Ludwig-Maximilians-Universität München
Institut für Informatik
PD Dr. Peer Kröger, Dr. Arthur Zimek

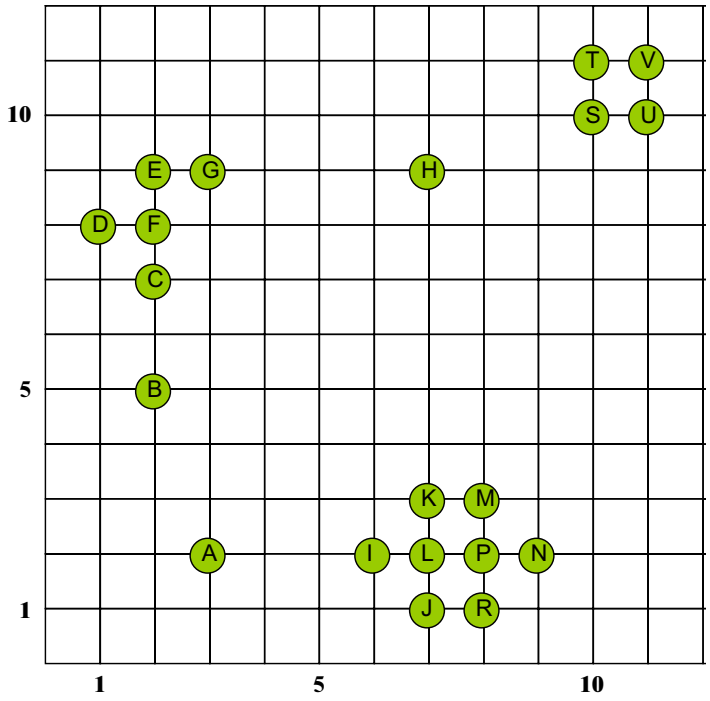
Knowledge Discovery in Databases

WS 2010/11

Übungsblatt 10: Hierarchisches Clustering, Outlier Detection

Aufgabe 10-1 *Single-Link*

Gegeben sei der folgende Datensatz:



Als Distanzfunktion zwischen den Punkten dient Ihnen jeweils wieder die Manhattan-Distanz (L_1-Norm):

$$L_1(x, y) = |x_1 - y_1| + |x_2 - y_2|$$

Berechnen Sie zwei Dendrogramme für diesen Datensatz. Als Distanzfunktion zwischen Mengen von Objekten verwenden Sie

(a) den Single-Link Ansatz,

(b) den Average-Link Ansatz.

Tipp: Innere Knoten müssen nicht binär sein, d.h. sie können mehr als zwei Söhne haben.

Aufgabe 10-2 *Drei-Sigma-Regel*

In der Statistik findet man oft die sogenannte "Drei-Sigma-Regel", auch als "68-95-99.7-Regel" bekannt. So spricht man ab einem Wert von mehr als 95% von "schwach signifikant (*)", ab 99% von "stark signifikant (**)" und ab 99.9% von "sehr stark signifikant (***)". Bezieht man diese Werte auf eine Normalverteilung, so entspricht dies etwas einer Abweichung von $\pm 2\sigma$ ($\approx 95\%$) bzw. $\pm 3\sigma$ ($\approx 99.7\%$). Eine Abweichung von mehr als 2σ ist also nach dieser Sprechweise "schwach signifikant" und ab 3σ "stark signifikant".

Jedoch finden diese Regeln in der Statistik normalerweise dann Anwendung, wenn es sich nur ein paar hundert Datensätze handelt. Die Anwendung einer solchen Regel in der automatischen Datenanalyse führt aber zu Problemen:

Berechnen Sie dazu einen einfachen Erwartungswert, wie viele Elemente auf einem Standardnormalverteilten Datensatz von einer Million Werte um mehr als 3σ vom Mittelwert abweicht.

Braucht ein statistisches Verfahren (z.B. EM-Outlier Detection), dass Ausreißer nach einer solchen Verteilung bewertet, daher vielleicht eine Korrektur? Wie kann man dies korrigieren?