Karlsruhe Institute of Technology

Bachelorarbeit

Automatische Auswahl von maschinellen Lernverfahren für kausale Inferenz

von

Atanas Dimitrov

Pervasive Computing Systems / TECO
Institut für Telematik
Fakultät für Informatik

Abgabadatum: 02.09.2019

Verantwortlicher Betreuer: Prof. Dr. Michael Beigl
Betreuerin: Ployplearn Ravivanpong

Erklärung

Hiermit erkläre ich, dass ich die vorliegende Bachelorarbeit selbstständig verfasst und keine anderen als die angegebenen Hilfsmittel und Quellen benutzt habe, die wörtlich oder inhaltlich übernommenen Stellen als solche kenntlich gemacht und weiterhin die Richtlinien des KIT zur Sicherung guter wissenschaftlicher Praxis beachtet habe.

Karlsruhe, 02.09.2019

Abstract

Diese Bachelorarbeit beschäftigt sich mit dem Vergleich von Methoden für kausale Inferenz und mit der automatischen Auswahl von dem besten von denen abhängig von dem vorhandenen Datensatz. Dazu benutzen und erweitern wir Synth-Validation - ein Verfahren, mit dem von den echten Daten synthetische Daten mit einem gewünschtem durchschnittlichen Behandlungseffekt erstellt und dann ausgewertet werden. Dabei haben die Datensätze, auf denen wir unsere Experimente durchführen, unterschiedliche Natur - echten Rohdaten mit größeren oder kleineren Zahl von Kovariaten, von Rohdaten synthetisch generierten Daten und zufällig generierten Daten. Die kausale Inferenz Verfahren, von denen Synth-Validation auswählt, benutzen ausschließlich Algorithmen aus dem maschinellem Lernen. Es wird die Fähigkeit von Synth-Validation gemessen, den Verfahren zu wählen, der die nächste Schätzung von dem durchschnittlichen Behandlungseffekt hat. Das wird unter unterschiedlichen Konstellationen unterstellt - nach der Art der Daten, nach der Anzahl der Elementen in der Stichprobe usw.

Inhaltsverzeichnis

Abbildungsverzeichnis

1 Einführung

Um die Welt besser zu verstehen, haben die Menschen immer die Antwort auf der folgenden Frage gesucht: Was passiert, wenn eine bestimmte Tat oder Handlung durchgeführt wird? Manche Handlungen kann man als „einfach“ qualifizieren und bei denen ist diese Antwort leicht zu erreichen. Wenn man einen Apfel nach oben wirft, fällt er wieder nach unten. Es kostet (fast) nichts, diese Tatsache zu prüfen. Fast jeder kann dieses Ergebnis vorhersagen, weil man irgendwann mal ein fallendes Objekt beobachtet hat, unabhängig davon, ob man weiß, warum die Objekte fallen. Handlungen können aber deutlich komplexer sein wie z.B. Einnahme einer neuen Steuerpolitik oder die medizinische Behandlung mit einem Medikamenten. Diese Handlungen sind schwerer durchzuführen, sind von mehreren Faktoren oder Naturgesetzen betroffen, sind „abhängiger“. Auf Basis früherer Erfahrung im Bereich können Experten Rahmen davon setzen, was passieren wird. Sicher kann man nur dann sein, wenn die Handlung durchführt ist und man beobachtet und bemesst, was passiert hat. Die letzte Schlussfolgerung gilt sowohl für die „einfacheren“, als auch für die „komplexeren“ Handlungen. Also um eine sichere Kenntnis zu erschaffen, brauchen wir Erfahrung - in dem Fall mit dem Apfel entweder werfen wir den alleine oder beobachten wir jemanden, der das macht oder der das irgendwann mal gemacht hat. Bei den „komplexeren“ Handlungen ist meistens die Antwort auf der Frage „Was?“ nicht ausreichend - man braucht „Wie viel?“. Das führt uns langsam zu dem Sachverhalt der kausalen Inferenz.

1.1 Kausale Inferenz

Wenn wir über die Kausalität in ihrem wissenschaftlichen Sinn sprechen, muss es klar sein, dass „$\boldsymbol{A}$ verursacht $\boldsymbol{B}$“ nicht bedeutet, dass $\boldsymbol{A}$ der Hauptgrund ist, warum $\boldsymbol{B}$ passiert hat. Es bedeutet, dass $\boldsymbol{A}$ den Unterschied oder die Differenz gemacht hat, die zu $\boldsymbol{B}$ geführt haben. Es existieren möglicherweise noch weitere Gründe z.B. $\boldsymbol{C}$ und $\boldsymbol{D}$, ohne deren Existenz und ohne deren verursachten Differenz $\boldsymbol{B}$ nicht möglich war oder nicht den Wert halten würde, den es hält. $\boldsymbol{A}$ soll man nicht als eine Zuschreibung von $\boldsymbol{B}$ anschauen, sondern als einen Beitrag[Mac]. Dabei sollen wir unter **kausale Inferenz** ähnlich wie bei der statistischen Inferenz den Prozess und die Methodiken zum Finden von Kenntnissen über die Kausalität von irgendeiner Handlung, insbesondere das Finden von der Differenz, die sie verursacht.

Die kausale Inferenz hat aber ein fundamentales Problem. Wenn wir den Effekt von einer Handlung wissen wollen, führen wir diese Handlung aus und bemessen den Wert von irgendeine Variable, die uns interessiert. Der echte Effekt ist aber die Differenz zwischen dieser Variable bedingt, dass wir die Handlung durchgeführt haben und die gleiche Variable, bedingt, dass wir die Handlung nicht durchgeführt haben. Natürlich ist nur eine der beiden Optionen direkt beobachtbar. Formal kann man es so dargestellt:

$$\boldsymbol{TE} = (\boldsymbol{Y}|\boldsymbol{W} = \mathbf{1}) - (\boldsymbol{Y}|\boldsymbol{W} = \mathbf{0}) \tag{1.1}$$

Hier bezeichnet $\boldsymbol{TE}$ das Behandlungseffekt, $\boldsymbol{Y}$ ist die beobachtete Variable und $\boldsymbol{W}$ ist eine binäre Variable, die bezeichnet, ob die Handlung durchgeführt wurde oder nicht.

Obwohl wir den Behandlungseffekt von einer Handlung individuell nicht beobachten können, können wir schätzen, wie hoch durchschnittlich diesen Effekt ist, wenn wir die Handlung auf mehreren Objekten durchführen beziehungsweise auch auf mehreren nicht durchführen und die Differenz zwischen diesen beiden Ergebnissen als **durchschnittliches Behandlungseffekt** oder auf Englisch **average treatment effekt(ATE)** bezeichnen. Meistens können wir die Behandlung nicht auf allen (oder eigentlich auf der Hälfte von allen) Objekten durchführen, sondern müssen wir eine Stichprobe ziehen. Formal stellen wir das so dar:

$$\boldsymbol{\tau} = \boldsymbol{E_Y}[\boldsymbol{Y}|\boldsymbol{W} = \mathbf{1}] - \boldsymbol{E_Y}[\boldsymbol{Y}|\boldsymbol{W} = \mathbf{0}] \tag{1.2}$$

In dieser Modelldarstellung sind $\boldsymbol{W}$ und $\boldsymbol{Y}$ Zufallsvariablen. $\boldsymbol{W}$ ist binär, wo $\mathbf{1}$ für eine Behandlung und $\mathbf{0}$ für keine Behandlung steht. $\boldsymbol{Y}$ ist die Zufallsvariable von einem gewünschten Wert nach der potenziellen Behandlung. $\boldsymbol{E_Y}$ steht für den Erwartungswert von $\mathbf{Y}$. Schon in dieser Darstellung treffen wir allen Problemen, die man sonst in der statistischen Inferenz trifft - damit unsere Schätzung über die Stichprobe näher zu der echten ATE der Gesamtpopulation liegt, müssen bestimmte Regeln eingehalten werden.

Gleichung 1.2 ist immer noch kein gutes Model der Wirklichkeit, weil es oft der Fall ist, dass die Objekten, die wir behandeln, sich durch einigen Merkmalen unterscheiden. Diese andere Differenzen können Einfluss auf $\boldsymbol{Y}$ sowohl bedingt $\boldsymbol{W}$, als auch unbedingt $\boldsymbol{W}$ haben und diese Tatsache soll in der Gleichung einbezogen werden:

$$\boldsymbol{\tau} = \boldsymbol{E_{Y,X}}[\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{W} = \mathbf{1}] - \boldsymbol{E_{Y,X}}[\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{W} = \mathbf{0}] \tag{1.3}$$

$\boldsymbol{X}$ ist eine Zufallsvariable für die so genannten **Kovariaten**. $\boldsymbol{X}$ ist mehrdimensional und stellt den Zustand, in dem sich ein Objekt oder allgemeiner gesagt eine Welt befindet. Die Kovariaten sollen nicht, aber können einen Einfluss auf das Ergebnis $\boldsymbol{Y}$ haben. Diejenigen mit Einfluss nennen wir **Confounders** oder **confounding Variablen**. Auf Deutsch kann man den Begriff am nächsten mit dem Wort Verwirrungsvariablen übersetzen. Die Existenz von Confounders erschwert unsere Aufgabe, den Effekt zu finden, den die Behandlung verursacht.[VS13].

In der Tat können Confounders auch indirekten Einfluss auf die Behandlung haben. Beispiel: Sei $\boldsymbol{X}$ das Alter sein, $\boldsymbol{W}$ - ob man raucht oder nicht und $\boldsymbol{Y}$ - eine Gesundheitsmaß. Es ist klar, dass das Alter an sich einen Effekt auf der Gesundheit hat. Die empirische Daten zeigen

aber auch, dass das Rauchen unter älteren Menschen verbreiteter ist als unter jüngeren, d.h das Alter bewirkt sowohl das Rauchen, als auch die Gesundheit, die an sich vom Rauchen bewirkt wird. Diese Abhängigkeiten werden in Abbildung 1 veranschaulicht.

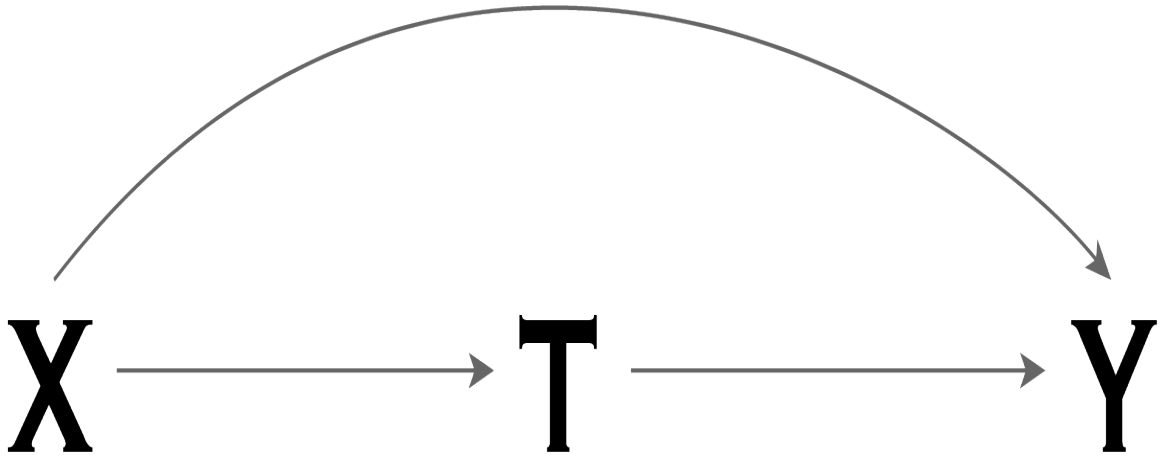

Abbildung 1: Einfluss von Confounder auf Behandlung und Ergebnis. Hier ist $\boldsymbol{X}$ der Confounder, $\boldsymbol{T}$ - die Behandlung und $\boldsymbol{Y}$ - das Ergebnis. Es wird gezeigt, dass das Ergebniss nicht nur von der Behandlung abhängt, sondern auch vom Confounder. Der Confouder hat auch eine Auswirkung auf die Behandlung.[Reb]

Wir schließen diesen Abschnitt mit der Gedanke von dem Philosophen David Lewis über die Kausalität, der sie folgendermaßen beschreibt: „etwas, was einen Unterschied verursacht und der Unterschied, den dieses etwas verursacht, soll der Unterschied davon sein, was ohne dieses etwas passieren würde“ [Lew74].

1.2 Motivation

Es ist wichtig zu wissen, was genau eine Handlung verursacht oder welche Differenz sie macht. Mit diesen Kenntnissen können wir besser die passenden Handlungen und deren genauen Maßen zum Erreichen unseres Ziels einschätzen und dann auswählen. Diese Behauptungen sind allgemeingültig und wir betrachten konkret drei Beispiele von unterschiedlichen Bereichen, wo das Kennen von dem durchschnittlichen Behandlungseffekt von großer Bedeutung ist.

Beispiel 1 In der Pharmazie spielt die Kausalität eine besonders große Rolle. Wenn man einen Medikamenten entwickelt, soll man ganz genau wissen, welchen durchschnittlichen Effekt dieses Medikament auf unterschiedenen variablen körperlichen Werten hat - sowohl für wohltuende, als auch für schädliche Effekte. Hier ist das Nutzen, das die kausale Inferenz bringt, die Gesundheit der Menschen.

Beispiel 2 Weitere Bedeutung hat die kausale Inferenz im Marketing und in der Bewertung von Werbungen. Eine Werbung kann eine Kette von unterschiedlichen Zielen haben, aber am Ende steht die Erhöhung von tatsächlichen Einkäufen von dem geworbenen Produkt oder Dienstleistung. Natürlich ist die Werbung nicht der einzige Treiber des Umsatzes. Also man muss nur den Effekt der Werbung einschätzen. Dann abhängig von diesem durchschnittlichen Effekt und dem Preis der Werbung stellt man fest, ob eine Werbung Nutzen bringt bzw. welche Werbung am nützlichsten ist.

Beispiel 3 Die kausale Inferenz ist bedeutend für die Auswahl vom eigenen Verhalten sein. Zum Beispiel kann man den Einfluss vom regelmäßigen Rauchen von Rauschgift auf die mentalen Fähigkeiten untersuchen. Wenn man diesen durchschnittlichen Effekt kennt, sollte es wahrscheinlicher sein, dass man auf diese schlechte Sucht verzichtet oder am besten gar nicht damit anfängt. Diese deutliche wissenschaftliche Daten können durch den Gesetzgeber als Rechtfertigung für einen Verbot von diesem Stoff benutzt werden.

1.3 Finden von ATE

Wir gehen davon aus, dass unser Problem schon klar definiert ist: Welchen durchschnittlichen Effekt(ATE) verursacht eine Behandlung? In diesem Abschnitt erläutern wir, wie man genau mit diesem Problem in den unterschiedlichen Situationen umgeht.

Grundsätzlich muss man Gleichung 1.3 lösen, die wir zur Bequemlichkeit hier noch mal hinbringen.

$$\boldsymbol{\tau = E_{Y,X}[Y|X, W = 1] - E_{Y,X}[Y|X, W = 0]} \tag{1.4}$$

Dazu gibt es zwei Vorgehensweisen, die sich durch die Art von Datensammlung unterscheiden - Durchführung von einem Experimenten und Durchführung von einer Beobachtungsstudie. Die Daten aus den unterschiedlichen Studien wertet man dann anders aus.

1.3.1 Experimentdurchführung

Es gilt, dass der durch Daten von Experimenten geschätzte Behandlungseffekt weniger Bias hat und deswegen näher zu dem echten durchschnittlichen Effekt der Population liegt. Der Grund dafür ist, dass die Objekten für die beiden Gruppen zufällig ausgewählt sind und somit wie die Population verteilt sind. Man soll immer diese Art von Datensammlung für kausale Inferenz wählen, wenn sie möglich ist. Von den in Abschnitt 1.2 erwähnten Beispielen ist es bei Beispielen 1 und 2 möglich.

Das Experiment läuft folgendermaßen durch: zwei Gruppen von Objekten werden gebildet. Die erste Gruppe heißt Behandlungsgruppe und die zweite - Kontrollgruppe. Die Objekten von der ersten Gruppe werden mit irgendetwas z.B. einem Medikamenten behandelt, diese in der zweite jedoch nicht. Alle anderen Konditionen, auf die die Objekten in der Kontrollgruppe unterlegt sind, sollen sich nicht unterscheiden. Natürlich bleiben aber die einzelnen Objekten unterschiedlich, was zu unterschiedlichen Ergebnissen bei den individuell gemessenen Werten führt. Uns interessiert aber der durchschnittliche Wert von jeder Gruppe. Deswegen muss sichergestellt werden, dass die Objekte in jeder Gruppe gleichmäßig verteilt sind. Es gibt unterschiedliche Strategien das zu erreichen. Eine der besten, wenn sie richtig durchgeführt wird, ist die zufällige Auswahl von Objekten aus der Population für jede Gruppe. Am Ende wird für jedes Objekt der Wert von Bedeutung gemessen. Die Mittelwerten von diesen Werten werden für jede Gruppe berechnet und die Differenz dazwischen ist der erwartete durchschnittliche Behandlungseffekt.

Man kann auch Daten von den beiden Gruppen benutzen, um signifikanten Aussagen über den durchschnittlichen Behandlungseffekt mit einer bestimmten Konfidenz zu treffen. Dazu führt man meistens einen statistischen $\boldsymbol{t}$-test durch. Man erstellt erstmal eine Nullhypothese und prüft, ob diese abzulehnen ist. Dafür berechnet man erstmal die Teststatistik $\boldsymbol{t}$, die hier dargestellt ist:

$$\boldsymbol{t} = \sqrt{\frac{\boldsymbol{n_0 n_1}}{\boldsymbol{n_0 + n_1}}} \frac{\overline{\boldsymbol{Y_1}} - \overline{\boldsymbol{Y_0}} - \boldsymbol{\omega}}{\boldsymbol{s}} \tag{1.5}$$

$$\boldsymbol{s} = \sqrt{\frac{(\boldsymbol{n_0} - \boldsymbol{1})\boldsymbol{s_0^2} + (\boldsymbol{n_1} - \boldsymbol{1})\boldsymbol{s_1^2}}{\boldsymbol{n_0 + n_1 - 2}}} \tag{1.6}$$

$\boldsymbol{n_0}$ und $\boldsymbol{n_1}$ sind die Anzahl von Objekten in den beiden Gruppen, $\overline{\boldsymbol{Y_1}}$ und $\overline{\boldsymbol{Y_0}}$ - die Mittelwerte von den Ergebnissen, $\boldsymbol{s_0^2}$ und $\boldsymbol{s_1^2}$ - die Varianzen, $\boldsymbol{s}$ - die gewichtete Abweichung, $\boldsymbol{\omega}$ - den Wert von dem durchschnittlichen Behandlungseffekt, den man für die Nullhypothese genommen hat. Dann abhängig davon, was man testen will und mit welcher Konfidenz die Behauptung gültig sein soll, vergleicht man den Wert von $\boldsymbol{t}$ mit dem Wert von einem Quantil von der studentischen $\boldsymbol{t}$-Verteilung. Am Ende hat man eine Behauptung, die mit irgendeiner Signifikanz wahr ist.[Wik]

1.3.2 Beobachtungsstudiedurchführung

Wir wiederholen noch mal, dass man lieber ein Experiment anstatt einer Beobachtungsstudie durchführen muss, um die Daten zu sammeln, wenn es möglich ist. Und natürlich ist es wegen unterschiedlichen Gründen nicht immer so.

Grundsätzlich kann ein Experiment viel mehr als eine Beobachtungsstudie kosten, weil man die Objekten (die Leuten) in experimentellen Bedingungen haben möchte. Man hat (höhere)

Kosten für:

- Vergütung der Leute
- die Behandlung
- Mitarbeiter, die das Experiment durchführen
- usw.

Bei manchen Studien kann der Prozess der Behandlung sogar länger dauern, was die obengenannten Kosten multipliziert.

Es gibt Behandlungen, wo es klar ist, dass sie einen negativen Effekt haben; man weiß doch nicht genau, wie hoch er ist. In diesem Fall ist es unethisch, Leuten auf diese negative Behandlung in Rahmen eines Experiments zu unterlegen. Vielleicht ist die Behandlung auch durchs Gesetz verboten. Genauso das ist die Situation im Beispiel 3 vom Abschnitt 1.2. Man darf nicht die Gesundheit der Leuten opfern, um etwas zu erfahren. In diesem Fall führt man eine Beobachtungsstudie unter Menschen durch, die auf die Behandlung unterlegt sind und zur Kontrolle auch unter solchen, die nicht unterlegt sind.

Und hier kommt ein großes Problem. Die Objekten, die behandelt sind, können anders verteilt sein als die ganze Population. Beispielsweise gibt es unter den Behandelten viel mehr Männer als Frauen. Die Behandlung auf Männer könnte aber einen anderen Effekt haben als diesen auf Frauen. Also wenn man seine Stichprobe durch zufälliges Ziehen bildet, bekommt man einen geschätzten durchschnittlichen Effekt, der von der höheren Anwesenheit von Männern gestreut ist und nicht an dem durchschnittlichen Effekt der Population entspricht. Man kann sich die Abbildung 1 noch mal anschauen und feststellen, dass die Behandlung $\boldsymbol{T}$ von den Confounders $\boldsymbol{X}$ abhängt. In einem Experiment ist das ausgeschlossen, da es bedingt ist, dass am Anfang niemand behandelt war.

Weil die Daten aus Beobachtungsstudien viel Bias haben, würde eine bloße Berechnung von den Mittelwerten von den beiden Gruppen nicht ausreichend gut sein. Deswegen hat man unterschiedliche Methoden entwickelt, die wir ab hier **Methoden für kausale Inferenz** nennen, die grundsätzlich das Ziel haben, die Behandlugs- und die Kontrollgruppe zu normaliesieren, damit die vergleichbar sind und damit das Ergebnis näher am Ergebnis der Population liegt. In dieser Bachelorarbeit haben wir über einige dieser Methoden in Abschnitt 2 erzählt. Es ist natürlich zu erwarten, dass diese Methoden unterschiedliche Ergebnisse liefern. Dabei gibt es keine allgemein Beste - in unterschiedlichen Situationen funktioniert die eine besser als die anderen. Deswegen wurde ein Verfahren entwickelt, den die Methode für einen Datensatz automatisch wählt und auf diesen Verfahren basieren wir diese Bachelorarbeit.

1.4 Ziele und Methodik

Nachdem wir in das Thema „Kausale Inferenz“ eingestiegen sind, werden wir im folgenden Abschnitt die Arbeitsmethodik dieser Bachelorarbeit erläutern. Dabei nennen wir auch die Ziele und Aufgaben, die wir uns setzen.

Wir basieren unsere Arbeit auf den Synth-Validation Verfahren[SJT+17]. Synth-Validation ist ein Verfahren, der für einen bestimmten Datensatz, zu dem wir den durchschnittlichen Behandlungseffekt schätzen wollen, versucht, von einer Menge von Methoden für kausale Inferenz, die Beste auszuwählen und setzt diese ein. Im Abschnitt 3 erzählen wir ausführlicher darüber. Wir arbeiten grundsätzlich an den Aufgaben, die im Abschnitt **Future work** im Artikel über Synth-Validation[SJT+17] genannt sind.

Das erste Ziel, die wir uns setzen, ist Synth-Validation auf R zu implementieren. Dazu verfügen wir über einen großen Teil des Codes. Der ist auf Julia geschrieben und wir haben den persönlich von einem der Authoren von Synth-Validation bekommen. Außer dem Übersetzen von Julia auf R müssen einige Teile von Synth-Validation, die im Artikel beschrieben sind, aber im vorgegebenen Julia Code jedoch fehlen, neu geschrieben werden. Darunter sind das Lesen von echten Daten, die Wahl von synthetischen Effekten, die finale Auswahl von einer Methode für kausale Inferenz von den synthetischen Daten und den ganzen Benchmarkprozess von Synth-Validation. Wir erzählen mehr darüber im Abschnitt 4.

Wir wählen R als Implementierungssprache, weil sie sehr verbreitet und ein Standard für statistikbasierte Ausarbeitungen ist. Es gibt eine große Gemeinschaft (auch unter Wissenschaftlern), die sie benutzt und deswegen sind auch viele Bibliotheken verfügbar. Dabei bietet die Scriptsprache einen einfachen Syntax, der erlaubt schnell einzusteigen und ein Basis für kürzere Implementierungszeit ist. Wir kennen natürlich auch die potentiellen Nachteilen, die die Sprache hat und die dazu geführt haben, dass man am Anfang entschieden hat, Synth-Validation auf Julia anstatt auf R zu implementieren. Diese besprechen wir im Abschnitt 5.

Eine weitere Aufgabe ist zu testen, wie Synth-Validation funktioniert, wenn sie auf echten Daten eingesetzt ist. Im Artikel von Synth-Validation wird beschrieben, dass sie nur mit zufälliggenerierten Daten getestet wurde. Das Finden von solchen Daten ist keine triviale Aufgabe, weil der echte Behandlungseffekt mit einer großer Sicherheit bekannt sein muss, damit wir Synth-Validation benchmarken können. Bei den automatischgenerierten Daten kann man natürlich immer einen beliebigen Effekt haben, aber bei den echten ist das Finden von diesem Effekt eigentlich das, was wir von Anfang an erzielen wollen.

Die Daten, über die wir verfügen, kommen aus Wettbewerben für kausale Inferenz, deswegen haben sie einen bekannten durchschnittlichen Behandlungseffekt. Dabei unterscheiden wir zwischen Daten, die roh oder gar nicht bearbeitet wurden und solche, die durch echten Daten generiert wurden. Die Datensätze haben auch eine unterschiedliche Anzahl von Beobachtungen und eine unterschiedliche Anzahl von Kovariaten. Neben diesen echten Daten testen wir auch

mit solchen, die zufälliggeneriet sind und vergleichen unsere Ergebnisse mit diesen von den Authoren von Synth-Validation, soweit es möglich ist. Mehr über die Ergebnisse, ihre Auswertung und den Sachverhalt der Daten erzählen wir im Abschnitt 5.

Eine weitere Aufgabe ist Synth-Validation mit neuen Methoden zu testen. Seitdem Synth-Validation entwickelt und den Artikel darüber geschrieben wurde, sind einige neuen Methoden für kausale Inferenz entstanden, die bessere Ergebnisse im Allgemein aufweisen. Diese Methoden benutzen Algorithmen aus dem maschinellen Lernen. Es wird nützlich zu wissen, ob Synth-Validation immer noch erfolgreich unter denen wählt oder ob eine von denen sich besser als Synth-Validation vorstellt.

2 Methoden für kausale Inferenz

2.1 Lineare Verfahren

2.1.1 Covariate Matching

2.1.2 Propensity Score Matching

2.1.3 Inverse Probability Weighting

2.2 Maschinelle Lernverfahren

2.2.1 Gradient Boosting

2.2.2 Lasso

2.2.3 Kausale Wälder

2.2.4 Targeted Maximum Likelihood Estimation

3 Synth-Validation

3.1 Generierung von synthetischen Daten

3.1.1 Auswahl von synthetischen Effekten

3.1.2 Schätzung von bedingten Erwartungswerten

3.2 Methodenauswahl

4 Implementierung

4.1 Fremde Bibliotheken

4.2 Lesen/Schreiben von Daten

4.3 Ziehen von Stichproben

4.4 Methoden für kausale Inferenz

4.5 Synth-Validation

4.5.1 Datenstrukturen

4.5.2 Schätzung

4.5.3 Constrained Boosting

4.5.4 Methodenauswahl

4.6 Benchmark von Synth-Validation

4.7 Erstellung von Abbildungen

4.8 Experimenten

4.9 Anderer Code

5 Ergebnisse und Evaluation

5.1 Methodik und Daten

5.2 TODO

6 Schlussfolgerung

6.1 Zusammenfassung

6.2 Diskussion

Anhang

Literatur

[ATW+19] Susan Athey, Julie Tibshirani, Stefan Wager, et al. Generalized random forests. *The Annals of Statistics*, 47(2):1148–1178, 2019. Keine Zitierungen.

[Lew74] David Lewis. Causation. *The journal of philosophy*, 70(17):556–567, 1974. (Seite 3).

[Mac] Macartan Humphreys. 10 things to know about causal inference. `https://egap.org/methods-guides/10-things-you-need-know-about-causal-inference`. [Online; accessed 14-August-2019]. (Seite 1).

[Reb] Rebecca Barter. Confounding in causal inference: what is it, and what to do about it? `http://www.rebeccabarter.com/blog/2017-07-05-confounding/`. [Online; accessed 16-August-2019]. (Seite 3).

[SJT+17] Alejandro Schuler, Ken Jung, Robert Tibshirani, Trevor Hastie, and Nigam Shah. Synth-validation: Selecting the best causal inference method for a given dataset. *arXiv preprint arXiv:1711.00083*, 2017. (Seite 7).

[VS13] Tyler J VanderWeele and Ilya Shpitser. On the definition of a confounder. *Annals of statistics*, 41(1):196, 2013. (Seite 2).

[WA18] Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018. Keine Zitierungen.

[Wik] Wikipedia Authoren. Zweistichproben-t-test. `https://de.wikipedia.org/wiki/Zweistichproben-t-Test`. [Online; accessed 16-August-2019]. (Seite 5).