MACHINE LEARNING IN DER AMTLICHEN STATISTIK – ERGEBNISSE UND BEWERTUNG EINES INTERNATIONALEN PROJEKTS

Florian Dumpert

Schlüsselwörter: maschinelles Lernen – Klassifikation – Plausibilisierung – Imputation – Bildanalyse – Qualität

ZUSAMMENFASSUNG

Die High-Level-Group for the Modernization of Official Statistics der Wirtschaftskommission der Vereinten Nationen für Europa hatte für die Jahre 2019 und 2020 ein Machine-Learning-Projekt initiiert. Das Projekt ermöglichte die Durchführung von Pilotstudien, um den Mehrwert maschinellen Lernens in der amtlichen Statistik zu zeigen, sowie Arbeiten an einem Qualitätsrahmenwerk. Die Potenziale von maschinellem Lernen wurden dabei ebenso deutlich wie die Fallstricke und Hindernisse für dessen Einführung. Der Aufsatz beleuchtet die verschiedenen Arbeitsgebiete des Projekts und ordnet die Ergebnisse für die deutsche amtliche Statistik ein.

Keywords: *machine learning – classification – data editing – imputation – imagery – quality*

ABSTRACT

The machine learning project of the High-Level Group for the Modernization of Official Statistics of the United Nations Economic Commission for Europe facilitated pilot studies in 2019 and 2020 to demonstrate the added value of machine learning in official statistics, as well as work on a quality framework. The potential of machine learning became clear, as did the pitfalls and barriers to its adoption. The article highlights the different work areas of the project and discusses the results for German official statistics.

Dr. Florian Dumpert

ist als Referent im Statistischen Bundesamt mit Verfahren des maschinellen Lernens und der Imputation im Referat „Künstliche Intelligenz, Big Data“ befasst. Der Diplom-Mathematiker beschäftigt sich unter anderem mit methodischen Fragestellungen beim Einsatz dieser Verfahren und vertritt das Themengebiet in internationalen Projekten.

1

Einleitung

Das maschinelle Lernen (Machine Learning – ML) ist ein Teilgebiet des Forschungsfelds „Künstliche Intelligenz“ und gehört unbestritten zu den Zukunftstechnologien unserer Zeit. Davor kann und will sich die amtliche Statistik weltweit nicht verschließen. Einen frühen umfassenden Schritt dazu machte der Proof of Concept Machine Learning (Beck und andere, 2018). Methodisch am Puls der Zeit zu sein, ist ein Qualitätsaspekt amtlicher Statistik, der so auch Eingang in die Qualitätshandbücher fand. Beispielsweise schreibt das Qualitätshandbuch der Statistischen Ämter des Bundes und der Länder, dass „[d]ie statistischen Prozesse zur Erhebung, Aufbereitung und Verbreitung von Statistiken [...] internationalen Standards und Leitlinien in vollem Umfang genügen und zugleich dem aktuellen Stand der wissenschaftlichen Forschung entsprechen [sollen]. Dies gilt sowohl für die eingesetzte Methodik als auch für die angewendeten statistischen Verfahren.“ (Statistische Ämter des Bundes und der Länder, 2021, hier: Seite 19)

Um diesem Anspruch gerecht zu werden, sind Verfahren des maschinellen Lernens in den Werkzeugkasten der amtlichen Statistik als „neue“ statistische Methoden aufzunehmen. Wie immer bei der Einführung neuer Ansätze und Ideen ist – neben der internen Validierung – eine Orientierung erforderlich, einerseits an der Wissenschaft und dem aktuellen Forschungsstand, andererseits an vergleichbaren Einrichtungen, im vorliegenden Fall also an anderen nationalen statistischen Ämtern. Zum Austausch mit letzteren hat die High-Level-Group for the Modernization of Official Statistics (HLG-MOS) der Wirtschaftskommission der Vereinten Nationen für Europa (United Nations Economic Commission for Europe – UNECE) im Jahr 2019 ein Machine-Learning-Projekt ins Leben gerufen, das im folgenden Kapitel in seinen Grundzügen vorgestellt wird. Die weiteren Kapitel gehen auf verschiedene Spezialaspekte des Projekts ein; der Aufsatz schließt mit einer Zusammenfassung und einer Einordnung für das Statistische Bundesamt.

2

Das UNECE-HLG-MOS-Machine-Learning-Projekt

Die HLG-MOS beschreibt sich als „Gruppe engagierter Leitungen von nationalen statistischen Ämtern, die die Modernisierung statistischer Organisationen aktiv steuern“ (HLG-MOS, 2021); diese Gruppe hat das Machine-Learning-Projekt 2019 ins Leben gerufen. Die Herausforderung des Projekts bestand im Vorantreiben der Forschung, Entwicklung und Anwendung von Machine Learning in der amtlichen Statistik mit dem Ziel, dort einen Mehrwert zu schaffen.

Drei inhaltliche Arbeitspakete waren schließlich vorgesehen: Pilotstudien, Qualität und Integration. Das Arbeitspaket Pilotstudien umfasste dabei die Unterpunkte Klassifizierung und Kodierung, Plausibilisierung und Imputation sowie die Nutzung von Bilddaten (Imagery). Hinsichtlich seiner Durchführung gliederte sich das Projekt in zwei Arbeitsjahre (2019 und 2020). Während der Arbeitsjahre gab es gemeinsame (Präsenz-)Tagungen und Arbeiten in den Arbeitspaketen über Distanz. Insgesamt waren 124 Vertreterinnen und Vertreter aus 23 Staaten, 33 nationalen statistischen Ämtern und vier internationalen Organisationen vertreten.

Die Resultate des Projekts wurden in einer öffentlichen Online-Veranstaltung vorgestellt und diskutiert (UNECE, 2020a).

3

Klassifikation und Kodierung

Der Abschnitt „Classification and Coding“ (C&C) behandelte Fragestellungen zum Mehrwert von Machine Learning im Teilprozess 5.2 „Daten klassifizieren und kodieren“ des Geschäftsprozessmodells Amtliche Statistik (GMAS; Statistische Ämter des Bundes und der Länder, 2021). Beispielsweise werden in diesem Teilprozess Textantworten einem numerischen Kode zugeordnet. Die im Zuge des Projekts durchgeführten Pilotstudien befassten sich auch sämtlich mit dieser Fragestellung, allerdings in unterschiedlichen Kontexten: berufliche Tätigkeit (von Personen), wirtschaftliche Tätigkeit (von

Firmen), Unfallbeschreibung (bei Arbeitsunfällen), Produktbezeichnungen (bei Preisen), Stimmungserkennung (bei Twittereinträgen). Die zugehörigen Klassifikationsschemata führen in ihrer großen Mehrheit zu sogenannten Multi-Class-Problemen, das heißt zu Klassifikationen mit (meist deutlich) mehr als zwei möglichen Klassen. Lediglich die Klassifikation der Twittereinträge wurde als binäres Klassifikationsproblem behandelt (positive gegenüber negative Grundstimmung des Eintrags).

Im Unterschied zu bereits tabellarisch vorliegenden Daten bedürfen Textdaten vorbereitender Arbeiten, sogenanntem Preprocessing. Die Pilotstudien zeigten, dass ein nicht zu unterschätzender zeitlicher wie auch fachlicher Aufwand für diese vorbereitenden Arbeiten zu erbringen ist, um schließlich gute Klassifikationsergebnisse zu erzielen. Im Preprocessing wurden verschiedene Techniken eingesetzt und untersucht: Entfernung von Stop-Wörtern (zum Beispiel Artikel und Präpositionen), Stemming (das heißt die Reduktion von Wörtern auf ihren Wortstamm), Lemmatisierung (das heißt das Ersetzen von Wörtern durch über definierte Wörterbücher festgelegte Basisformen), das Ersetzen von Groß- durch Kleinbuchstaben, das Ersetzen von Umlauten und allgemein diakritischen Zeichen sowie die Bildung von n-Grammen auf Wort- und Zeichenebene, Bag-of-Words- und Bag-of-Features-Ansätze.[1]

Diese vorbereitenden Arbeiten bereiten die textlichen Eingaben unter anderem derart auf, dass sie in Tabellenform vorliegen. Prinzipiell kann nun jedes Klassifikationsverfahren eingesetzt werden, um die Zuordnung der Texte zu den Klassen des Klassifikationssystems vorzunehmen. In den Pilotstudien wurden hierfür Random Forests, Logistische Regression, Nächste Nachbarn, Naive Bayes, XGBoost, neuronale Netze (Multi-Layer-Perceptron, Convolutional Neural Network, Recurrent Neural Network) und Support Vector Machines (SVMs)[2] sowie direkt das Programm FastText eingesetzt.

Als Gütemaße zur Evaluation wurden die für die Klassifikation üblichen Größen (Gesamtgenauigkeit, Sensitivität und Vorhersagewert sowie Kombinationen daraus) herangezogen. Der Umstand, dass einige Klassen deutlich schwächer in den Daten vertreten waren als andere, dass also ein Imbalanced-Data-Problem vorlag, wurde dabei berücksichtigt.

Die Details zu den einzelnen Pilotstudien sind in Sthamer (2020) und den dort zitierten Pilotstudienberichten zu finden.

Der Mehrwert für die amtliche Statistik im Falle von Textklassifikation besteht vorrangig darin, langwierige, zeit- und arbeitsintensive, bislang manuell durchzuführende Prozessschritte zu automatisieren. Bei der Bemessung des Mehrwerts ist neben der schnelleren Bearbeitung auch die Qualität der Ergebnisse zu beurteilen. Die Klassifikationsgüten sollten dabei mit den Güten verglichen werden, die auf Basis der bisherigen manuellen oder algorithmischen Arbeiten erzielt werden. Dabei stellt sich die Frage nach der wahren Klassenzugehörigkeit einer Texteingabe. Untersuchungen im Rahmen einer Pilotstudie haben ergeben, dass auch händische Klassifikationen durch Fachleute nicht einheitlich und mithin fehleranfällig sind. Die Ergebnisse des ML-Verfahrens nur mit den händisch erzielten Ergebnissen zu vergleichen, kann die Güteschätzung daher verzerren. In der genannten Pilotstudie wurden dabei zunächst viel Zeit und Arbeitskraft dafür investiert, einen fehlerfreien Trainings- und (vor allem) einen fehlerfreien Testdatensatz zu erzeugen. Gegen diesen konnten dann die automatisierten und die händisch erzeugten Ergebnisse gehalten werden. Dieser Vergleich zeigte den Mehrwert des automatisierten Verfahrens auch in der Dimension Genauigkeit auf (Measure, 2020, und darin genannte Quellen).

Statt einer Vollautomatisierung wurden in einer anderen Pilotstudie die Ergebnisse des ML-Verfahrens nur als Vorschlag unterbreitet, der oder die zuständige Beschäftigte konnte diesen Vorschlag annehmen oder überschreiben. Eine Mischung aus beiden Ansätzen ist ebenfalls denkbar und auch wünschenswert: Ist das ML-Verfahren hinreichend sicher in der Klassifikation, so wird diese übernommen. Unterhalb eines Schwellenwerts jedoch werden die Texteingaben einer Sachbearbeiterin oder einem Sachbearbeiter zur gegebenenfalls maschinell assistierten manuellen Kodierung übergeben.

1 Für einen ersten, praxisorientierten Überblick zu Methoden der Textklassifikation siehe Lane und andere (2019).

2 Für Details zu den hier und im Weiteren beschriebenen Methoden siehe James und andere (2013) sowie Goodfellow und andere (2016).

Einordnung für die deutsche amtliche Statistik

Auch in der deutschen amtlichen Statistik stellt sich immer wieder – so auch in aktuellen Arbeiten – die Frage nach einer effizienten Bearbeitung von Freitexteingaben. Dabei können die Erkenntnisse aus den C&C-Pilotstudien des Machine-Learning-Projekts bestätigt und ergänzt werden:

- Gutes Datenmaterial zum Trainieren und Testen der ML-Verfahren ist essenziell, einerseits für die erfolgreiche Modellbildung und -testung, andererseits aber auch für den fairen Vergleich mit dem bisher eingesetzten (gegebenenfalls manuellen) Verfahren. Was „gutes Datenmaterial" bedeutet, ist dabei problemabhängig, bezieht sich jedoch grundsätzlich auf die in den Daten enthaltene Information. Je mehr unterschiedliche Klassen beispielsweise voneinander unterschieden werden sollen, desto mehr sich angemessen auf die Klassen verteilende Trainingsdaten werden benötigt.
- Zwar ist es häufig möglich, schnelle Ergebnisse auch ohne intensives Preprocessing zu erreichen, den Qualitätsanforderungen der amtlichen Statistik genügende Ergebnisse sind jedoch nur mit einem solchen zu erzielen.
- Neuronale Netze erscheinen im Hinblick auf die Textklassifikation anderen ML-Verfahren überlegen, stellen aber auch erhöhte Anforderungen an die Hardware, insbesondere in Form spezieller Grafikkarten. Steht diese nicht zur Verfügung, ist – qualitätseinschränkend – auf andere ML-Verfahren auszuweichen.
- Gerade bei der Arbeit mit internationalen Klassifikationssystemen ergeben sich drei bislang nicht geklärte methodische Fragen für die amtliche Statistik:
  - Ist es zielführend, die hierarchische Gliederung vieler dieser Systeme zu nutzen, beispielsweise um mehrstufig Modelle zu trainieren? Falls ja: Wie sollte dies idealerweise geschehen?
  - Wie ist sicherzustellen, dass auch Klassen, deren Anteil im Trainingsmaterial oder in der Grundgesamtheit im Verhältnis zu anderen Klassen sehr gering ist, adäquat vom ML-Verfahren berücksichtigt werden?
  - Wie kann es gelingen, einen zukunftssicheren Bestand an Trainingsdaten auch in deutscher Sprache zu konzipieren, aufzubauen und aktuell zu halten?

Hinsichtlich der – kontextabhängig zu beantwortenden – Frage, ob ein ML-Verfahren eher Vorschläge unterbreiten oder eher vollautomatisieren sollte, erscheint auch eine Dreiteilung denkbar: 1. Sicher klassifizierte Texteingaben werden vollautomatisiert verarbeitet. 2. „Mittelsicher" verarbeitete Eingaben werden automatisiert verarbeitet, verpflichtend jedoch stichprobenartigen Prüfungen größeren Ausmaßes durch Sachbearbeiterinnen und Sachbearbeiter unterzogen. 3. Unsicher klassifizierte Eingaben werden vollständig manuell klassifiziert. Für eine solche Dreiteilung ist es erforderlich, ein Maß für die Unsicherheit einer Klassifikation zu vereinbaren. Unabhängig davon sind im Nachgang und über die Perioden der Fachstatistik hinweg immer wieder stichprobenartige Überprüfungen aller (auch der sicher klassifizierten) Eingaben erforderlich, um

- die Qualität der Klassifikation zu sichern,
- das für künftiges Training herangezogene Datenmaterial zu validieren und
- Veränderungen im Meldeverhalten oder der Struktur der Eingaben im Allgemeinen (zum Beispiel bei Einzug neuer, noch nicht standardisierter Berufsbezeichnungen in den Sprachgebrauch) wahrzunehmen.

Die hier dargelegten Einordnungen lassen sich analog auch auf die Kapitel 4 und 5 übertragen.

4

Plausibilisierung und Imputation

Der zweite Untersuchungsbereich „Editing and Imputation" (E&I) beschäftigte sich mit dem Mehrwert von Machine Learning im Bereich von Plausibilisierung (PL) und Imputation, also vornehmlich mit dem GMAS-Teilprozess 5.4 „Daten plausibilisieren und imputieren", jedoch mit engem Bezug zum GMAS-Teilprozess 5.3 „Daten prüfen und validieren". Dabei waren verschiedene Konstellationen denkbar: Machine Learning könnte bestehende Verfahren ersetzen, verbessern oder ergänzen.

Die am Projekt Beteiligten verständigten sich darauf, die Untersuchungen auf zwei Fälle zu fokussieren: einerseits das (nicht rein regelbasierte) Erkennen fehlerhafter oder als solche verdächtiger Werte, andererseits das Abändern dieser und das Ersetzen fehlender Werte mit durchaus unterschiedlichen Zielen. Wie schon bei Klassifikation und Kodierung ist ein Ziel von Machine Learning im E&I-Prozess, bisher manuell durchgeführte Arbeiten zu unterstützen oder teilweise zu automatisieren und damit zu genaueren oder schnelleren Ergebnissen beizutragen.

Im Folgenden werden die Ergebnisse der Pilotstudien zusammengefasst, Details finden sich bei Dumpert (2020) und den dort zitierten Pilotstudienberichten. Traditionell werden regelbasierte Vergleiche von beobachteten Werten mit (schwachen oder starken) Plausibilitätsbedingungen, Verteilungsuntersuchungen (zum Beispiel zur Ausreißererkennung) und Vergleiche mit externen und/oder früheren Datensätzen angewendet. Dabei kommen mehrere Schritte zum Tragen, bei denen sowohl die Automatisierung (durch Plausibilitätsregeln) als auch die Fachleute (durch interaktives Bearbeiten) eine wichtige Rolle bei der Erkennung problematischer Daten spielen. Eine vollständige Automatisierung sollte nicht das unbedingte Ziel des Einsatzes von Machine Learning bei der Plausibilisierung sein. Ebenso wie beim Klassifizieren und Kodieren ist eine zumindest stichprobenartige Überprüfung (auch von vermeintlich sehr zuverlässig erkannten Fehlern im Datenbestand) angeraten. Die Pilotstudien zur Plausibilisierung ergaben, dass ein Lernen von Mustern aus früheren Fehlererkennungsrunden möglich ist. Liegen also Trainingsdaten vor, die plausible und unplausible Werte, die als solche auch gekennzeichnet sind, enthalten, so kann daraus ein ML-Verfahren ein Modell erlernen und neu ankommende Daten hinsichtlich der Plausibilität klassifizieren. Eine daraus prinzipiell mögliche Ableitung von PL-Regeln leidet aber daran, dass eine gute Performanz im Sinne einer guten Detektion fehlerhafter Werte nur möglich ist, wenn die Regeln sehr komplex sind, also sehr viele Fallunterscheidungen enthalten. Solche Regeln wiederum bieten keinen unbedingten Mehrwert. Die Durchführung der Fehlererkennung wurde in der Pilotstudie durch Einsatz des ML-Verfahrens deutlich beschleunigt und lieferte im Vergleich zur manuellen Fehlerdetektion konsistentere Resultate.

Hinsichtlich der Imputation lieferten die Pilotstudien die Einsicht, dass Machine Learning häufig, aber nicht immer plausible Einsetzungen vornimmt. Auch kann es durch das Überspringen einiger Preprocessing-Schritte (zum Beispiel Berücksichtigung von Korrelationen, statistische Transformationen von Variablen und so weiter) zu einer schnelleren Statistikproduktion beitragen. Dennoch erfordert maschinelles Lernen viele einschlägige Simulationen und mithin eine genaue Beschäftigung mit den Daten. Außerdem erlaubt Machine Learning, die Anzahl menschlicher Eingriffe zu reduzieren (wenn es beispielsweise die Variablenauswahl für das Imputationsmodell automatisch vornimmt).

Zum Einsatz kamen beim Plausibilisieren und Imputieren von Daten die folgenden ML-Verfahren: Klassifikations- und Regressionsbäume, Random Forests, neuronale Netze, (regularisierte) lineare Modelle, Nächste Nachbarn, bayesianische Netze, SVMs und Kombinationen daraus.

Einordnung für die deutsche amtliche Statistik

Es gibt kaum Möglichkeiten, schnelle Erfolge mit dem Machine Learning im Bereich Plausibilisierung und Imputation zu erzielen; die sorgfältige methodische Arbeit ist hier von besonders großer Bedeutung. Ein anderer, bei der Diskussion des Projekts aufgeworfener, aus Zeitgründen jedoch bislang nicht näher untersuchter Aspekt der Plausibilisierung ist die Frage, wie die Sachbearbeiterinnen und Sachbearbeiter bei manueller Plausibilisierung zu ihren Einsetzungen kommen. Das mag in manchen Fällen offensichtlich sein, in anderen Fällen wiederum stammt das nötige Wissen aus Zeitungsartikeln, Branchenreports, von Unternehmenswebsites und ähnlichen Quellen. Die Unterstützung der manuellen Plausibilisierung mithilfe solcher Informationen setzt einerseits ihre Beschaffung, andererseits ihre Verarbeitung voraus. Hier kann Machine Learning gegebenenfalls hilfreich sein.

5

Bildanalyse

Mit zunehmender Verfügbarkeit von Fernerkundungsdaten stellt sich auch für die amtliche Statistik die Frage nach deren Nutzung. In diesem Zusammenhang auftretende Fragen und Probleme wurden in den Pilotstudien zur Nutzung von Bilddaten (Imagery) erörtert. Ähnlich wie bei den Untersuchungsbereichen Klassifikation und Kodierung sowie Plausibilisierung und Imputation ist ein Ziel von Machine Learning im Kontext von Fernerkundungsdaten eine automatisierte Verarbeitung der in den Daten enthaltenen Informationen. Um welche Informationen es sich handelt, ist kontextabhängig. In den Pilotstudien handelte es sich um die Nutzung landwirtschaftlicher Flächen, um die Erfassung von Armut oder von Siedlungswachstum oder auch um ganz konkrete Fragestellungen im Zuge der Aktualisierung von Adressregistern. Die Arbeiten sind den GMAS-Teilprozessen 5.2 „Daten klassifizieren und kodieren“ und 5.5 „Neue Merkmale und Einheiten ableiten“ zuzuordnen.

Die in den verschiedenen Pilotstudien genutzten Fernerkundungsdaten hatten Auflösungen von 23 cm bis 30 m. Der Prozess, die Daten zu labeln, das heißt dem Bild die entsprechende Klassenzugehörigkeit anzufügen und somit Trainings- und Testdaten erst zu generieren, war jeweils zeit- und arbeitsintensiv. Es standen in der Regel auch noch keine vorgelabelten Bilddaten zur Verfügung, sodass nicht auf bereits vorhandene Bestände aufgebaut werden konnte. Bei der Erstellung der Trainings- und Testdaten war zu beachten, dass verschiedene Situationen (Stadt/Land, gutes Wetter/schlechtes Wetter, Regionalspezifika in Bau und landwirtschaftlicher Nutzung und Ähnliches) hinreichend repräsentiert sind. Die eingesetzte Anzahl der gelabelten Bilder betrug dabei das 1 000- bis 20 000-Fache der unterschiedenen Klassen. Zusätzlich zu den reinen unmittelbar im Bild enthaltenen Daten wurden noch weitere Informationen als erklärende Variablen eingesetzt, beispielsweise die Geokoordinaten des aufgenommenen Objekts oder dessen Lage über dem Meeresspiegel.

Wichtige Schritte in der Datenaufbereitung umfassen neben der technischen Anpassung zur weiteren Bearbeitung auch die Variation des vorhandenen Materials (Drehungen, Spiegelungen, Skalierungen), um einerseits mehr Trainingsmaterial zu erhalten, andererseits aber auch der Gefahr einer Überanpassung an das Originalmaterial entgegenzuwirken.

Außerdem spielt die Variablenselektion bei der Bildanalyse eine herausgehobene Rolle, um die Bilder mit der vorhandenen Hardware in vertretbarer Laufzeit im Lernprozess auswertbar zu machen.

Als ML-Verfahren kamen Convolutional Neural Networks sowie Random Forests, SVMs und Extreme Randomized Trees auf gewöhnlichen Rechenkernen und Grafikkarten zum Einsatz. Sofern auf der vorhandenen Hardware berechenbar, erwiesen sich meist die Convolutional Neural Networks als überlegene Verfahren.

Alle durchgeführten Pilotstudien wurden positiv evaluiert und sind zum Teil bereits in die Produktion überführt. Der Mehrwert von Machine Learning bestand letztlich darin, dass die Fernerkundungsdaten überhaupt nutzbar gemacht wurden und somit die Statistikproduktion bereichern konnten, sei es durch zusätzliche Informationen oder die Entlastung der Sachbearbeiterinnen und Sachbearbeiter in sicher klassifizierten Fällen (Coronado/Juárez, 2020).

Einordnung für die deutsche amtliche Statistik

Die deutsche amtliche Statistik greift die Idee, Fernerkundungsdaten für die amtliche Statistikproduktion zu nutzen, immer wieder auf (Arnold/Kleine, 2017; Statistisches Bundesamt, 2019). Auch im Zuge der Einführung des Registerzensus (Modul Gebäude und Wohnungen) sollen gemeinsam mit Kooperationspartnern Verfahren entwickelt werden, mit denen Fernerkundungsdaten zur Validierung von Daten aus Registern eingesetzt werden können. Gemeinsam mit den im Projekt durchgeführten Pilotstudien werden Potenzial und praktischer Nutzen von Fernerkundungsdaten offensichtlich. Neben hinreichend verlässlichen und möglichst alle auch künftigen Aufnahmen abdeckenden Trainings- und Testdaten ist aber insbesondere für die Bildverarbeitung eine ausreichend dimensionierte IT-Ausstattung erforderlich.

6

Qualität

Neben Pilotstudien, die den Mehrwert von Machine Learning in der amtlichen Statistik prinzipiell untersuchen sollten, befasste sich ein weiteres Arbeitspaket des Machine-Learning-Projekts mit Fragen der Qualität. Die Arbeiten in diesem Arbeitspaket orientierten sich einerseits an bestehenden Qualitätsrahmen der amtlichen Statistik im Allgemeinen, beispielsweise dem European Statistics Code of Practice (Eurostat, 2017), präzisierten und ergänzten diese. Andererseits wurde nicht ausschließlich Machine Learning betrachtet, sondern ein Qualitätsrahmen für alle statistischen Algorithmen, alle statistischen (maschinellen Lern)Verfahren, entworfen, eine Mischung aus Handlungsanweisung und Rahmengebung. Der Qualitätsrahmen „Quality Framework for Statistical Algorithms (QF4SA)“ (Yung und andere, 2020) umfasst fünf Dimensionen.

Ein statistischer Algorithmus (als Umschreibung für statistische Verfahren, seien sie „klassisch“ oder im Sinne des maschinellen Lernens eingesetzt oder verstanden) hat in diesen fünf Dimensionen die Tauglichkeit für den Einsatz in der amtlichen Statistik unter Beweis zu stellen:

1. Erklärbarkeit ist definiert als die Möglichkeit zu verstehen, welche Zusammenhänge der Algorithmus nutzt, um Vorhersagen zu treffen oder Analysen auszuweisen. Es geht also darum, den gegebenenfalls nur lokalen Zusammenhang zwischen Eingabe- und Ausgabevariablen (sofern vorhanden) darlegen zu können.
2. Genauigkeit ist definiert als der Grad eines statistischen Outputs, zu dem dieser in der Lage ist, das zu messende Phänomen korrekt zu beschreiben. Es geht also letztlich um den geeignet zu messenden Abstand zwischen Schätzung und wahrem Wert.
3. Reproduzierbarkeit ist im Basisniveau definiert als die Fähigkeit, gleiche Ergebnisse zu erzielen, sofern mit den gleichen Daten und dem gleichen Algorithmus gearbeitet wird. Im höheren Niveau ist darunter zu verstehen, dass verschiedene Zufallsstichproben aus einer Grundgesamtheit bei gleichem Algorithmus regelmäßig zu im Wesentlichen gleichen Ergebnissen führen sollen.
4. Rechtzeitigkeit ist definiert als die Fähigkeit, innerhalb der geforderten Zeit den Algorithmus konzipieren, trainieren und anwenden zu können. Dabei können gegebenenfalls die Rechtzeitigkeit von Konzeption und Training und die Rechtzeitigkeit der Anwendung getrennt voneinander betrachtet werden.
5. Wirtschaftlichkeit schließlich ist definiert als Genauigkeit oder Schnelligkeit oder Erklärbarkeit je Kosteneinheit. Dabei enthalten die Kosten eines Algorithmus fixe Kosten (zum Beispiel für Personal und IT-Infrastruktur) und variable Kosten (zum Beispiel für Schulungsmaßnahmen, das Monitoring des Algorithmus oder gegebenenfalls notwendiges Nachtrainieren).

Einordnung für die deutsche amtliche Statistik

Die qualitätsgesicherte Einführung von Machine Learning ist eine wichtige Aufgabe für das Statistische Bundesamt und den Statistischen Verbund[3]. Der Begriff der Qualität von Machine Learning ist valide zu fassen und weiterzuentwickeln. Dazu werden – basierend auf Vorwissen und Vorerfahrungen – im Austausch im Statistischen Verbund und im Kontakt mit der Wissenschaft bereits verschiedene Anstrengungen unternommen. Der Qualitätsrahmen „Quality Framework for Statistical Algorithms“ wird zu dieser Debatte in der deutschen amtlichen Statistik einen wichtigen Beitrag leisten.

7

Integration

In diesem Arbeitspaket ging es um die Frage der Integration von Machine Learning in die amtliche Statistik. Schwierigkeiten konnten in allen beteiligten nationalen statistischen Ämtern beobachtet werden. Es wurde herausgearbeitet, dass auf organisatorischer Seite die Koordination zwischen den internen Stakeholdern sowie Widerstände interner Stakeholder die größten Hindernisse für die Einführung von Machine Learning darstellen, gefolgt von Unklarheit über Verantwortlich-

3 Den Statistischen Verbund bilden die Statistischen Ämter des Bundes und der Länder.

keiten in einem Projekt. Auf technischer Seite bestehen Schwierigkeiten darin, geeignetes Personal für diese Aufgaben zu finden, geeignete Hardware nutzen zu können sowie geeignetes Trainings- und Testmaterial zur Verfügung zu haben. Als förderlich wurden hingegen die Zusammenarbeit mit anderen statistischen Ämtern, die Kooperation mit der Wissenschaft, interne und externe Schulungsmaßnahmen sowie die klare Festlegung von Rollen und Verantwortlichkeiten in den Ämtern genannt.

Auf die Frage, wo Machine Learning organisatorisch verortet ist, gab es keine einheitliche Antwort. Einige statistische Ämter sehen Machine Learning im Bereich der Methodik, andere in eigenen Exzellenzzentren. In weiteren Fällen tritt Machine Learning rein dezentral oder explizit multidisziplinär in Erscheinung.

Bei der Frage schließlich, wie sich der Mehrwert von Machine Learning in der amtlichen Statistik zeigen kann, wurden drei Aspekte herausgearbeitet:

1. Es bedarf eines fairen Vergleichs zwischen bisherigem Vorgehen und Machine Learning – fair hinsichtlich der Testdaten, des Arbeits- und Zeitaufwands und der statistischen Qualitätsindikatoren.
2. Machine Learning muss Bestehendes nicht vollständig ersetzen. Auch die Unterstützung oder Ergänzung der bisherigen Vorgehensweise stellt einen Mehrwert dar.
3. Der Mehrwert von Machine Learning zeigt sich nicht zuletzt beim Einsatz für Aufgaben, die sonst gar nicht zu bearbeiten wären. Darunter fallen beispielsweise Aufgaben im Bereich der Bilderkennung.

Einordnung für die deutsche amtliche Statistik

Die aufgezeigten Schwierigkeiten sind auch in Deutschland bekannt. Engpässe zeigen sich regelmäßig bei gutem Trainings- und Testdatenmaterial sowie bei der vorhandenen IT-Infrastruktur. Zwar stehen für die Programmiersprachen R und Python Serverlösungen zur Verfügung, ein gewinnbringender Einsatz von neuronalen Netzen ist auf diesen aber aktuell nicht möglich. Darüber hinaus ist bereits jetzt zu beobachten, dass die Kapazitäten der vorhandenen Hardware häufig nicht ausreichen, um mit – für Machine Learning üblichen – großen Datensätzen valide methodische Untersuchungen durchzuführen. Diese sind jedoch unbedingte Voraussetzung für einen qualitätsgesicherten Einsatz von Machine Learning im späteren Produktivbetrieb.

Hinsichtlich der im Projekt genannten begünstigenden Faktoren ist das Statistische Bundesamt auf einem guten Weg. Es gibt Angebote für Schulungsveranstaltungen zu Machine Learning und die Kooperation mit der Wissenschaft in diesem Bereich wird etabliert. Auch der Austausch mit anderen statistischen Ämtern ist über internationale Projekte (wie diesem) und über den Arbeitskreis Maschinelles Lernen (für den Statistischen Verbund) gewährleistet.

8

Zusammenfassung des Projekts

Als konsolidierte Fassung und Bewertung der Teilberichte legt der Abschlussbericht (Julien, 2020) die wesentlichen Aspekte dar, die für die Akzeptanz von Machine Learning in der amtlichen Statistik erforderlich sind: Der Einsatz von Machine Learning muss

- sich an den Bedürfnissen der Ämter als Datenproduzenten ausrichten,
- durch ein Qualitätsrahmenwerk begleitet werden,
- faktischen Mehrwert bieten,
- über die Zeit stabile Resultate liefern,
- im Einklang mit ethischen und rechtlichen Vorgaben eingesetzt und
- auf Basis wissenschaftlicher Erkenntnisse ausgewählt und weiterentwickelt werden.

Um die Einführung von Machine Learning in die Statistikproduktion zu ermöglichen, bedarf es der engen Zusammenarbeit von Fachstatistikerinnen und Fachstatistikern, Methodikerinnen und Methodikern und (spätestens beim Übergang in den Produktivbetrieb) IT-Expertinnen und -Experten. Notwendig sind darüber hinaus eine leistungsfähige IT-Infrastruktur, Forschung und Entwicklung, der Austausch mit anderen statistischen Ämtern, die Unterstützung der Führungsebenen sowie letztlich die Unterstützung durch alle Beschäftigten.

Die im dargestellten Machine-Learning-Projekt durchgeführten Arbeiten warfen weitere Fragen auf und verdeutlichten die Notwendigkeit des weiteren internationalen Austauschs. Das Nachfolgeprojekt „Machine Learning Group 2021“ (UNECE, 2021) setzt sich mit den an den aufgekommenen Fragen orientierenden Arbeitspaketen „Von der Idee zur Lösung“, „Von der Lösung zur Produktion“, „Datenethik und -governance“, „Qualität von Trainingsdaten“ und „QF4SA – Qualitätsrahmen für statistische Algorithmen“ auseinander.

9

Einordnung für die Bundesstatistik

In der Gesamtschau lässt sich festhalten, dass die Bundesstatistik den internationalen Vergleich nicht scheuen muss. Digitalabteilung und Fachbereiche arbeiten zusammen, um Bedarfe für den Einsatz von Machine Learning zu identifizieren, konkrete Projekte werden gemeinsam geprüft, durchgeführt und evaluiert. Einschlägige Fortbildungen finden statt, die Zusammenarbeit mit der Wissenschaft wird ausgebaut. Hinsichtlich der eingesetzten Verfahren folgt die Bundesstatistik – im Rahmen der technischen Möglichkeiten – der aktuellen Forschung.

Die IT-Infrastruktur stellt – hier wie in Ämtern anderer Staaten – sowohl in der Entwicklungs- und Testphase als auch im späteren Produktivbetrieb regelmäßig einen Engpass dar. Das bezieht sich einerseits auf die reine Quantität (Arbeitsspeicher, Rechenkerne), andererseits auch auf die Qualität (für neuronale Netze sind Grafikkarten das Mittel der Wahl). Dass auch bei diesem Engpass der Vergleich nicht gescheut werden muss, liegt weniger an der überdurchschnittlichen Ausstattung im Statistischen Verbund als vielmehr an einem weltweiten Niveau, das unter seinen Möglichkeiten bleibt. Zuletzt steht und fällt der Erfolg von Machine Learning mit dem Vorhandensein qualitativ hochwertiger Trainings- und Testdaten. Diese stehen in der benötigten Form nicht immer direkt aus den bestehenden Produktionsprozessen zur Verfügung, sondern müssen in der Regel noch aufbereitet werden. Besonders bei Fragen der Zuordnung von statistischen Einheiten zu einer Klasse (Unternehmen zu wirtschaftlicher Tätigkeit, Luftbild zu Bodennutzung und so weiter) ist es entscheidend, dass die Trainings- und Testdaten korrekt gelabelt, also richtig klassifiziert sind. Dies geschieht idealerweise nicht nur durch eine einzige Beschäftigte oder einen einzigen Beschäftigten. Außerdem muss sichergestellt sein, dass die Trainings- und Testdaten die künftig zu klassifizierenden Daten gut widerspiegeln. Diese Aufbereitung und Qualitätssicherung benötigt Zeit und Arbeitskraft; ohne sie ist ein Projekt nicht zu realisieren. Gleiches gilt für den Fall, dass notwendige methodische Untersuchungen oder Weiterentwicklungen nicht durchgeführt oder vorangetrieben werden können.

Das qualitätsgesicherte Etablieren maschinellen Lernens in den Prozessen der amtlichen Statistikproduktion ist somit keine einmalige, sondern vielmehr eine fortlaufende Aufgabe. Sie ist stets an die aktuellen wissenschaftlichen Entwicklungen, die vorhandenen IT-Kapazitäten und die individuellen Bedürfnisse der Fachbereiche anzupassen.

LITERATURVERZEICHNIS

Arnold, Stephan/Kleine, Sarah. *Neue Wege der Geodatennutzung: Perspektiven der Fernerkundung für die Statistik*. In: WISTA Wirtschaft und Statistik. Ausgabe 5/2017, Seite 31 ff.

Beck, Martin/Dumpert, Florian/Feuerhake, Jörg. *Proof of Concept Machine Learning: Abschlussbericht*. 2018. [Zugriff am 30. Juni 2021]. Verfügbar unter: www.statistischebibliothek.de

Coronado, Abel/Juárez, Jimena. *UNECE-HLG-MOS Machine Learning Project Imagery Theme Report*. 2020. [Zugriff am 30. Juni 2021]. Verfügbar unter: statswiki.unece.org

Dumpert, Florian. *UNECE-HLG-MOS Machine Learning Project Edit and Imputation Theme Report*. 2020. [Zugriff am 30. Juni 2021]. Verfügbar unter: statswiki.unece.org

Eurostat. *European statistics code of practice*. 2017. [Zugriff am 30. Juni 2021]. Verfügbar unter: ec.europa.eu

Goodfellow, Ian/Bengio, Yoshua/Courville, Aaron. *Deep Learning*. Cambridge 2016.

HLG MOS. *HLG-MOS Strategy*. 2021. [Zugriff am 30. Juni 2021]. Verfügbar unter: statswiki.unece.org

James, Gareth/Witten, Daniela/Hastie, Trevor/Tibshirani, Robert. *An Introduction to Statistical Learning*. New York 2013.

Julien, Claude. *Machine Learning Project Report*. 2020. [Zugriff am 30. Juni 2021]. Verfügbar unter: statswiki.unece.org

Lane, Hobson/Howard, Cole/Hapke, Hannes Max. *Natural Language Processing in Action*. Shelter Island 2019.

Measure, Alexander. *Automatic classification of work-related injury and illness narratives*. 2020. [Zugriff am 30. Juni 2021]. Verfügbar unter: statswiki.unece.org

Statistische Ämter des Bundes und der Länder. *Qualitätshandbuch der Statistischen Ämter des Bundes und der Länder*. 2021. [Zugriff am 30. Juni 2021]. Verfügbar unter: www.destatis.de

Statistisches Bundesamt. *Nutzen von Satellitendaten für die amtliche Statistik*. 2019. [Zugriff am 30. Juni 2021]. Verfügbar unter: www.destatis.de

Sthamer, Claus. *UNECE-HLG-MOS Machine Learning Project Classification and Coding Theme Report*. 2020. [Zugriff am 30. Juni 2021]. Verfügbar unter: statswiki.unece.org

UNECE. *HLG-MOS ML Project webinar*. 2020a. [Zugriff am 30. Juni 2021]. Verfügbar unter: statswiki.unece.org

UNECE. *HLG-MOS Machine Learning Project*. 2020b. [Zugriff am 30. Juni 2021]. Verfügbar unter: statswiki.unece.orgt

LITERATURVERZEICHNIS

UNECE. *Machine Learning Group 2021*. 2021. [Zugriff am 30. Juni 2021]. Verfügbar unter: statswiki.unece.org

Yung, Wesley/Tam, Siu-Ming/Buelens, Bart/Dumpert, Florian/Ascari, Gabriele/Rocci, Fabiana/Burger, Joep/Chipman, Hugh/Choi, InKyung. *A Quality Framework for Statistical Algorithms*. 2020. [Zugriff am 30. Juni 2021]. Verfügbar unter: statswiki.unece.org

Herausgeber
Statistisches Bundesamt (Destatis), Wiesbaden

Schriftleitung
Dr. Daniel Vorgrimler
Redaktion: Ellen Römer

Ihr Kontakt zu uns
www.destatis.de/kontakt

Erscheinungsfolge
zweimonatlich, erschienen im August 2021
Ältere Ausgaben finden Sie unter www.destatis.de sowie in der Statistischen Bibliothek.

Artikelnummer: 1010200-21004-4, ISSN 1619-2907