Karlsruher Institut für Technologie

Institut für Programmstrukturen und Datenorganisation (IPD)
Lehrstuhl für Systeme der Informationsverwaltung, Prof. Böhm

Bachelorarbeit/Masterarbeit

Bachelorarbeit/Masterarbeit

Entwicklung und Evaluation von Ensemble-Learning Techniken basierend auf kontrastreichen Teilräumen

Gegenwärtig will man in eigentlich allen Lebensbereichen Vorhersagemodelle rein datenbasiert aufbauen. Oft sind die zugrunde liegenden Datenbestände hochdimensional, d. h. haben viele Attribute. Dies wirkt sich i. d. R. negativ auf die Güte des Vorhersagemodells aus. Häufig sind jedoch nicht alle Dimensionen gleich wichtig. Oft lässt sich ein konkretes Verhalten gut aus einer Kombination weniger Dimensionen herauslesen. Diese Kombination nennt man dann einen **kontrastreichen Teilraum**. In vielen Fällen hat ein Datenbestand mehrere kontrastreiche Teilräume, die unterschiedlich viele Dimensionen und unterschiedlichen Kontrast haben können. - **Ensemble Learning** ist eine allgemeine Bezeichnung für Methoden, die mehrere Vorhersagemodelle kombinieren. Ensemble Learning hat in Experimenten oft zu sehr guten Ergebnissen geführt, besser als mit nur einem Modell. In dieser Aufgabe sollen Sie deshalb die beiden genannten Gebiete verbinden: Sie werden sich überlegen, wie Ensemble-Learning Techniken aussehen sollten, die Vorhersagen über unterschiedliche kontrastreiche Teilräume kombinieren. Daraus ergeben sich folgende Aufgabenteile:

- Sichtung aktueller Techniken des Ensemble Learnings (wie Bagging, Boosting und Stacking) durch Literaturrecherche und Inbetriebnahme von Standard-implementierungen auf verfügbaren Benchmark-Daten.
- Entwicklung neuer Ensemble Learning Verfahren unter Berücksichtigung kontrastreicher Teilräume unter Verwendung bestehender Verfahren für die Entdeckung kontrastreicher Teilräume.
- Planung und Durchführung von Qualitäts- und Performance-Experimenten auf hochdimensionalen Datenbeständen. Ihre Evaluation soll systematisch sein und sowohl unterschiedliche Datensätze (die verfügbar sind) als auch unterschiedliche zugrunde liegende Verfahren für die Modellbildung betrachten. Neben Genauigkeit der Vorhersagen und Performance gibt es noch weitere Kriterien wie zum Beispiel die Einfachheit der verwendeten Modelle.

Sie erwerben mit der Bearbeitung eine ausgeprägte Kompetenz (sowohl in theoretischer als auch in sehr praktischer Hinsicht) im Bereich ‘Big Data Analytics‘. Dabei lernen Sie konkrete Ensemble Learning Techniken kennen sowie Möglichkeiten, Vorhersagemodelle zu konstruieren. Kenntnisse aus einer einschlägigen Vorlesung zu Datenanalyse sind keine notwendige Bedingung für die Bearbeitung. – Der Umfang dieser Arbeit lässt sich variieren, dadurch ist sowohl die Bearbeitung als Bachelor- als auch als Masterarbeit möglich.

Ansprechpartner

Georg Steinbuß | georg.steinbuss@kit.edu | +49 721 608-43911 | Raum: 363

Am Fasanengarten 5 | 76131 Karlsruhe | Gebäude: 50.34