MEHR ZAHLEN, BESSERE ENTSCHEIDUNGEN?

Neue digitale Daten und Methoden in der empirischen Analyse und Beratung

27. Wissenschaftliches Kolloquium

gemeinsam mit der Deutschen Statistischen Gesellschaft am 22. und 23. November 2018 in Wiesbaden

Kurzfassung:
Machine Learning in der amtlichen Statistik

Martin Beck

hat Wirtschaftswissenschaften in Gießen studiert. Nach Stationen in der Sozial- und Bildungsstatistik ist er seit 2007 im Statistischen Bundesamt als Gruppenleiter für die Verdienststatistiken und seit 2010 darüber hinaus für das Unternehmensregister, die Klassifikationen und wirtschaftsbereichsübergreifende Unternehmensstatistiken zuständig. Er befasst sich u. a. mit der effizienteren Gestaltung der Datengewinnung und -analyse durch die Einführung neuer Methoden.

Florian Dumpert

ist wissenschaftlicher Mitarbeiter am Lehrstuhl für Stochastik der Universität Bayreuth. Der Diplom-Mathematiker forscht im Bereich maschineller statistischer Lernverfahren, insbesondere Support Vector Machines.

Machine Learning und Digitalisierung – diese Schlagworte sind aktuell aus Statistik und Datenanalyse kaum wegzudenken. Auch der amtlichen Statistik bietet die Digitalisierung vielfältige Möglichkeiten, ihre Aufgaben noch besser, schneller, flexibler und effizienter zu erfüllen. Das Statistische Bundesamt hat dazu Anfang 2018 eine Digitale Agenda mit 59 Maßnahmen aufgelegt, darunter auch das Leuchtturmprojekt „Proof of Concept Machine Learning“.

Das sogenannte statistische maschinelle Lernen zeichnet sich dadurch aus, dass auf Grundlage endlich vieler Beobachtungen ein Zusammenhang zwischen Eingabevariablen und Ausgabevariable erlernt wird, der anschließend auf neue, ggf. noch nicht bekannte Eingabewerte angewendet werden kann, um den unbekannten Ausgabewert zu schätzen. Im Unterschied zu Methoden der klassischen Statistik steht dabei die Prädiktion im Vordergrund; die Bedeutung der Erklärung, welche Faktoren wie Einfluss auf die Ausprägung der zu erklärenden Variable nehmen, rückt in den Hintergrund. Das Machine Learning kann die klassische Statistik also nicht ersetzen, sondern nur mit seinen Stärken und nur in geeigneten Fragestellungen ergänzen. Hinsichtlich des grundsätzlichen Vorgehens unterscheidet man zwischen überwachtem (supervised learning) und unüberwachtem (unsupervised learning) maschinellen Lernen. Das überwachte Lernen wird für Klassifikation und Regression eingesetzt und ist im Falle des maschinellen Lernens i. d. R. nichtparametrisch. Benötigt werden geeignete Trainings- und Testdaten, für die das wahre Ergebnis der Klassifikation (bzw. der Regression) bekannt ist und deren Strukturen durch die Methode erlernt werden. Unüberwachtes Lernen wird beispielsweise für Ausreißeridentifikation oder Clustering verwendet. Gängige, auch in der Praxis statistischer Institutionen verbreitete Methoden des maschinellen Lernens sind Support Vector Machines, Random Forests und Neuronale Netze.

Der „Proof of Concept Machine Learning“ baut auf bereits durchgeführten Projekten in der amtlichen Unternehmensstatistik auf und wurde Mitte des Jahres 2018 abgeschlossen. Ziel war es, die Einsetzbarkeit von maschinellem Lernen in den Prozessen der Fachstatistiken zu untersuchen. Ein erster Schritt bestand darin, sich einen Überblick zu verschaffen, welche Methoden des maschinellen Lernens in nationalen und internationalen Statistikinstitutionen bereits eingesetzt wurden und welche Verwendungszwecke dabei im Fokus standen. In die Abfrage wurden die 14 Statistischen Landesämter, weitere 18 nationale Statistikorganisationen sowie 39 internationale Statistikinstitutionen, insbesondere aller Mitgliedstaaten der Europäischen Union, einbezogen. Danach wurde überprüft, ob und auf welche Aufgabenstellungen des Statistischen Bundesamtes diese Erfahrungen übertragen werden können. In dem Beitrag werden die Vorgehensweise bei der Durchführung des „Proof of Concept Machine Learning“, dessen Ergebnisse sowie die Konsequenzen für die fachstatistischen Arbeiten der amtlichen Statistik, insbesondere in Datenerhebung und -aufbereitung, sowie die abgeleiteten Handlungsempfehlungen für das Statistische Bundesamt vorgestellt.