Ludwig-Maximilians-Universität München
Institut für Informatik
PD Dr. Arthur Zimek
Dr. Erich Schubert

Knowledge Discovery in Databases
SS 2014

Übungsblatt 11: Evaluation von Algorithmen

Aufgabe 11-1 Evaluierung von Clusterings

Gegeben seien zwei Clusterings $\{A_1, \ldots, A_5\}$ und $\{B_1, \ldots, B_4\}$,
und folgende Matrix mit den Größen der Schnittmengen $|A_i \cap B_j|$:

	A_1	A_2	A_3	A_4	A_5
B_1	2	1	11	0	13
B_2	0	10	0	3	1
B_3	15	0	4	0	0
B_4	0	0	2	8	1

- Bewerten Sie das Clustering, indem Sie einmal nur die Zeilenmaxima (als "Precision") und einmal nur die Spaltenmaxima (als "Recall") betrachten. Kombinieren Sie diese Werte mit dem F-Measure.
- Berechnen Sie in jeder Zeile eine "Precision" und jeder Spalte einen "Recall", indem sie den größten Wert als "true positive" annehmen. Berechnen Sie das F-Measure aus der mittleren Precision der Zeilen und dem mittleren Recall der Spalten.
- Berechnen Sie für jede Zelle ein F-Measure (indem sie die "Precision" bzgl. der Zeilensumme und den "Recall" bezüglich der Spaltensumme berechnen). Suchen Sie für jede Zeile und Spalte das Maximum, und berechnen Sie daraus je einen mittleren F-Score für alle Zeilen und alle Spalten, sowie den Mittelwert aus diesen beiden.
- Pair Counting: um nicht immer genau einen Wert aus einer Zeile oder Spalte in Betracht zu ziehen, werden alle Objektpaare betrachtet die in beiden Clusterings zusammen in einem Cluster sind. Ein Objekt bilde dabei kein Paar mit sich selbst (also keine Paare der Form (x, x)), d.h. Paare existieren in einem Clustering A:

  $$(x, y) \in P(A) \Leftrightarrow \exists_{A_i \in A} x \in A_i \wedge y \in A_i \wedge x \neq y$$

  Berechnen Sie Precision und Recall der Paare und das F-Measure daraus. Berechnen Sie auch den Rand Index, Adjusted Rand Index und Jaccard Index mit den Formeln aus der Vorlesung.

Aufgabe 11-2 Evaluation von Outlier-Detection-Algorithmen

Auf einen Datensatz mit bekannten Ausreißern $+$ wurden zwei Verfahren S_1 und S_2 angewendet.
Die Ergebnisse der Verfahren finden Sie in folgender Tabelle:

Object	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9	A_{10}
Label	$-$	$-$	$-$	$-$	$+$	$-$	$-$	$-$	$+$	$-$
S_1	1.0	1.1	1.1	1.3	3.0	2.0	1.5	0.9	1.4	1.2
S_2	.80	.80	.10	.81	.89	.50	.50	.91	.90	.20

Bewerten Sie die beiden Ausreißerverfahren S_1 und S_2 mit den folgenden Metriken:

- Precision, Recall und F-Measure,
  unter der Annahme, dass die größten $k = 2$ Werte als Ausreißer klassifiziert werden.
- Average Precision für $k = [1 \dots 4]$,
  unter der Annahme, dass die größten k Werte als Ausreißer klassifziert werden.
- Zeichnen Sie die ROC Kurve, und berechnen sie die Fläche unter dieser Kurve (AUC).
- Normalisieren Sie die Werte auf $[0; 1]$, und berechnen Sie die Kostenfunktion aus der Vorlesung.