Unit Testing Data for Machine Learning

Bachelorprojekt für 2018/2019
Fachgebiet Informationssysteme
Prof. Dr. Felix Naumann

Datenbanken großer Unternehmen können Tausende von Tabellen umfassen, die teilweise Hunderte von Spalten enthalten. Nutzer dieser Datenbanken benötigen daher Unterstützung bei der Analyse und Auswahl von Daten für Experimente zur Automatisierung und Optimierung von Unternehmensabläufen, etwa mittels *Machine Learning*. Im Rahmen dieses Projekts werden Werkzeuge entwickelt, die einen breiten Einsatz im Unternehmen finden sollen, etwa beim Amazon-Produktkatalog oder bei Preis- und Verkaufsdaten.

Die Arbeit dieses Bachelorprojekts baut auf einem aktuellen Projekt von Amazon Research im Bereich Datenqualität auf. So wie *Unit Testing* von *Source Code* bereits Best Practice ist, werden hierbei Verfahren für *Unit Testing for Data* entwickelt, die eine automatisierte Verifikation der Verteilung von Daten ermöglichen. Im Rahmen dieses Bachelorprojekts soll eine webbasierte Plattform zur Verwaltung, Ausführung und Visualisierung der Datenverifikation entwickelt werden.

Projektbeschreibung

Dieses Projekt umfasst die folgenden Aufgabenpakete, wobei der genaue Umfang jedes Pakets mit den Teilnehmern im Laufe des Projekts abgestimmt wird:

- **Webplattform zur Verwaltung und Visualisierung**: Als Benutzerschnittstelle zur Verwaltung der Datenverifikationsschritte und Visualisierung von Statistiken soll eine webbasierte Plattform entwickelt werden. Hierüber verbindet sich ein Nutzer mit einem Datenbanksystem, wählt Tabellen zur Verifikation aus, konfiguriert Verifikationsparameter und analysiert die Ergebnisse mithilfe von geeigneten Diagrammen und Statistiken.

- **Berechnung von Statistiken in Datenbanken**: Die von Amazon entwickelte Software zur Berechnung von Statistiken nutzt Apache Spark. Im Rahmen dieses Projekts soll eine Schnittstelle zu weiteren Datenbanksystemen (etwa AWS Redshift) entwickelt werden. Durch die Nutzung von SQL können dadurch deutlich mehr Datensätze ohne zusätzliche Ladevorgänge analysiert werden. Eine Herausforderung ist die hoch-performante Ausführung der Berechnungen.

- **Tests von Veränderungen von Datensätzen**: Viele Datensätze sind dynamisch; einzelne Operationen können erhebliche Änderungen in der Datenverteilung zur Folge haben. So wird etwa der Produktkatalog von Amazon durch das Hinzufügen neuer Produkte und Attribute von existierenden Produkten stetig erweitert. Im Rahmen dieses Arbeitspakets sollen Verfahren zur Erkennung von systematischen Änderungen über mehrere Versionen von Datensätzen entwickelt werden.

- **Generierung von Vorschlägen zur Datenzusammenfassung**: Zur Beurteilung, welche Tests für einen Datensatz verwendet werden sollen und für die potenzielle Nutzung eines Datensatzes in einem Experiment ist es wichtig, sich schnell einen Überblick über den Inhalt eines Datensatzes zu verschaffen. Data Scientists setzen hierfür ihre Erfahrung im Umgang mit großen Datensätzen und verschiedenen Datentypen ein. In diesem Aufgabenpaket sollen Methoden zur automatischen Vorhersage von *relevanten* Statistiken entwickelt werden, etwa Verteilungskennzahlen für kontinuierliche Daten oder ein Clustering von diskreten Daten. Hierbei sind Methoden der explorativen Datenanalyse (EDA) hilfreich.

- **Visualisierung der Datenzusammenfassung**: Vorgeschlagene Statistiken sollten zum einfachen Verständnis visualisiert werden. Für verschiedene Datentypen sollen automatisch geeignete Plots und Skalierungen der Daten generiert werden.

Da Datensätze bei Amazon oft eine sehr hohe Speicherkomplexität aufweisen, müssen die entwickelten Verfahren hocheffizient und parallelisierbar sein. Amazon stellt hierfür Rechenkapazität in Amazon Web Services (AWS) zur Verfügung. Das Bachelorprojekt kann die Weiterentwicklung eines Open-Source-Projekts umfassen.

Projektpartner

Amazon Research in Berlin entwickelt neuartige Methoden im Bereich Machine Learning und Data Management. Ein Team aus international anerkannten Experten befasst sich mit Algorithmen und Modellierungstechniken, die die Skalierung sehr großer Daten bewältigen und den aktuellen Stand der Forschung vorantreiben. Die entwickelten Verfahren werden in Produktionssysteme von Amazon integriert und verbessern die Einkaufserfahrung von Millionen Kunden weltweit.

Amazon wird von vier Grundprinzipien geleitet: Fokus auf den Kunden statt auf den Wettbewerb, Leidenschaft fürs Erfinden, Verpflichtung zu operativer Exzellenz und langfristiges Denken. Kundenrezensionen, 1-Click-Shopping, personalisierte Empfehlungen, Prime, Versand durch Amazon, AWS, Kindle Direct Publishing, Kindle, Fire Tablets, Fire TV, Amazon Echo und Alexa sind nur einige der Produkte und Services, für die Amazon Pionierarbeit geleistet hat. Mehr Informationen zu Amazon unter www.aboutamazon.de.

Das Projekt beginnt im Oktober 2018 und wird durch Prof. Dr. Felix Naumann und sein Team betreut. Fragen können gerne an felix.naumann@hpi.de gerichtet werden.