MACHINE LEARNING IN DER AMTLICHEN STATISTIK

27. Wissenschaftliches Kolloquium, Wiesbaden, 22./23. November 2018

Martin Beck
Statistisches Bundesamt

Florian Dumpert
Universität Bayreuth

Einführung

- Digitale Agenda des Statistischen Bundesamtes seit Anfang 2018
- Ziel: Noch besseres Ausfüllen der Rolle als führender Anbieter qualitativ hochwertiger statistischer Informationen über Deutschland
- Weg dahin: Eine ganzheitliche, an den Bedürfnissen der Nutzerinnen und Nutzer ausgerichtete und durch nahtlose elektronische Abläufe und agiles Arbeiten unterstütze Transformation
- Zur Einordnung: Vision „Towards Self-Driving Data Curation“ (Qatar Computing Research Institute)
- Thema heute: Integration von Machine-Learning-Methoden in den Prozess der Statistikproduktion
  - Proof of Concept Machine Learning *abgeschlossen*
  - Machine-Learning-Methodik *andauernd*
  - Proof of Concept automatisierte Plausibilisierung (in den Verdienststatistiken) *andauernd*

Was ist maschinelles Lernen?

- Es gibt keine einheitliche Definition.
- Erkennen von und Lernen aus Datenstrukturen: Aus Beobachtungen wird ein Zusammenhang zwischen Eingabevariablen und Ausgabevariable erlernt, der anschließend auf neue Eingabewerte angewendet werden kann, um den unbekannten Ausgabewert zu schätzen.
- Varianten sind
  - überwachtes Lernen (supervised learning): Klassifikation und Regression. Benötigt werden geeignete Trainings- und Testdaten, für die das wahre Ergebnis der Klassifikation (bzw. der Regression) bekannt ist;
  - unüberwachtes Lernen (unsupervised learning): Ausreißeridentifikation oder Clustering .
- Gängige, auch in der Praxis statistischer Institutionen verbreitete Methoden des maschinellen Lernens sind Support Vector Machines, Random Forests und Neuronale Netze.

Was ist maschinelles Lernen?

Proof of Concept Machine Learning

- Überprüfung der Einsetzbarkeit von maschinellem Lernen in den Prozessen der Fachstatistiken
  - Künstliche Intelligenz und Big Data sowie Anwendungen in der Verwaltung wurden aus der Analyse ausgeklammert
- Ziel: Überblick über Anwendungsmöglichkeiten in den Fachstatistiken im Statistischen Bundesamt; Umsetzung erst danach
- Basis u.a. bereits im Statistischen Bundesamt durchgeführte Projekte in der Unternehmens-/Verdienststatistik

Bereits durchgeführte Projekte

Statistik	Problem	Methode	Stand	Ergebnis
Unternehmensregister	Zuordnung von Unternehmen zum 3. Sektor	Support Vector Machine (SVM)	abgeschlossen	+
Handwerkszählung	Erkennung nicht relevanter Unternehmen in den Handwerksstatistiken	Random Forest Support Vector Machine (SVM)	abgeschlossen	++
Verdienststrukturerhebung	Schätzung einer Erwerbsunterbrechung von Frauen in der Verdienststrukturerhebung	Support Vector Machine (SVM)	abgeschlossen	+/-
Verdienststrukturerhebung	Schätzung der Staatsbürgerschaft von Arbeitnehmern in der Verdienststrukturerhebung	Support Vector Machine (SVM)	abgeschlossen	(-)
Verdienststrukturerhebung	Anreicherung der Integrierten Erwerbsbiografien (IEB) von BA/IAB um Informationen aus der Verdienststrukturerhebung	Random Forest	läuft	+/-
Einführung des EU-Unternehmensbegriffs in den Strukturstatistiken	Identifikation von Ausreißern im Spenderdatenpool für die Imputation	Isolation Forest	läuft	+

Dabei aufgetretene Schwierigkeiten [und Lösungsansätze]

- Begrenzte Rechnerkapazitäten
  [Hardwarebeschaffung]
- Zu viele potentielle erklärende Variablen
  [feature selection]
- Lernen des Modells und spätere Anwendung in verschiedenen Statistiken
  [Mustererkennung auf Basis einer nur kurzen, nicht immer ausreichenden Liste gemeinsamer Merkmale]
- Zu wenige der interessierenden Datenpunkte in den Trainingsdaten vorhanden
  [Methoden für imbalanced data]
- Kooperationen mit Partnern außerhalb des Statistischen Bundesamtes
  [Beschränkung auf Modelle, die keine Rückschlussmöglichkeiten auf Einzeldaten bieten]

Weitere Aspekte

- Einrichtung einer Informations- und Austauschplattform
- Bestandsaufnahme der Anwendung von Machine-Learning-Verfahren im Statistischen Bundesamt sowie in nationalen und internationalen Statistikinstitutionen
- Durchführung von hausinternen Informationsveranstaltungen
- Hausabfrage zu Anwendungsmöglichkeiten
- Abschlussbericht mit Handlungsempfehlungen

Abfragen zum Einsatz maschineller Lernverfahren

- Adressaten: 14 StLÄ, 18 weitere nationale und 39 internationale Statistikinstitutionen
- Ziel: Anwendungen von Machine Learning in der Statistik zu identifizieren. Ergebnis:
  - StLÄ: eine Anwendung
  - Destatis: 6 Anwendungen
  - National: 36 Anwendungen in 6 Statistikinstitutionen
  - International: 119 Anwendungen in 29 Institutionen
- Projekte, die maschinell Merkmale klassifizieren, Werte imputieren und Einheiten identifizieren und dazu Random Forests oder ähnliche baumbasierte Verfahren, Support Vector Machines oder Neuronale Netze einsetzen, sind zurzeit weit verbreitet. Anwendungsfälle können in fast allen Fachstatistiken gefunden werden. In der Regel werden Prozesse in der Statistikkonzeption, der Datengewinnung und -analyse sowie der Statistikverbreitung und -evaluation mit Machine-Learning-Verfahren unterstützt.

Hausumfrage zum Potenzial maschineller Lernverfahren

- Adressaten: 29 Gruppen;
  - 16 Fehlanzeigen (überwiegend keine Anwendungsmöglichkeiten)
  - 13 Gruppen
    - meldeten 31 Projektideen, darunter 6 in der Experimentier- bzw. Testphase,
    - die auf maschinelle Klassifikation von Merkmalen oder die Identifikation von Einheiten (Dubletten, Ausreißer) abzielen,
    - und benötigen (gruppen-)externe Expertise bei der eventuellen Umsetzung
- Zusammenfassend:
  - Viele vielversprechende Ansätze und Ideen für den Einsatz von Machine Learning
  - Engpass beim Aufbau bzw. der Bereitstellung von Expertise

Anwendungsmöglichkeiten im Statistischen Bundesamt

- Qualitätsverbesserung durch automatisierte Plausibilitätskontrollen
  - Ziel: Schnellere und effizientere Plausibilisierung der Rohdaten sowie Imputation (data editing and cleaning) mit Hilfe von Machine Learning (z.B. HoloClean)
  - Test in Verdienststatistik und Kostenstrukturerhebung
- Zuordnung von Klartext (z.B. Berufsangaben) zu Positionen einer Klassifikation
- Verbesserte Analysemöglichkeiten durch „Ergänzung“ neuer Merkmale (z.B. aus Angaben aus Vorperioden oder anderen Datenquellen)
- Abgrenzung von Berichtskreisen
- Identifikation von Ausreißern in den Merkmalen

Erkenntnisse

- Es gibt ein großes Potential für den Einsatz von maschinellen Lernverfahren in den Fachstatistiken.
- Machine Learning ist aber nicht omnipotent und nicht immer die geeignete Methode.
- Fehlschläge und Enttäuschungen sind möglich, aber ebenso deutliche Verbesserungen (in verschiedenen Dimensionen gemessen).
- Führungskräfte sowie Mitarbeiterinnen und Mitarbeiter sollten für neue Methoden und Veränderungen, die durch Machine Learning angestoßen werden, offen sein.
- Es besteht Bedarf an angemessener Infrastruktur (insb. IT) und (Weiter-)Qualifizierung.
- Es besteht Bedarf an Austausch von Wissen und Kooperation mit Hochschulen, Forschungs- und Statistikinstitutionen.
- Es besteht Bedarf an einem Kompetenzzentrum zu maschinellem Lernen im Bereich mathematisch-statistischer Methoden.

Machine-Learning-Methodik

- Im Bereich mathematisch-statistische Methoden angesiedelt
- Ziele
  - Evaluierung von Methoden des maschinellen Lernens für Plausibilisierung und Imputation (zzgl. Ergebnisvalidierung)
  - Beschaffung externen Wissens bzgl. Softwarelösungen, IT-Infrastruktur, aktuelle Forschungsstände, rechtliche und ethische Fragestellungen

Machine-Learning-Methodik

- Bereich Plausibilisierung / data editing
  - Beteiligung an der Untersuchung von HoloClean und möglicher Alternativen
  - Bislang hauptsächlich theoretische Betrachtungen
  - Schwerpunkt: Ausreißererkennung
- Bereich Imputation
  - Erste Studie zu verschiedenen ML-Methoden bei einfacher Imputation (k-NN, CART, SVM); Referenzwert: Predictive Mean Matching
  - Zwischenergebnis: Resultate differieren in Abhängigkeit von den Gütemaßen
  - Darüber hinaus: Theoretische Betrachtung der verschiedenen Verfahren

Proof of Concept automatisierte Plausibilisierung

- Ziel: Überprüfung der Einsatzbarkeit von spezialisierter Software zur automatisierten Plausibilisierung (Fehlererkennung und Imputation)
- Hintergrund:
  - Neue digitale Verdiensterhebung ab 2021 in Vorbereitung
  - Ungefähr 7 Millionen Datensätze pro Monat
  - Keine „händische“ Plausibilisierung in diesem Umfang möglich
- Aktuell: Test der Software „HoloClean“ (*framework for holistic data repairing driven by probabilistic inference*)
- Alternative: CANCEIS

Literatur

- Dumpert, F.; Beck, M. (2017): Einsatz von Machine-Learning-Verfahren in amtlichen Unternehmensstatistiken, in: AStA Wirtschafts- und Sozialstatistisches Archiv, Band 11, Heft 2, S. 83–106.
- Rekatsinas, T.; Chu, X.; Ilyas, I. F.; Ré, C. (2017): HoloClean: Holistic data repairs with probabilistic inference, in: Proceedings of the VLDB Endowment, Vol. 10, No. 11, S. 1190–1201.
- Riede, T.; Tümmler, T. ; Wondrak, S. (2018): Die Digitale Agenda des Statistischen Bundesamtes. WISTA Wirtschaft und Statistik, (1), 102–111.
- Sa, C. D.; Ilyas, I. F.; Kimelfeld, B.; Ré, C.; Rekatsinas, T. (2018): A formal framework for probabilistic unclean databases. arXiv, 1801.06750.
- Spies, L.; Lange, K. (2018): Implementation of artifcial intelligence and machine learning methods within the Federal Statistical Office of Germany. https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.44/2018/T4_Germany_LANGE_Paper.pdf.
- Thirumuruganathan, S.; Tang, N.; Ouzzani, M. (2018): Data curation with deep learning [vision]: Towards self driving data curation. arXiv, 1803.01384.

Martin Beck, Statistisches Bundesamt

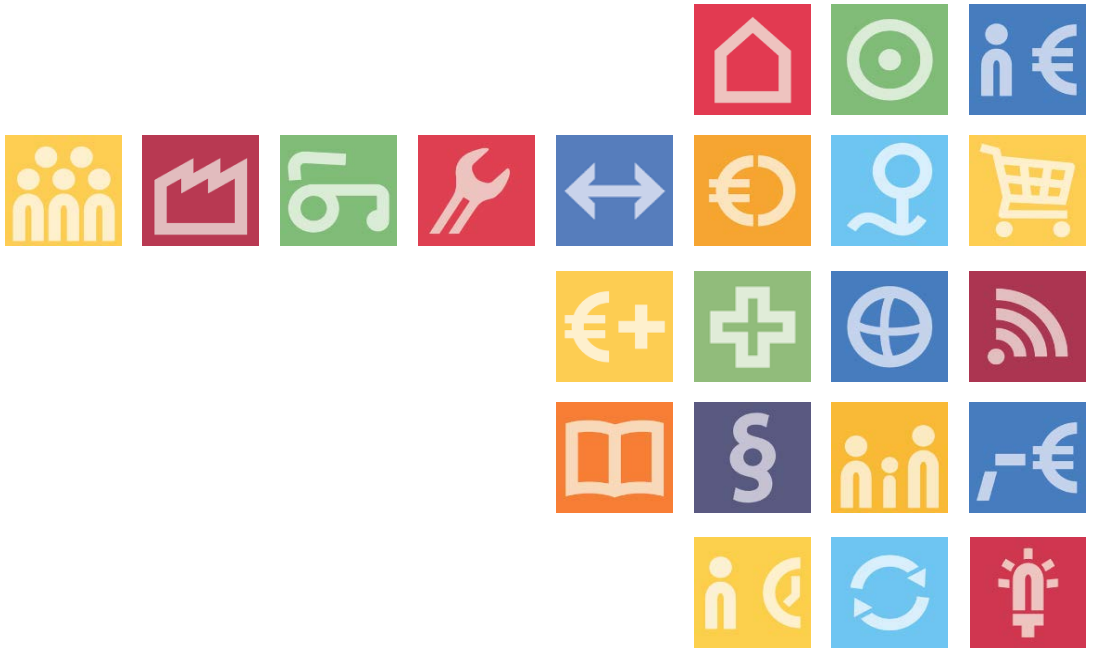
+49/(0) 611 / 75 44 60

martin.beck@destatis.de

Florian Dumpert, Universität Bayreuth

+49/(0) 921 / 55 32 74

florian.dumpert@uni-bayreuth.de



www.destatis.de