Statistisches Bundesamt • 65180 Wiesbaden • Deutschland

Technische Universität Dortmund
Fakultät für Statistik

Bearbeiter/-in: Florian Dumpert
Telefon: +49 (0)611 / 75-3887
Telefax: +49 (0)611 / 75-4000
florian.dumpert@destatis.de

Geschäftszeichen: C1/2001-ML

Wiesbaden, 17.12.19
Seitenanzahl: 3

Wissenschaftliche Abschlussarbeiten (ggf. in Verbindung mit einem Praktikum)

in Zusammenarbeit mit dem Lehrstuhl für Mathematische Statistik und industrielle Anwendungen (Prof. Dr. Markus Pauly)

Sehr geehrte Damen und Herren, liebe Studentinnen und Studenten,

haben Sie Interesse, Ihre Masterarbeit mit einem aktuellen Forschungsthema aus dem Bereich der amtlichen Statistik zu verfassen? Das Referat für Maschinelles Lernen und Imputationsverfahren des Statistischen Bundesamtes bietet hierzu gemeinsam mit Prof. Dr. Markus Pauly die Gelegenheit mit folgenden Themen:

1. Optionen zur Bemessung des Abstandes zweier Verteilungen in der Praxis

Vom theoretischen Standpunkt aus betrachtet ist es möglich, Metriken auf Räumen von Wahrscheinlichkeitsmaßen zu definieren. Solche Metriken werden für theoretische Beweisführungen benötigt. In der Praxis muss jedoch auf Basis endlich vieler Beobachtungen beurteilt werden, ob zwei Verteilungen übereinstimmen oder nicht. Im Eindimensionalen stehen hierfür Tests zur Verfügung, z. B. der Kolmogorov-Smirnov-Test. In höheren Dimensionen sind die Möglichkeiten deutlich begrenzter; Ansätze sind jedoch vorhanden, z. B. [1], [2].

Die Frage, ob zwei (empirische) Verteilungen übereinstimmen, tritt in der amtlichen Statistik immer wieder auf: Im Bereich des maschinellen Lernens z. B. wenn sichergestellt werden soll, dass Trainings- und Testdaten die gleiche Verteilung aufweisen, oder auch, um zu klären, ob ein früher gelerntes Modell heute noch anwendbar ist (also ob der datenerzeugende Prozess unverändert ist). Auch kann auf diese Weise in Simulationen geprüft werden, ob eine in Unkenntnis der späteren Inferenz erfolgte Imputation erfolgreich war.

Ziel der Arbeit ist die Auseinandersetzung mit der einschlägigen Literatur. Bewertet werden soll, ob verschiedene dort vorgeschlagene Verfahren für die oben genannten Fragestellungen geeignet sind. Die Arbeit soll eine Simulationsstudie enthalten, die die praktische Einsetzbarkeit der gewählten Verfahren belegt.

Literatur:

[1] Fasano, G., & Franceschini, A. (1987). A multidimensional version of the Kolmogorov-Smirnov test. Monthly Notices of the Royal Astronomical Society, 225(1), 155-170.

[2] Justel, A., Peña, D., & Zamar, R. (1997). A multivariate Kolmogorov-Smirnov test of goodness of fit. Statistics & Probability Letters, 35(3), 251-259.

2. Baumbasierte Verfahren zur Imputation – surrogate splits

In dem grundlegenden Buch [1] führten Breiman et al. die sogenannten Klassifikations- und Regressionsbäume in die Welt der Statistik ein. Sie wählten einen nicht modellbasierten Ansatz, der darüber hinaus (bzgl. der Laufzeit) effizient implementiert werden konnte. Verschiedene Nachteile konnten in späteren Jahren durch Weiterentwicklungen (Bagging/Random Forests oder Boosting) marginalisiert werden. Der Erfolg bei Klassifikations- und Regressionsaufgaben ist empirisch sehr gut belegt und auch die Theorie weist inzwischen einige Resultate zu statistischen Eigenschaften baumbasierter Methoden auf. In der Regel nicht so sehr im Fokus ist jedoch ein Detail der Implementierung, die sogenannten surrogate splits. Die Idee dahinter ist die Folgende: Man definiert ein Ähnlichkeitsmaß zwischen allen denkbaren Splits an einem Knoten. Angenommen im Fall der Klassifikation ist der optimale Split s an einem Knoten bezüglich Variable v und eben diese Variable v fehlt nun bei einem zu klassifizierenden Datenpunkt. Die Entscheidung, ob dieser zu klassifizierende Datenpunkt nun im Baum nach links oder nach rechts weitergereicht wird, wird dann anhand des ersten surrogate splits getroffen, also dem Split, der s am ähnlichsten ist.

Das Ziel der Arbeit besteht in der theoretischen und praktischen Auseinandersetzung mit Bäumen und deren surrogate splits. Insbesondere soll die Arbeit aufzeigen, wie die in den surrogate splits enthaltenen Informationen nutzbar gemacht werden können. Zusätzlich soll geprüft werden, ob und wie diese surrogate splits genutzt werden können, um eine Imputation fehlender Werte im Datensatz vorzunehmen. Diese Untersuchung soll außerdem eine kleine Simulationsstudie dazu enthalten.

Literatur:

[1] Breiman, L., Friedman, J., Olshen, R. A., & Stone, C. J. (1984). Classification and regression trees. Chapman & Hall. New York.

3. Theoretische Untersuchungen zur baumbasierten Regressionsimputation

Das Statistische Bundesamt führt seine Aufgaben [...] auf der Grundlage der jeweils sachgerechten Methoden durch (§ 2 Abs. 3 BStatG), was zunächst erfordert, sachgerechte Methoden als solche zu identifizieren. Im Rahmen dessen wurden verschiedene Methoden zur Imputation betrachtet, die häufig im KI/ML-Kontext genannt werden, u. a. baumbasierte Verfahren. Dabei ergaben sich Ergebnisse, die auf den ersten Blick kontraintuitiv sind: Beim Einsatz von Random Forests (ebenso wie beim Einsatz von k-nearest-neighbour-Verfahren) zur Regressionsimputation wird die Varianz der zu imputierenden Variable nicht reduziert. Je höher die Ausfallrate und bei Missingmechanismen jenseits von MCAR wird das Ergebnis erwartungsgemäß zunehmend schlechter. Interessanterweise wurden ähnliche Ergebnisse auch von einem weiteren statistischen Amt in der Europäischen Union erzielt. Auf Basis der früheren Ergebnisse werden dieses Amt und das Statistische Bundesamt im

kommenden Jahr eine größere Studie (auf weit mehr amtlichen Datensätzen als bislang, ggf. auch mit gesonderten Simulationen, um Grenzfälle zu testen) durchführen. Aber selbst wenn die Studie die bisherigen Resultate bestätigt, lägen nur empirische Belege für das überraschend gute Verhalten von Random Forests bei Regressionsimputation vor.

Ziel der Arbeit ist es, das Phänomen unter theoretischen Gesichtspunkten zu untersuchen und das Zustandekommen des bislang kontraintuitiven Ergebnisses zu erklären.

Hintergrundinformationen zur Imputation

Ein Feld der Statistik, insbesondere in den Wirtschafts- und Sozialwissenschaften sowie der amtlichen Statistik, ist der Umgang mit fehlenden Werten durch sogenannte Imputationsmethoden. Imputation bezeichnet eine Herangehensweise, die es ermöglicht, unvollständige Fälle in die Datenanalyse einzubeziehen (anstelle sie vollständig zu ignorieren). Tatsächlich ist das Hauptziel der Imputation nicht die Wiederherstellung der Informationen in den fehlenden Werten. Imputation hat stattdessen das Ziel, die wesentlichen Outputs einer auf einem vollständigen Datensatz basierenden statistischen Analyse im Fall von fehlenden Werten hinreichend gut zu reproduzieren.

Weitere Hinweise

Das Schreiben der Abschlussarbeit kann bei Vorliegen der persönlichen Voraussetzungen gegebenenfalls mit einem Praktikum im Statistischen Bundesamt kombiniert werden. Ein Anspruch auf ein Praktikum im Statistischen Bundesamt begründet sich durch die Wahl eines der oben genannten Themen jedoch nicht.

Bei Fragen zu den Abschlussarbeiten oder zum Praktikum können Sie sich gerne an mich wenden (Tel.: 0611/75-3887; E-Mail: florian.dumpert@destatis.de).

Hinweis: Weitere Abschlussarbeiten beim Statistischen Bundesamt sind im Internet zu finden:

https://www.destatis.de/DE/Service/Statistik-Campus/Wissenschaftliche-Abschlussarbeiten/abschlussarbeiten.html.

Mit freundlichen Grüßen
Im Auftrag
gez. Florian Dumpert