



基于格兰杰因果关系贝叶斯网络的大规模无线局域网流量预测方法

王 浩, 吕云飞, 陈源宝, 彭云飞

(武汉第二船舶设计研究所 武汉 430064)

摘 要:研究了大规模无线局域网内的流量特性,发现不同接入点间的流量存在格兰杰因果关系。流量的格兰杰因果关系说明,可以通过多个存在因果关系的接入点的历史流量,提高对目标接入点的当前流量预测的准确性。通过贝叶斯网络对存在因果关系的接入点流量进行建模,并利用多个接入点的历史流量对目标接入点的流量进行预测,提高了预测的准确性。最后,通过接入点数量大于 100 个的无线局域网的实际流量数据,验证了该方法的有效性及准确性,建立了一套完整的数据特征分析、建模及预测的流量数据处理流程。

关键词:无线局域网;流量预测;流量特性;格兰杰因果关系;贝叶斯网络

doi: 10.11959/j.issn.1000-0801.2015201

Predicting Large-Scale WLAN Traffic via Granger Causality Based Bayesian Network

Wang Hao, Lv Yunfei, Chen Yuanbao, Peng Yunfei

(Wuhan Second Ship Design and Research Institute, Wuhan 430064, China)

Abstract: Granger causality existed between traffic at different access points of large-scale wireless LANs was discovered. The Granger causality illustrates that the historical traffic of access points that exist causality within target access points help predict the future of target access points with better accuracy than when considering information from the past of target access point alone. Bayesian network to model the causal relationship between access points and adopted a Gaussian mixture model (GMM) was used, as well as a weighted combination of several normal distribution functions in order to approximate the joint probability distribution in Bayesian networks. Finally, the traffic data in large-scale wireless LANs was imported, having hundreds of access points, to verify the accuracy of the proposed method, and a processing flow of analysis, modeling and prediction of traffic flow data was established.

Key words: WLAN, traffic prediction, traffic characteristics, Granger causality, Bayesian network

收稿日期:2015-04-26;修回日期:2015-07-31

论文引用格式:王浩,吕云飞,陈源宝等. 基于格兰杰因果关系贝叶斯网络的大规模无线局域网流量预测方法. 电信科学, 2015201
Wang H, Lv Y F, Chen Y B, *et al.* Predicting large-scale WLAN traffic via Granger causality based Bayesian network. Telecommunications Science, 2015201

1 引言

近年来,随着手机、平板电脑、笔记本电脑等移动式无线设备的普及,无论是家庭、企业还是电信运营商,都在城市里大量地布设无线接入点(access point, AP),大规模无线局域网的优化问题亟待解决。无线局域网真实流量特性的分析总结与建模预测是研究网络优化问题的基础。本文的研究结果可以为大规模网络规划设计以及网络优化提供依据。

本文分析及建模的流量数据采集自国内一家电信运营商在某个城市布设的无线局域网。在该运营商的网络管理平台上采集了大规模无线局域网的流量信息。选取以上场景的原因是该场景是目前公共无线局域网用户数量及用户流量最大的场景。本文以 1 h 的间隔连续地采集了这些场景长达 4 个星期的流量数据。

目前,很多相关工作研究的无线局域网的规模较小,无法充分体现大规模无线局域网中流量的潜在特性。本文通过大量的数据分析,发现在 AP 的二阶邻居节点中存在大量的格兰杰因果关系,说明这些存在因果关系的节点的历史信息可以用来为目标节点流量进行预测。本方法利用格兰杰因果邻居节点及目标节点自身历史数据建立贝叶斯网络,对目标节点当前值进行预测。通过贝叶斯网络与高斯混合模型对存在格兰杰因果关系的 AP 流量进行了建模,将存在格兰杰因果关系的 AP 的历史流量作为原因节点,将被预测 AP 的流量作为目标节点,构建贝叶斯网络,并通过混合高斯模型作为预测的输入与输出模型,通过输入历史流量实现对目标流量的预测。

为了更好地服务于网络资源分配与网络优化,对流量的分析、建模及预测成为研究的重点^[1,2]。目前的研究工作,主要采用基于时间序列的建模方法,如单变量的 ARMA 模型及其衍生模型的流量建模方法被广泛地用于流量建模^[3-6]。实际无线局域网流量数据的分析,在一些小规模场景被研究过,相关参考文献^[6-8]通过研究从 IETF2005 会议收集到的 3 个不同时间段的无线网络流量对流量的特性进行了分析,建立了自相似性模型,并且将流量的特性运用到了无线局域网的流量控制中,建立了一个较为完整的分析体系。但是以上工作都没有对真实的无线局域网系统进行大规模的数据分析、总结流量分布特性、建立流量建模并对流量进行预测。

2 格兰杰因果关系与贝叶斯网络预测理论

在时间序列情形下,若在包含了变量 X 、 Y 的历史信息的条件下,对变量 Y 的预测效果要优于只单独由 Y 的历史信息对 Y 进行的预测效果,即变量 X 有助于解释变量 Y 的将来变化,则认为变量 X 是引致变量 Y 的格兰杰原因。

格兰杰因果关系检验要求估计以下回归模型:

$$\begin{aligned} Y(t) &= \sum_{i=1}^p A_{11,i} X(t-i) + \sum_{i=1}^p A_{12,i} Y(t-i) + \varepsilon_1(t) \\ X(t) &= \sum_{i=1}^p A_{21,i} X(t-i) + \sum_{i=1}^p A_{22,i} Y(t-i) + \varepsilon_2(t) \end{aligned} \quad (1)$$

其中, $X(t)$ 、 $Y(t)$ 表示 X 、 Y 原始序列当期值,即 t 时刻两个 AP 的流量值, $X(t-i)$ 、 $Y(t-i)$ 表示 X 、 Y 原始序列滞后 i 期的值, A_{11} 、 A_{12} 、 A_{21} 、 A_{22} 表示回归系数, $\varepsilon_1(t)$ 、 $\varepsilon_2(t)$ 表示误差项。

格兰杰因果关系检验是通过构造 F 统计量,利用 F 检验完成的。如针对 X 不是 Y 的格兰杰原因这一假设,即针对式(1)中 X 滞后项前的参数整体为零的假设,分别做包含与不包含 X 滞后项的回归,记前者的残差平方和为 RSS_U ,后者的残差平方和为 RSS_R ,再计算 F 统计量:

$$F = (RSS_R - RSS_U) \times (N - 2n - 1) / RSS_U \times n \quad (2)$$

其中, n 表示 X 的滞后项的个数, N 表示样本容量。如果计算的 F 大于给定显著水平 α 下 F 分布的响应的临界值 $F_\alpha(n, N-2n-1)$,则拒绝原假设,即认为 X 是 Y 的格兰杰原因。

由于预测过程是通过为目标 AP 有因果关系的 AP 的历史流量信息对目标 AP 的流量进行估计,所以在构造贝叶斯网络时,网络的边既要考虑空间因素,还要考虑时间因素。在贝叶斯网络进行预测前,需要推导输入与输出值之间的联合概率分布。本文采用高斯混合模型(多个加权组合的正态分布, GMM)来近似贝叶斯网络的联合概率分布。

令 x 表示一个随机变量或多维随机向量,通过 M 个正态分布组成的 GMM 来表示它的概率:

$$p(x|\Theta) = \sum_{l=1}^M \alpha_l p_l(x|\theta_l) \quad (3)$$

其中,参数 $\Theta = (\alpha_1, \dots, \alpha_M, \theta_1, \dots, \theta_M)$, M 满足 $\sum_{l=1}^M \alpha_l = 1$ 。每个 $p_l(\cdot)$ 是参数为 $\theta_l = (\mu_l, \Sigma_l)$, $l=1, \dots, M$ 的高斯概率密度函数。



参数估计采用最大期望算法(expectation-maximization algorithm)对参数 α_i, θ_i 进行估计。

设 X 为某种分布产生的观测数据集。假设完备数据集 $Z=(X,Y)$, 于是变量的联合分布 $z=(x,y)$ 形式如下:

$$p(z|\Theta) = p(x,y|\Theta) = p(y|x,\Theta)p(x|\Theta) \quad (4)$$

其中, Y 表示隐藏数据集, Θ 表示分布 $p(z|\Theta)$ 参数集。通过 EM 算法的 E 步和 M 步的迭代后, 参数集为:

$$\Theta^{(i)} = \arg \max_{\Theta} Q(\Theta, \Theta^{(i-1)}) \quad (5)$$

其中, $Q(\Theta, \Theta^{(i-1)}) = E_y[\lg p(X,Y|\Theta)|X, \Theta^{(i-1)}]$ 。各参数为:

$$\begin{aligned} \alpha_i^{\text{new}} &= \frac{1}{N} \sum_{l=1}^N p(l|x_i, \Theta^{(i-1)}) \\ \mu_i^{\text{new}} &= \frac{\sum_{l=1}^N x_l p(l|x_i, \Theta^{(i-1)})}{\sum_{l=1}^N p(l|x_i, \Theta^{(i-1)})} \\ \Sigma_i^{\text{new}} &= \frac{\sum_{l=1}^N p(l|x_i, \Theta^{(i-1)})(x_l - \mu_i^{\text{new}})(x_l - \mu_i^{\text{new}})^{\dagger}}{\sum_{l=1}^N p(l|x_i, \Theta^{(i-1)})} \end{aligned} \quad (6)$$

本文中, 流量的预测被归结为一个贝叶斯网络的推断问题。贝叶斯网络推断的主要目的就是在给定观测节点值的情况下估计目标节点的值。本文利用高斯混合模型在给定观测值概率的情况下, 计算未知值的条件概率。

令 (E, F) 分别表示贝叶斯网络中的观测数据集与未知数据集, 而 (x_E, x_F) 分别表示这两个数据集对应的随机变量/向量。 x_E 的边缘概率为:

$$p(x_E) = \sum_{x_F} p(x_E, x_F) \quad (7)$$

依据贝叶斯理论, 条件概率 $p(x_F|x_E)$ 为:

$$p(x_F|x_E) = \frac{p(x_F, x_E)}{p(x_E)} = \frac{p(x_F, x_E)}{\sum_{x_F} p(x_E, x_F)} \quad (8)$$

于是, 可以利用观测变量 x_E 即历史流量值来预测未知变量 x_F 。依据最小均方差准则, x_F 的最佳估计为:

$$\hat{x}_F = E(x_F|x_E) \quad (9)$$

为了在 GMM 条件下, 求解式(9), 首先引入参考文献[10]中证明的引理, 如下。

引理 1 令 $G(x; \mu, \Sigma)$ 表示多维正态密度函数, 其均值与方差分别为 μ, Σ 。在 $x^{\dagger} = (x_1^{\dagger}, x_2^{\dagger}), \mu^{\dagger} = (\mu_1^{\dagger}, \mu_2^{\dagger})$ 与

$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$ 条件下, 概率密度函数 $p(x)$ 为:

$$p(x) = G(x_1; \mu_1, \Sigma_{11})G(x_2; \mu_{x_2|x_1}, \Sigma_{x_2|x_1}) \quad (10)$$

其中:

$$\begin{aligned} \mu_{x_2|x_1} &= \mu_2 - \Sigma_{21}\Sigma_{11}^{-1}(\mu_1 - x_1) \\ \Sigma_{x_2|x_1} &= \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} \end{aligned} \quad (11)$$

依据引理 1, 定义 $x^{\dagger} = (x_F^{\dagger}, x_E^{\dagger}), \mu^{\dagger} = (\mu_{IF}^{\dagger}, \mu_{IE}^{\dagger})$,

与 $\Sigma_I = \begin{pmatrix} \Sigma_{IFF} & \Sigma_{IFE} \\ \Sigma_{IEF} & \Sigma_{IEE} \end{pmatrix}$ 。式(12)为:

$$\begin{aligned} p(x_F, x_E) &= \sum_{l=1}^M \alpha_l G(x; \mu_l, \Sigma_l) \\ &= \sum_{l=1}^M \alpha_l G(x_E; \mu_{IE}, \Sigma_{IEE})G(x_F; \mu_{IF|E}, \Sigma_{IF|E}) \end{aligned} \quad (12)$$

依据式(7), 条件概率 $p(x_F|x_E)$ 可以表示为:

$$p(x_F|x_E) = \sum_{l=1}^M \beta_l G(x_F; \mu_{IF|E}, \Sigma_{IF|E}) \quad (13)$$

其中:

$$\begin{aligned} \beta_l &= \frac{\alpha_l G(x_E; \mu_{IE}, \Sigma_{IEE})}{\sum_{l=1}^M \alpha_l G(x_E; \mu_{IE}, \Sigma_{IEE})} \\ \mu_{IF|E} &= \mu_{IF} - \Sigma_{IFE}\Sigma_{IEE}^{-1}(\mu_{IE} - x_E) \\ \Sigma_{IF|E} &= \Sigma_{IFF} - \Sigma_{IFE}\Sigma_{IEE}^{-1}\Sigma_{IEF} \end{aligned} \quad (14)$$

于是, 最小均方差约束下的流量的最佳预测 $\hat{x}_F$ 为:

$$\begin{aligned} \hat{x}_F &= E(x_F|x_E) \\ &= \int x_F p(x_F|x_E) dx_F \\ &= \sum_{l=1}^M \beta_l \int x_F G(x_F; \mu_{IF|E}, \Sigma_{IF|E}) dx_F \\ &= \sum_{l=1}^M \beta_l \mu_{IF|E} \end{aligned} \quad (15)$$

3 预测方法及流程

结合所述理论将 AP_k 的流量预测作为实例来展示模型建立及流量预测的过程。 $AP_k(t)$ 表示 t 时刻 AP_k 的预测流量。由式(1)、式(2)所示的格兰杰因果关系检验, 分析出 $AP_1, \dots, AP_n$ 为与 AP_k 有因果关系的所有 AP。考虑时间因素, 为了预测 $AP_k(t)$ 的流量值, 还要考虑 AP_k 的 $1, \dots, m$ 滞后的历史流量值, 即 $AP_k(t-1), \dots, AP_k(t-m)$, 它们可以反映 AP_k 的变化趋势。因此, 在构建贝叶斯网络时, $AP_1, \dots, AP_n$ 与 AP_k 的历史流量值, 都是 $AP_k(t)$ 的原因节点。从流量预测的观点来看, 将所有的原因节点的流量值都作为 GMM 的输入, 代入式(15), 而预测值 $AP_k(t)$ 则作为输出。

预测流程如下:

- 将依据格兰杰因果关系分析出的原因节点及其历史值作为预测输入, 预测目标值为预测输出, 建立

贝叶斯网络;

- 利用 EM 算法, 估计已建立的贝叶斯网络中的 GMM 的各条件概率分布及参数;
- 依据式(6)计算预测目标值的最佳估计。

4 实际场景验证结果

利用真实的无线局域网流量数据验证本方法对真实数据的有效性及其准确性。本文将基于格兰杰因果关系贝叶斯网络的预测结果与目前常用的基于小波分析与整合自回归平滑的预测模型的预测结果进行对比, 说明本方法的有效性。

福州长乐机场总共有 432 个在用 AP, 其中格兰杰关系节点的关系为: 一阶邻居中存在比例为 54%, 二阶邻居中存在比例为 83.79%。这说明大部分的因果关系发生在二阶邻居节点内, 并且在无线局域网中邻居节点存在较强的格兰杰因果关系。

为了说明方法在实际场景中的有效性, 通过实际数据分析以长乐机场二楼 13#AP 作为分析对象(13#AP 流量统计数据略, 具体参见网络版 <http://www.telecomsci.com.cn>), 该区域总共有 18 个 AP, 它们的格兰杰因果关系如图 1 所示。

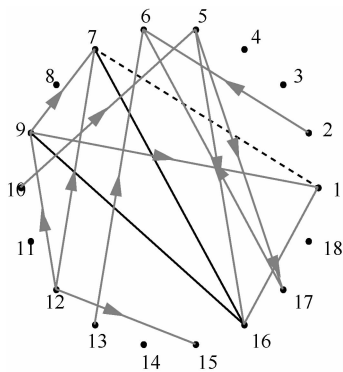


图 1 AP 间的格兰杰因果关系

根据 AP 间的格兰杰因果关系, 选择 AP13#1、AP13#7 及 AP13#16 作为预测对象, 将 AP13#1、AP13#5、AP13#7、AP13#9、AP13#12 及 AP13#16 的历史流量值作为原因节点, 这 3 个 AP 的预测值作为目标节点, 建立贝叶斯网络。

以 AP13#1 为例, 对比基于贝叶斯网络与 WARMA 模型的预测结果, 如图 2 所示。

从图 2 可以看出, 贝叶斯网络的预测结果在趋势与数量级上都与真实值接近, 可以很准确地反映真实值的变化。

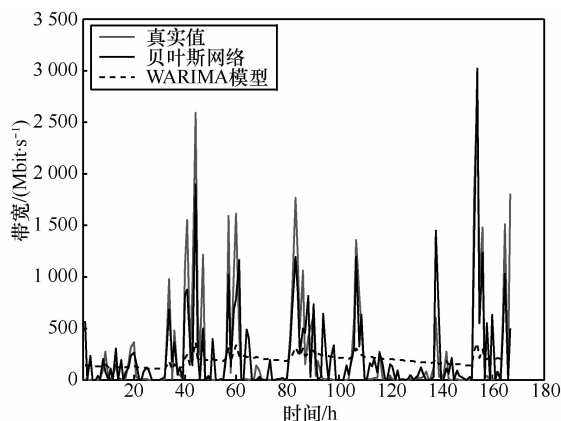


图 2 贝叶斯网络与 WARMA 模型的预测结果对比

为了说明贝叶斯网络与 WARIMA 模型预测的准确性, 通过统计长乐机场所有 AP 流量预测的结果与真实值的绝对百分误差, 对比两种方法预测的准确性。从统计结果可以得到贝叶斯网络预测的准确性明显优于 WARIMA 混合模型。基于贝叶斯网络的预测结果, 90% 的绝对百分误差低于 5%, 而 WARIMA 的预测结果只有 80% 的绝对百分误差低于 10%, 还有 10% 的绝对百分误差大于 30%, 准确性较本方法低。

5 结束语

本文研究了两个大规模无线局域网内的流量特性, 发现 AP 间广泛存在格兰杰因果关系, 并且存在格兰杰因果关系的接入点位置集中在一阶与二阶邻居 AP 中。然后, 本文通过贝叶斯网络利用格兰杰因果关系特性, 对存在因果关系的接入点流量进行建模, 并通过贝叶斯网络模型利用多个接入点的历史流量对目标接入点的流量进行预测, 提高了预测的准确性。最后, 通过接入点数量大于 100 个的无线局域网的实际流量数据, 验证了该方法相较于 WARIMA 模型的准确性, 并验证了该方法在不同场景下的有效性, 建立了一套完整的数据特征分析、建模及预测的流量数据处理流程。

参考文献

- 1 Petkov V, Obraczka K. The case for using traffic forecasting in schedule-based channel access. Proceedings of Consumer Communications and Networking Conference (CCNC), Las Vegas, NV, USA, 2011: 208~212
- 2 Petkov V, Obraczka K. Collision-free medium access based on traffic forecasting. Proceedings of 2012 IEEE International Symposium on a World of Wireless, Mobile and Multimedia



Networks (WoWMoM), San Francisco, CA, USA, 2012:1~9

- 3 Chen C, Pei Q, Ning L. Forecasting IEEE 802.11 traffic using seasonal ARIMA model. Proceedings of International Forum on Computer Science-Technology and Applications, Chongqing, China, 2009:347~350
- 4 Suarez C A H, Parra O S, Martinez L F P. Traffic model based on time series to forecast traffic future values within a Wi-Fi data network. Proceedings of 4th International Conference on Wireless Communications, Networking and Mobile Computing, Dalian, China, 2008:1~4
- 5 Xu Y. A network traffic model based on fractal. Proceedings of International Conference on Wireless Communications, Networking and Mobile Computing, Shanghai, China, 2007: 1921~1924
- 6 Xiang L, Ge X, Zhang K, *et al.* A self-similarity frame traffic model based on the frame components in IEEE 802.11 networks. Proceedings of Conference on Computational Science and Engineering, Vancouver, BC, Canada, 2009: 955~960
- 7 Zhang K, Ge X, Liu C, *et al.* Analysis of frame traffic characteristics in IEEE 802.11 networks. Proceedings of Conference on Communications and Networking in China, Troy, New York, USA, 2009: 1~5
- 8 Liu C, Ge X, Xiang L, *et al.* A frame traffic control algorithm for IEEE 802. 11 WLAN. Proceedings of 2010 IEEE/ACM Int'l Conference on & Int'l Conference on Cyber, Physical & Social Computing (CPSCoM) , Hangzhou, China, 2010: 479~483
- 9 Sun S, Zhang C, Yu G. A Bayesian network approach to traffic flow forecasting. IEEE Transactions on Intelligent Transportation Systems, 2006, 7(1): 124~132
- 10 Rao C R. Linear Statistical Inference and Its Applications, 2nd Ed. New York: Wiley, 1973

[作者简介]



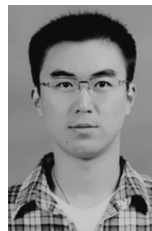
王浩,男,博士,武汉第二船舶设计研究所工程师,主要研究方向为船舶信息化。



吕云飞,男,博士,武汉第二船舶设计研究所高级工程师,主要研究方向为船舶信息化。



陈源宝,男,博士,武汉第二船舶设计研究所工程师,主要研究方向为船舶信息化。



彭云飞,男,武汉第二船舶设计研究所工程师,主要研究方向为船舶总体技术。