

This is a postprint version of the following published document: Javier Mata; Ignacio de Miguel; Ramón J. Durán; Juan Carlos Aguado; Noemí Merayo; Lidia Ruiz; Patricia Fernández; Rubén M. Lorenzo; Evaristo J. Abril; Ioannis Tomkos. *Supervised Machine Learning Techniques for Quality of Transmission Assessment in Optical Networks*. In: 2018 20th International Conference on Transparent Optical Networks (ICTON). DOI: <https://doi.org/10.1109/ICTON.2018.8473819>

© 2018 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

Supervised machine learning techniques for quality of transmission assessment in optical networks

Javier Mata, Ignacio de Miguel, Ramón J. Durán, Juan Carlos Aguado, Noemí Merayo, Lidia Ruiz, Patricia Fernández, Rubén M. Lorenzo, Evaristo J. Abril, Ioannis Tomkos*

Universidad de Valladolid, ETSI de Telecomunicación, Campus Miguel Delibes, 47011 Valladolid, Spain

* Athens Information Technology (AIT) center, Greece

e-mail: javier.mata@ribera.tel.uva.es, {ignacio.miguel, rduran}@tel.uva.es

ABSTRACT

We propose and compare a number of machine learning models to classify unestablished lightpaths into high or low quality of transmission (QoT) categories in impairment-aware wavelength-routed optical networks. The performance of these models is evaluated in long haul communication networks and compared to previous proposals. Results show that, especially random forests and bagging trees approaches, significantly reduce the required computing time to classify the QoT of a given lightpath, while accuracy remains around 99.9%.

Keywords: Machine learning, quality of transmission, lightpath, impairment-aware optical networks.

1. INTRODUCTION

Lightpath Quality of Transmission (QoT) estimation prior to deployment is crucial in impairment-aware optical network operation, not only in terms of accuracy of the estimates, but also in terms of computing time, in order to be able to handle lightpath requests in real-time. Different approaches have been proposed so far to address this issue. Analytical or semi-analytical models are able to provide accurate results in terms of QoT estimation (e.g., Q-Tool [1][1]). However, this type of approaches may incur heavy computational load and increased computing time, thus compromising its practical implementation. Margined formulas have also been used to reduce computational load, but lead to network over-dimensioning [2]. Recent studies aim at reducing these margins by proposing schemes based on active monitoring [3], considering actual physical and traffic conditions in multi-period scenarios [4] or using monitored physical parameters in a learning process [5]. In fact, the adoption of artificial intelligence techniques, and in particular Machine Learning (ML), to estimate lightpath QoT is a very promising trend, as they may provide fast and accurate estimates, and may be able to adapt to changing conditions [6], [7], [8]. A Case-Based Reasoning (CBR) approach, relying on a historical set of data to predict the QoT of new lightpaths, is proposed in [7]. This approach is indeed several orders of magnitude faster than the above mentioned semi-analytical tool, Q-Tool [1], without undermining classification accuracy. A recent study [9] also makes use of ML classifiers (based on k-nearest neighbours and random forests) to predict whether the bit error rate (BER) of unestablished lightpaths meets requirements. In addition, a QoT estimator based on support vector machines (SVM) is proposed in [10]. It significantly reduces the computing time to estimate the QoT of a given lightpath, and even slightly improves the accuracy with respect to the CBR approach previously mentioned [7]. We now extend that work, analysing additional ML-based QoT estimators (based on logistic regression, classification and regression trees, bagging trees and random forests), and comparing with the SVM approach proposed in [10]. We incorporate a cross-validation procedure to analyse a grid of tuning and regularization parameters for each ML technique considered with the aim of determining the best configuration.

2. DESCRIPTION OF THE ML-BASED QoT ESTIMATORS

We have developed different ML-based QoT estimators to classify lightpaths into high and low quality categories based on a user-defined Q-factor threshold, an indicator of the QoT which is inversely related to BER. As stated in [10], this threshold is set to 16.9 dB, corresponding to a BER of 10^{-12} . Thus, lightpaths labelled with a Q-factor above the threshold, i.e., with low BER, are considered to comply with QoT requirements (high quality category), and those below the threshold are considered non-viable lightpaths.

The ML-based QoT estimators make use of supervised learning mechanisms, i.e., they are provided with a set of instances (examples) composed of a number of input variables (features) and their associated output variable (target value) in order to learn the corresponding mapping function. Particularly, in this case, the dataset consists of a collection of 10 Gb/s OOK lightpaths that are obtained by running off-line simulations, varying traffic load conditions, over the network under study. The attributes that describe the lightpath (features) consist of: the source and destination nodes, the set of links it traverses (represented by the percentage of its individual contribution to the total length of the lightpath), the associated wavelength, the total length, the sum of co-propagating lightpaths per link and the standard deviation of this number. The output variable is the associated Q-factor, calculated also off-line by means of the Q-Tool [1], which is transformed in a target binary variable according to the threshold mentioned above.

Two different approaches are proposed for the ML-based QoT estimators: the *hybrid approach* and the *plain approach*. The hybrid estimator takes advantage of the fact that the total length of the lightpath has a decisive impact on QoT [7]. Lightpaths shorter than a certain total length limit can be directly considered as having high quality, while lightpaths longer than another length limit can be considered as having low quality, regardless other features of the lightpaths. Thus, an uncertainty area, in which the quality category cannot be determined *a priori*, can be defined between these two threshold limits. Therefore, in the *hybrid approach*, a simple decision maker classifies cases outside the uncertainty area, while a ML-based module does so with all those cases within that uncertainty area. The limits of the uncertainty area are chosen so that the probability of successful classification of lightpaths outside this area is higher than 99.99% (like in [7], [10]). In contrast, the *plain approach* forces the ML-based modules to learn also the relevance of the total length by removing the decision maker, letting these modules to classify every lightpath regardless its total length.

2.1 Training, validation and testing phases

The QoT classifiers are trained by means of a training dataset, i.e., a set of example lightpaths whose class is known, in order to learn the corresponding mapping function between the characteristics of the lightpaths and the quality category. Depending on the approach, this training dataset will contain only cases belonging to the uncertainty area (*hybrid approach*) or cases throughout the whole length range (*plain approach*).

One of the most critical aspects that needs to be taken into account during the training and validation phases is the selection of the best model, avoiding *overfitting* (i.e., a model fitting the training data too closely, resulting in poor generalization) and *underfitting* (i.e., a model not complex enough so as to approximate well the data). Thus, a process consisting on finding the model and its tuning parameters with the best performance on a given metric (e.g., accuracy, area under the receiver operating characteristic curve or AUC, etc.) is run by performing 10-fold cross-validation (CV). On the one hand, 10-fold CV divides the training dataset into 10 folds and performs 10 iterations. In each iteration, a fold is removed and used to test the model obtained after training with the remaining folds. On the other hand, and at the same time, all possible combinations of tuning parameters are tested in each iteration (note that the initial number of iterations, 10, escalates with the number of tuning parameters to be tested). These tuning parameters may introduce regularization (e.g., penalization terms) in order to avoid overfitting and improve the performance of the model when executed in new (previously unseen) data. Finally, the performance is averaged over the 10 iterations that conform each possible combination of tuning parameters, allowing to select the best configuration in terms of the given optimization metric.

Table 1 shows the techniques employed in this study: SVM, logistic regression, classification and regression trees (CART), bagging trees (TREEBAG) and random forests (RF), together with their tuning parameters, as defined in the *caret* package of the R software environment, that we have used in this 10-fold CV process.

Table 1. Tuning parameters for the CV process. (TREEBAG has no tuning parameters)

Algorithm	Tuning parameter	Range	# of values in the range
SVM	C	$[2^{-5}, 2^{15}]$	11
	Γ	$[2^{-15}, 2^3]$	10
Logistic regression	α	$[0.1, 1]$	10
	λ	Values selected by the <i>caret</i> package	10
CART	<i>Complexity parameter</i>	Values selected by the <i>caret</i> package	8
RF	<i>MTRY</i>	Number of predictors (features)	8

Once the models are trained and tuned thanks to CV, the performance is then tested with new (previously unseen) data subsets containing random lightpaths throughout the whole length range. For the *hybrid approach*, instances from these subsets are split into cases belonging to the uncertainty area, which are classified by the ML-based modules and cases outside this area, which are classified by means of a simple decision maker. As output variables, the accuracy in classification of lightpaths into high and low QoT categories carried out by the joint action of the decision maker and the ML-based modules and the prediction time per lightpath are obtained. For the *plain approach*, the ML-based modules classify all instances and the same metrics are obtained.

3. SIMULATION SCENARIOS AND RESULTS

The same dataset as in [10] is analysed here, considering the same assumptions regarding the network under study (the Deutsche Telekom -DT- network) and assuming 32 and 64 wavelengths per link (although only results for the latter are shown here). Like in [10], the length range of the uncertainty area is [975 km, 2050 km], and simulations are run in the same computer using the same number of training/testing cases. However, the experiments are transferred to a new programming environment, R, using its *caret* package, in order to take advantage of its capabilities for ML. During the CV phase, the models have been evaluated and selected according to accuracy and AUC metrics, although differences were not significant and we have opted for showing the results corresponding to accuracy optimisation. The results shown in the following figures represent average results together with 95% confidence intervals after repeating the experiments 20 times.

3.1 Hybrid approach

Simulations have been carried out in this scenario by training the ML-based estimators with training sets containing only random cases belonging to the uncertainty area, within the range of 500 to 5000 cases. After the CV process, the performance of the system was analysed by classifying 6000 new (not used in the training phase) lightpaths randomly chosen regardless their total length.

Fig. 1 (left) shows the evolution of the percentage of successful classification of lightpaths into high and low quality categories when the training cases are increased, while Fig. 1 (right), presents the required time to classify a single lightpath. As expected, increments of the number of training cases involve, in turn, improvements of the performance in terms of accuracy in classification for every estimator but for CART, the simplest model, which reaches a saturation point in around 3000 training cases (~99.65% of accuracy). This simplicity allows CART to be one of the fastest models (Fig. 1, right) when the time comes to classify lightpaths, turning it into an excellent alternative when the accuracy requirements are not as restrictive as the computing time. Accuracy results for the remainder models are very similar to those of SVM, especially in the case of TREEBAG and RF, capable of reaching around 99.9% of successful classification of lightpaths. And what is even more important, all of them are significantly faster than SVM. Let us highlight RF, which is approximately 2.75 times faster than SVM, while reaching same levels of accuracy.

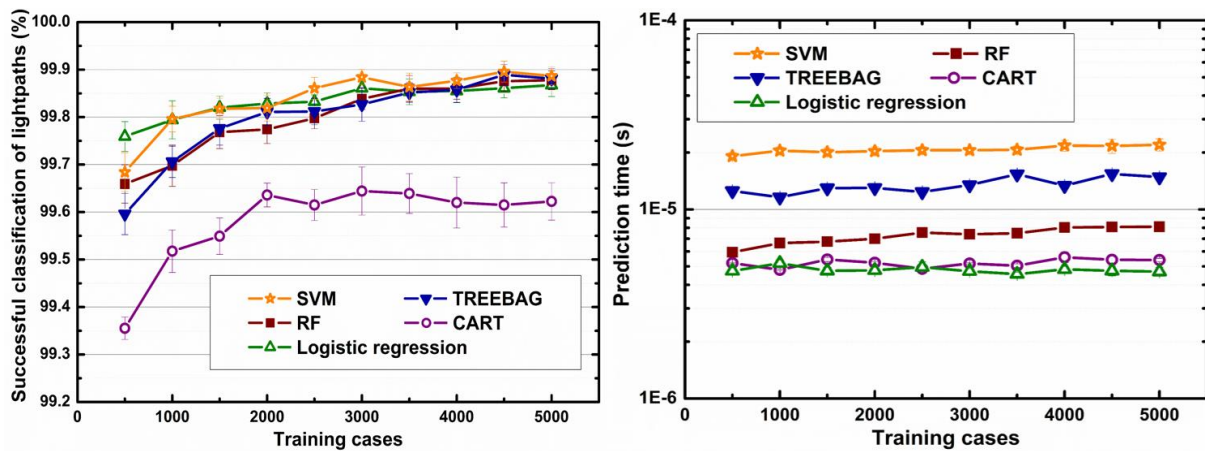


Figure 1. Successful classification of lightpaths into high and low *QoT* categories (left) and computing time required to classify a given lightpath (right), for DT network (64 lambdas) for the hybrid approach.

3.2 Plain approach

Simulations have been carried out in this scenario, where the simple decision maker (based on lightpath length) has been removed. Now, the training datasets also include lightpaths of lengths not belonging to the uncertainty area. For the sake of a fair comparison with the previous scenario, the training datasets contain the same cases belonging to the uncertainty area as before and new cases outside this area are added in such a number that the proportion of lightpaths inside and outside the uncertainty area in the whole database is kept (i.e., the number of lightpaths in the training dataset shared with the previous scenarios will represent only approximately 8% of the total training dataset, while the remaining 92% will be randomly selected among the lightpaths outside the uncertainty area). Thus, the training dataset of 500 cases in the previous scenario will turn into 6500 cases in this new one. Again, after finishing the CV process, the performance of the system is analysed by classifying 6000 new lightpaths (not used in the training and validation phases) randomly chosen regardless their total length.

As in the previous approach, Fig. 2 (left) shows the evolution of the percentage of successful classification of lightpaths into high and low quality categories when the training cases are increased, while Fig. 2 (right) represents the required time to classify a single lightpath. Firstly, it is worth mentioning that the dramatic increment of the number of training cases involved, in turn, a significant growth of the required time to train the models. In the case of SVM, which is a particularly complex model and whose CV process implies a very dense grid of tuning parameters (i.e., a high number of iterations), the necessary time to train it led us to discard it for this approach. However, as shown in Fig. 2 (left), RF and TREEBAG are able to reach high percentages of successful classification of lightpaths, especially RF, reaching up to a 99.92% level of accuracy. Logistic regression and CART, although faster than RF and TREEBAG, reach saturation points at 99.75% and 99.65% levels of accuracy, respectively. Analysing Fig. 2 (right), an increase in the prediction time is observed with respect to the corresponding graph in Fig. 1, fact that shows the influence of the simple decision maker on this parameter. However, when comparing the figures corresponding to the accuracy, what is revealed is that ML models are perfectly capable of learning the patterns hidden in the training datasets for this particular problem. Finally, although we do not have enough space to present results in relation to AUC, sensitivity and specificity, which are especially relevant metrics that take into account the impact of false negatives/positives within the

performance of ML models, we do not want to miss the opportunity to briefly comment a few results. Let us consider as an example the *plain approach*. CART exhibited the lowest performance in relation to these metrics, moving in the ranges of (depending on the number of training cases) [0.914, 0.978], [0.64, 0.79] and [0.9975, 0.9987] for AUC, specificity and sensitivity, respectively, while RF undoubtedly offered the best specificity-sensitivity trade-off, moving in the ranges of [0.996, 0.9999], [0.72, 0.94] and [0.9981, 0.9997] for AUC, specificity and sensitivity, respectively. Hence, RF provides excellent results, very close to the ideal value, 1.

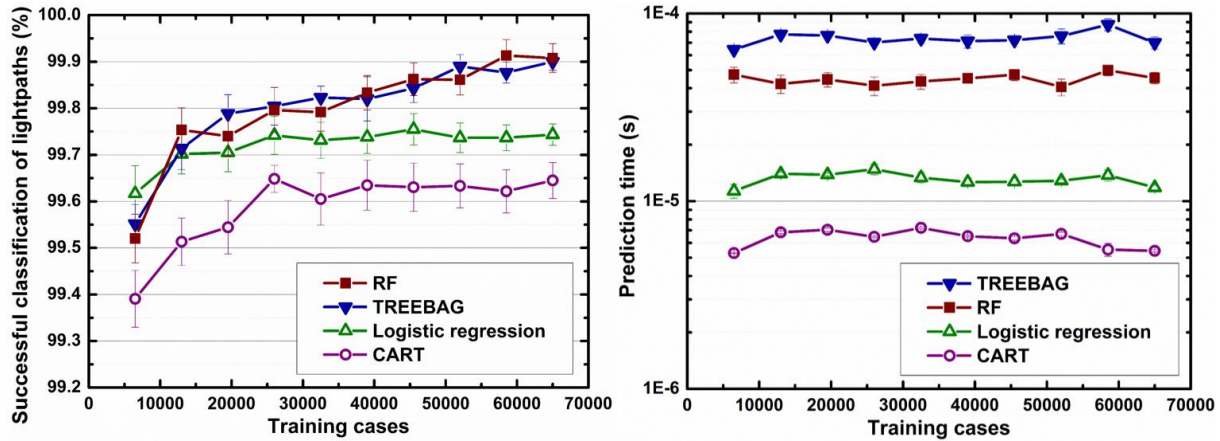


Figure 2. Successful classification of lightpaths into high and low QoT categories (left) and computing time required to classify a given lightpath (right), for DT network (64 lambdas) for the plain approach.

4. CONCLUSIONS

We have extended our previous work in QoT estimation [7], [10], by introducing new ML models (RF, TREEBAG, logistic regression and CART). All of them have proven to be very efficient in solving this problem, especially RF and TREEBAG, which are able to further reduce the required computing time to classify a single lightpath with respect to a SVM approach [10], without undermining the accuracy. Specifically, RF proved to be 2.75 times faster than SVM while maintaining the accuracy around 99.9%. Moreover, we have also evaluated the performance of these new ML-based QoT estimators in a new scenario where the decision maker of the *hybrid approach* was removed. RF and TREEBAG were perfectly capable of learning the patterns hidden in the training data along the entire range of study, reaching accuracy levels up to above 99.9%.

ACKNOWLEDGEMENTS

This work has been partially funded by the Spanish Ministry of Science and Innovation TEC2014-53071-C3-2-P, TEC2017-84423-C3-1-P and TEC2015-71932-REDT.

REFERENCES

- [1] S. Azodolmolky *et al.*: Experimental demonstration of an impairment aware network planning and operation tool for transparent/translucent optical networks, *Journal of Lightwave Technology*, vol. 29, no. 4, pp. 439–448, 2011.
- [2] Y. Pointurier: Design of low-margin optical networks, *Journal of Optical Communications and Networking*, vol. 9, no. 1, pp. A9-A17, Jan. 2017.
- [3] I. Sartzetakis, K. Christodouloupoulos and E. Varvarigos: Improving QoT estimation accuracy through active monitoring, in *Proc. ICTON*, Girona, 2017, pp. 1-4.
- [4] P. Soumplis *et al.*: Multi-period planning with actual physical and traffic conditions, *Journal of Optical Communications and Networking*, vol. 10, no. 1, pp. A144-A153, Jan. 2018.
- [5] E. Seve *et al.*: Learning process for reducing uncertainties on network parameters and design margins, *Journal of Optical Communications and Networking*, vol. 10, no. 2, pp. A298-A306, Feb. 2018.
- [6] J. Mata *et al.*: Artificial intelligence (AI) methods in optical networks: A comprehensive survey, *Optical Switching and Networking*, vol. 28, pp. 43-57, 2018. <https://doi.org/10.1016/j.osn.2017.12.006>
- [7] T. Jiménez *et al.*: A cognitive quality of transmission estimator for core optical networks, *Journal of Lightwave Technology*, vol. 31, no. 6, pp. 942–951, 2013.
- [8] F. Musumeci *et al.*: A survey on application of machine learning techniques in optical networks, *ArXiv e-prints*, [arXiv:1803.07976v2](https://arxiv.org/abs/1803.07976v2) [cs.NI].
- [9] C. Rottondi *et al.*: Machine-learning method for quality of transmission prediction of unestablished lightpaths, *Journal of Optical Communications and Networking*, vol. 10, no. 2, pp. A286-A297, Feb. 2018.
- [10] J. Mata *et al.*: A SVM approach for lightpath QoT estimation in optical transport networks, in *Proc. IEEE Big Data*, Boston, MA, 2017, pp. 4795-4797.