Machine Learning models for the prediction of Wi-Fi links performance using a CityLab testbed

Paulo Marques

Machine Learning models for the prediction of Wi-Fi links performance using a CityLab testbed

Paulo Marques[1]
[1] Instituto Politécnico de Castelo Branco, Portugal
paulomarques@ipcb.pt

Abstract

The Wi-Fi links performance depends in a highly complex way on the actual topology, channel qualities, spectral configurations, etc. Existing Wi-Fi radio link performance models usually adopt explicit and bottom-up approaches in order to predict throughput figures based on Markov chains and SINR levels. In this work we have validated a new approach for predicting the performance of Wi-Fi networks. Based on data measurements from the outdoor Wi-Fi CityLab testbed in Antwerp we have tested four different supervised learning algorithms. We observed that abstract "black box" models built using supervised machine learning techniques – without any deep knowledge of the complex interference dynamics of IEEE 802.11 networks – can estimate the link throughput with very good accuracy, reaching a value of R2-score of 90% for the case of the Gradient Boosting Regressor.

1 Introduction

Accurate prediction of wireless performance links can be very useful to optimize the Wi-Fi radio planning and resources allocation. However, the vast variety of possible wireless configurations and propagation scenarios make it hard to de-sign explicit\theoretical models to forecast the performance of a specific link. Wi-Fi networks are notoriously hard to model in multi node scenarios. They exhibit several performance intricacies due to complex interactions between the PHY and MAC layers, which manifest themselves in frequency, spatial and time domains.

Existing radio link performance models for Wi-Fi networks, such as the model proposed in [1], usually adopt explicit and bottom-up approaches; they model the actual mechanics of the protocol (for example, the CSMA/CA procedure of the MAC layer) in order to predict throughput figures based on Markov chains.

Due to the difficulty of predicting performance in the presence of complex interference patterns, most works proposing models or optimizations for the PHY layer (e.g.,[2],[3]) are reduced to using SINR-based models and ideal AWGN channels. Although SINR models can provide a characterization

of the Shannon capacity at the PHY layer, they are not meant to capture IEEE 802.11 performance and they can fail to capture important CSMA/CA performance patterns.

In this experiment we did an experimental validation of a different approach for predicting the performance of Wi-Fi radio links. Rather than manually fitting analytical models to capture complex dependencies, we have directly learned the models themselves, using Machine Learning techniques with a limited set of observed measurements. In fact, we do not attempt to seed a pre-existing model (such as SINR based or Markov-based) with measurements. Rather, we learn and build the model itself from a limited set of measurements (state parameters) as illustrated in Fig.1.

We treat Wi-Fi links as black boxes with potentially unknown internal mechanics. Such a black box takes some input parameters and it outputs the estimated throughput value.

The main objective of this work is the experimental validation of machine learning algorithms for predicting the performance of Wi-Fi radio links in multi node scenarios.

This paper is organized as follows: Section II describes the setup of this experiment, Section III describes the collected measurements and do a correlation analysis, Section IV proposes four Machine Learning algorithms to forecast the Wi-Fi link throughput, Section V shows the performance analysis and finally section VI concludes the paper and hints at future work.

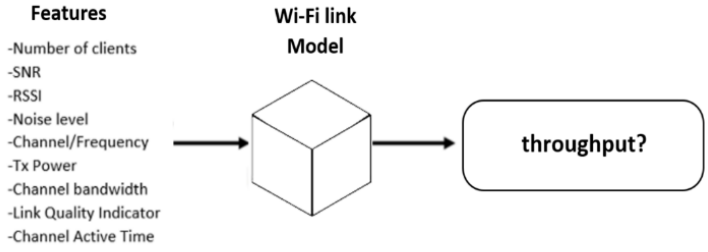


Fig. 1. Prediction of a link throughput based on a "black-box" model.

2 Setup of the Experiment

In this experiment we have used the CityLab (part of City of Things) testbed which is a smart cities FIRE testbed federated through the Fed4FIRE federation, operated by imec [4]. It is intended for large-scale wireless networking experimentation at a city neighbourhood level in the unlicensed spectrum. CityLab is in the city center of Antwerp, Belgium. The testbed can be found in the streets in and around the city campus of the University of Antwerp, in an area of about 0.5km by 0.5km. This testbed is a realistic environment where experiments typically face a lot of external radio interference from nearby equipment (e.g. Wi-Fi networks, IoT devices, …). Hardware is installed at 50 locations, each with its own gateway attached to houses in the street or installed on a pole on a roof. Each gateway houses multiple radios with full low-level access for experimenters, including Wi-Fi at 2.4GHz and 5GHz.

Fig. 2 illustrates two outdoor nodes from the CityLab testbed and Fig. 4 shows the area of the CityLab testbed where this experiment was remotely carried through the jFed toolkit (Fig. 3). In order to test different deployment scenarios and configurations, a gateway acts as experiment's controller which can change the configuration of all the nodes on the fly.

Fig. 2. Example of gateway deployment in the city of Antwerp available for remotely wireless experimentation.

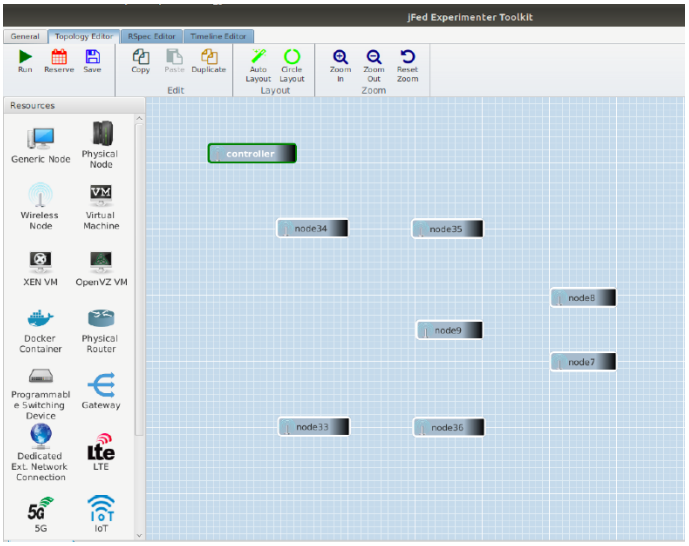


Fig. 3. jFed toolkit used to remotely setup the experiment.

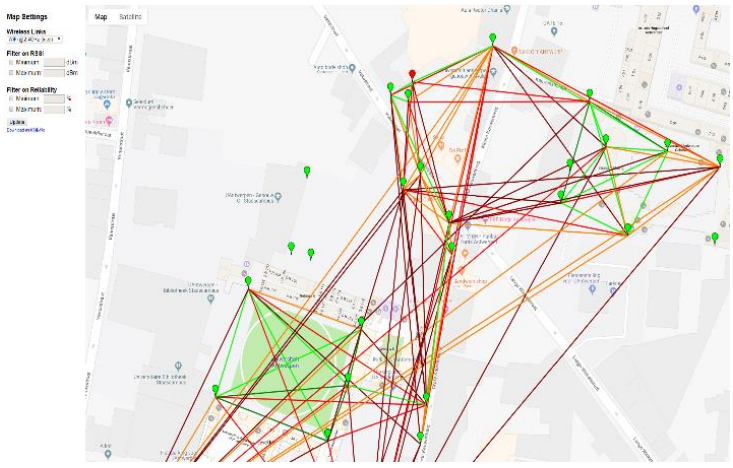


Fig. 4. Layout of the CityLab wireless testbed used in this experiment.

3 Measurements and Correlation Analysis

In this work we have performed N short-duration controlled experiments in the CityLab Wi-Fi outdoor testbed. Considering the "black box" representation of Fig. 1, each experiment consists in measuring the throughput (t) of a given link (l), for each combination of features. Those features are: number of clients, SNR, RSSI, noise level, channel, txPower, and the link quality in percentage. The goal is to expose the learning procedure to a wide variety of possible configurations. In total we did 3851 different tests.

In this experiment the throughput prediction is a multivariable regression problem with seven input features and one output to be estimated. Priory to build the Machine Learning models is important to understand the variables interdependencies and therefore the correlation level between them was computed according to the equation 1.

$$\boldsymbol{\rho} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\cdot\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} = \frac{\boldsymbol{cov(x,y)}}{\sqrt{\boldsymbol{var(x)}\cdot\boldsymbol{var(y)}}} \quad (1)$$

Fig. 5 shows the measured correlation matrix where ρ=1 means a perfect positive correlation between the variables; ρ= -1 means a perfect negative correlation between the variables and ρ=0 indicates that the variables don't have linear dependencies between then. Based on these results we can see that there is a strong positive correlation between the txPower and the throughput and a strong negative correlation between the number of clients and the throughput. These dependence between variables indicate that linear regression models can be used in the throughput estimation process.

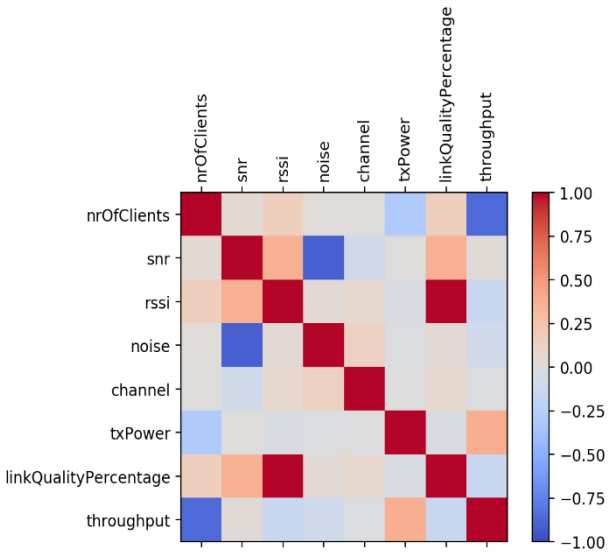


Fig. 5. Correlation matrix between all the measured features of the Wi-Fi links.

4 Machine Learning Models

Let us consider $(X,t)_{i=1}^{N}$ the set of N measurements. X is a matrix of multiple independent input variables $(x_1, x_2, \ldots, x_7)_{i=1}^{N}$, i.e., number of clients, SNR, RSSI, noise, channel, txPower and link Quality Percentage. The goal is to find a function $f: X \to t$ that maps x_i to a value close to t_i for each measurement i. This is an instance of a regression problem where we the function f is learned directly from the observed data.

The estimate $\hat{f}(X)$, minimizes the loss function $\Psi(t,f)$ given by equation (2):

$$\hat{f}(X) = t \Leftrightarrow \hat{f}(X) = arg \min_{f(X)} \Psi(t, f(X)) \qquad (2)$$

There are several supervised learning methods in the literature to solve multiple regression problems (e.g. [5]). In this experiment we are going to test the following four Machine Learning algorithms: Gradient Boosting Regressor, Linear Regression, kNN (k-Nearest Neighbors) and Decision Tree. We have used the Python machine learning package scikit-learn [6] to implement the various models.

5 Performance Analysis

The objective of this experiment is to test the performance of the predictive algorithms of Wi-Fi throughput with unknown combinations of features. As such, we only predict throughputs for data points that do not appear in the N measurements used for learning (or training). To this end, we split our total set of measurements into a training set and a test set. The training set consists in the actual N measurements used for learning the models and their parameters, whereas the test set is used only once, for measuring the final accuracy. We compute the root mean squared error (RMSE) for each algorithm:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(t_i - \hat{t}_i)^2} \quad (3)$$

where t_i is the actual measured throughput and $\hat{t}_i$ is the estimated value. We also compute the R^2-score given by:

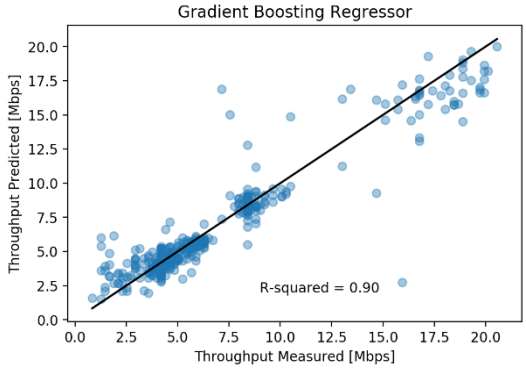
$$R^2 = 1 - \frac{\sum_i (t_i - \hat{t}_i)^2}{\sum_i (t_i - \bar{t})^2} \quad (4)$$

where $\bar{t}$ is the average throughput. Concretely, the R2-score quantifies how well a predictor does, compared to the simplest baseline strategy, which always predicts the mean throughput. It is equal to 1 if there is a perfect match between predicted and measured throughputs.

For each algorithm we have computed the RMSE and the R2-score (Table 1), moreover, in order to visualize the actual predictions in detail, we also show a scatter plot of the predicted throughputs, against the actual measured throughputs as illustrated in Fig. 6. On these plots, the closer the points are to the diagonal, the better the prediction accuracy. The Gradient Boosting Regressor outperforms the other methods and produce fewer outlying predictions.

Method	RMSE (Mbps)	R2-score [%]
Gradient Boosting Regressor	1.48	90%
Linear Regression	2.39	73%
kNN	1.53	89%
Decision Tree	1.73	86%

Table 1. Performance analysis of the machine learning algorithms



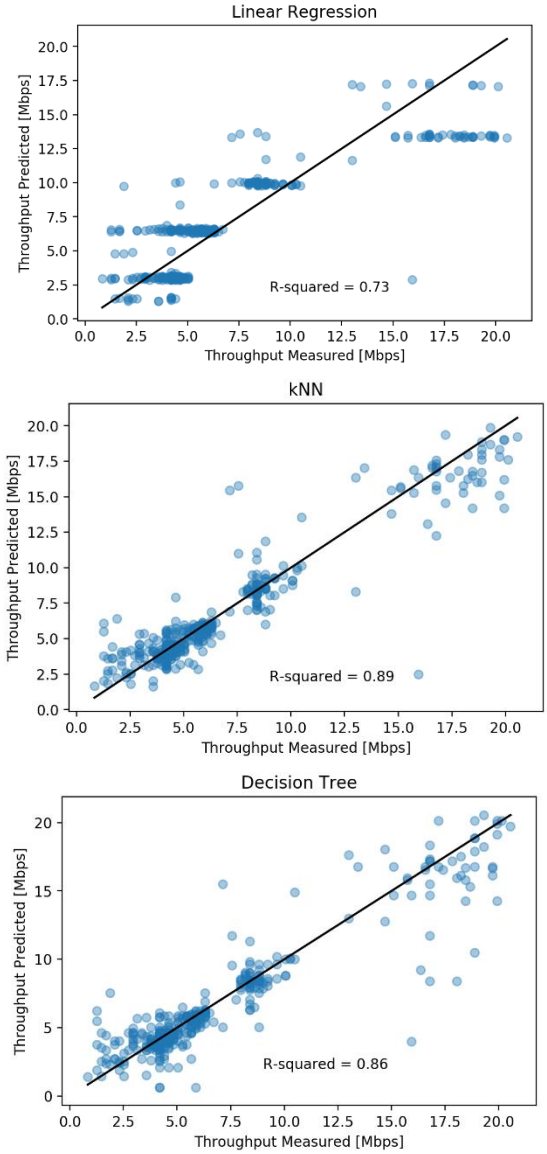


Fig. 6. Predicted versus measured throughput for 4 Machine Learning algorithms.

6 Conclusions and future work

The Wi-Fi links performance depends in a highly complex way on the actual topology, channel qualities, spectral configurations, etc. It is especially hard to predict in quantitative terms how a given configuration will perform.

In this experiment we advocate an approach of "learning by observation" that can remove the need for designing explicit and complex performance models. We use machine learning techniques to learn implicit performance models, from a limited number of real-world measurements. These models do not

require to know the internal mechanics of interfering Wi-Fi links.

In this work we investigated and validated a different approach for predicting the performance of Wi-Fi links. Rather than manually fitting complex models to capture complex dependencies, we have shown that it is possible to directly learn the models themselves, from a limited set of observed measurements. This approach bypasses the usual analytical modelling process, which requires deep knowledge, and yet often yields models that are either too restricted or too inaccurate [7].

Based on data measurements from the outdoor Wi-Fi CityLab testbed in Antwerp (imec) we have tested four different supervised learning algorithms. Using supervised machine learning techniques, it is possible to generalize the observations made on this limited subset of measurements, while still capturing the com-plex relationships between the inputs. We build such implicit models using real-world measurements and we test them systematically, by asking them to predict the throughput for links and configurations that have never been observed during the initial measurement phase

We observed that abstract "black box" models built using supervised machine learning techniques – without any deep knowledge of the complex interference dynamics of IEEE 802.11 networks – can estimate the link throughput with very good accuracy, reaching a value of R2-score of 90% for the case of the Gradient Boosting Regressor.

A scientific level, the results obtained on the modelling of multi-node Wi-Fi networks have potential to help on the developing of better resource management algorithms and help provide guidance to radio network planners.

A possible follow-up of this work is the extension of the "black box" approach to forecast the QoE (Quality of Experience) delivered by the Wi-Fi link for specific applications such as video or web browsing, taking as inputs QoS parameters. Another interesting follow-up is the extension of the Machine Learning models to the forecast the capacity of LTE radio links without using active transmission over the mobile network.

Acknowledgment

This work was funded from the Fundo Europeu de Desenvolvimento Regional (FEDER) through the Programa Operacional Regional do Centro (CENTRO2020) [Project Nr. 17711].

References

1. G. Bianchi. Performance analysis of the IEEE 802.11 distributed coordination function. IEEE Journal on Selected Areas in Communications.
2. S. Rayanchu, V. Shrivastava, S. Banerjee, and R. Chandra. FLUID: Improving through-puts in enterprise wireless LANS through flexible channelization. In ACM MobiCom, 2011.
3. V. Mhatre, K. Papagiannaki, and F. Baccelli. Interference mitigation through power control in high density 802.11 WLANs. In IEEE INFOCOM, 2007.
4. J. Struye, B. Braem, S. Latré and J. Marquez-Barja, The CityLab testbed — Large-scale multi-technology wireless experimentation in a city environment: Neural network-based interference prediction in a smart city, IEEE INFOCOM 2018
5. Friedman, J. (2001). Greedy boosting approximation: a gradient boosting machine. Ann. Stat. 29, 1189–1232.
6. F. Pedregosa, G. Varoquaux, A. Gramfort, and al. Scikit-learn: Machine learning in Py-thon. Journal of Machine Learning Research, 12:2825– 2830, 2011.
7. J. Herzen, H. Lundgren and N. Hegde "Learning Wi-Fi Performance", 12th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON), 2015.