**Automatische Auswahl von maschinellen Lernverfahren für kausale Inferenz**

|  |  |
| --- | --- |
| Start | End |
| 15.04.2019 | 15.08.2019 |

**PROBLEMSTELLUNG**

Um die Entscheidung zu treffen, ob eine Handlung getan werden muss oder nicht, werden seine potenziellen Folgen geschätzt – bringt es ein Nutzen oder ein Schaden. Beispiele für solche Handlungen findet man in der Medizin, in der Werbungplanung, in der Politik usw. Mit eine datengetriebene Analyse dieses Prozesses beschäftigt sich die kausale Inferenz.

Es wurden schon viele verschiedene Verfahren zur kausale Inferenz entwickelt, die nicht immer gleiche Ergebnisse liefern. Für die richtige Auswahl des Verfahrens muss man schon gute Einschätzungen für den echten Treatmenteffekt haben. Das ist aber keine triviale Aufgabe, denn es sind dafür gute Domain- und Statistikkenntnisse notwendig.

**IDEE**

Für die automatische Auswahl von einem Verfahren zur kausalen Inferenz bei vorgegebenen Daten wurde schon die Methode der „Synth-Validation“ entwickelt. Seitdem das passiert ist, wurden aber einige neuen maschinellen Lernverfahren zur kausale Inferenz entworfen.

In der Bachelorarbeit wird die automatische Auswahlmethode auch mit den neu entworfenen maschinellen Lernverfahren eingesetzt werden. Dabei streben wir eine bessere Vorhersagbarkeit von dem echten Treatmenteffekt einer Handlung an, die durch die größere Vielfalt von Verfahren und durch die größere Effektivität der neuen Verfahren in allgemeinem Fall geliefert werden soll.

**METHODIK**

Erstens werden die unterschiedlichen (maschinellen) Verfahren zur kausalen Inferenz identifiziert und beschrieben. Dabei werden seine Stärken, seine Schwächen und hauptsächlich seine Einschränkungen untersucht.

Von den identifizierten Methoden werden diejenigen genommen, die sich für eine automatische Auswahl von der „Synth-Validation“ Methode eignen. Sowohl die einzelnen Verfahren, als auch die Methode zur automatischen Auswahl müssen implementiert werden.

Um die implementierten Verfahren produktiv einzusetzen und ihren Effektivität zu messen bzw. zu vergleichen, benötigen wir geeignete Datensätze, von denen kausale Inferenz herausgezogen werden kann. Sie müssen möglichst als unterschiedlichen Bereichen kommen, damit die Unabhängigkeit der Methode sichergestellt werden kann.

Die Ergebnisse von den einzelnen maschinellen Lernverfahren sollen mit dem „Synth-Validation“ Methode verglichen werden. Dazu sollen passende Grafiken erstellt werden.

Schließlich soll auf Basis der Ergebnisse eine Stellung genommen werden, ob bzw. wann die automatische Auswahlmethode den echten Treatmenteffekt einer Handlung besser als die Baseline vorhersagt.

**BEITRAG**

Eine existierende Methode zur automatischen Auswahl von Lernverfahren soll mit bis jetzt nicht eingesetzten maschinellen Lernverfahren implementiert werden. Durch die Methode und durch die Evaluation der Ergebnisse wird noch mal bestätigt, dass es kein universelles Verfahren zur kausale Inferenz für beliebigen Datensätzen existiert.

**ARBEITSPLAN**

* Identifikation und Beschreibung von (maschinellen) Lernverfahren zur kausalen Inferenz, Auswahl für passende Verfahren für die „Synth-Validation“(2 Wochen)
* Systematische Suche/Auswahl von passenden Datensätze(2-3 Wochen, parallel zum ersten Punkt)
* Implementierung von zwei Lernverfahren und von den ganzen automatischen „Synth-Validation“ Auswahlverfahren (4 Wochen)
* Implementierung von den weiteren Lernverfahren und Anfang mit der Ausarbeitung (4 Wochen)
* Analyse der Ergebnisse und Anfertigung der Ausarbeitung (4 Wochen)
* Vorbereitung für den Vortrag (1 Woche)

**Voraussetzungen für die Bearbeitung**

* Allgemeine Kenntnisse über kausale Inferenz
* Kenntnisse über maschinelle Lernverfahren und deren Evaluierung
* Programmierkenntnisse in R

**BEKANNTE RISIKEN**

* Nicht alle maschinellen Lernverfahren können für automatische Auswahl bereitgestellt werden
* Unzureichende Menge an Datensätzen
* Auswahl von ungeeigneten Daten und Not zu deren Ersatz