Fraunhofer IPA

Fraunhofer-Institut für Produktionstechnik und Automatisierung IPA

Abbildung 1:
Einsatz einer lokalen Erklärmethode zur Bestimmung der Bildbereiche, die für die Erkennung der Hand maßgeblich sind.

xAI – Erklärbarkeit maschineller Lernverfahren

Marktanforderungen

Verfahren des Maschinellen Lernens kommen immer häufiger in verschiedenen Anwendungsbereichen der Produktion, Medizin oder dem Dienstleistungswesen zum Einsatz. Beispielsweise kann über Bildverarbeitung oder Sensordatenanalyse Ausschuss erkannt und frühzeitig aus dem Prozess genommen werden. Auch für die Programmierung von Robotern werden vermehrt tiefe Neuronale Netze erprobt. Es gibt jedoch viele Szenarien, in denen nicht hochgenaue Vorhersagen alleine von größter Bedeutung sind und stattdessen Vertrauen, Akzeptanz oder Konformität mit Regularien entscheidend sind. Kritische Entscheidungen müssen hier von Erklärungen begleitet werden, sodass Nutzer die Ergebnisse oder das generelle Verhalten des Algorithmus nachvollziehen können. Durch die Schaffung von Erklärbarkeit kann einerseits die korrekte Funktionsweise der Modelle überprüft, aber auch mögliche Diskrepanzen zwischen menschlichen Entscheidungen und algorithmisch getroffenen Entscheidungen untersucht werden. So lässt sich beispielsweise die Frage, wieso eine bestimmte Stellgröße für einen Regler ausgegeben wurde, beantworten. Auch für sicherheitskritische Bereiche, wie das autonome Fahren oder Mensch-Roboter-Kollaborationen, können Erklärungen einen erheblichen Mehrwert bieten. Mithilfe von Erklärungen für die Entscheidungen eines Modells können Experten ihr Verständnis für das Modell verbessern und Risiken bewerten.

Unsere Lösung

Grundsätzlich gibt es zwei Ansätze, Erklärbarkeit Maschineller Lernverfahren herzustellen. Bei der Modellerklärbarkeit (global) steht das Verständnis des Modells als Ganzes im Fokus: Das Ziel ist, eine möglichst gute Nachvollziehbarkeit der inneren Entscheidungswege eines Black-Box-Modells zu liefern. Im Gegensatz dazu werden mittels lokaler Erklärmethoden einzelne Entscheidungen erklärt (Abbildung 1). Das Fraunhofer IPA bietet für Ihre bestehenden Machine Learning-Modelle (ML-Modelle) die Herstellung von Erklärbarkeit durch Erklärungsgenerierung – sowohl in lokaler, als auch globaler Form. Auch im Zuge von Neuentwicklungen bietet das Fraunhofer IPA die Möglichkeit, Erklärbarkeit von Beginn an als elementaren Bestandteil des Prozesses und Endproduktes zu berücksichtigen.

Der Stand der Technik im Forschungsfeld »Explainable AI« umfasst eine Vielzahl verschiedener Methoden. Da nicht jede Methode für jeden Anwendungsfall gleich gut geeignet ist, ist die Wahl der optimalen Methode zeit- und rechercheintensiv. Vor diesem Hintergrund entwickelt das Fraunhofer IPA zudem eine Software-Toolbox, die existierende Erklärmethoden in Einklang bringt. In diesem Zusammenhang werden auch eigene, am Fraunhofer IPA entwickelte, Verfahren integriert. Unter Einsatz dieser universalen Toolbox kann ein umfassendes Modellverständnis mit Hilfe einer raschen Erklärungserzeugung sowie einer gegenüberstellenden Betrachtung verschiedener Techniken ermöglicht werden.

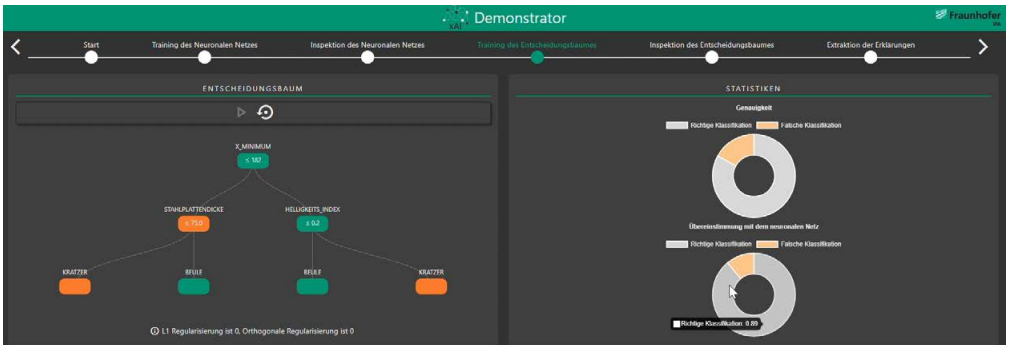

Abbildung 2:
Approximation eines neuronalen Netzes mit Hilfe eines Entscheidungsbaums.

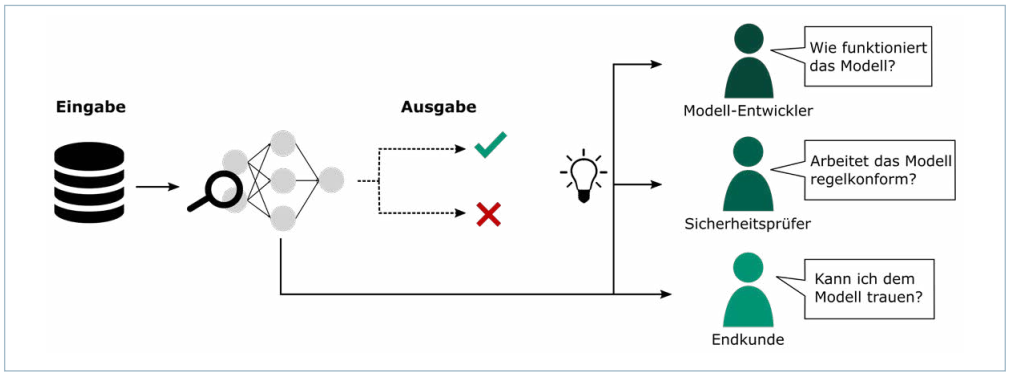

Abbildung 3:
Erklärbarkeit maschineller Lernverfahren ist für verschiedene Interessensgruppen von Bedeutung.

Neben der Erklärung des ML-Modells spielen auch die dem Modell zugrundeliegenden Daten eine zentrale Rolle. Die Software-Toolbox stellt Methoden bereit, welche den Einfluss der Trainingsdaten auf das gelernte Modell quantifizieren. So kann bestimmt werden, welche Daten besonders wertvoll sind und in welchen Bereichen keine ausreichende Datengrundlage vorhanden ist.

Ihr Nutzen

Für Sie als Anwender ergeben sich folgende Vorteile:

Modellvalidierung
Überprüfung der erwartungskonformen Funktionsweise Ihrer eingesetzten ML Modelle. Besonders relevant ist dieser Aspekt für sicherheitskritische Anwendungen, die beispielsweise von einer internen oder externen Prüfstelle abgenommen werden müssen.

Modell-Debugging
Werden Fehlentscheidungen des Modells festgestellt, kann für diese detailliert untersucht werden, auf Basis welcher Merkmale die Entscheidung getroffen wurde (Abbildung 2). Durch das Debugging können Ursachen für Fehlfunktionen des Modells identifiziert und anschließend behoben werden.

Datenwert
Unterstützung einer datenzentrischen Sicht, die Ihnen erlaubt die Qualität und den Nutzen Ihrer Daten zu bewerten. Sie erhalten Einblicke darüber, welche Daten das ML-Modell besonders beeinflussen. Zudem bekommen Sie aufgezeigt, welche Daten überflüssig und welche Daten für eine hohe Qualität des Modells noch zu erheben sind.

Akzeptanz und Vertrauen
Ein fehlendes Verständnis für die Entscheidungsfindung hochkomplexer ML-Algorithmen kann oftmals eine hohe Einstiegshürde für einen Einsatz selbiger Algorithmen sein. Erklärungen für algorithmisch getroffene Entscheidungen können das Vertrauen der Anwender in die Systeme stärken und zu einer höheren Akzeptanz führen.

Erkenntnisgewinn
Die Untersuchung des gelernten ML-Modells ermöglicht das Aufdecken allgemeiner Zusammenhänge, beispielsweise welche Eingabedaten besondere Bedeutung haben und wie einzelne Eingabedaten miteinander interagieren.

Kontakt

Prof. Dr.-Ing. Marco Huber
Telefon +49 711 970-1960
marco.huber@ipa.fraunhofer.de

M.Sc. Danilo Brajovic
Telefon +49 711 970-3647
danilo.brajovic@ipa.fraunhofer.de

Fraunhofer-Institut für Produktionstechnik und Automatisierung IPA
Nobelstr. 12, 70569 Stuttgart
www.fraunhofer.de