ERCIM fellowship Programme

Final scientific report

ERCIM Alain Bensoussan Fellowship Programme

ERCIM
European Research Consortium
for Informatics and Mathematics

Fellow	Sezer Kutluk
Host Organisation	Fraunhofer Institute for Mechatronic Systems Design
Scientific coordinator	Steven Koppert

I – SCIENTIFIC ACTIVITY DURING YOUR FELLOWSHIP

This ERCIM fellowship took place in the Trusted Machine Intelligence Group at Fraunhofer Institute for Mechatronic Systems Design (IEM) in Paderborn, Germany, from 1 April 2022 to 31 March 2023.

The main research theme of this fellowship was explainable AI methods with a focus on time series models, particularly for classification and anomaly detection tasks.

Recent developments in machine learning research resulted in complex models which can be trained with large datasets. These models have achieved great predictive performance, even surpassing human-level accuracy. However, complex models are harder to interpret, and accuracy as a performance metric is not enough to evaluate the models in terms of trustworthiness. Explainable AI methods aim at generating explanations to understand the reasoning behind the model's decision for a specific input.

In this fellowship, the objective was to generate instance-based post-hoc explanations for black box predictive models, specifically classification and anomaly detection models for time series data. The aim here is to generate explanations only by evaluating the model response to certain inputs, without knowing or modifying the inner structure and dynamics of the model.

The explanations generated in this study are called *counterfactual* explanations. Here we define a counterfactual as a modified time series for which the classifier changes its decision and predicts a class label that is equal to the desired class label, while the modifications are as minimal as possible. Therefore, for a wrongly classified instance we can generate counterfactuals that are classified correctly. When generating counterfactuals, we consider the following:

- the classifier must decide that the counterfactual is in the desired class,
- the distance between the original time series instance and the generated counterfactual must be minimal,
- the time steps where the modifications occur must be as sparse as possible for better human interpretability,
- the generated counterfactuals should be within the same distribution as the original data.

The proposed method generates counterfactuals by making modifications to the explained instance by using instances from the training set and making predictions with the unmodified classifier. This is an iterative optimization procedure that finds the point that the classifier changes its decision.

As the secondary theme, an initial study of uncertainty quantification was conducted. Uncertainty quantification is another important research topic for the trustworthiness of machine learning models. Another idea of generating explanations is by using better calibrated classification scores and quantified uncertainties. The follow-up work is planned to include these connections between explainable AI, uncertainty quantification, and model calibration methods.

A manuscript explaining the developed method of counterfactual explanations is in preparation.

II – PUBLICATION(S) DURING YOUR FELLOWSHIP

Kutluk, S., Koppert, S., Henke, C., & Trächtler, A. (2023). *Counterfactual explanations for time series classifiers*. [Manuscript in preparation]. Fraunhofer Institute for Mechatronic Systems Design.

III – ATTENDED SEMINARS, WORKHOPS, CONFERENCES

- RISE Learning Machines Seminars: "Communication Complexity of Federated Learning" by Giulia Fanti, 19 January 2023.
- RISE Learning Machines Seminars: "Interpretability in NLP" by Lena Voita, 26 January 2023.

IV – RESEARCH EXCHANGE PROGRAMME (REP)

I visited the Deep Learning Research Group, which is led by Dr. Olof Mogren, at Research Institutes of Sweden (RISE) in Gothenburg between 16-20 January 2023. This visit was mostly focused on knowledge exchange and discussions over research ideas and possible collaborations. I also attended the Learning Machines Seminar with the Deep Learning Research Group.