Research Seminar 1: Task 1 - Review

Markus Deutschl

Paper: Algorithms for Mining Distance-Based Outliers in Large Datasets

Autoren: Knorr, Edwin M. und Ng, Raymond T.

Erscheinungsjahr: 1998

Struktur

Zu Anfang des Papers findet sich eine kurze Zusammenfassung des behandelten Themengebietes, gefolgt von einer Einführung und Beschreibung von „outlier detection“, wobei auch die Problemstellung klar dargelegt wird. Im Zuge dessen werden auch bereits durchgeführte Arbeiten in diesem Gebiet aufgezählt, um das Themengebiet und den Nutzen des Papers noch genauer zu spezifizieren. Des Weiteren werden praktische Anwendungsgebiete genannt, wo die im Paper erarbeiteten Lösungen eingesetzt werden können.

Im Anschluss daran werden vier Algorithmen vorgestellt und analysiert, die das Problem auf unterschiedliche Weise lösen sollen. Darauf folgend werden Messungen beschrieben und ausgewertet, um die Algorithmen in Hinblick auf ihre Laufzeit unter verschiedenen Bedingungen vergleichen zu können. Vergleiche mit bereits bestehenden Lösungen wurden ausgelassen, da bereits in zuvor publizierten Arbeiten bewiesen wurde, dass diese für Daten mit mehr als zwei Dimensionen nicht praktikabel sind.

Kritik

Wie bereits im vorigen Abschnitt beschrieben, sind alle groben Punkte, die für ein wissenschaftliches Paper nötig sind, vorhanden. Es gab zum Zeitpunkt der Publikation zwar schon einige Arbeiten, die sich mit dem Thema auseinander setzten, jedoch wurden im vorliegenden Paper Algorithmen präsentiert, die auch für mehr als zwei Dimensionen Ergebnisse in praktikabler Rechenzeit liefern. In Anbetracht dieser Tatsachen waren die Ideen und Resultate für die damalige Zeit neu und auch schwierig zu erreichen.

Nun soll die inhaltliche Aufbereitung unter die Lupe genommen werden. Zu Anfang werden das Themengebiet und damit zusammenhängende Definitionen besprochen. Diese Informationen werden gut verständlich erklärt und geben dem Leser einen guten Überblick über die Hintergründe und auch die Probleme des Forschungsgebiets. Des Weiteren wird auch klar gemacht, dass hinter dem zentralen Problem bei der „outlier detection“ ein wichtiger Nutzen für die praktische Anwendung steht.

Die vorgestellten Algorithmen werden textuell und mathematisch beschrieben und sind in sich schlüssig. Der begleitende Pseudo-Code ist jedoch etwas verwirrend, da er sehr textlastig ist und somit von gewohnten Code-Strukturen abweicht. Hier wäre eine C-artige Syntax wünschenswert, da die meisten Menschen auf dem Gebiet der Informatik damit vertraut sind. Sehr hilfreich sind hingegen die genauen Komplexitätsbetrachtungen der Algorithmen und der Vergleich dieser in

Hinblick auf die Dimensionsanzahl der Daten, da so ein relativ gutes Bild der Effizienz dieser vermittelt wird.

Die durchgeführten Messungen ergänzen die zuvor genannten Komplexitätsbetrachtungen und deren Resultate scheinen stimmig zu sein. Es wurden außerdem die Programmiersprache und das Testsystem genannt, weshalb eine Überprüfung der Ergebnisse relativ einfach möglich wäre. Das Datenset (NHL-Statistiken) der Messungen war mit lediglich 855 realen Datensätzen etwas zu klein dimensioniert. Es wurden zwar 2 Millionen Datensätze mittels der Verteilung der Originaldaten generiert, jedoch kann dies niemals so aussagekräftig sein wie Daten aus der realen Welt. Des Weiteren ist auch die Herkunft der Daten und deren Format unbekannt und können nicht nachvollzogen werden.

Conclusio

Die vorliegende Publikation ist in sich stimmig und ist auch von fachlicher Kompetenz gezeichnet. Die vermittelten Inhalte werden gut verständlich beschrieben und erklärt. Eine solide Kurzeinführung in das Themengebiet und die Problemstellung gibt dem Leser einen guten Überblick und dient als gute Vorbereitung für die weiteren Ausführungen.

Im Großen und Ganzen finde ich das Paper sehr gut geschrieben und einfach zu verstehen. Außerdem hat es mir eine solide Basis im Hinblick auf „distance-based outlier detection“ gegeben sowie auch Lösungsansätze zu diesem Problem. Es wäre außerdem sehr interessant, Vergleiche mit heutigen Algorithmen zu diesem Thema nachzuforschen, um zu sehen, wie stark sich dieses Forschungsgebiet weiterentwickelt hat.