Bachelorarbeit/Masterarbeit

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

Bachelorarbeit/Masterarbeit

Effiziente Bestimmung kontrastreicher Teilräume für unterschiedliche Kontrastmaße

Gegenstand der Betrachtung sind hochdimensionale Datenbestände. In solchen Daten manifestieren sich Auffälligkeiten (z. B. Cluster oder Ausreißer) meist nur in einer Teilmenge der Dimensionen/Attribute. Diese Teilräume sind i. d. R. die mit ausgeprägtem Kontrast, d. h. die Datenverteilung weicht stark von der ab, die man aufgrund der Datenverteilungen in Unterräumen erwarten würde. Die Zahl der Teilräume eines hochdimensionalen Merkmalsraums wächst exponentiell, das Finden der kontrastreichsten Teilräume ist deshalb schwierig. In dieser Aufgabe geht es darum, neue Lösungen für das Finden dieser Teilräume zu entwickeln, die die folgenden Punkte berücksichtigen:

• Es gibt unterschiedliche Kontrast-Maße in der Literatur, d. h. unterschiedliche Möglichkeiten, Kontrast zu quantifizieren. Die Begründungen, warum ein bestimmtes Maß verwendet werden sollte, sind abstrakt und nicht immer eindeutig. Klar ist jedoch, dass weniger gute Maße sich oft viel schneller berechnen lassen als andere. Für viele Kontrast-Maße sind außerdem Approximationen möglich, d. h. ihre Berechnung lässt sich beispielsweise abbrechen, weil absehbar ist, dass der zu berechnende Kontrast niedrig ist.
• Es fehlt eine Untersuchung mehrerer realer Datenbestände, wie der Kontrast eines Teilraums mit den Kontrast-Werten seiner Unterräume in Zusammenhang steht. Dies sowohl für direkte Unterräume, d. h. der Unterraum hat nur eine Dimension weniger als der Teilraum, der gerade betrachtet wird, als auch für indirekte Unterräume. Betrachtet werden sollte auch der Zusammenhang zwischen dem Kontrast eines Teilraums gemäß eines Maßes und dem von Unterräumen gemäß anderer (leicht zu berechnender) Maße.

Daraus ergeben sich die folgenden Teile der Aufgabenstellung:
• Datengetriebene Entwicklung eines Modells zur Vorhersage des Kontrasts eines d-dimensionalen Teilraums auf Grundlage der Kontraste seiner Unterräume (aller Unterräume bzw. eines Teils der Unterräume, unter Verwendung des gleichen Kontrastmaßes oder eines anderen Kontrastmaßes usw.) und Evaluierung dieses Modells.
• Entwicklung von Algorithmen zum schnellen Finden kontrastreicher Teilräume, unter Verwendung dieses Vorhersagemodells, und systematische Durchführung einschlägiger Experimente.
• Ggf. weitere Untersuchungen, wie sehr die so gefundenen kontrastreichen Teilräume bei der Lösung klassischer Data Mining Aufgaben, z. B. Klassifizierung oder Histogramm-Konstruktion, helfen können.

Sie erwerben mit der Bearbeitung eine ausgeprägte Kompetenz (sowohl theoretisch als auch in sehr praktischer Hinsicht) im Bereich Big Data Analytics sowie beim Entwurf und der Bewertung von Algorithmen. 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

Prof. Dr. Klemens Böhm	klemens.boehm@kit.edu	+49 721 608-43968	Raum: 366
Am Fasanengarten 5	76131 Karlsruhe	Gebäude: 50.34	