

Zusammenfassung
When Is "Nearest Neighbor" Meaningful?
Beyer, Goldstein, Ramakrishnan, Shaft.

Gruppe X

November 9, 2013

Abstract

Diese Zusammenfassung ist im Zuge der Übung zur Einführung in die Mustererkennung WS 2013 entstanden. Das zusammenzufassende Werk trägt den Titel "When Is 'Nearest Neighbor' Meaningful?" von Beyer, Goldstein, Ramakrishnan, Shaft und wurde an der University of Wisconsin-Madison veröffentlicht.

Gruppe X
Florian Groh, 1168186
Felix Ledochowski, 1028318
Daniel Witurna, 1125818

1 Zusammenfassung des Papers

Der Artikel behandelt das Problem von hohen Dimensionen bei der Verwendung eines Nearest Neighbor Algorithmus. Es werden Datensätze vorgestellt, die empirisch zeigen, dass schon ab 10-15 Dimension eine Anwendung von NN-Verfahren kein wertvolles Ergebnis liefern kann. Dies wird insbesondere zum Problem, da immer mehr versucht wird, schwierige und komplexe Daten durch hoch-dimensionale Merkmalsvektoren anzunähern.

Um die Wertigkeit von Datensätzen und deren Eignung für NN zu bewerten, führen Beyer et al. die Definition der Instabilität ein. Instabiles Verhalten entsteht, wenn innerhalb einer durch ein $\epsilon > 0$ erweiterten Umgebung der NN-Anfrage der Großteil aller Punkte zu finden ist. In die andere Richtung wird von stabilem Verhalten gesprochen, falls wenige Punkte in diese erweiterte Umgebung fallen und dadurch die NN-Anfrageumgebung von den anderen Daten sinnvoll getrennt ist. Diese Definition wird verwendet, um zu zeigen, dass in vielen Situationen, in denen die Dimensionalität ansteigt, die Wahrscheinlichkeit der Instabilität einer NN-Anfrage gegen 1 konvergiert. Diese eben erwähnten Situationen treten nur unter Zutreffen der von Bayer et al. aufgestellten Theoreme auf.

Auch auf die Frage der sinnvollen Anwendung von hoch-dimensionaler Indexierung gehen die Autoren ein. Beispielsweise macht eine Klassifizierung sehr wohl Sinn, falls Punkte existieren, die genau der Anfrage entsprechen und dadurch die minimale Distanz 0 wird. Eine andere Möglichkeit bietet sich, wenn Punkte der Anfrage nicht genau übereinstimmen müssen sondern auch in einer kleinen Entfernung eines Datenpunkts erlaubt sind. Durch dieses Vorgehen werden die Anforderungen gelockert, weswegen eine sinnvolle Nachbarschaftsanfrage möglich wird. Weitere Ausführungen beschäftigen sich mit der Bildung von sogenannten Clustern, in welche die Anfrage fallen muss.

Als Ergebnis dieses Papers werden mehrere Punkte angeführt. Beispielsweise wurde mehrere Szenarien der Anwendung gefunden, in denen die Entfernungen von Punkten und deren nächsten Nachbarn vernachlässigbar klein werden. In diese Anwendungsgebiete fällt auch die übliche und vielgeprüfte Verwendung von NN als Heuristik. Weiters war die Frage in welchen Bereichen die Dimensionalität so hoch ist, dass kein aussagekräftiges Ergebnis mehr gefunden wird. Die Antwort konnten die Autoren durch Simulationen bestimmen und meinen, dass in den ersten 20 Dimensionen der Abstand der Punkte sehr stark abnimmt und über die 20. Dimension sehr schnell einen Punkt erreicht, ab dem die Entfernung zu gering wird. Der praktische Nutzen dieser Arbeit wird in den Bereichen der Evaluierung eines Nearest Neighbor - Aufgabenbereich sowie einer Nearest Neighbor - Vorgehensweise angegeben.