Richtlinie zum „Software-Sprint“

causy – Kausalitätsalgorithmen mit ausführlichen Erklärungen und interaktiven Visualisierungen

Schlussbericht

Zuwendungsempfänger:

Sofia Faltenbacher

Das diesem Bericht zugrundeliegende Vorhaben wurde mit Mitteln des Bundesministeriums für Bildung und Forschung unter dem Förderkennzeichen 01IS24S12 gefördert. Die Verantwortung für den Inhalt dieser Veröffentlichung liegt beim Autor.

Kurze Darstellung der Aufgabenstellung und Motivation

Warum stornieren Menschen ihre Hotelbuchungen? Wie beeinflusst Rauchen die Zahngesundheit? Welche Auswirkungen haben Waldbrände auf das Schmelzen des Meereises? Kausale Fragestellungen treten überall auf: in der Wissenschaft, in Unternehmen oder in Organisationen. In vielen Situationen kann man Kausalzusammenhänge nicht durch Experimente überprüfen, weil es unmoralisch, zu teuer oder unmöglich ist. Man kann etwa eine Stadt nicht aufwärmen oder herunterkühlen, man kann die Temperatur nur messen. Doch auch dann gibt es statistische Methoden auf beobachteten Daten, mit denen unter bestimmten Annahmen Kausalitäten ausgeschlossen oder abgeleitet werden können. Um diese Methoden zu verstehen, braucht es allerdings viel Vorwissen.
causy möchte Einstiegshürden in das Thema Kausale Inferenz mit Hilfe von Komfort-Features, Dokumentation, niedrigschwelligen Erklärungen sowie interaktiven Visualisierungen senken und die Methoden so für eine breitere Zielgruppe zugänglich machen. Dazu gehört auch, klar zu machen, für welche Datensätze sich die Algorithmen eignen und für welche nicht. Dieses Aufzeigen der Modellannahmen und Grenzen als fester Bestandteil des Prozesses ermöglicht einen verantwortungsvollen Umgang mit den Ergebnissen. Die Meilensteine waren:

- Die Implementierung der Algorithmen auf Basis von flexibel zusammenstellbaren Modulen.
- Das Tracking der einzelnen Ausführungsschritte; die Erstellung der Dokumentation; die Visualisierungen von Ergebnissen der Algorithmen.
- Das Testen und Implementieren von weitergehenden Lösungen zur interaktiven Unterstützung der Konfiguration der Algorithmen sowie Interpretation der Auswertungen.

Beitrag des Projektes zu den Zielen der Förderinitative „Software-Sprint“

causy leistet einen Beitrag zur Förderung der Data Literacy, indem es Anwendern ermöglicht, Kausalitätsalgorithmen zu verstehen und anzuwenden, ohne tief in die Implementierung oder mathematischen Details eintauchen zu müssen. Nutzern werden in der Kommandozeile schrittweise durch den Initialisierungsprozess geführt, ohne eigenen Code schreiben zu müssen. Anschließend können sie mit der *causy* UI im Browser interaktiv mit den Ergebnisgraphen arbeiten und die Teilschritte der Algorithmen übersichtlich einsehen. Zudem lassen sich die Resultate mit nur zwei Klicks komfortabel mit Kollegen teilen. Dieser intuitive Zugang erleichtert es nicht nur Data Scientists ohne Spezialisierung auf Kausalität, sondern auch Forschenden aus anderen Disziplinen wie Chemie, Psychologie oder Soziologie, Kausalitätsanalysen durchzuführen.
Darüber hinaus ermöglicht *causy* durch seine Komfortfeatures und Automatisierungen Forschenden im Bereich der Kausalen Inferenz, Robustheitstests effizient und systematisch durchzuführen. Die modulare Architektur sowie die Pre- und Post-Hooks bieten die Flexibilität, neue Methoden – beispielsweise die Kombination von Machine Learning mit Kausalitätsalgorithmen – schnell und unkompliziert zu testen. Dies kann indirekt zur Data Literacy beitragen, indem es erleichtert, kontextspezifische Methoden, präzisere Fehlerabschätzungen und praxisnahe Leitlinien zu entwickeln.

Ausführliche Darstellung der Ergebnisse

***causy* Workspaces**

Mithilfe von *causy* Workspaces können User Datensätze, Algorithmen und Konfigurationen verwalten und ausführen, ohne selbst Code zu schreiben. Damit kann ein User etwa auf denselben Datensatz (etwa über Hotelbuchungen) verschiedene Kausalitätsalgorithmen mit verschiedenen p-Werten für die statistischen Tests anweden (um etwa herauszufinden, warum Menschen ihre Buchung canceln). Die Unterschiede in den Ergebnissen der verschiedenen Algorithmen und Konfigurationen können User sich übersichtlich zusammengestellt in der Kommandozeile anzeigen lassen.

***causy* UI**

causy UI bietet darüber hinaus die Möglichkeit, die Ergebnis-Graphen im Browser anzeigen zu lassen und die einzelnen Schritte der Algorithmen einzusehen und nachzuvollziehen. Statische Graphen, die etwa mit networkx generiert werden können, haben das Problem, dass mit steigender Anzahl an Variablen immer öfter Anpassungen jedes einzelnen Outputs für die Lesbarkeit nötig werden. *causy* UI löst diese Problem durch die Möglichkeit, die Position der Knoten einfach in der UI zu verschieben. So kann sich der User seinen Graphen individuell arrangieren, ohne für jedes Beispiel manuelle 2-d Koordinaten für die Knoten auszuprobieren und den Algorithmus immer wieder laufen zu lassen, bis die Ergebnisse gut erfassbar sind.

Methoden und Data Generators

Es sind aktuell der PC Algorithmus, eine parallelisierte Version der PC Algorithmus und eine stable version des PC Algorithmus verfügbar. Außerdem wurde eine causal effect estimation für den linearen Fall implementiert. Der FCI Algorithmus ist in Arbeit. Es gibt Datengeneratoren für lineare Structural Causal Models (SCMs), welche von unabhängigen, identisch verteilten Zufallsvariablen oder stationären stochastischen Prozessen generiert sind.

Dokumentation und einfache Erklärungen

Neben der Dokumentation und der API Dokumentation gibt es auf der Website ein Video-Tutorial,

einfache Erklärungen zu den Voraussetzungen der Algorithmen, eine Liste hilfreicher Bücher und Quellen sowie einen Troubleshooting Guide.

Bilder: Workspaces, UI

```
➜  causy_demo causy workspace execute
Executing experiment: config_1
Loading Pipeline: pc
Using variables: {'threshold': 0.01}
Loading Data: hotel_data
Executing experiment: config_2
Loading Pipeline: pc
Using variables: {'threshold': 0.005}
Loading Data: hotel_data
➜  causy_demo causy workspace diff config_1 config_2
```

Edge	config_1_latest	config_2_latest
IsCanceled – LeadTime	-->	-->
IsCanceled – IsRepeatedGuest	-->	-->
IsCanceled – PreviousCancellations	-->	-->
IsCanceled – RequiredCarParkingSpaces	-->	-->
IsCanceled – TotalOfSpecialRequests	-->	-->
IsCanceled – total_members	-->	-->
IsCanceled – BookingChanges	-->	-->
StaysInWeekNights – BookingChanges	-->	-->
Children – BookingChanges	-->	-->
BookingChanges – RequiredCarParkingSpaces	-->	-->
Meal_SC – BookingChanges	-->	-->
MarketSegment_Offline TA/TO – BookingChanges	-->	-->
DepositType_Non Refund – BookingChanges	-->	-->
IsCanceled – DaysInWaitingList	-->	-->
LeadTime – DaysInWaitingList	-->	-->
arrival_date_month – DaysInWaitingList	-->	-->
MarketSegment_Groups – DaysInWaitingList	-->	-->
DepositType_Refundable – DaysInWaitingList	-->	-->
Meal_Undefined – DaysInWaitingList		-->
IsCanceled – ADR	-->	-->
ArrivalDateWeekNumber – ADR	-->	-->
Children – ADR	-->	-->
ADR – IsRepeatedGuest	-->	-->
ADR – RequiredCarParkingSpaces	-->	-->
TotalOfSpecialRequests – ADR	-->	-->
arrival_date_month – ADR	-->	-->
total_members – ADR	-->	-->
Meal_SC – ADR	-->	-->

Bild 1: Ausführen eines Workspaces und Vergleich zweier Konfigurationen in der Kommandozeile.

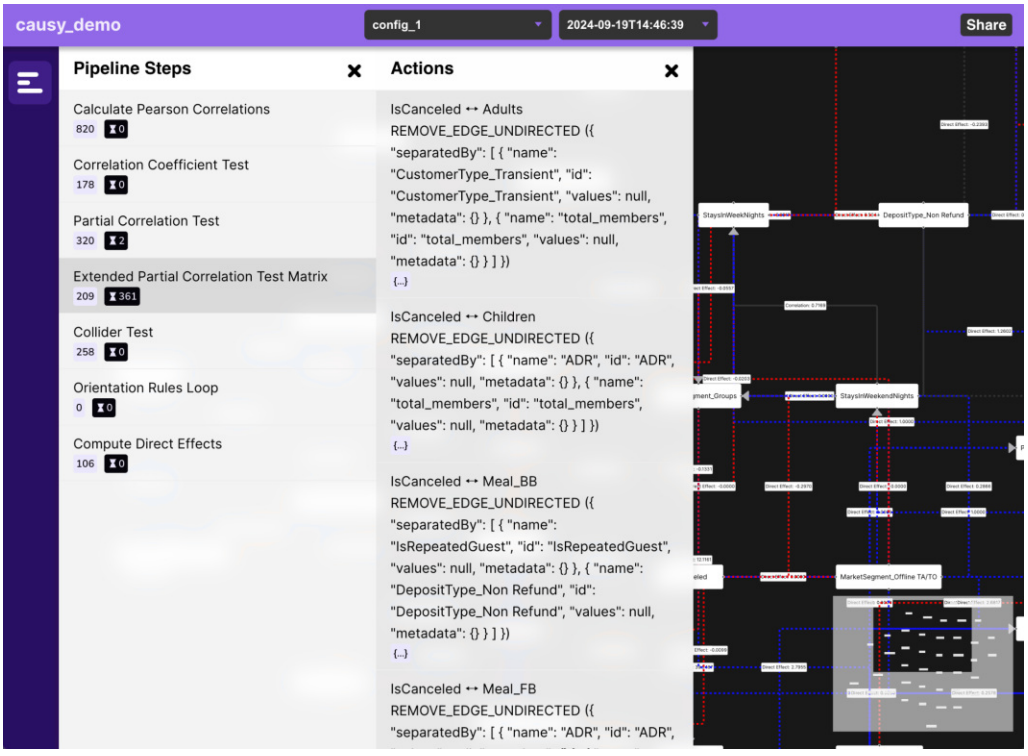


Bild 2: Ansicht der einzelnen Pipeline Steps mit *causy* UI.

Zielgruppe, Nutzen und mögliche Weiterentwicklungen

causy bietet interessierten Nutzern einen Einstieg in Methoden der Kausalen Inferenz. Aktuell sind das vor allem mit Forschende – sowohl aus dem Feld Kausale Inferenz als auch anderen Fachgebieten (z. B. angewandte Mathematik oder Soziologie). Darüber hinaus wird es von Softwareentwicklern und Datenjournalisten getestet. Die modulare Architektur bietet eine gute Grundlage, um Algorithmen individuell anzupassen oder Machine Learning Methoden vor- oder nachzuschalten. Dies wird von Usern positiv zurückgemeldet.
Mich persönlich hat causy in meiner wissenschaftlichen Arbeit auf vielen Ebenen vorangebracht. Zum einen ganz simpel, weil man wichtige Details noch besser versteht, wenn man Algorithmen und ihre verschiedenen Optimierungen from scratch implementiert. Dabei sind mir einige theoretische Herausforderungen noch klarer geworden, welche ich in der Forschungsarbeit weiter untersuchen werde. Zum anderen bin ich überzeugt, dass die Automatisierungen, welche *causy* bereitstellt, mir in meiner weiteren wissenschaftlichen Karriere – und hoffentlich vielen meiner Kolleg:innen – helfen werden, etwa bei systematischen Robustheitstests oder um neue Ideen schnell zu implementieren, zu testen und zu visualisieren.

Kurze Darstellung der Arbeiten, die zu keiner Lösung geführt haben

Die Konzeption von Erklärvideos in dem Format “Erkläre es mir, als wäre ich fünf / ein Schüler / ein Student / ein Experte” hat nicht wie geplant zu YouTube-Videos geführt. Es stellte sich heraus, dass das Konzept recht aufwendig ist (entsprechende Personen finden, Animationen, ansprechendes Videosetting, Nachbearbeitung, etc.) und kurze Tutorials als Videos besser umzusetzen waren. Die Beispiele aus dem Videokonzept konnten für Erklärtexte auf der Website genutzt und für Vorträge über *causy* weiterverwendet werden.

Kurze Angabe von Präsentationsmöglichkeiten für mögliche Nutzer

- Projekt Website: https://causy.dev/
    - Dokumentation und Video-Tutorial: https://causy.dev/use/
    - Einführung, Troubleshooting, Leseliste, Tipps: https://causy.dev/learn/
- GitHub Repository: https://github.com/causy-dev/causy
- API Dokumentation: https://causy-dev.github.io/causy/causy.html

Kurze Erläuterung zur Einhaltung der Arbeits- und Kostenplanung

Insgesamt konnte ich alle Features bauen, welche in den Milestones festgelegt waren. Ich habe etwas mehr Zeit darauf verwendet, den Workflow noch angenehmer zu gestalten als weitere Algorithmen zu implementieren. Deswegen gibt es eine Reihe an Komfort-Features, die nicht alle geplant waren, dafür ist der FCI Algorithmus noch in Arbeit.

Kurze Darstellung von etwaigen Ergebnissen bei anderen Stellen

Ich führe eine Liste von Kausalitäts-Packages auf GitHub wobei besonders *dowhy*, *CausalPy* und *tigramite* (für time series data) hervorzuheben sind. Zudem beobachte ich interessiert den Fortschritt webbasierter und proprietärer Anwendungen in dem Bereich, wie etwa *causaLens*. Mit einigen der Entwicklern stehe ich im Austausch, mit manchen habe ich zusätzlich User Interviews geführt. Ich denke, hier ergänzen sich die Angebote gut: Wer etwa eine orientation conflict resolution strategy

aus *dowhy* ausprobieren möchte, aber gleichzeitig *causy* workspaces und *causy* UI nutzen möchte, kann aufgrund der Modularität mit wenigen Änderungen die default orientation conflict resolution strategy in *causy* mit der gewünschten Methode ersetzen.