基于非完备大数据的业务预测

Traffic Prediction with Incomplete Big Data

李建东/LIJiandong，盛敏/SHENG Min，文娟/WEN Juan
（西安电子科技大学，陕西 西安 710071）
(Xidian University, Xi'an 710071, China)

DOI：10.12142/ZTETJ.201901010
网络出版地址：http://kns.cnki.net/kcms/detail/34.1228.TN.20190131.1032.004.html
收稿日期：2018-11-26
网络出版日期：2019-01-31

摘要：高效、精准预测无线网络业务数据，例如业务的到达率、用户数以及吞吐量等，将为网络提供用户的实时需求，是实现无线网络智能化的关键。然而，由于无线网络传输的不可靠性、采集设备故障、采样率低等原因，使得无线大数据具有不可避免的非完备性。将使系统丢失大量有用信息，从而给无线网络业务预测带来巨大挑战。为了应对该挑战，提出了基于非完备数据集的业务预测架构，从缺失值补充以及空时信息挖掘2个维度高效利用非完备数据集，提升预测精度，助力无线网络的智能化。

关键词：业务预测；智能无线网络；非完备数据

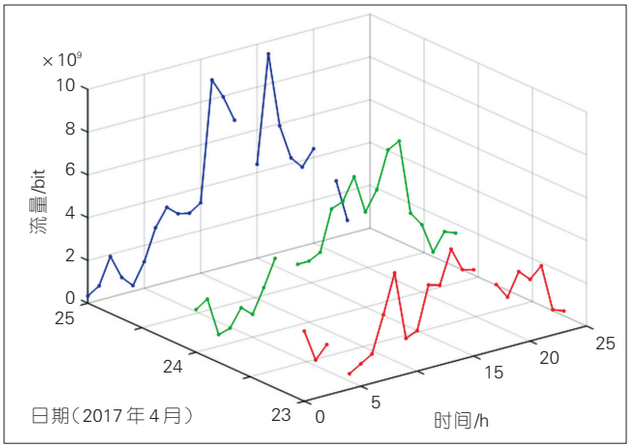
Abstract: High efficient and accurate wireless traffic prediction, such as arrival rate, user account, and throughput, will provide users' real demand for network providers, which is the key for intelligent wireless networks. However, there exists incomplete nature for wireless big data because of the unreliable wireless transmission, the failure of data acquisition and low sample rate. This unique feature may make wireless networks lose massive useful information and bring great challenge for accurate traffic prediction. To meet this challenge, an incomplete data-based traffic prediction framework is proposed, leveraging the incomplete data set efficiently via filling the missing data and digging the temporal-spatial information.

Key words: traffic prediction; intelligent wireless networks; incomplete data set

无线网络正经历着从基于信息论的可靠传输到基于智能的高效通信的巨大变革[1]。智能无线资源管理是无线网络智能化的核心，其目标是通过无线资源的动态调配使网络资源与用户需求精准适配。具体来讲，网络将根据用户业务需求的时空分布，在相对较大的时间尺度内，合理地配置各区域网络资源，使网络资源结构最优化；在小时间尺度内，动态地为各用户分配网络资源，使系统资源利用率以及用户体验最大化[2]。由此可见，高效、准确预测无线网络业务需求，是无线网络实现智能化的重要基础。

然而，网络结构的异构化和密集化使网络干扰异常复杂，加剧了无线网络传输的不可靠性，使得无线业务信息在传输过程中产生不可避免的丢失[3]。此外，数据采集设备故障或供电不足都将导致业务数据在收集过程中的缺失。最后，由于硬件设备限制导致的低采样率往往无法准确获得业务变化的重要信息。这些原因都将使无线网络业务数据具备如图1所示的非完备特性，即数据缺失或者无法反应业务变化趋势，从而给无线网络业务预测带来巨大挑战。

目前，对于缺失数据处理方法主要是根据已有数据的统计特性，如均值、中位数等，对缺失值进行补充[4]。此类方法对于统计规律比较强的数据有很好的作用；但是，当数据统计规律较弱时，其统计特性无法较好地反应数据本身特点。此时用其统计特性进行缺失值填充会引入大量噪声，从而影响数据预测效果。如表1所示，传统业务预测方法主要从时间以及空间2个维度，采用时间序列分析[5]、机器学习[6]，以及深度学习[8-9]等方法，对收集到的业务数据进行预测，但是基本没有考虑数据集



◀图1
无线业务非完备特性示例

▼表1 典型业务预测算法分类总结

分类	典型预测算法	特点描述
时间序列分析法	ARIMA、AR、MA等	适用于线性、平稳过程，计算复杂度低
机器学习	Ridge、RF、Light GBM等	不局限于线性、平稳过程，计算复杂度适中
深度学习	LSTM、RNN等	所需训练集样本大，模型可解释性弱，计算复杂度高

AR：自回归
ARIMA：差分整合滑动平均自回归模型
Light GBM：轻量级梯度提升机
LSTM：长短期记忆网络
MA：滑动平均
RF：随机森林
RNN：循环神经网络

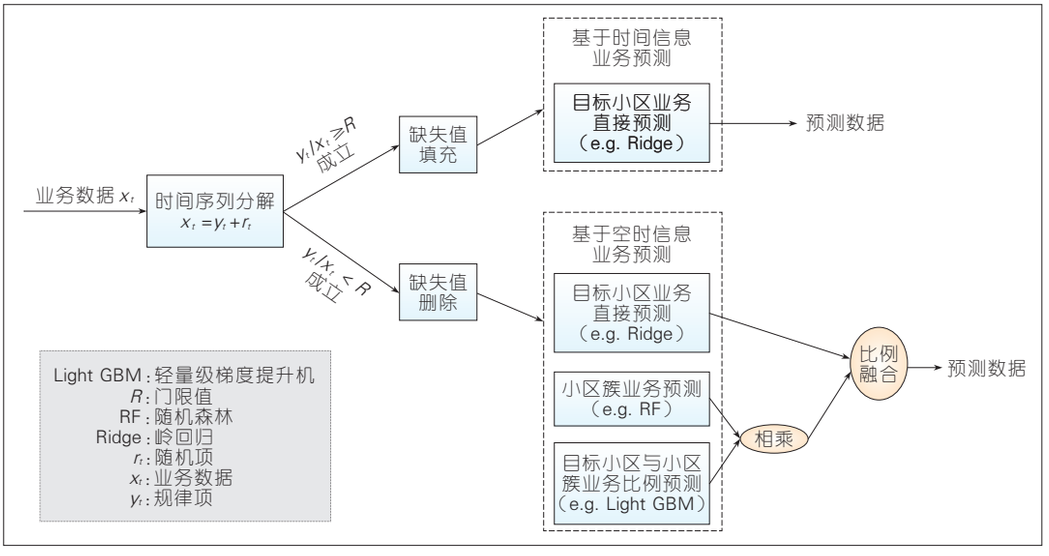
的非完备特性对业务预测带来的影响。

1 非完备海量数据业务预测

为了应对非完备数据给业务预测带来的挑战，本文中我们提出了如图2所示的基于非完备无线大数据的业务预测架构，从缺失值填充、时空信息挖掘2个维度，高效利用非完备数据集，助力无线网络智能化。

简单来讲，当预测数据规律性较强时，例如办公楼以及住宅区域的业务数据呈现明显的“潮汐现象”，我们根据待预测数据的统计特性对缺失值进行补充，然后选取合适的预测算法对待预测数据直接进行预测。当预测数据规律性较弱时，例如交通枢纽区域等业务数据流动性强、规律弱，如果仍根据其统计特性对缺失值补充，将会引入大量噪声；因此我们直接将缺失值丢弃，并充分利用空间维度信息进行数据挖掘，对待预测数据进行间接预测。

首先，我们采用时间序列分解法，将待预测数据分解为规律项和随机项，并根据规律项占业务量比值的大小，将待预测数据分为规律性强或弱2种情况。具体做法为：将待预测数据 $x=\{x_1,x_2,\cdots x_n\}$（其中 $x_t,1\leqslant t\leqslant n$ 表示第 t 时刻待预测的业务量，例如用户数、流量等）分解为周期项 p_t、趋势项 m_t 以及随机项 r_t，并且将周期项与趋势项的和称为规律项 $y_t=p_t+m_t$。当规律项与业务量的比值高于某一门



◀图2
基于非完备无线大数据的业务预测架构

限值 R 时，待预测数据规律性较强；反之，当规律项与业务量的比值低于某一门限值 R 时，待预测数据规律性较弱。

当待测数据规律性较强时，其历史数据的统计特性例如均值，可较好地反映待测数据规律；因此，我们可利用缺失值历史信息的均值，对其进行填充，扩充训练样本数。我们可以将填充好后的数据直接输入所选择的预测算法，对待预测数据直接进行预测。

当待测数据规律性较弱时，其历史数据的统计特性无法较好地反映待测数据规律。如果对缺失值进行强行填充，将会引入较多噪声，从而影响预测精度。此时，我们将缺失值直接删除，确保使用数据的真实性。经过研究发现，在无线网络中，即使单小区的业务规律性较弱，由多个小区构成的小区簇的业务规律性一般都很强。因此，可以充分挖掘相邻小区的空间信息，先对小区簇的业务总量进行预测，然后再对目标小区业务与小区簇业务比值进行预测，最后将这2部分的预测值相乘，即可得到基于空间信息获得的目标小区待测业务量。为了进一步提升预测精度，我们采用“提升”（boosting）算法的基本思想，即设计多个好而不同的预测方法对同一问题进行预测，并将其结果进行融合，通过模型和数据的分集增益提升预测精度。为此，我们采用与基于空间信息预测模型不同的预测方法对删除缺失值后的数据直接进行预测。最后，将预测结果与基于空间信息的预测结果进行有机融合，便可利用模型和数据的分集增益提升预测精度。

2 仿真设计与分析

为了验证本文提出的基于非完备大数据业务预测架构的有效性，我们采用校园网实测数据对各个区域各时间段的用户数进行预测。

为了判断待预测数据规律性的强弱，我们假设门限值 R=0.8。对于规律性比较强的业务数据，我们采用均值对相应缺失值进行补充，并使用补充后的数据集作为训练数据集，采用岭回归（Ridge）方法对其进行直接预测。图3对比了对缺失值进行均值填充和缺失值删除后的预测效果。为此，我们将获得的相对完整的数据看做实验中的“完备”数据集，然后在人为随机删除部分数据进行验证。从图3中可以看出，当待测数据规律性较强时，当缺失值比例不大时，采用均值补充可以有效提升预测精度。此外，删除某些数据时，例如异常值，也可提升预测精度。因此，我们在对数据进行预测前，要先分析数据的特性，并根据数据的特性进行相应的处理。

对于规律性相对较弱的业务数据，我们先将缺失值删除，然后使用删除缺失值后的数据集作为训练数据集，并采用基于时空信息的预测方法对其预测。具体来讲，分别采用Ridge、随机森林（RF）以及轻量级梯度提升机（Light-GBM）方法对目标小区用户数、小区簇用户数以及目标小区和小区簇用户数的比例进行预测，然后将预测出的小区簇用户数和相应比例相乘，所得结果与直接预测的用户数以合适比例融合，得出最后的目标小区用户数预测值。图4对比了仅基于时间信息的业务预测与基于时空信息的业务预测精度。从图中我们可以看出，借助于空间信息可有效提高业务预

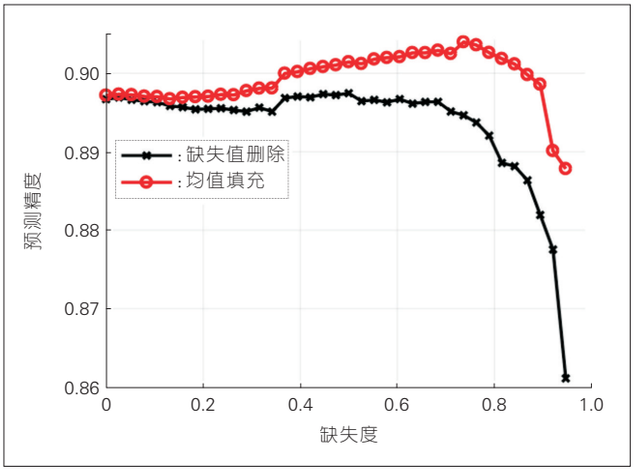
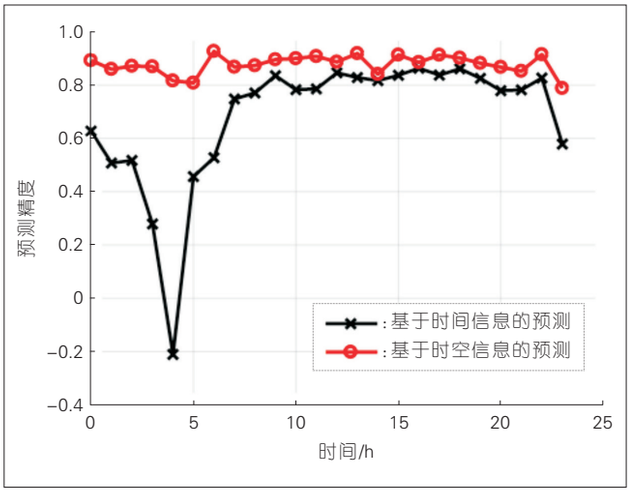


图3 ▶ 缺失值填充与删除对业务预测影响对比图



◀图4
基于时间与时空信息业务预测对比图

测精度。

3 结束语

基于无线大数据，结合人工智能算法，将使无线网络的设计、管理与优化更加自动化、智能化与智慧化。然而，无线网络信道传输的不可靠性、业务多样性以及网络结构密集化、异构化等特点使得无线大数据呈现非完备性、空时大尺度变化等特点，为无线大数据挖掘以及人工智能算法应用与结合带来巨大挑战。本文中，我们提出了基于非完备无线大数据业务预测的基本架构，对非完备无线大数据的应用具有重要启发意义。在无线网络智能化的关键时期，仍需要我们不断探索如何针对无线网络特异性，最大限度地挖掘无线大数据的价值并设计相应的智能算法。

参考文献

[1] LI R P, ZHAO Z F, ZHOU X, et al. Intelligent 5G: When Cellular Networks Meet Artificial Intelligence [J]. IEEE Wireless Communications, 2017, 24(5): 175–183. DOI: 10.1109/mwc.2017.1600304wc
[2] 张琰, 盛敏, 李建东. 大数据驱动的“人工智能”无线网络[J]. 中兴通讯技术, 2018, 24(2): 2–5
[3] LIU J Y, SHENG M, LIU L, et al. Interference Management in Ultra–Dense Networks: Challenges and Approaches [J]. IEEE Network, 2017, 31(6): 70–77. DOI:10.1109/mnet.2017.1700052
[4] SESSA J, SYED D. Techniques to Deal with Missing Data[C]//2016 5th International Conference on Electronic Devices, Systems and Applications (ICEDSA). United Arab Emirates:ICEDSA, 2016: 1–4. DOI:10.1109/ICEDSA.2016.7818486
[5] XU F L, LIN Y Y, HUANG J X, et al. Big Data Driven Mobile Traffic Understanding and Forecasting: A Time Series Approach [J]. IEEE Transactions on Services Computing, 2016, 9(5): 796–805. DOI:10.1109/tsc.2016.2599878
[6] ZARE MOAYEDI H, MASNADI–SHIRAZI M A. Arima Model for Network Traffic Prediction and Anomaly Detection[C]//2008 International Symposium on Information Technology. Malaysia, 2008: 1–6. DOI: 10.1109/ITSIM.2008.4631947
[7] WANG X, ZHOU Z M, YANG Z, et al. Spatio–Temporal Analysis and Prediction of Cellular Traffic in Metropolis[C]//2017 IEEE 25th International Conference on Network Protocols (ICNP). Canada: ICNP, 2017: 1–10. DOI:10.1109/ICNP.2017.8117559
[8] WANG J, TANG J, XU Z, et al. Spatiotemporal Modeling and Prediction in Cellular Networks: A Big Data Enabled Deep Learning Approach[C]//IEEE INFOCOM. USA: IEEE, 2017:1–9

作者简介

李建东，西安电子科技大学教授、博士生导师，教育部长江学者特聘教授，国家杰出青年科技基金获得者；主要研究方向为智能宽带无线通信、认知无线网络、大规模自组织网络以及无线网络的干扰管理等；先后主持和参加基金项目20余项，获得国家技术发明奖二等奖2项；已发表论文200余篇，其中被SCI/EI检索100余篇。

盛敏，西安电子科技大学教授、博士生导师，教育部长江学者特聘教授，国家杰出青年科技基金获得者；主要研究方向为智能宽带无线通信、认知无线网络、大规模自组织网络等；先后主持和参加基金项目20余项，获得国家技术发明奖二等奖2项；已发表论文100余篇，其中被SCI/EI检索80余篇。

文娟，西安电子科技大学讲师、硕士生导师；主要研究方向为智能无线网络、信息智能处理与传输技术以及异构无线网络容量研究等；先后主持和参加国家基金项目10余项；获陕西省科学技术一等奖1项；已发表SCI/EI检索论文10余篇。