Datenaugmentationsverfahren für das Training von Algorithmen für die Erkennung von fehlerhaften Datenpunkten

(Bachelorarbeit)

ADAM HAMAN

Motivation

Um bei medizinischen Daten feststellen zu können, ob eine Anomalie durch einen Messfehler oder durch eine Zustandsveränderung des Patienten entstanden ist, werden Fehlererkennungsalgorithmen angewendet. Diese Fehlererkennungsalgorithmen für Zeitreihendaten werden häufig erst auf Daten trainiert, bevor sie angewendet werden. Die Menge an vorhandenen Daten ist jedoch begrenzt und mangelt an Diversität. Auch sind Zeitreihendaten häufig mit Rauschen belastet. Um Algorithmen, die diese Daten verwenden besser trainieren und Überanpassung reduzieren zu können, sind Datenaugmentationsverfahren hilfreich.

Stand der Technik

Datenaugmentationsverfahren wurden in verschiedenen Bereichen, beispielsweise Computer Vision und Natural Language Processing, schon erfolgreich angewendet und umfassend erforscht. Im Bereich Computer Vision können Daten beispielsweise durch einfache Verfahren wie Spiegeln, und Einfügen eines künstlichen Rauschens, oder auch durch komplexere generative Verfahren wie durch Nutzung eines GAN augmentiert werden. Für Zeitreihendaten können die grundlegenden Verfahren ähnlich wie zu Bilddaten verwendet werden, es existieren aber auch Zeitreihenspezifische Datenaugmentationsverfahren wie zum Beispiel Magnitude Warping. Für Zeitreihendaten existieren ebenfalls komplexere Verfahren, wie generative Modelle durch GAN oder VAE. Eine Anwendung von Datenaugmentationsverfahren auf medizinischen Zeitreihendaten für Fehlererkennungsalgorithmen ist bisher jedoch nicht erfolgt.

Zielsetzung

Ziel dieser Bachelorarbeit ist es zu untersuchen, ob die Performanz von Fehlererkennungsalgorithmen durch Datenaugmentationsverfahren verbessert wird. Hierfür werden Datensätze aus mehreren medizinischen Datenbanken als Ausgangspunkt für die Datenaugmentation genutzt. Durch die Verwendung einfacher Augmentationsverfahren wie das Einfügen von Rauschen oder eine Skalierung der Daten, aber auch durch komplexere Verfahren wie AEs und anderen generativen Modellen, sollen neue Datensätze für das Training der Fehlererkennungsalgorithmen generiert werden.

Geplante Vorgehensweise

Zuerst erfolgt eine Literaturrecherche zu verschiedenen Augmentationsverfahren. Danach sollten die Verfahren ausgewählt werden, die sinnvoll auf die gegebenen Daten anzuwenden sind. Die ausgewählten Verfahren werden soweit möglich aus bestehenden Bibliotheken und öffentlich zugänglichen Code auf die Datensätze angewandt. Anschließend wird analysiert, wie diese Verfahren die Performance der Fehlererkennungsalgorithmen beeinflussen. Dabei wird getestet, wie groß der Anteil von augmentierten Daten sein soll und welche Parameter der Verfahren für eine bestmögliche Performance gewählt werden. Eventuell aufkommende schlechte Performance von bestimmten Verfahren wird ebenfalls untersucht.

Betreuer: Alexander Kruschewsky
14.02.2024

Informatik 11
Embedded Software
RWTH AACHEN UNIVERSITY