Review of Master thesis

Using Shapley Value for Interpretive Artificial Intelligence based on Feature Interactions Graph

Xu Feiran

The primary aim of this paper is to integrate Shapley values with artificial intelligence models, in order to establish interpretive methods and optimize them for models with high-dimensional inputs, thus improving computational efficiency.

This paper examines the issue of interpretability in machine learning models, highlighting that as the number of features increases, the computation time required to calculate Shapley values also increases. Existing sampling methods continue to face the challenge of lengthy computation times. To tackle this problem, the authors propose a novel approach that utilizes a "high-impact" coalition of participants to calculate Shapley values, thereby reducing computation time and proposing a more meaningful way of measuring the validity of interpreting the results.

In the course of this work, the following results were achieved:

- The causes of the lengthy calculation time of the sampling method were precisely analysed, and solutions were proposed.
- Detailed explanations of the principles of artificial neural networks and their application to solving anomaly detection problems are provided, and experimental results demonstrate the method's effectiveness.
- New convergence metrics are proposed to corroborate the convergence of sampling and graph-based sampling algorithms from two perspectives.
- The graph-based sampling method reduces the computation time by 40% compared to the sampling method and does not differ significantly from the results of the sampling method in terms of final settlement results.

During his studies, XuFeiran has achieved success in the field of interpretable artificial intelligence, having contributed to a number of papers published in reputable journals and conferences.

Article for Conference:

- Zhang, Y., Xu, F., Zou, J., Petrosian, O. L., & Krinkin, K. V. (2021, June). XAI evaluation: evaluating black-box model explanations for prediction. In *2021 II International Conference on Neural Networks and Neurotechnologies (NeuroNT)* (pp. 13-16). IEEE.
- Zou, J., Xu, F., & Petrosian, O. (2020). Explainable AI: using Shapley value to explain the anomaly detection system based on machine learning approaches. *Процессы управления и устойчивость*, *7*(1), 355-360.

Article for Journal:

- Zou, J., Xu, F., Zhang, Y., Petrosian, O., & Krinkin, K. (2021). High-dimensional explainable AI for cancer detection. *International Journal of Artificial Intelligence*, *19*(2), 195.

Overall, the work successfully achieved its objective and provided valuable insights into improving interpretability in machine learning models. The proposed approach shows promise in reducing computation time and producing more meaningful interpretations. The research deserves an "Excellent" rating.

Science Supervisor:
Professor of Department of Mathematical
Modelling of Energetic Systems
Dr. Sc. Ovanes Petrosian