

Identifikation und Evaluation von Patientenclustern in intensivmedizinischen Datenbanken

(Bachelorarbeit)



ELLA PETZOLD

Motivation

In der Medizininformatik werden intensivmedizinische Datenbanken verwendet, um u. a. KI Methoden zu trainieren. Hierbei handelt es sich häufig um Sekundärdaten, also Daten, die zuvor für einen anderen Zweck aufgezeichnet wurden und nun weiter verwendet werden. Aus diesem Grund sind viele dieser Datenbanken nicht angemessen für den neuen Zweck dokumentiert. Um sich einen besseren Überblick über die Daten zu verschaffen und Erkenntnisse über enthaltene Patientengruppen zu gewinnen, können Cluster-Algorithmen angewandt werden. Die Identifikation von ähnlichen bzw. unterschiedlichen Patientenklassen kann die Erstellung von KI-Methoden unterstützen.

Stand der Technik

Um Patientengruppierungen zu erkennen, können Cluster-Algorithmen bzw. unüberwachtes Lernen eingesetzt werden. Hierzu existieren verschiedene etablierte Methoden wie zum Beispiel DBScan, k-means oder k-medoids. Einige dieser Ansätze wurden in der Literatur bereits zum Clustern von intensivmedizinischen Daten eingesetzt. Im Kontext des akuten Lungenversagens wurden z. B. Clustern-Algorithmen verwendet, um klinische Phänotypen des Krankheitsbildes (ARDS) zu identifizieren oder die Mortalität von unterschiedlichen Patientengruppen mit ARDS zu untersuchen. In einer weiteren Arbeit wurde ein Autoencoder entwickelt, mit dem Daten aus der MIMIC III-Datenbank geclustert werden können. Zur Validierung von resultierenden Clustern existieren etablierte Bewertungsmaße, wie z.B. „Silhouette“, „Davies-Bouldin“ oder der „Dunn Score“. Nach aktuellem Wissensstand existiert keine Clusteranalyse von den am Lehrstuhl vorliegenden intensivmedizinischen Datenbanken.

Zielsetzung

Ziel der Arbeit ist es, das unüberwachte Clustern von intensivmedizinischen Daten auf andere dem Lehrstuhl vorliegende Datenbanken zu erweitern. Dazu soll nach dem Vorbild des bereits bestehenden Autoencoders die Datenbankschnittstelle angepasst und weitere Methoden auf diese Datensätze angewandt sowie resultierende Cluster verglichen werden. So sollen Patientengruppen mit ähnlichen Symptomen, klinischen Bedürfnissen und Krankheitsverläufen identifiziert werden.

Geplante Vorgehensweise

Zuerst erfolgt eine Einarbeitung in die Thematik, Methodik des Clusters und die zur Verfügung stehenden Daten mitsamt Literaturrecherche zu bereits bestehenden Ansätzen. Danach wird der bestehende Autoencoder adaptiert und auf die intensivmedizinischen Daten der heterogenen Datenbanken angewandt. Außerdem soll das Clustern der Daten mit weiteren Methoden wie k-means und DBSCAN erfolgen und die Ergebnisse zum Schluss mit denen des Autoencoders verglichen und evaluiert werden. Dazu werden die Cluster in den einzelnen Datenbanken durch verschiedene Bewertungsmaße validiert.