Informationstheoretische Verfahren zur visuellen Analyse von Klima und Strömungsdaten

Heike Jänicke

Zusammenfassung

Alljährlich verdoppeln sich die in der Wissenschaft erzeugten Datenmengen. Aufgrund der Verfügbarkeit leistungsstarker Rechner, vergrößert sich dabei nicht nur die Anzahl der simulierten und gemessenen Datensätze, sondern auch deren Größe. Da Wissenschaftler Daten mit immer feineren Auflösungen in Zeit und Raum und einer Vielzahl gekoppelter Variablen erzeugen, wächst der Speicherbedarf eines einzigen Datensatzen schnell auf mehrere Giga- oder Terabyte (1.000 Gigabyte) an. Obwohl schon die Verarbeitung und Analyse dieser Datensätze vergleichsweise schwierig ist, entstehen zunehmend Daten im Petabyte-Bereich (1.000 Terabyte). Die jährliche Erzeugung solch großer Datensätze wird immer alltäglicher, wie man am Beispiel des geplanten Large Synoptic Survey Teleskops (LSST) sieht. Um der enormen Datenmenge Herr zu werden, sind Kompressionsverfahren nötig, die die Daten auf das Wesentliche reduzieren und somit Wissenschaftler dabei unterstützen, den Gesamtzusammenhang zu verstehen.

In der wissenschaftliche Visualisierung wurden in den letzten Jahrzehnten vielfältige Techniken entwickelt, die es ermöglichen, mit wenig Aufwand selbst von komplizierten Datensätzen aussagekräftige Bilder zu erzeugen. Bei der Darstellung von Millionen von Punkten mit multivariaten Korrelationen geraten die meisten Standardverfahren jedoch an ihre Grenzen. Bei der Illustration moderner zeitabhängiger multivariater Daten muss deshalb auf eine effektive Reduktion der darzustellenden Daten geachtet werden. Aus der Fülle an Informationen müssen die Teile extrahiert werden, die den höchsten Informationsgehalt aufweisen. Häufig wird dazu der Datensatz auf einige wichtige Merkmale reduziert. Diese Klasse von Verfahren hat jedoch den Nachteil, dass Merkmale oft schwer zu definieren sind und teilweise auch von der jeweiligen Anwendung abhängen. Im Gegensatz zu den gängigen merkmalsbasierten Algorithmen stellen wir in dieser Arbeit Ansätze vor, die auf der Informationstheorie basieren, und mit deren Hilfe automatisch relevante Strukturen und Regionen extrahiert werden können.

Das erste hier präsentierte Verfahren dieser Art erweitert die Idee der lokalen statistischen Komplexität von endlichen Automaten auf diskretisierte multivariate Daten. Die lokale statistische Komplexität ist ein Maß, das der Extraktion von informationsreichen Strukturen im Datensatz dient. Hierdurch können relevante Bereiche rein mathematisch extrahiert und hervorgehoben werden. Da der ursprüngliche Algorithmus sehr rechenintensiv ist, wird im Anschluss ein Verfahren zur schnelleren und robusteren Berechnung präsentiert, mit dem auch große dreidimensionale Datensätze verarbeitet werden können. Um relevante Regionen in Daten auf beliebig strukturierten Gittern extrahieren zu können, wird anschließend die Theorie der lokalen statistischen Komplexität zur linearen lokalen statistischen Komplexität erweitert. Um veranschaulichen zu können, warum bestimmte Bereiche als relevant klassifiziert wurden, werden daraufhin ϵ-Maschinen eingeführt, die der statischen Visualisierung dynamischer Systeme dienen. Wir beginnen die Ausführungen mit der Visualisierung von Attributwolken, die der explorativen Analyse multivariater Daten dienen. Attributwolken stellen multivariate Daten in zwei verschiedenen Ansichten dar. Die eine Ansicht illustriert die hochdimensionale Verteilung der Daten in einer zweidimensionalen Visualisierung, während die andere Ansicht den physikalischen Raum abbildet und dazu genutzt wird um Positionen mit bestimmten Wertekombinationen hervorzuheben. Da beide Ansichten miteinander interagieren, kann der Nutzer eine der beiden Ansichten manipulieren und korrespondierende Strukturen in der anderen Ansicht werden automatisch dargestellt. Der Nutzen der verschiedenen Verfahren wird anhand verschiedener aktueller Datensätze aus dem Bereich der Klima- und Strömungssimulation illustriert.