Internship Subject (2025)

Title: **Feature Selection using Tree-based Models**

Chung Shue (Calvin) Chen, Research Scientist (DMTS)
Nokia Bell Labs
chung_shue.chen@nokia-bell-labs.com

Context

Consider a data set with many features of tens or hundreds, these can be some system or network parameters. Sometimes, many of the features are highly correlated. In a communication system, the signal attenuations in the radio link and the free space optical channel rely on the environment including the weather [1-2]. In a weather data set, one may have the average temperature, the humidity and wind speed, also the median, the minimum and the maximum of such values, measured for a time period. These are the input to a telecom system. The question or research problem here is how to build a model for the channel attenuation using the weather parameters as input. Besides, to avoid overfitting, we do not want to use all the features which are often or known correlated. Instead, we want to know how to select an important subset.

Description

Consider a data set that contains many features, including both continuous and categorical variables. To avoid overfitting, it is preferable to select a small subset of important features that can be used to construct an accurate model. For this purpose, machine learning methods [3] such as lasso regression, gradient boosting regression, support vector regression and neural networks have disadvantages for our data. Lasso regression assumes a linear relationship between the features and the target variable, which may not be true for some datasets. Gradient boosting regression is more prone to overfitting than random forest because it uses all the data and features for each tree. It also requires more hyperparameter tuning and computational resources than random forest. Support Vector Machines (SVMs) and Neural Networks (NNs) do not provide simple feature selection or measure feature importance. SVMs require feature scaling and kernel parameter tuning. NNs are computationally expensive to train and difficult to interpret and explain.

This project explores how Random Forest can be used to select the important features from a set of correlated features. Random forest can handle both continuous and categorical features without any pre-processing that might affect the data. It can also model both linear and non-linear relationships. Random forest provides a measure of feature importance that can help us identify the relevant features and eliminate the redundant or noisy ones. In [1], the importance of the features is ranked using the out-of-bag (OOB) information together with the wrapper methodology in feature selection [4]. Random forest is relatively easy to train and computationally inexpensive because it can be parallelized by growing each decision tree independently. This is an important property in producing the feature selection results.

There are some recent results on feature importance using tree-based models. Reference [5] explored methods for correcting biases in feature importance measures, particularly for tree-based models such as random forests, where features with more potential splits may be favoured. Ensuring that feature importance scores are fair and do not inadvertently introduce or perpetuate bias was explored in [6]. This involves developing methods to interpret tree-based models in a way that is both fair and understandable.

This project aims to further develop the method of feature selection using tree-based models. Further research is needed to fully understand the theory behind this approach and its limitations. In particular, a valuable direction for future research would be to automate and systematise the process of determining the threshold for selecting a feature.

In the following, we consider a wireless optical communication system (OWC) but the application is not limited to. We wish to develop channel models for OWC systems. Accurate channel modelling is crucial for achieving high speed communication and accurate indoor positioning [7]. The model will be derived from a dataset containing multiple correlated time series of sensor measurements. One challenge is to avoid overfitting when using these correlated measurements. Besides, we have to address the difficulty of inferring latent variables in the system, such as interference from ambient light sources in the environment. The project will focus on developing a causal discovery technique that can construct a reliable model from the correlated measurement data. We would also use machine learning algorithms to explore the correlation coefficient and mutual information in the empirical data, which measures both linear and non-linear correlation. Besides, causal discovery method can help us understand how to select the features correctly. Results can be applied to various system and network data science problems.

There are techniques in the literature. Linear regression (LRM) is commonly used to construct predictive models that describe the relationship between a response variable and one or more predictor variables [8]. LRM can be extended to linear regression model with interaction (LRI) for high accuracy. Non-linear regression models have higher complexity but often can outperform traditional linear regression models. For example, one can use binary decision tree (BDT) through supervised machine learning [9]. Considering overfitting that can occur when using individual BDTs, we would use bootstrap-aggregated decision trees, which mitigate overfitting effects by combining the results of an ensemble of BDTs such that each BDT within the ensemble is grown on an independently drawn bootstrap replica of the data, using random forest. A method for ranking the importance of predictor variables in the construction of random forest models can be also derived. **We aim to automate and systematise the process of predictor variable selection and to study methods of constructing better models using empirical data.** We would implement and compare various methods, to analyze and optimize their performance for a system or network. Suitable results can be submitted to a distinguished conference or journal or can be considered as a preliminary work for a potential PhD programme.

Keyword: Machine learning methods, Random Forest, important features, causal discovery

References

[1] S.-W. Ho, L. Mitchell, and V. Wang, “Modelling channel attenuation in hybrid optical/E-band system,” 2024. Online available: http://dx.doi.org/10.36227/techrxiv.170492384.40277580/v1

[2] A. Khatoon, W. G. Cowley, and N. Letzepis, “FSO/RF correlation measurement and hybrid system hidden Markov model,” in Australian Communications Theory Workshop, 2013, pp. 93–98.

[3] I. H. Sarker, “Machine learning: Algorithms, real-world applications and research directions,” SN Computer Science, vol. 2, no. 3, 2021.

[4] I. Guyon and A. Elisseeff, “An introduction to variable and feature selection,” Journal of Machine Learning Research, vol. 3, no. pp. 1157–1182, 2003.

[5] Z. Zhou and G. Hooker, “Unbiased measurement of feature importance in tree-based methods,” ACM Transactions on Knowledge Discovery from Data, 2020. https://doi.org/10.1145/3429445

[6] C. O. Little, D. H. Lina, and G. I. Allen. “Fair feature importance scores for interpreting tree-based methods and surrogates." 2023, arXiv preprint arXiv:2310.04352.

[7] Nokia Digital Automation and Internet-of-Things (IoT), “High Accuracy Indoor Positioning,” 2024. Online available https://www.dac.nokia.com/applications/high-accuracy-positioning/

[8] A. Afifi, S. May, and V. A. Clark, Practical Multivariate Analysis. Chapman and Hall/CRC, 2011.

[9] H. Trevor, T. Robert, and F. Jerome, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Spinger, 2009.