**Ludwig-Maximilians-Universität München**
**Institut für Informatik**
Priv. Doz. Dr. Matthias Schubert, Dr. Arthur Zimek
Erich Schubert

München, den 13.1.2010

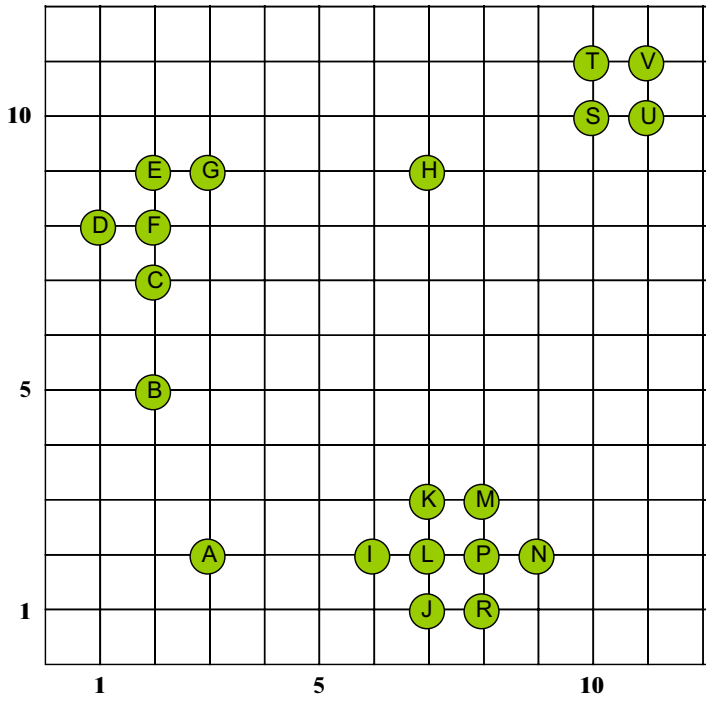
**Knowledge Discovery in Databases**
WS 2009/10

## Übungsblatt 10: Outlier Detection

Besprechung am 15.1.2010

### Aufgabe 10-1 *Outlier Scores*

Gegeben sei der folgende 2-dimensionale Datensatz:




Verwenden Sie als Distanzfunktion zwischen den Punkten wieder die Manhattan-Distanz ($L_1$-Norm)

Berechnen Sie (unter Ausschluss des Anfragepunktes bei der Berechnung der $k$NN):

- Den LOF-Wert für $k = 2$ für die Punkte $H$, $L$ und $B$.
- Den LOF-Wert für $k = 4$ für die Punkte $H$, $L$ und $B$.
- Die $k$NN-Distanz für $k = 2$ für alle Punkte.
- Die $k$NN-Distanz für $k = 4$ für alle Punkte.
- Die aggregierten $k$NN-Distanzen für $k = 2$ und $k = 4$ für alle Punkte.

Diskutieren Sie die Wahl von $k$ für diesen Datensatz.

**Lösungsvorschlag:**

$k$NN-Nachbarschaften:

| $p$ | 2NN | 2dist | 4NN | 4dist | $1/\text{lrd}_2$ | $1/\text{lrd}_4$ | a2NN | a4NN |
|---|---|---|---|---|---|---|---|---|
| $H$ | $G\ S$ | **4** | + $E\ T\ U$ | **5** | $(4+4)/2$ | $(4+4+5+5+5)/5$ | **8** | **18** |
| $L$ | $I\ K\ P\ J$ | 1 | = | 1 | $(2+1+1+1)/4$ | $(2+2+1+2)/4$ | 2 | 4 |
| $B$ | $C\ F$ | **3** | + $A\ D\ E$ | 4 | $(2+3)/2$ | $(2+3+5+4+4)/5$ | **5** | **13** |
| $A$ | $I\ B\ L$ | **4** | + $K\ P\ J$ | **5** | $(3+4+4)/3$ | $(3+4+4+5+5+5)/6$ | **7** | **16** |
| $C$ | $F\ D\ E\ B$ | 2 | = | 2 | $(1+2+2+3)/4$ | $(2+3+2+4)/4$ | 3 | 7 |
| $D$ | $F\ E\ C$ | 2 | + $G$ | 3 | $(1+2+2)/3$ | $(2+2+2+3)/4$ | 3 | 8 |
| $E$ | $F\ G$ | 1 | + $C\ D$ | 2 | $(1+2)/2$ | $(2+3+2+3)/4$ | 2 | 6 |
| $F$ | $C\ D\ E$ | 1 | + $G$ | 2 | $(2+2+1)/3$ | $(2+3+2+3)/4$ | 2 | 5 |
| $G$ | $E\ F$ | 2 | + $C\ D$ | 3 | $(1+2)/2$ | $(2+2+3+3)/4$ | 3 | 9 |
| $I$ | $L\ K\ P\ J$ | 2 | = | 2 | $(1+2+2+2)/4$ | $(1+2+2+2)/4$ | 3 | 7 |
| $J$ | $L\ R$ | 1 | + $I\ K\ P$ | 2 | $(1+1)/2$ | $(1+2+2+2+2)/5$ | 2 | 6 |
| $K$ | $L\ M$ | 1 | + $I\ J\ P$ | 2 | $(1+1)/2$ | $(1+2+2+2+2)/5$ | 2 | 6 |
| $M$ | n.b. | 1 | n.b. | 2 | n.b. | n.b. | 2 | 6 |
| $N$ | n.b. | 2 | n.b. | 2 | n.b. | n.b. | 3 | 7 |
| $P$ | $L\ M\ N\ R$ | 1 | = | 1 | $(1+1+2+1)/4$ | $(1+2+2+2)/4$ | 2 | 4 |
| $R$ | n.b. | 1 | n.b. | 2 | n.b. | n.b. | 2 | 6 |
| $S$ | $T\ U$ | 1 | + $V\ H$ | 4 | $(1+1)/2$ | $(5+5+6+5)/4$ | 2 | 8 |
| $T$ | $V\ S$ | 1 | + $U\ H$ | **5** | $(1+1)/2$ | $(6+4+5+5)/4$ | 2 | 9 |
| $U$ | $V\ S$ | 1 | + $T\ H$ | **5** | $(1+1)/2$ | $(6+4+5+5)/4$ | 2 | 9 |
| $V$ | n.b. | 1 | n.b. | **6** | n.b. | n.b. | 2 | 10 |

Wir formen LOF wie folgt um: $LOF(p) = \frac{\sum_{o \in kNN(p)} \frac{lrd_k(o)}{lrd_k(p)}}{|kNN(p)|} = \frac{\sum_{o \in kNN(p)} lrd_k(o)}{|kNN(p)|} / lrd_k(p)$

Damit ergibt sich:

- LOF $k = 2$:

$$H = \frac{(\frac{2}{3} + \frac{2}{2})}{2} / \frac{2}{8} \approx 3.333$$

$$L = \frac{(\frac{4}{7} + \frac{2}{2} + \frac{4}{5} + \frac{2}{2})}{4} / \frac{4}{5} \approx 1.054$$

$$B = \frac{(\frac{4}{8} + \frac{3}{5})}{2} / \frac{2}{5} \approx 1.375$$

- LOF $k = 4$:

$$H = \frac{(\frac{4}{10} + \frac{4}{21} + \frac{4}{10} + \frac{4}{20} + \frac{4}{20})}{5} / \frac{5}{23} \approx 1.279$$

$$L = \frac{(\frac{4}{7} + \frac{5}{9} + \frac{4}{7} + \frac{5}{9})}{4} / \frac{4}{7} \approx 0.986$$

$$B = \frac{(\frac{4}{11} + \frac{4}{10} + \frac{6}{26} + \frac{4}{9} + \frac{4}{10})}{5} / \frac{5}{18} \approx 1.324$$

Wählt man $k$ zu groß, werden bei beiden Verfahren die Ergebnisse schlechter. Insbesondere wird bei LOF die Score von $H$ sogar niedriger als von $B$, bei $k$NN mit $k = 4$ wird der Punkt $V$ zum deutlichsten Ausreißer etc.

Keines der Ergebnisse ist jedoch als "falsch" zu bezeichnen – entsprechend der Definition sind die Punkte dann Outlier. Die Definition passt nur nicht zu unserer "Intention" und unseren Zielen.

**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 $2\sigma$ ist also nach dieser Sprechweise "schwach signifikant" und ab $3\sigma$ "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 $3\sigma$ 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?

**Lösungsvorschlag:**

$99.7\%$ der Punkte, genauer gesagt das $0.9973002$ Quantil ist erwartungsgemäß unter $3\sigma$. Bei einer Million Punkte sind daher erwartungsgemäß etwa $2700$ Punkte weiter entfernt.

Ob oder wie man hier etwas korrigieren kann ist fraglich: bei einer derartigen Datenmenge gibt es einfach statistisch diese Ausreißer. Definiert man Ausreißer über ihre Standardabweichung, so muss man auch mit einer entsprechenden Anzahl rechnen. Man kann natürlich auch nur die $k$ "unwahrscheinlichsten" Ausreißer nehmen, aber jeder solche Ansatz ist mit ähnlichen – statistischen – Fehlern behaftet. Im Gegenteil: findet man erheblich weniger (oder mehr) als $2700$ Punkte die derartig stark abweichen, so legt dies sogar nahe, dass einfach das Modell einer Normalverteilung nicht zu den Daten passt.