Wireless Communication Data Sets for Machine Learning

Matthias Herlich and Stefan Farthofer <firstname.lastname@salzburgresearch.at>

***Abstract*—Applying machine learning techniques to wireless networks is a hot research topic. One of the biggest problems in this area is accessing a suitable data set. Because creating such data sets is expensive, reusing them is important. To provide more researchers access to data for machine learning in wireless networking we surveyed data sets. Our overview supports researchers that want to apply machine learning to wireless networking in finding a suitable data set.**

I. Introduction

Big data has many applications in wireless networking [1]. One is artificial intelligence for 5G management [2], which will supposedly also play a key role in 6G networks [3].

Wireless communication data sets range from low level physical layer measurements to social network analyses. Due to the inherent stochastic nature of wireless communication the collection of data sets always played a crucial role. For example, channel sounding has a long tradition to validate theoretical channel models in real environments and to determine the model parameters. Increased application of signal processing techniques and advances in machine learning broadened the attention towards the analysis of higher layers. In this paper we focus on data sets suitable, for example, for network planning and traffic prediction.

II. Data sets

Table I provides an overview of selected data sets which might be suitable for machine learning applications in wireless networks. It is not a complete list and we have no inclusion criteria. Nevertheless, the list provides a starting point for researchers looking for data sets. We group data sets into finished one-time experiments and ongoing data-collection efforts.

A. Finished one-time experiments

Experiments with a fixed duration have been run to determine the characteristics of wireless networks. They have been run at laboratories, locations with characteristics expected to be representative of a wide variety of locations and in locations which were expected to be exceptional. Examples are general urban areas and factories. Most data sets are generated by universities (e.g., [4]) or government agencies (e.g., NIST [5]). Data sets based on one-time experiments usually focus on lower layer (PHY) data.

B. Ongoing data-collection efforts

An alternative are data-collection efforts with an open time window. Such efforts are usually not based on specifically installed hardware, but turn to the public to generate data. That is, they provide software tools that allow every interested person to use their hardware to add data points to the data set. Many developed countries have agencies which support this, but also independent efforts (both commercial and non-commercial) exist. A recent white paper provides definitions, use cases and challenges for crowd-sourced measurements [6]. Data sets based on ongoing data collection usually focus on higher layer (or application) data.

III. Applications

Possible applications and challenges of applying machine learning (ML) to wireless networks have been discussed [7], [8], [9].

Due to the expected heterogeneity and complexity of wireless networks the traditional model-based approach in development and operation will no longer be feasible in the future [7]. Data-driven approaches will complement traditional design techniques based on mathematical models [7]. This shifts the goal of acquired data from model parameterization to machine learning or even continuous optimization for Self-Organzing Networks (SON). Along with Software Defined Networking (SDN) this data-driven approach enables network operators to apply ML-based methods in a variety of network locations: On the physical layer machine learning can be used for power control or spectrum management,

This research is partly funded by the Federal Ministry for Climate Action, Environment, Energy, Mobility, Innovation and Technology (BMK) and the Austrian state Salzburg.

Table I
SELECTED DATA SETS FOR WIRELESS NETWORKS

Source	Link	Time	Access	Area	Technology	Size/Entries	Data
Finished one-time experiments							
CRAWDAD		-	Open	Depends	Depends	Depends	Depends
IEEE ComSoc		-	Open	Depends	Depends	Depends	Depends
COSINE		2020	Open	Lab	Cellular	55 GiB[a]	SNR, ...
KU Leuven		2017?	Partly	University	LTE	?	SINR, PHY info
RICE		2016?	Open	Lab	Experimental	37 GiB[a]	Channel response matrix
NIST		2016?	Open	Factory	Low level	176 GiB	Impulse responses
Ongoing data-collection efforts							
OECD[b]		-	Depends	Country	Cellular	Depends	Data rate, location, ...
RTR		2012 - ...	Open	Austria	Cellular	1.5×10^6	Data rate, latency, location
Bundesnetzagentur		2015 - ...	Open	Germany	Cellular	1×10^6	Data rate, latency, location
Ookla		2006 - ...	Closed	Global	Cellular	30×10^9	Data rate, latency, location
OpenSignal		2010 - ...	Partly	Global	Cellular	Depends[c]	Data rate, latency, location
CellMapper		2010 - ...	Open	Global	Cellular	?	Base station location, configuration
WiGLE		2001 - ...	Open	Global	WiFi[d]	9×10^9	Location, SSID, channel

[a] in compressed form [b] list of speed tests [c] data sets $> 10^9$ seem to be the norm [d] also some data for Bluetooth and Cellular

or on higher layers for backhaul, cache, and resource management [9].

The training data that is needed depends on the specific application; either training data is individually generated and parametrized by the configuration or collected in an online fashion [9]. Typically, the former approach is used for lower layer problems whereas the later is used for higher layers.

Besides applications where machine learning directly interacts with the network itself, machine learning can be applied for analytic predictive tasks. The inherent structure of the collected data can be exploited by ML models to predict network metrics (e.g., data rate, latency, or reliability). These predictions span over time and space and their quality depends on the density of the data and the variability in time (primarily caused by user behavior) and space (primarily caused by network topology and geography). Technically, this is a problem of reconstructing undersampled data where the undersampling rate is defined by the data set. In its generality, a Nyquist-Shannon-based perfect unique reconstruction is impossible. Thus, due to the ability of implicitly exploiting the underlying data structure ML-based approaches might perform better in such scenarios.

IV. CONCLUSION

Machine learning is already used in telecommunication, especially in the core network, and its significance will most likely grow in the near future. However, ML-based network planning and traffic prediction lags behind due to the previous lack of data sets. Selecting a suitable data set depends on the use case, but our table and the accompanying categorization supports researchers in finding such data sets.

REFERENCES

[1] L. Qian, J. Zhu, and S. Zhang, "Survey of Wireless Big Data," *Journal of Communications and Information Networks*, 2017.

[2] R. Li, Z. Zhao, X. Zhou, G. Ding, Y. Chen, Z. Wang, and H. Zhang, "Intelligent 5G: When cellular networks meet artificial intelligence," *IEEE Wireless communications*, 2017.

[3] K. B. Letaief, W. Chen, Y. Shi, J. Zhang, and Y.-J. A. Zhang, "The roadmap to 6G: AI empowered wireless networks," *IEEE Communications Magazine*, 2019.

[4] D. Block, N. H. Fliedner, D. Toews, and U. Meier, "Wireless channel measurement data sets for reproducible performance evaluation in industrial environments," in *IEEE Conference on Emerging Technologies & Factory Automation*, 2015.

[5] R. Candell, C. A. Remley, J. T. Quimby, D. R. Novotny, A. Curtin, P. B. Papazian, G. H. Koepke, J. Diener, and M. T. Hany, "Industrial wireless systems: Radio propagation measurements," tech. rep., NIST, 2017.

[6] T. Hoßfeld and S. Wunderer, "White Paper on Crowdsourced Network and QoE Measurements–Definitions, Use Cases and Challenges," *doi: 10.25972/OPUS-20232*, 2020.

[7] A. Zappone, M. Di Renzo, and M. Debbah, "Wireless networks design in the era of deep learning: Model-based, AI-based, or both?," *IEEE Transactions on Communications*, 2019.

[8] U. Challita, H. Ryden, and H. Tullberg, "When machine learning meets wireless cellular networks: Deployment, challenges, and applications," *IEEE Communications Magazine*, 2020.

[9] Y. Sun, M. Peng, Y. Zhou, Y. Huang, and S. Mao, "Application of machine learning in wireless networks: Key techniques and open issues," *IEEE Communications Surveys & Tutorials*, 2019.