# First Demonstration of Imbalanced Data Learning-Based Failure Prediction in Self-Optimizing Optical Networks with Large Scale Field Topology

Boyuan Yan, Yongli Zhao*, Yajie Li, Xiaosong Yu, and Jie Zhang
Beijing University of Posts and Telecommunications
Beijing, China
{yanboyuan, yonglizhao, yajieli, xiaosongyu, lgr24}@bupt.edu.cn

Yilin, Haomian Zheng
Huawei Technologies Co., Ltd.
Shenzhen, China
{yi.lin, haomian.zheng} @huawei.com

***Abstract*—Machine learning is a promising solution to address some issues in large scale optical networks. Failure prediction is a typical application which can be completed by the aid of machine learning. However, the constraint of massive imbalanced data distribution limits the performance of machine learning in large scale optical networks. The paper first demonstrate the imbalanced data learning based failure prediction in self-optimizing optical networks (SOON). The large scale field topology with 274 nodes and 487 links is considered. Experimental results show that the proposed imbalanced data learning based failure prediction solution can get high accuracy effectively.**

***Keywords—optical networks, machine learning, failure prediction, imbalanced data, large scale***

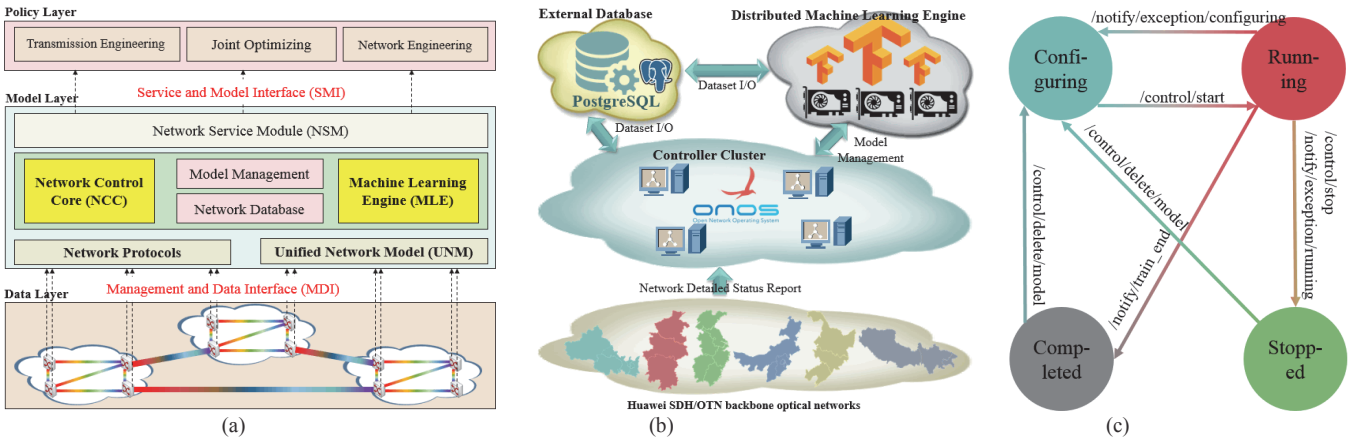
## I. INTRODUCTION

As the most important transmission media, optical fibers carry tons of data. One simple fiber link failure may lead to huge economic losses. Therefore, accurate failure prediction becomes more and more important for network operators. Recently, machine learning (ML) becomes a hot topic in various prediction tasks because of its capability about non-linear feature extraction. Some researchers have made some contributions in the area of failure prediction. Z. Wang proposed a failure prediction method by using support vector machine and double exponential smoothing [1]. L. Barletta investigated a ML technique to predict whether the bit-error-rate of unestablished lightpaths is within the normal range based on traffic, routing, and modulation format [2]. T. Panayiotou proposed a Gaussian process classifier to predict the failure probability of each optical link [3]. All these ML algorithms reaches high prediction accuracy on well-handled train dataset and test dataset. However, network failures could be usually classified into multiple reasons, for example, fiber cut, or low output power. In reality, the sizes of network failure records caused by different reasons are vastly imbalanced, which may be unsatisfied for ML training. Because the imbalance may make direct learning pose unsatisfying results over-focusing on the accuracy of prediction and deriving a suboptimal model [4]. The imbalanced learning problem is concerned with the performance of ML algorithms in the presence of under-represented data and serve class distribution skews [5]. Especially in the large scale optical networks, how to address the imbalanced learning problem is pressing and important for network operators.

In this paper, we build a ML-enabled network platform, named self-optimizing optical network (SOON). Based on SOON, an imbalanced-data-learning failure prediction (IFP) algorithm is proposed to address the imbalance issue of dataset for performance improvement. The rest of this paper is organized as follows. Section II shows the architecture of SOON. Section III describes the procedure of IFP algorithm. Section IV gives the large scale network scenario and experimental results. The conclusion is drawn in section V.

## II. SOON ARCHITECTURE

Software defined optical networks (SDON) could provide logically centralized control mechanism, flexible network function via open northbound application program interface (API), and etc. Based on SDON, a new network architecture named self-optimizing optical networks (SOON) is proposed, which integrates ML technology to SDON to improve intelligence capability of optical networks. SOON architecture is composed of three layers, i.e., data layer, model layer, and policy layer as Fig. 1(a) shows. Data layer contains physical/virtual optical networks. Each element of network connects to model layer through Management and Data Interface (MDI). Model layer is the responsible to




(a) (b) (c)

Fig.1. (a) SOON architecture, (b) SOON deployment scenario, (c) state machine of ML model's lifecycle



978-1-5386-6158-1/18/$31.00 ©2018 IEEE

control network and lifecycle of ML models. Model layer receives instruction from policy layer through service and model interface (SMI). Policy layer builds various policies for different applications of optical networks based on trained ML models in model layer. ***(1) Data Layer:*** In data layer, the element of optical networks, named self-optimizing network element (SNE), contains intelligence board on its slots to enable on-board ML. The core of intelligence board is powerful artificial intelligence chip, which is able to execute compute-intensive tasks of ML algorithms. SNE could not only support traditional protocols like GMPLS and PCEP, but also report detailed physical parameters, complete data process and help model layer train ML models. ***(2) Model Layer:*** In model layer, unified network model reformats big data from data layer into unified data structure, and network service module (NSM) provide basic services for policy layer including traditional services like topology discovery, and ML-related services like model creation. Network control core (NCC) and machine learning engine (MLE) are two key parts of model layer. NCC is implemented based on software-defined network (SDN) controller, and designed to control optical network with flexibility and scalability, but without intelligence. And MLE is implemented based on ML platform, and designed to train, evaluate, and apply ML models for specific network applications. Besides, network database is specially aimed to provide efficient storage and access for massive data collected from data layer. Model management module processes data from network database to build available train datasets and test datasets following specific policies, and controls model lifecycle to provide ML-related services for policy layer. ***(3) Policy Layer:*** The policies used in the policy layer could be divided into three parts. Policies for transmission engineering and network engineering focus on performance improvement on physical layer and network layer, respectively. Joint optimizing policies consider network optimization for specific target by considering both physical layer and network layer, such as physical impairment-aware routing.

The deployment of SOON has been shown in Fig. 1 (b). The data layer is carried by Huawei backbone optical network in different area of China. The model layer is implemented based on SDN controller (ONOS) and distributed machine learning engine (TensorFlow). An external databased is built based on PostgreSQL to store the training dataset and test dataset. Fig. 1(c) shows the state machine of ML model's lifecycle implemented by a designed websocket-based protocol between ONOS and TensorFlow. There are four states for a ML model, i.e., Configuring, Running, Stopped and Completed. In Configuring state, ML model could be configured with the detailed parameters. Running state indicates that ML model is running on train dataset. Stopped state is an abnormal state indicting that some mistake happens while running, or training process is being interrupted. Completed state indicates that ML model is ready for application. Both the conversion between two states and available operations inner a state are achieved by SOON message, which use URI prefix to express operations.

## III. Imbalanced-Data-Learning Failure Prediction

As mentioned above, failure data collected from model layer are imbalanced in commercial networks. Based on SOON, we propose imbalanced-data-learning failure prediction (IFP) algorithm, which is composed of three steps: data-preprocess, balanced distribution adjustment, and artificial neural network (ANN) as Fig. 2 shows. ***(1) Data-preprocess:*** it is a basic step to handle original dirty data by using data completion, data merging, data competition, data validity check, and other methods. Dirty data include missing data, wrong data, redundant data, and non-standard representation of the same data. The results of analyzing dirty data can be damaged or unreliable. Dataset becomes cleaner and smaller after data-preprocess. ML model trained based on it would not learn wrong features, but skip some important features. In a network failure dataset whose input is correlated performance data in time series, and the output is whether failure will happen in future. If the failure data caused by specific reason plays the main part of dataset, it is easy for ML model to reach high precision on this failure prediction, but difficult to learn the key features on other failures efficiently. Generally, the distribution is mean on various failure reasons, so the possibilities for ML model to learn the features from different failures are same. ***(2) Balanced distribution adjustment:*** in order to balance the dataset for mean distribution by changing data size of each failure, we use random deletion to decrease dataset size, and introduce Gaussian noise to increase size. Random deletion removes data randomly, and Gaussian noise is used to generate new trusted data without changing key features. ***(3) ANN:*** based on the balanced train dataset, ANN is designed to execute classification to predict whether network failure would happen. Data preprocess is implemented by UNM. Balanced distribution adjustment and ANN are implemented in MLE. All these three steps are controlled by failure prediction policy in policy layer through SMI.

## IV. Field Trial Scenario and Results

SOON has been implemented on distributed servers. The training dataset is collected from the large scale commercial optical networks. Fig. 3(a) shows the field trial scenario including telecommunication room of the network and servers that includes GPU server, computing server, and storage server where MLE, NCC and network database are running. 56901 alarms signals are collected from April 18th

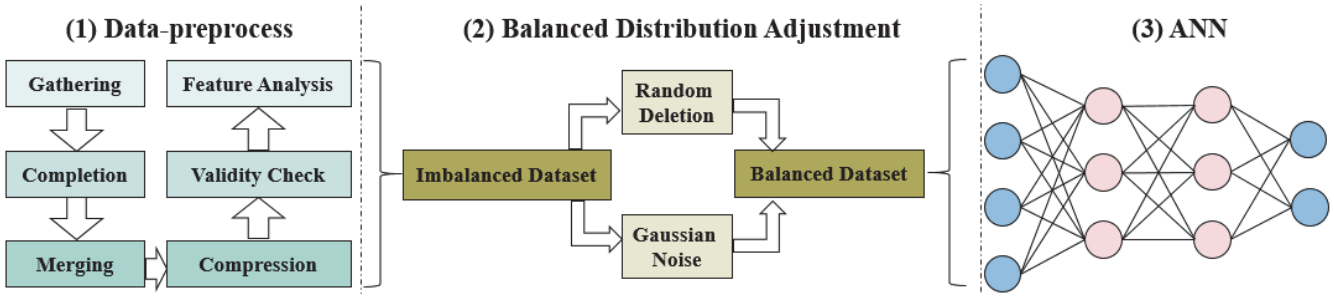



Fig. 2. Flow diagram of IFP algorithm

to May 13th on large-scale optical networks with 274 nodes and 487 links as Fig. 3(b) shows. The state change of IFP model is captured by Wireshark tool as Fig. 3(c) shows. After configuration for model parameters and dataset transmission, state changes to Running from Configuring. Then this model is applied after training ends. We gather IN_PWR_LOW, OUT_PWR_ABN and R_LOS as predicted alarms, and related performance as input. The ratio of data size among these three alarms is 53:9:1. The input size is 13, whose first element indicates the alarm type, and last 12 elements indicate a time series in 3 hours. The output is probability of alarm happen event in future 15 minutes. In order to evaluate the performance of IFP, we use simple ANN algorithm without balanced distribution adjustment process as benchmark. We extends each size of different alarm type to 53, and select 13 of them to form test dataset. Then, by randomly selecting parts of 40:40:40 dataset, IFP(25%), IFP(50%), IFP(75%) and IFP(100%) mean 10:10:10, 20:20:20, 30:30:30, and 40:40:40 respectively. Both IFP and ANN use three hidden layers with 64, 64, and 96 neurons. Besides, MSE loss function, 5e-3 fixed learning rate, SGD optimizer, and other detailed parameters are set.

Fig. 3(d) shows the loss variation when iteration increases. All five algorithms convergence to less than 0.1. The loss of IFP(25%) is relatively high because size of train dataset is the smallest. Fig. 3(e) shows the prediction precision. Compared with ANN whose highest precision is 0.8281, IFP(50%) and IFP(75%) are higher than ANN with 0.8594 and 0.875 precision respectively, that means balanced data is easily learned by ML algorithm. However, IFP(25%) and IFP (100%) are relatively low, because IFP(25%) cut off much features by randomly deletion, and IFP(100%) introduces excessive noise by Gaussian noise. Further, Fig. 3(f) analyzes the reason of performance improvement of IFP. There is no mis-prediction of IN_PWR_LOW in ANN, which means ANN learns features of this alarm perfectly, but skips feature learning of other alarms. And in IFP, the learning for each type of alarms is relatively balanced.

## V. Conclusions

In this paper, we propose SOON architecture to integrate ML technology into control and management of large scale commercial optical networks. Based on SOON, we propose IFP algorithms to overcome the influence of imbalance distribution in dataset. Experimental results show that IFP could improve 4.69% accuracy at most.

## Acknowledgment

This work has been supported in part by China State Grid Corp Science and Technology Project (No. 5210ED180047), NSFC Project (Nos. 61571058 and 61601052), and State Key Lab of Advanced Optical Communication Systems Networks, China.

## References

[1] Z. Wang, et al., "Failure prediction using machine learning and time series in optical network," Optics Express, vol. 25, pp. 18553-18565, 2017.

[2] L. Barletta, et al., "QoT Estimation for Unestablished Lighpaths using Machine Learning," in OFC, Th1J.1, 2017.

[3] T. Panayiotou, et al., "Leveraging Statistical Machine Learning to Address Failure Localization in Optical Networks," J. Opt. Commun. Netw, vol. 10, pp. 162-173, 2018.

[4] Q. Wang, et al., "A Novel Ensemble Method for Imbalanced Data Learning: Bagging of Extrapolation-SMOTE SVM," Computational Intelligence and Neuroscience, vol. 2017, 2017.

[5] H. He, et al., "Learning from Imbalanced Data," IEEE Transactions on Knowledge & Data Engineering, vol. 21, pp. 1263-1284, Sept. 2009.

[6] B. Yan, et al., "First Demonstration of Machine-Learning-based Self-Optimizing Optical Networks (SOON) Running on Commercial Equipment," in ECOC, TuDS.3, 2018.

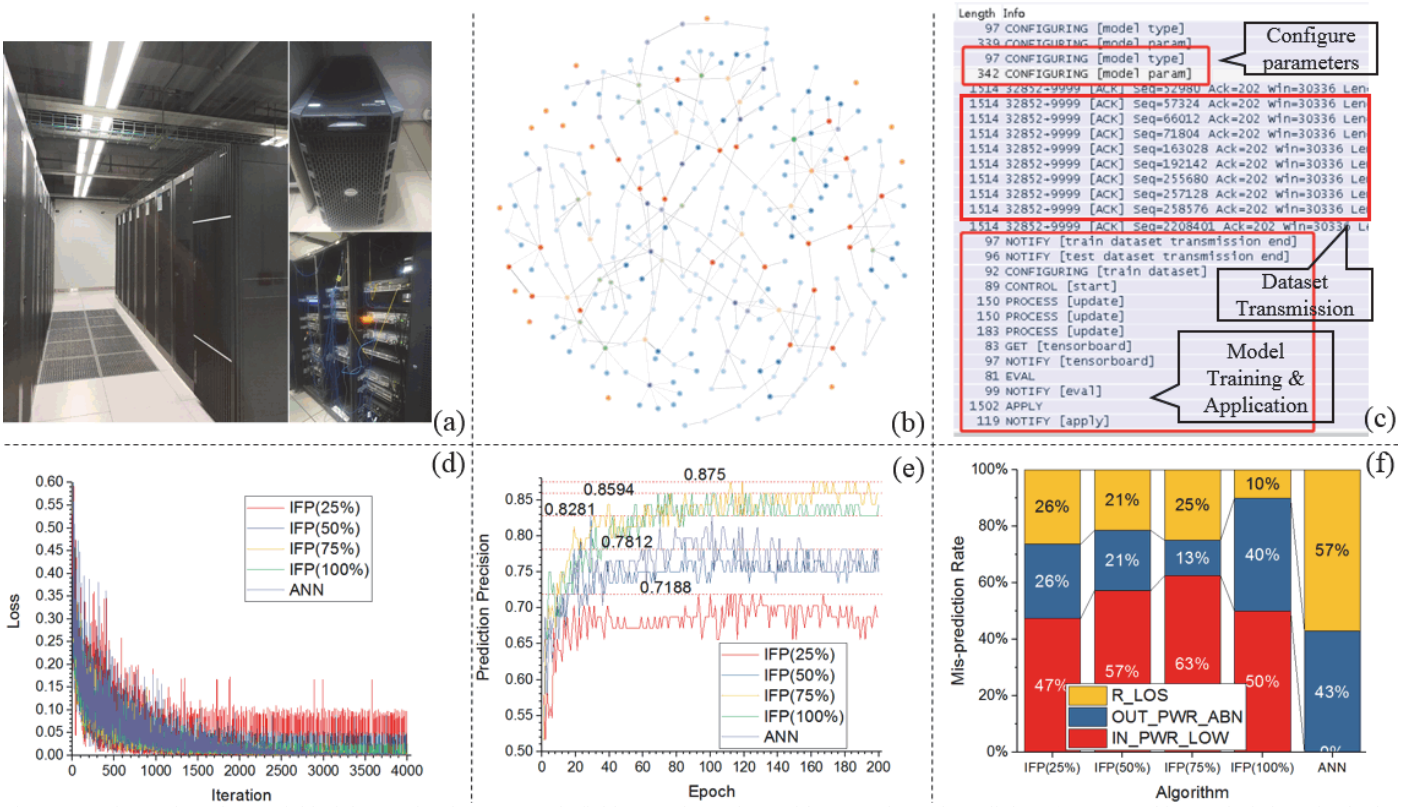



Fig. 3. Experimental result: (a) Field trial scenario; (b) Large-scale field network topology with 274 nodes and 487 links; (c) State machine analysis by Wireshark; (d) Loss variation with iteration; (e) Prediction precision with epoch; (f) prediction error analysis.

978-1-5386-6158-1/18/$31.00 ©2018 IEEE