inAIco

Interpretierbarkeit von KI-getriebenen Korrekturen

Programm / Ausschreibung	Bridge, Bridge_NATS, Bridge_NATS 2019	**Status**	abgeschlossen
Projektstart	01.10.2020	**Projektende**	30.09.2024
Zeitraum	2020 - 2024	**Projektlaufzeit**	48 Monate
Keywords	Explainable AI (XAI); Interpretable AI Error-Modeling; Base-Model Induced Interpretations; Local and Global Intrinsic Surrogate Models; Startup Success Prediction		

Projektbeschreibung

Ausgangssituation/Problem: Das Thema Explainable Artificial Intelligence (XAI, Erklärbare KI) stellt die zentrale, für die Praxis äußerst wichtige Frage, ob moderne KI-Lernalgorithmen nicht nur die Genauigkeit von Prognosen verbessern, sondern auch Interpretationen dieser Korrekturen zu deren allgemeinem Verständnis erzeugen können.

Ziele/Innovation: In inAIco wird die Erklärbarkeit im Sinne von Interpretierbarkeit von jenen KI-Algorithmen erforscht, die als Korrektur zu bestehenden Lösungen von Problemen des überwachten Lernens wirken. Die Interpretation wird dabei in unserem Ansatz BAPC ("Before and After correction Parameter Comparison") mit Hilfe der Lösung eines bestehenden Grundmodells formuliert. Mit diesem Ansatz wird XAI mathematisch definiert und die Voraussetzungen für eine systematische, sog. modellagnostische Untersuchung der Interpretierbarkeit verschiedenster KI-Modelle geschaffen. Bezeichnenderweise werden Interpretationen als "effektive" Änderungen der Parameter (Gewichte der Features) des Grundmodells aufgefasst. Die mathematische Innovation besteht in der definierenden Eigenschaft unseres Interpretierbarkeitsbegriffs von KI-Modellen, sich auf die Existenz eines Grundmodells zu stützen. Dies macht BAPC zu einem intrinsischen Surrogat (das Grundmodell wird zum sog. Explainer Model), das man sowohl global (d.h. in der gleichen Weise für den gesamten Datensatz), als auch lokal (für eine geeignete Umgebung eines gegebenen Datenpunktes) anwenden kann. Charakteristisch und innovativ gegenüber bekannten lokalen Surrogaten (z.B. LIME) ist die Auffassung des zu interpretierenden KI-Systems als additive Korrektur des Grundmodells. Der Grundproblematik von lokalen Surrogaten, nämlich die Wahl der hinreichend kleinen Umgebungen des jeweils zu interpretierenden Datenpunktes ("local fidelity"), wird durch die Forderung nach der Beschränktheit der Korrektur garantiert. Diese übersetzt sich in eine Schranke an die maximalen Parameteränderungen, welche wiederum die maximale Ausdehnung der Umgebung für hinreichende Modellangepasstheit begrenzen. inAIco macht damit für das aktuell beforschte XAI Thema des lokalen Surrogats einen allgemeinen, rigoros fundierten und experimentell überprüfbaren Lösungsvorschlag.

Erwartete Ergebnisse: Die Verifikation von BAPC wird anhand eines durch eine KI-Korrektur unterstütztes Grundmodell zur Einschätzung von Erfolgschancen verschiedener Startup Unternehmen durchgeführt. Dabei wird der Datenvorrat des Firmenpartners Speedinvest Heroes Consulting GmbH über die marktwirtschaftliche Kompetenz verschiedener Firmen

einerseits und Persönlichkeitsprofile der Firmengründer und Mitarbeiter andererseits verwendet. Zunächst ergibt sich dadurch als Mehrwert die verbesserte Prognosegenauigkeit der Erfolgswahrscheinlichkeiten für jede einzeln betrachtete Startup Firma durch das um die KI-Korrektur verbesserte Grundmodell, welche Erkenntnisse über den Einfluss von psychologischen Eigenschaften der GründerInnen und MitarbeiterInnen liefert. Noch wertvoller und von zentraler Bedeutung ist jedoch die Interpretation der Korrekturen, also die Erkenntnisse darüber, warum die Grundmodellvoraussagen noch verbessert werden können.

Konsortium: Im Team von inAIco sind sowohl die theoretischen, mathematischen Kompetenzen als auch fundierte Expertise über Maschine Learning und Deep Learning durch das Software Competence Center Hagenberg vorhanden. Dies wird durch die Expertise des Verwertungspartners Speedinvest Heroes zur Bewertung von Startups ergänzt.

Abstract

Initial Situation/Research Problem: The field of Explainable AI (XAI) poses the central and for many applications relevant question if, in addition to improving predictions, modern learning algorithms can provide interpretations of these improvements for the benefit of them being understood in a more general way.

Goals/levels of innovation beyond state-of-the-art: The project inAIco explores explainability in the sense of interpretability of AI-algorithms which act as corrections to existing solutions of supervised-learning-problems. Specifically, the definition of interpretability is formulated with the help of our approach BAPC ("Before and After correction Parameter Comparison") using the solution of a presupposed existing base-model. With this approach, XAI is defined mathematically and the prerequisites for a systematic, so-called model-agnostic interpretation of interpretability of various AI-models are given. Characteristically, interpretations are conceived as "effective" changes of parameters (weights of features).
The mathematical innovation consists of the defining property of our idea of interpretability of AI-models, which is to presuppose the existence of an interpretable base-model. This makes BAPC a so-called intrinsic surrogate (the base-model becomes the explainer model). It can be applied in a global way (i.e. in the same way for the entire data set), or as a local surrogate (for a specific neighbourhood of a single instance). The characteristic and innovative aspect of our approach compared to known local surrogates (e.g. LIME) consists in regarding the AI-system as an additive correction of the base-model. The fundamental problem of local surrogates is the choice of a sufficiently small local neighbourhood of single instances for the model to be sufficiently exact (local fidelity). BAPC solves this problem by requiring a bound on the correction size, which translates into a bound on the maximal parameter changes. This, in turn, limits the extension of the local neighbourhood on which fidelity is guaranteed. In this way, inAIco suggests a general, rigorous, and experimentally verifiable solution of the currently actively investigated concept of the local surrogate.

Expected results: The verification of BAPC is conducted using a suitable model pair consisting of a base-model and an AI-correction, to estimate the likeliness of success of several start-up companies. In doing so, the data supply of the company partner Speedinvest Heroes Consulting GmbH is used, which characterises the economic success of several start-up companies as well as the personal profiles of founders and employees. On one hand, the benefit emerges in terms of improved predictions of success probabilities, delivering insights about the influence of psychometric data on economic well-being of start-ups. On the other hand, and more importantly, the interpretation of the corrections is of particular interest, i.e., what influenced the corrections of the base-model.

Consortium: The inAIco team includes both theoretical and mathematical competencies as well as in-depth expertise in machine learning and deep learning from the Software Competence Center Hagenberg. This is complemented by the expertise of the exploitation partner Speedinvest Heroes for the assessment of start-ups.

Projektkoordinator

- Software Competence Center Hagenberg GmbH

Projektpartner

- Speedinvest Heroes Consulting GmbH