Automated Concept-based Explanation als Erklärverfahren für die Fehlererkennung auf Zeitreihendaten

(Bachelorarbeit)

FINNEGAN KRÜMPELMANN

Motivation

In der Intensivmedizin können Machine Learning Algorithmen unverzichtbare Vorteile bei der Analyse von großen Datenmengen liefern. Da diese Algorithmen potenziell lebenswichtige Entscheidungen treffen können, wurden verschiedenste Erklärverfahren entwickelt, um ihre Vorgehensweise möglichst nachvollziehbar zu begründen. Ein solches Erklärverfahren stellt „Automated Concept-based Explanation“ (ACE) dar, welches sich bei der Erklärung des Outputs nicht nur auf einzelne Input-Features, sondern, ähnlich wie beim menschlichen Entscheidungsprozess, auf höhere Konzepte im Input beruft. Oft fassen solche Konzepte somit mehrere Input-Features in einer interpretierbaren Einheit zusammen. Dieses Vorgehen wurde bisher vorwiegend auf Bilddaten angewendet und konnte dort überzeugende Ergebnisse erzielen.

Stand der Technik

„Automated Concept-based Explanation“ kann in zwei wesentliche Schritte unterteilt werden. Zunächst werden etwaige in den Eingabedaten vorliegende Konzepte automatisch extrahiert. Anschließend werden diese Konzepte basierend auf Ihrem Einfluss auf den Output gewichtet. Auf Bilddaten wurde für den ersten Schritt bisher eine Kombination aus Segmentierung und Ähnlichkeits-Klassifikation genutzt. Für den zweiten Schritt wurden „Concept Activation Vectors“ (CAV) eingesetzt, welche die jeweiligen Neuronenaktivierungen des Netzwerkes für ein spezifisches Konzept repräsentieren. Auf Zeitreihendaten ist ACE bisher unerforscht, weshalb sich ein „state-of-the-art“ schwer festlegen lässt, jedoch bieten „Shapelets“ eine attraktive Methode für den ersten Schritt. Für den zweiten Schritt lässt sich eine auf Zeitreihendaten adaptierte Version von CAVs anwenden.

Zielsetzung

Ziel dieser Arbeit ist es, das Erklärverfahren „Automated Concept-based Explanation“ (ACE) für Zeitreihendaten zu adaptieren und somit in das Lehrstuhleigene Novelty Detection Analysis System (NDAS) zu integrieren, um getroffene Entscheidungen für Nutzer möglichst nachvollziehbar zu begründen. Das neue Erklärverfahren soll im Anschluss mit Hilfe einer Umfrage mit den bestehenden Erklärverfahren auf dessen Verständlichkeit verglichen werden.

Geplante Vorgehensweise

Zunächst muss das Erklärverfahren ACE für Zeitreihendaten adaptiert werden. Dafür müssen sowohl die automatische Konzepterkennung, als auch die anschließende Gewichtung der Konzepte auf Zeitreihendaten angepasst werden. Hierfür optimale Verfahren werden mit Hilfe einer Literaturrecherche identifiziert und daraufhin implementiert. Im Anschluss wird eine Umfrage bezüglich der Effektivität des Erklärverfahrens ACE im Vergleich zu den bestehenden Erklärverfahren durchgeführt.

Betreuer: Alexander Kruschewsky
14.02.2024

Informatik 11
Embedded Software
RWTH AACHEN UNIVERSITY