Analyse von Maschinen- und Prozeßzuständen mit Methoden der Nichtlinearen Dynamik
FKZ 13 N 7038/9
Projektleiter Prof. Dr. Lauterborn
Drittes Physikalisches Institut
Bürgerstraße 42–44
37073 Göttingen

Effiziente Nachbar-Suchalgorithmen und deren Anwendung in der nichtlinearen Zeitreihenanalyse

J. Wichard, C. Merkwirth, Dr. U. Parlitz, Prof. Dr. W. Lauterborn,
Drittes Physikalisches Institut der Universität Göttingen

Einleitung

Das Problem der Suche benachbarter Punkte in normierten Räumen taucht in vielen Bereichen der Datenanalyse auf. In der nichtlinearen Zeitreihenanalyse kann z.B. das Studium der Nachbarschaftsverhältnisse eines Systems Rückschlüsse auf wichtige Systemgrößen zulassen. Um solche Untersuchungen auch noch bei großen Datenzätzen durchführen zu können, wird hier ein effizientes Verfahren zur Nächsten-Nachbarn-Suche (NN-Suche) vorgestellt.

Probleme der Nachbarsuche

Ein großes Problem der Nachbarsuche ist der Ressourcenbedarf, denn die Komplexität des Problems wächst mit der Anzahl der Datenpunkte und bei großen Datensätzen benötigen einfache Realisierungen der NN-Suche zuviel Zeit.

Hierarchische Clusterung

Das Herzstück der effizienten NN-Suche ist ein hierarchisches Clusterverfahren. Der Algorithmus beruht in Teilen auf dem ANNA Algorithmus, der von James McNames vorgeschlagen wurde /1/. In der *Vorverarbeitungsphase* wird ein Suchbaum aufgebaut, der während der eigentlichen *NN-Suche* das Auffinden benachbarter Punkte beschleunigt.

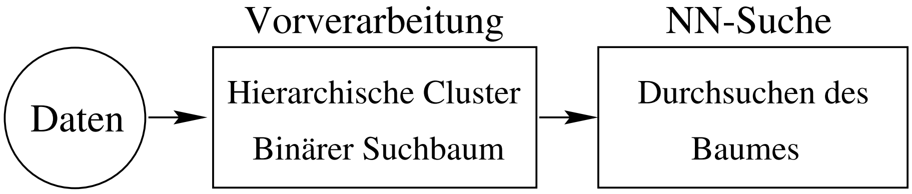

Abbildung 1: Schema der Verarbeitungsschritte

Der Aufbau des Suchbaumes

Der Aufbau des Suchbaumes geschieht durch rekursive Einteilung des Datensatzes in *Cluster*. Bei Erreichen der kritischen Clustergröße (ca. 60-200 Datenpunkte) bricht die Rekursion ab. Jeder Cluster wird beschrieben durch sein Zentrum und seinen Durchmesser. Diese Information wird später beim Durchsuchen des Baumes genutzt, um all jene Cluster von der Suche auszuschliessen, die keine NN enthalten können /2/.

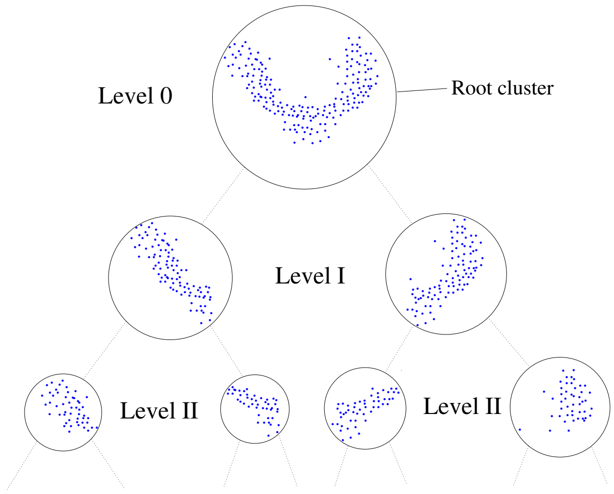

Abbildung 2: Der Aufbau eines typischen Suchbaumes

Sucheffizienz

Definiert man die Effizienz der Suche als:

$$\frac{\text{Anzahl aller Datenpunkte}}{\text{Mittl. Anz. d. Berechnungen pro Nachfrage}}$$

so ist die Effizienz dieses Algorithmus $O(N/\log(N))$, während ein „Brute-Force-Verfahren" immer die Effizienz 1 besitzt.

Die Wahl der Metrik

Beim Aufbau und bei der Durchsuchung des Baumes wird die Dreiecksungleichung verwendet, die in jeder Metrik gültig ist. Deshalb kann bei der NN-Suche eine beliebige Metrik zugrunde gelegt werden, z.B. die Metrik bzgl. der L^p-Norm.

Anwendungen der NN-Suche

Viele Probleme der nichtlinearen Zeitreihenanalyse lassen sich auf die Untersuchung von Nachbarschaftsverhältnissen zurückführen /3/, so zum Beispiel die:

- Modellierung und Vorhersage von Zeitreihen
- Schnelle Berechnung der Korrelationssumme

- Nichtlineare Rauschbefreiung
- Berechnung der Lyapunovexponenten
- Wiederkehrzeiten

Suche in hochdimensionalen Räumen

In hochdimensionalen Räumen ist es möglich, daß der Unterschied zwichen dem maximalen und dem minimalen Abstand eines Referenzpunktes zu den restlichen Punkten im Datensatz verschwindend gering wird. Dies ist ein grundsätzliches Problem hoher Dimensionen und im allgemeinen unahängig von der Wahl der verwendeten Metrik. Um die hiermit verbundenen

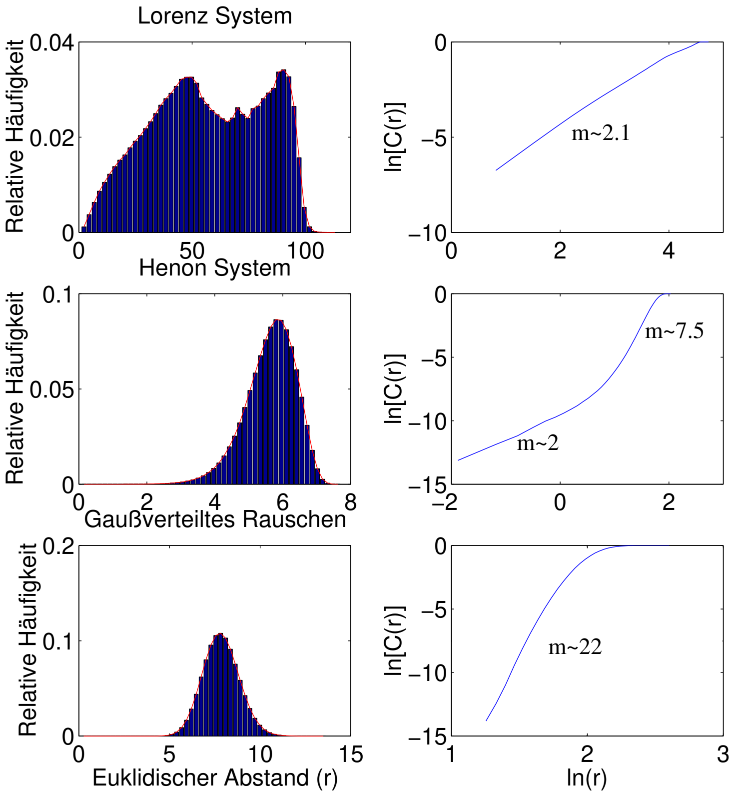

Abbildung 3: Die linke Spalte zeigt das Histogramm der euklidischen Abstände zwischen 100 Referenzpunkten und dem Rest des Datensatzes für das Lorenz-Systems, der Hénon-Abbildung und von Gaußverteiltem Rauschen. In der linken Spalte ist jeweils die Korrelationssumme geplottet und die lokale Steigung angegeben.

Schwierigkeiten an einem Beispiel zu verdeutlichen haben wir Zeitreihen des Lorenz-Systems, der Hénon-Abbildung und von Gaußverteiltem Rauschen (jeweils 100000 Werte) in 30 Dimensionen mit Delay-Zeit 1 eingebettet und die Histogramme der euklidischen Abstände zwischen

100 Referenzpunkten und dem Rest des Datensatzes errechnet. Zudem wurden die Korrelationssummen gebildet und in einem doppeltlogarithmischen Diagramm dargestellt.

Auf größeren Skalen wird die Dimension des Hénon-Attraktors überschätzt infolge der starken Faltung der hochdimensionalen Einbettung, die ein hochdimensionales Objekt vortäuscht (siehe Abbildung 3). Dieser Effekt tritt bei diskreten dynamischen Systemen unvermeidlich ein, da es hier eine untere Schranke der Delay-Zeit gibt. Bei kontinuierlichen Systemen läßt sich durch die Wahl einer kleinen Abtastzeit die Überfaltung bei der Einbettung verhindern. So wird beim Lorenzsystem die Dimension des niedrigdimensionalen Attraktors korrekt wiedergegeben.

Literaturverzeichnis

/1/ J. McNames, A Nearest Trajectory Strategy for Time Series Prediction, Proc. of the International Workshop on Advanced Black-Box Techniques for Nonlinear Modeling 1998, 112 ff, 1998

/2/ S. Berchtold, C. Böhm, D.A. Keim, H.P. Kriegl, A cost model for nearest neighbour search in highdimensional data space, PODS'97 (Tuscon, AZ), 78-86, 1997

/3/ T. Schreiber, Efficient neighbor searching in nonlinear time-series analysis, Int. J. Bifurcation and Chaos (5), 349 ff, 1995