Data Aggregation and Clustering for Traffic Prediction in Backbone Optical Networks

Aleksandra Knapińska, Piotr Lechowicz, Adam Włodarczyk, Krzysztof Walkowiak
Department of Systems and Computer Networks, Wrocław University of Science and Technology, Wrocław, Poland
aleksandra.knapinska@pwr.edu.pl

***Abstract*—We propose new, practical models for traffic prediction in backbone networks, balancing the forecast quality and system complexity. We show the benefits of data aggregation and clustering techniques with various ML algorithms.**

***Index Terms*—network traffic prediction, machine learning, application-aware network**

I. Introduction

The constantly increasing internet traffic volume and the rising popularity of various network-based services trigger the fast development of new technologies, using tremendous amounts of measured heterogeneous data. Machine learning (ML) techniques are gaining popularity and are applied for various tasks in the field of optical networks [1]. New, holistic approaches such as multilayer application-aware network optimization are promising for the existing and newly deployed backbone networks [2]. Furthermore, individual connections are aggregated into optical channels, thus creating visible trends and seasonality patterns in traffic measurements. Knowledge of the future volume of traffic in various network regions can increase the effectiveness of resource allocation and decrease bandwidth blocking [3], [4] and enable more thoughtful network upgrades [5].

The prediction of traffic for the whole network can be obtained, e.g., by creating dedicated models for network links, lightpaths or traffic types [4], [6] or using graph neural networks [3], [7], which are a powerful tool for capturing complex traffic patterns. However, simpler and thus easier-to-implement in real-world scenarios methods are recently explored by researchers in the optical networking domain, achieving similar or better performance and requiring significantly less training data and computational power [8], [9]. A recent study [8] has shown how linear regression (LR) can achieve comparably high accuracy to deep learning models, with a noticeable complexity reduction in optical systems. Furthermore, the streaming algorithm with LR as a base proposed in [9] outperformed different neural network models trained on 60 times more data. Additionally, graph neural networks are a powerful prediction tool for irregular structures [10], such as network links. However, having more generic knowledge about the traffic between each pair of nodes can help create a more effective routing algorithm [6], [11]. Note that such a resulting graph, where each node is connected to all the other ones, creates a very regular but colossal structure. To this end, there is a space for new traffic prediction models that balance their complexity and forecast quality. Therefore, in this paper, we propose new models for traffic prediction across the whole network, balancing the forecast quality and system complexity. The proposed methods can be easily deployed in practical applications of real-world backbone networks. We show the benefits of data aggregation and clustering techniques through experiments with various ML algorithms.

II. Proposed Models

In the considered problem, the traffic is predicted for each pair of nodes in a backbone optical network. Thus, the number of timeseries to forecast is significant. In a real-world topology, the total number of node pairs can be several hundred, e.g., in a 28-node network, that is 756 separate forecasts. Therefore, in this work, we propose mechanisms to decrease the number of required models and analyze the quality of created aggregate models.

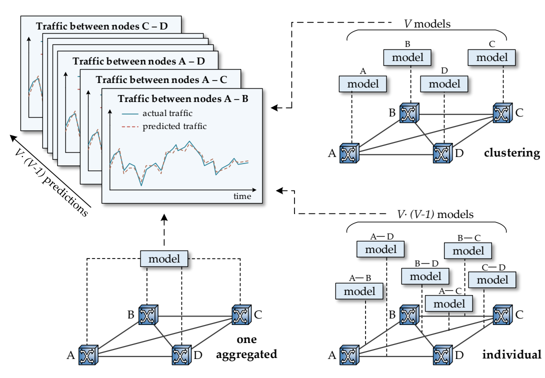

Fig. 1. Proposed prediction models.

In previous work [12], we showed how using other traffic types as features can enhance the prediction of a single traffic type. In a practical setting of a backbone network topology, we can extend this idea to efficiently predict the traffic between each pair of nodes. In more detail, in a traditional approach (e.g., [4]), we can train local models for each pair of nodes (*individual* model in Fig. 1). Instead, we propose to employ multioutput regression, thus creating one aggregated model for the entire network with features describing traffic in all node pairs (*one aggregated* model in Fig. 1). As suggested in [12] and [9], we base the forecasts on two past significant traffic

978-3-903176-54-6

measurements: a day and a week before the predicted period because of traffic seasonality in backbone networks. Thus, the created model uses $2k$ input features, with k representing the number of forecasted timeseries (pairs of nodes). However, in large real-world networks, this number can be very significant, which increases the complexity of the resulting model [13]. Moreover, the traffic in distant locations of the network can vary immensely. Therefore, the final proposed model balances the number of models and their quality with clustering. The idea is to cluster the predicted timeseries by location. In this setting, each network node has a dedicated multioutput prediction model for its incoming or outgoing traffic (*clustering* model in Fig. 1). Note that the number of required prediction models is equal to the number of network nodes, which is significantly smaller than the number of individual traffic timeseries. Moreover, we analyze the traffic between pairs of nodes instead of specific network links. That makes our analysis agnostic to the used routing algorithm and makes it possible to base dynamic routing decisions on the forecasts and choose the most appropriate lightpaths for upcoming connection requests.

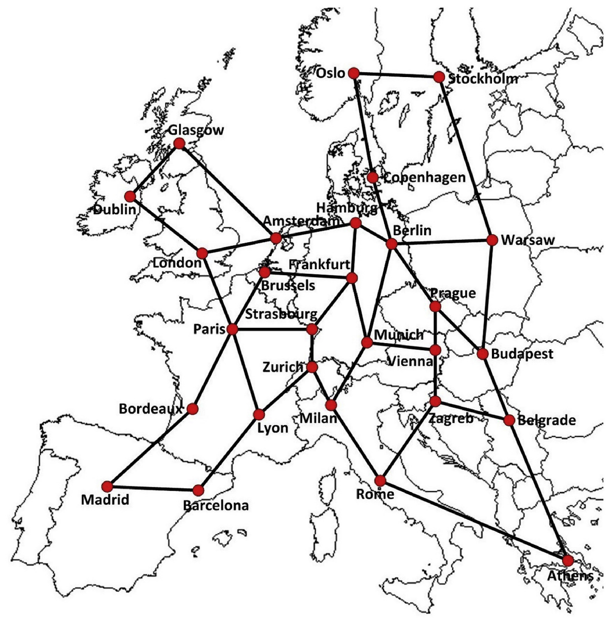

Fig. 2. Euro28 topology.

III. Numerical Experiments

For the experiments, we use semi-synthetic data. In more detail, the dataset used in this work was created using real measurements from the Seattle Internet Exchange Point, collected over six months with a 5-minute sampling. Using a custom generator from [11], the raw aggregated traffic data is differentiated into specific traffic types and spread across a backbone optical network. The resulting dataset contains aggregated traffic timeseries for each undirected pair of nodes. For our analysis, we use the Euro28 topology (see Fig. 2) with 28 nodes. Thus, the number of target timeseries to forecast is 756. A zoomed-in fragment of chosen traffic timeseries is provided in Fig. 3. Note the high level of fluctuations and differences in volume between them.

We use the first 5-months' worth of data to train the models and the final month for evaluation. The performance metric of choice is the mean absolute percentage error (MAPE), enabling direct comparison between traffic timeseries differing in volume. As a baseline solution from the literature (e.g., [4]), we train individual, dedicated prediction models for each traffic timeseries. These models only have information about the past measurements of the forecasted traffic pattern. In the following part, we analyze the models' prediction quality for all 756 timeseries.

As the proposed models are generic, they can be used with any ML algorithm as a base. To demonstrate the solution's versatility, in this paper, we analyze three diverse ones used in the literature for the network traffic prediction task [14]: linear regression (LR), decision tree regressor with MSE split criterion and max depth of five (CART), and multilayer perceptron with one hidden layer of twenty-five neurons, with the *ReLU* activation function and *adam* optimizer (MLP).

In Tab. I, we present the obtained results – average error from the 756 predicted timeseries. Although we tested both clustering versions (traffic outgoing/incoming to a specific node), we noted almost identical results in both of them. Therefore, for the sake of conciseness, we only report the results of the latter approach. The first conclusion is that the prediction quality of each time series can be significantly improved through data aggregation into one model. The prediction errors obtained from the aggregate model were lower than those of the dedicated models, regardless of the ML algorithm. Note that only one model was needed to make these predictions instead of 756 separate ones. In turn, using the aggregated multioutput regression model can improve the prediction quality and drastically decrease the number of needed models.

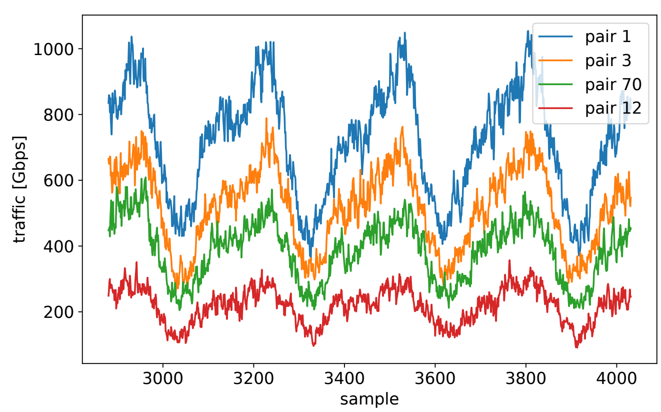

Fig. 3. Representative timeseries (zoomed-in fragment).

Furthermore, the models obtained through clustering enabled even more precise traffic forecasting. In particular, for both LR and CART, the prediction errors of the clustering model were lower than those of the one aggregated model.

TABLE I
MAPE VALUES OF THE PROPOSED MODELS FOR CONSIDERED ML ALGORITHMS.

	individual	**one aggregated**	**clustering**
LR	0.0530 – 0.1469, avg 0.0949	0.0535 – 0.1323, avg 0.0871	0.0491 – 0.1256, avg 0.0821
CART	0.0551 – 0.1483, avg 0.0964	0.0551 – 0.1288, avg 0.0867	0.0486 – 0.1290, avg 0.0833
MLP	0.0532 – 0.1491, avg 0.0947	0.0498 – 0.1253, avg 0.0833	0.0486 – 0.1290, avg 0.0833

For the MLP, the errors were as low as for the general model. Note that clustering balances the number of needed models and their complexity. By placing the prediction model in each network node, we only need 28 predictors instead of 756 in the considered case of the Euro28 topology. Moreover, by using the clustering techniques and creating local models instead of the general aggregate one, each model can be less complex, as it outputs much fewer predictions at once and uses much fewer features. Thus, deploying a prediction model for each network node for its incoming or outgoing traffic is simpler to build in practical settings. Our analysis shows that this approach enables obtaining forecasts of as good or better quality than the general models and significantly better than the individual models. As a summary, in Fig. 4 we present the MAPE values for all pairs of nodes for the LR as boxplots. A clear trend in the quality improvement of the prediction is evident for each subsequent model.

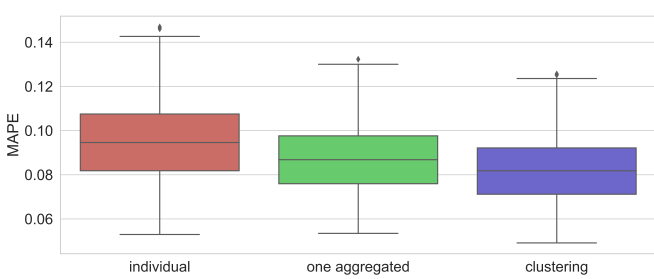

Fig. 4. MAPE values for all pairs of nodes as boxplots for the LR algorithm.

Finally, some patterns regarding each ML algorithm are noticeable in the obtained results. Individual prediction models for each timeseries were the most accurate using the MLP. However, we noted a more substantial improvement from dedicated models to the recommended clustering for the simpler CART and LR algorithms. With careful data aggregation and feature selection from proposed clustering techniques, a much more complex MLP was overhauled by CART and outperformed by LR.

IV. Conclusions

In this paper, we proposed new models for traffic prediction in backbone optical networks, employing data aggregation and clustering techniques. From the conducted experiments on various ML algorithms, it can be concluded that aggregating the network traffic data into a multioutput regression model improves the prediction quality of individual forecasts for traffic between pairs of network nodes. Furthermore, it can be significantly enhanced by using clustering techniques with simultaneous model complexity reduction. In turn, the proposed approach implies two clear benefits: fewer needed models and better quality forecasts. Finally, with the proposed aggregation and clustering techniques, simple ML algorithms outperform neural network models. In the future, we plan to explore further data aggregation and clustering techniques to improve traffic prediction in backbone optical networks.

Acknowledgment

This work was supported by National Science Center, Poland under Grant 2019/35/B/ST7/04272.

References

[1] F. Musumeci *et al.*, "An overview on application of machine learning techniques in optical networks," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 2, pp. 1383–1408, 2018.

[2] V. Lopez *et al.*, "On the benefits of multilayer optimization and application awareness," *Journal of Lightwave Technology*, vol. 35, no. 6, pp. 1274–1279, 2017.

[3] M. Aibin *et al.*, "On short-and long-term traffic prediction in optical networks using machine learning," in *International Conference on Optical Network Design and Modeling (ONDM)*, 2021.

[4] R. Goścień, "Traffic-aware dedicated path protection in elastic optical networks," in *International Workshop on Resilient Networks Design and Modeling (RNDM)*, 2022.

[5] P. Lechowicz *et al.*, "Greenfield gradual migration planning toward spectrally-spatially flexible optical networks," *IEEE Communications Magazine*, vol. 57, no. 10, pp. 14–19, 2019.

[6] H. Khodakarami *et al.*, "Quality of service provisioning and energy minimized scheduling in software defined flexible optical networks," *Journal of Optical Communications and Networking*, vol. 8, no. 2, pp. 118–128, 2016.

[7] D. Andreoletti *et al.*, "Network traffic prediction based on diffusion convolutional recurrent neural networks," in *IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, 2019.

[8] Y. Fan *et al.*, "Linear regression vs. deep learning for signal quality monitoring in coherent optical systems," *IEEE Photonics Journal*, vol. 14, no. 4, pp. 1–8, 2022.

[9] A. Knapińska *et al.*, "Long-term prediction of multiple types of time-varying network traffic using chunk-based ensemble learning," *Applied Soft Computing*, vol. 130, p. 109694, 2022.

[10] Z. Wu *et al.*, "A comprehensive survey on graph neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 1, pp. 4–24, 2021.

[11] A. Włodarczyk *et al.*, "An algorithm for provisioning of time-varying traffic in translucent sdm elastic optical networks," in *International Conference on Transparent Optical Networks (ICTON)*, 2020.

[12] A. Knapińska *et al.*, "Machine-learning based prediction of multiple types of network traffic," in *International Conference on Computational Science*, 2021.

[13] X. Hu *et al.*, "Model complexity of deep learning: A survey," *Knowledge and Information Systems*, vol. 63, no. 10, pp. 2585–2619, 2021.

[14] A. Knapińska *et al.*, "Prediction of multiple types of traffic with a novel evaluation metric related to bandwidth blocking," in *IEEE Global Communications Conference (GLOBECOM)*, 2022.