对样本 N 进行随机采集 m 个有效节点.令节点 O 相邻节点的自相似系数为 H,根据式(3)对具有线性关系的目标求得基于自相似的流量抖动率 φ.

$$\varphi(O) = \frac{1}{\frac{N}{m}}\sum_{k=1}^{\frac{N}{m}}(X^{(m)}(k))^2 - (\frac{1}{\frac{N}{m}}\sum_{k=1}^{\frac{N}{m}}(k))^2 H \quad (0 < H < 1) \tag{3}$$

如果通信量不断增大,流量输出形成的网络节点逐渐增多,流量监测难度加大,仅仅依靠传统模型无法有效刻画物联网数据流量的统计精度.这里基于贝叶斯网络结构,本文构建依赖于流量之间的因果关系的计算方法.假设具有 n 个节点的网络结构框架 f,计算在 i 个有效指令下的节点之间的延迟率 $f(n)$.

$$f(n) = \sum_{i=1}^{n}(-1)^{i+1}\cdot(\frac{n!}{i!}(n-i)!)\cdot 2^{i(n-i)}\cdot f(n-i)\prod_{i=1}^{n} f(n!\mid i!(n-i)!) \tag{4}$$

本文为了具体表现出物联网流量在应用时的变化,通过计算阻塞率、抖动率来描述数据流量,并且基于平均队列时延、节点之间的延迟率来建立流量预测模型.同时本文基于行为特征和因果岭回归分析网络量承载的内容,主要对内容进行深层次滤

扫描，实施内部监管和流量控制. 令 $(\overrightarrow{g},\overrightarrow{\varphi},\theta^2)$ 是时间序列的参数组，p、q 分别是定阶的最佳阶数，(h,k)是两个时间序列的两个随机点，建立初步模型 Ψ 分别对 $\hat{S}_t$(有效值)、$\hat{S}'_t$(参考值)进行流量预测：

$$\Psi(S,S') = -\sum_{k=1}^{\infty} g(k)\hat{S}_t(h-k) + \sum_{i=1}^{p}\varphi(i)\hat{S}'_t(h-i) + \sum_{i=1}^{q}\theta^2(i)a_{t+h-j} \tag{5}$$

3　预测模型

不同种类流量聚合时，易发生堵塞情况，即便聚合具有周期性，也很难进行预测；事件驱动性物联网流量在应用时容易离散，所以大多流量发出离散信号不便感知，且应用于物联网的流量变化速度快并且难以捕捉其特点，因此利用形态特征对物联网流量进行抽象定义. 岭回归法通过选择自变量的一个子集产生新的线性模型，能对多重共线性问题进行统一分析，根据流量输出的因果关系. 但是由于它是一个离散过程，使得子集选择方法表现出高方差，因此需要结合其他方法来降低整个模型的预测误差. 而贝叶斯可以通过选择先验分布，将岭回归作为后验分布均值输出，提高流量刻画精度. 因此，可以结合贝叶斯和因果岭回归来刻画数据流量传输特征. 该模型首先利用感知层对有效网络点进行采集，根据时间序列计算形态特征，并将计算的有效值和参照值进行对比. 同时根据薛定谔方程优化流量尺度特性，计算流量抖动率和延迟率，实施内部监管和流量控制，以此实现对形态特征更新预测.

本文为了刻画出流量变化抽象情况，在空间范围设置(k,t,y)坐标系，利用提出的时间序列 S(有效值)和 S' (参照值)，对比时间序列形态特征表示为

$$S = \{(k_1,t_1,y_1),(k_2,t_2,y_2),\cdots,(k_n,t_n,y_n)\} \tag{6}$$

$$S' = \{(k'_1,t'_1,y'_1),(k'_2,t'_2,y'_2),\cdots,(k'_n,t'_n,y'_n)\} \tag{7}$$

其中，i 代表样本计数，$i = 1,2,\cdots,n$，k_i 代表聚合周期链路数量；k'_i 代表事件驱动链路数目；y_i 表示为聚合周期时间序列的第 i 段曲线的幅度值；y' 表示为事件驱动时间序列第 i 段曲线的幅度值；t_i 是聚合周期时间序列的第 i 刻的时间点；t'_i 是事件驱动时间序列第 i 刻的时间点.

为了对预测节点采集的信息形态特征进一步计算，首先将时间序列进行平稳处理，根据延迟率 $f(n)$，基于因果强度关系提出关于不同时间序列的加权参数 D，n 是节点数；t 是时间序列的指数；W_i 是指聚合后的流量链路第 i 段的波动权值.

$$D(S,S') = \frac{1}{t_n}\sum_{i=1}^{n}\Delta t_i W_i |(k_i - k'_i)| \cdot f(n),\quad i = 1,2,\cdots,n \tag{8}$$

同时，通过时间序列参数对样本集合进行平稳化处理，逐渐增加模型叠加算法优化模型的预测精度. 令采集到的流量平稳化之后列为矩阵 $\begin{Bmatrix} x_1y_1 & \cdots & x_ny_1 \\ \cdots & \cdots & \cdots \\ x_1y_n & \cdots & x_ny_n \end{Bmatrix}$，在不同时间序列的加权参数 D 下，考虑到流量在时间序列上具有特定的尺度性，利用薛定谔方程[20]对公式(8)的流量序列进行尺度优化.

$$K(x,y) = \frac{1}{4\pi D}\int e\int \frac{i}{4\pi D}|x-y|^2\varphi(x)\mathrm{d}x\varphi(y)\mathrm{d}y \sum_{x\times y} f(x)\varphi(x)\mathrm{d}x\sum_{x\times y} f(y)\varphi(y)\mathrm{d}y \tag{9}$$

其中，ε 是静态缓冲系数；φ 是尺度函数；g 是多尺度量化处理函数. 经过尺度优化后的流量节点序列更加稳定，具有规律性.

通常，流量预测方法分为线性流量预测和非线性流量预测，基于物联网流量预测适用于具有非线性变化的对等计算，该计算方法不能处理具有一定自适应性的物联网流量变化，自适应性会影响模型的报错程序，对于一些病态应用流量无法捕捉，影响预测结果. 本文提出基于流量之间延迟率 $f(n)$ 的计算方法，令 i 和 j 分别是二维空间的参数，由此可得

$$|f_i(n) - f_j(n)| = ((f_i(n,1) - f_j(n,1)^2) + \cdots + (f_i(n,i) - f_j(n,j))^2)^{1/2} \tag{10}$$

根据上述提方式，这里给出物联网流量预测公式，如下.

$$Z(f_n) = \sum_{m=1,n=1}^{S\cup S'} P(S_m \mid S'_n) f(S'_n) = \sum_{m=1,n=1}^{S\cup S'} \frac{\sum_{m=1,n=1}^{\infty} C(\exists S_m \forall S'_n)}{\sum_{n=1}^{C} N(S'_n)} P(S'_n) \tag{11}$$

其中，C' 表示聚合后的链路数量，$m \in C'$，$n \in C'$；P 是全概率计算公式.

同时,这里结合基于贝叶斯和因果岭强度给出具体算法求解流程:

Step 1 感知层对有效网络点进行采集. 根据式(12)来提高采集数据精度和普遍度:

$$P = \sqrt{\sum_{i=1}^{a} w_{ij}^2 x_{k-j+1} \cdot \alpha^3 \left(\frac{a_{ij}^2 - b_j}{a_j}\right)} \tag{12}$$

同时,令 W_i 为第 i 段的波动权值;k 和 j 为二维空间参量;α 为优化参数;a 和 b 分别是对应的时间序列参数)进行平稳化计算,根据式(13)初始化参数 D.

$$D(a,b) = \frac{1}{t_n}\sum_{i=1}^{n} \Delta t_i W_i \left|(k_i - k_i')\right| \cdot f(n) \tag{13}$$

Step 2 在网络层空间(k, t, j)根据时间序列提出形态特征 S(有效值)和 S'(参照值),将计算的有效值和参照值进行对比达到预测的目的.

$$S = \{(k_1,t_1,j_1),(k_2,t_2,j_2),\cdots,(k_n,t_n,j_n)\} \cdot D(a,b) \tag{14}$$

$$S' = \{(k_1',t_1',j_1'),(k_2',t_2',j_2'),\cdots,(k_n',t_n',j_n')\} \cdot D(a,b) \tag{15}$$

同时根据提出的形态特征计算数据流量的阻塞率 e 和平均队列时延 E.

$$e(M_{i,j}) = \frac{f(S_M) - f(S_M')}{\sum_{i=1}^{k}(f(S_0) - f(S_{M_i}')) + \sum_{j=1}^{l} f(S) - f(S_{M_j}')} \cdot P(M) \tag{16}$$

$$E = 1 - \gamma\left[0.2\left(\frac{S(x)}{S_{\max}'}\right)^{\alpha} + 0.8\frac{\sum_{f\in P(x)} b^{o(p)}}{b^5 \left| P(x)\right.}\right] \tag{17}$$

Step 3 应用层根据薛定谔方程的思想优化流量特定的尺度性,根据公式(9)计算基于自相似系数 H 计算流量抖动率 φ, 并利用传递函数 N 实现频率的不变性,如下式.

$$\varphi(M) = \frac{1}{\frac{N}{m}}\sum_{k=1}^{\frac{N}{m}}(X^{(m)}(k))^2 - \left(\frac{1}{\frac{N}{m}}\sum_{k=1}^{\frac{N}{m}}(k)\right)^2 H \tag{18}$$

$$N = \frac{1}{2}K(x,y)(1 + e^{-ait\left|y\right|^2}) \tag{19}$$

Step 4 通过计算节点 n 之间的延迟率 $f(n)$ 解决网络节点之间形成的计算盲点,如下式.

$$f(n) = \sum_{i=1}^{n}(-1)^{i+1} \cdot \left(\frac{n!}{i!}(n-i)!\right) \cdot 2^{i(n-i)} \cdot f(n-i) \prod_{i=1}^{n} f(n! \mid i!(n-i)!) \tag{20}$$

同时实施内部监管和流量控制,分别对 $\hat{S}_t$、$\hat{S}_t'$ 在时间点 t 进行更新预测,

$$\Psi(S,S') = -\sum_{k=1}^{\infty} g(k)\hat{S}_t(h-k) + \sum_{i=1}^{p}\varphi(i)\hat{S}_t'(h-i) + \sum_{i=1}^{q}\theta(i)a_{t+h-j} \tag{21}$$

Step 5 输出预测结果;

Step 6 仿真结束.

4 数学仿真

为了验证上述基于贝叶斯与因果岭回归建立的预测模型,这里利用 Maltab 中进行仿真实验. 首先在终端上访问物联网应用程序,并用抓包软件捕获数据包 2 万个,根据前 1 万个样本数据,结合本文提出的预测模型以及标准岭回归模型分别来预测后 1 万个数据,同时和实际捕获的数据记性比较,结果如 1 所示. 从图 1 可以发现,本文提出的预测模型所预测的结果更加接近捕获的实际数据. 通过残差分析,本文所提预测模型残差大小为 13.5%,而标准岭回归模型的残差大小为 19.7%.

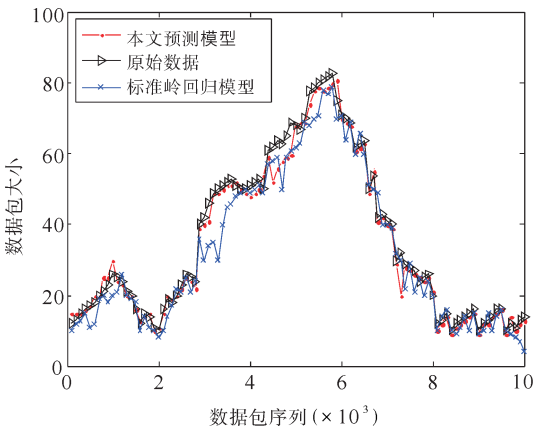

图 1 预测精度比较

Fig. 1 Compare of prediction accuracy

流量特征主要分为平稳和非平稳类,针对平稳型流量,本文利用提出的预测模型对数据特征进行预测. 这里仍然抽取前 1 万个数据作为样本,并去除异常数据使其基本满足平稳型流量特征. 图 2 给出了不同时间下这两种模型预测的平均队列比较结果. 平均队列作为数据流量的一个特征,涉及到时间序列的流量特性. 从图 2 可以看到,在初始阶段,两种算法的预测效果比较接近. 但是随着流量的加剧,标准岭回归模型预测效果逐渐下降,而本文提出的预测模型依然能够有较好适应性.

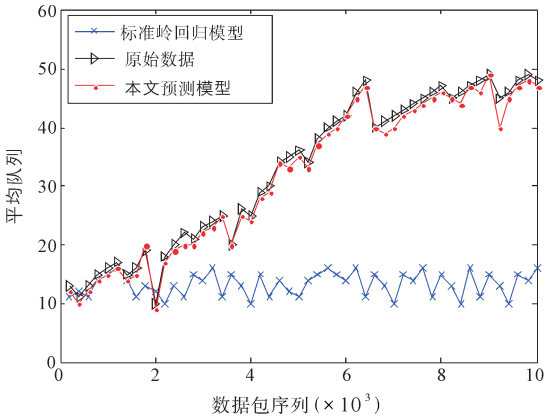

图 2　平均队列比较结果

Fig. 2　Compare of average queue

由于流量序列具有特定的尺度特征，在上述建立的流量预测模型时已对时间序列特定的尺度进行优化. 为了体现对尺度优化之后的性能优势，图 3 给出了本文预测模型、标准岭回归模型以及因果岭回归模型这三种模型的阻塞率比较结果.

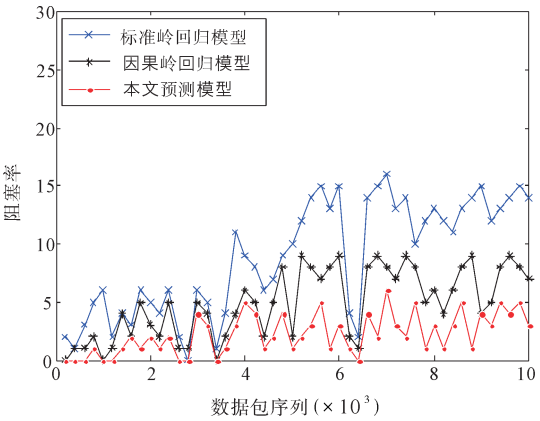

图 3　阻塞率比较结果

Fig. 3　Compare of blocking rate

从图 3 可以看出，阻塞率随着流量增大呈现逐渐递增趋势，标准岭回归模型增加趋势最为明显. 和标准岭回归模型相比，因果岭回归预测模型的阻塞率也会随着流量增加也迅速增加，当使用贝叶斯优化特定的尺度特性后，其阻塞率较为稳定，增长幅度较小，这也体现了本文预测模型的稳定性.

最后，图 4 给出了这三种模型延迟率的比较结果. 标准岭回归模型延迟率在大约数据包达到 2800 左右时开始急剧上升，因果岭回归模型的延迟率大约在 3000 时呈现缓慢上升趋势，而本文提出的预测模型的延迟率较为平稳，维持在 6% 以下. 因此，从延迟率比较结果来看，本文的预测模型具有较大优势.

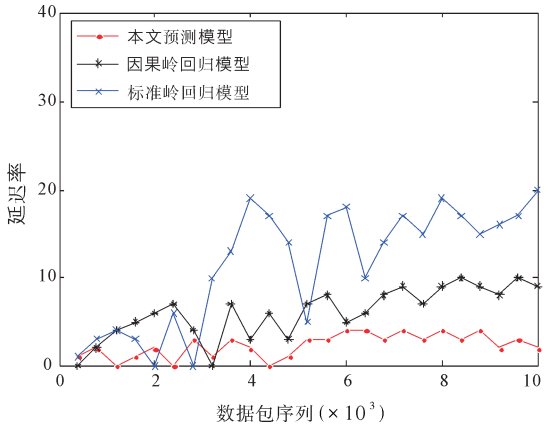

图 4　延迟率比较结果

Fig. 4　Compare of delay rate

5　结　论

为了解决传统模型难以有效预测物联网流量的问题，本文结合贝叶斯与因果岭回归提出了一种新的物联网流量预测模型. 该模型首先根据物联网流量传输波动影响链路变化等因果关系，并根据聚合周期性流量变化和事件驱动性流量变化，深入刻画物联网流量局部特征. 其次，本文通过计算阻塞率、抖动率来描述数据流量，并且基于平均队列时延、节点之间的延迟率来建立流量预测模型. 此外，本文提出基于流量之间延迟率的计算方法，并结合基于贝叶斯和因果岭强度建立预测模型. 最后，通过仿真实验研究了该模型与标准岭回归模型、因果岭回归模型之间的性能状况，给出了这三种算法的预测精度，并且深入分析了三种算法在平均队列、阻塞率和延迟率等方面的变化情况，发现本文所提模型具有较大优势. 在后续研究中，可以考虑结合流量的多重分形特性来完善预测模型.

参考文献：

[1]　钱志鸿，王义君. 面向物联网的无线传感器网络综述[J]. 电子与信息学报，2013，35：215.

[2]　赵东，韩晓艳，赵宏伟，等. 基于分类优化的物联网节点负载均衡策略[J]. 吉林大学学报：工学版，2015，45：926.

[3]　何秀青，王映辉. 物联网服务动态评价选择方法研究[J]. 电子学报，2013，41：117.

[4]　李勤，师维，孙界平，等. 基于卷积神经网络的网络流量识别技术研究[J]. 四川大学学报：自然科学版，2017，54：959.

[5]　张普宁，刘元安，吴帆，等. 物联网中适用于内容搜索的实体状态匹配预测方法 [J]. 电子与信息学报，

2015, 37: 2815.

[6] Ning H S, Liu H. Cyber-physical-social-thinking space based science and technology framework for the Internet of Things[J]. Science China, 2015, 58: 2.

[7] Laner M, Svoboda P, Nikaein N, *et al*. Traffic models for machine type communications[J]. Int Sympos Wirel Commun Syst , 2013, 76: 1.

[8] Centenaro M, Vangelista L. A study on M2M traffic and its impact on cellular networks[C]//Proceedings of the 2015 IEEE 2nd World Forum on Internet of Things. [s.l.]: IEEE,2015.

[9] Karagiannis T, Molle M, Faloutsos M, *et al*. A nonstationary Poisson view of Internet traffic[J]. IEEE Comput Comm Societies, 2004, 3: 1558.

[10] 朱婧, 李鹏飞, 高华. 基于射频识别技术的物联网空口安全体系[J]. 西安邮电大学学报, 2017, 22: 122.

[11] 胡永利, 孙艳丰, 尹宝才. 物联网信息感知与交互技术[J]. 计算机学报, 2012, 35: 1147.

[12] 毛燕琴, 沈苏彬. 物联网模型与能力分析[J]. 软件学报, 2014, 25: 1685.

[13] 朱洪波, 杨龙祥, 于全. 物联网的技术思想与应用策略研究[J]. 通信学报, 2010, 31: 2.

[14] 李健,王明月,姚汝婧,等. 大数据背景下混合层次包围盒碰撞检测算法的优化[J]. 吉林大学学报:理学版,2017,55:673.

[15] 王双成, 高瑞, 杜瑞杰. 小时间序列的动态朴素贝叶斯分类器学习与优化[J]. 控制与决策, 2017, 32: 163.

[16] 夏莘媛, 戴静, 潘用科, 等. 基于贝叶斯证据框架下SVM 的油层识别模型研究[J]. 重庆邮电大学学报: 自然科学版, 2016, 28: 260.

[17] 房丙午, 黄志球, 李勇, 等. 基于贝叶斯网络的复杂系统动态故障树定量分析方法[J]. 电子学报, 2016, 44: 1234.

[18] 吴多,刘来君,苗如松. 利用 B-TBU 模型评估桥梁状态的神经网络法[J]. 江苏大学学报:自然科学版, 2017,38:466.

[19] 马晓琴, 王浩, 李俊照. 基于形态特征与因果岭回归的股市预测算法[J]. 计算机工程, 2016, 42: 175.

[20] 许永红, 韩祥临, 石兰芳, 等. 薛定谔扰动耦合系统孤波的行波近似解法[J]. 物理学报, 2014, 63: 21.

引用本文格式:

中 文: 陈翔, 唐俊勇. 基于贝叶斯与因果岭回归的物联网流量预测模型[J]. 四川大学学报: 自然科学版, 2018, 55: 965.

英 文: Chen X,Tang J Y. The flow prediction model in internet of things based on Bayesian and causal ridge regression [J]. J Sichuan Univ: Nat Sci Ed, 2018, 55: 965.