Cluster Ensembles

Combining Multiple Partitions

Daniel Spierling

11. Oktober 2007

Inhaltsverzeichnis

Was sind eigentlich Cluster Ensembles?

Ausgangssituation

Was sind eigentlich Cluster Ensembles?

Ausgangssituation

- Menge von Objekten
  $\chi = \{x_1, x_2, \ldots, x_n\}$

Was sind eigentlich Cluster Ensembles?

Ausgangssituation

- Menge von Objekten
  $\chi = \{x_1, x_2, \ldots, x_n\}$
- Mehrere Cluster $\lambda^{(q)}$
  - Zuordnung der Objekte in Gruppen
  - Jedes Objekt gehört (maximal) zu einer Gruppe

Was sind eigentlich Cluster Ensembles?

Ausgangssituation

- Menge von Objekten
  $\chi = \{x_1, x_2, \ldots, x_n\}$
- Mehrere Cluster $\lambda^{(q)}$
  - Zuordnung der Objekte in Gruppen
  - Jedes Objekt gehört (maximal) zu einer Gruppe
- keine Kenntnis über den Algorithmus Φ der Clusterer (Erzeuger der Zuordnungen)

Was sind eigentlich Cluster Ensembles?

Ziel

Was sind eigentlich Cluster Ensembles?

Ziel

- Vereinigung der Cluster zu einem neuen Cluster
  - Effizienz des Algorithmus
  - Clustering, das die meisten Informationen wie die Originalen besitzen
  - Ermöglichung von verteilten Clustern

Was sind eigentlich Cluster Ensembles?

Ziel

- Vereinigung der Cluster zu einem neuen Cluster
  - Effizienz des Algorithmus
  - Clustering, das die meisten Informationen wie die Originalen besitzen
  - Ermöglichung von verteilten Clustern
- Wiederverwendbarkeit
  - Eigenschaften des Objektes werden ignoriert
  - Zusätzliche Informationen des Clusterer werden nicht berücksichtigt

Was sind eigentlich Cluster Ensembles?

Cluster Ensemble Funktion

Was sind eigentlich Cluster Ensembles?

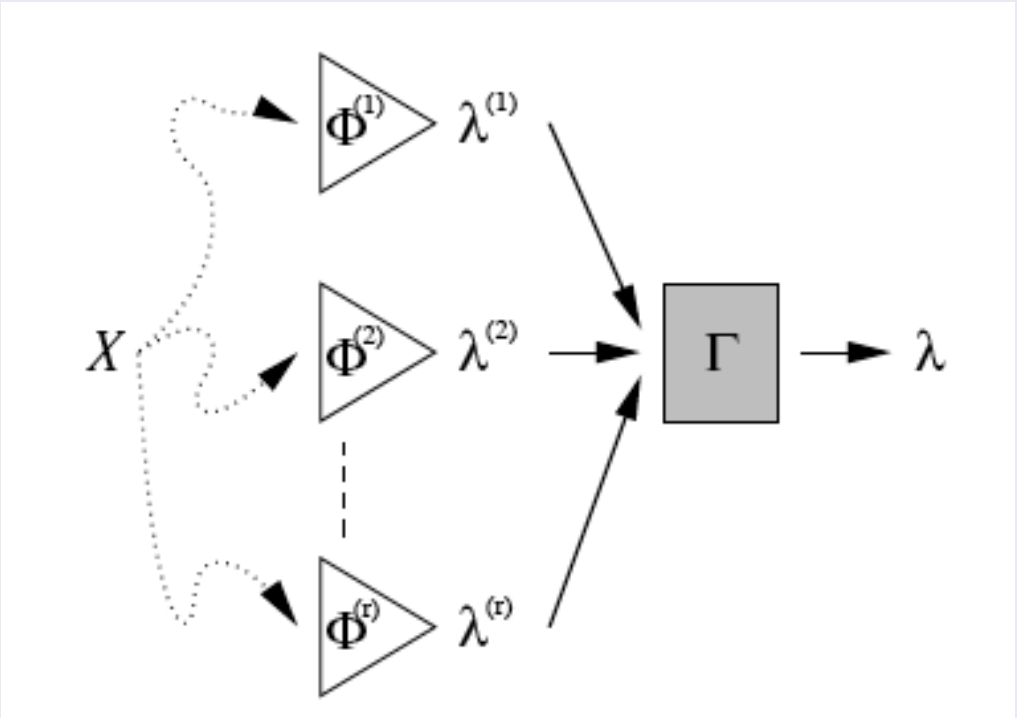
Cluster Ensemble Funktion

$$\Gamma : \{\lambda^{(q)} | q \in \{1, \ldots, r\} \rightarrow \lambda\}$$

Was sind eigentlich Cluster Ensembles?

Cluster Ensemble Funktion

$$\Gamma : \{\lambda^{(q)} | q \in \{1, \ldots, r\} \rightarrow \lambda\}$$



Wie können wir damit arbeiten?

Darstellung

Wie können wir damit arbeiten?

Darstellung

- Ordnung der Elemente (Reihenfolge)

Wie können wir damit arbeiten?

Darstellung

- Ordnung der Elemente (Reihenfolge)
- Jedem Element ist eine Gruppe zugeordnet
  $\longrightarrow$ Vektor mit k Elementen dargestellt werden

Wie können wir damit arbeiten?

Darstellung

- Ordnung der Elemente (Reihenfolge)
- Jedem Element ist eine Gruppe zugeordnet
  $\longrightarrow$ Vektor mit k Elementen dargestellt werden
- Eine Menge von n Clusterings
  $\longrightarrow$ n Vektoren oder n $\times$ k Matrix

Wie können wir damit arbeiten?

Beobachtungen

$$\lambda^{(1)} = (1;1;1;2;2;3;3)^T \quad \lambda^{(2)} = (2;2;2;3;3;1;1)^T$$
$$\lambda^{(3)} = (1;1;2;2;3;3;3)^T \quad \lambda^{(4)} = (1;2;?;1;2;?;?)^T$$

Wie können wir damit arbeiten?

Beobachtungen

$$\lambda^{(1)} = (1;1;1;2;2;3;3)^T \ \lambda^{(2)} = (2;2;2;3;3;1;1)^T$$
$$\lambda^{(3)} = (1;1;2;2;3;3;3)^T \ \lambda^{(4)} = (1;2;?;1;2;?;?)^T$$

- $\lambda^{(1)}$ und $\lambda^{(2)}$ sind logisch identisch
  (ignorieren möglicher zusätzlicher Informationen)

Wie können wir damit arbeiten?

Beobachtungen

$$\lambda^{(1)} = (1;1;1;2;2;3;3)^T \ \lambda^{(2)} = (2;2;2;3;3;1;1)^T$$
$$\lambda^{(3)} = (1;1;2;2;3;3;3)^T \ \lambda^{(4)} = (1;2;?;1;2;?;?)^T$$

- $\lambda^{(1)}$ und $\lambda^{(2)}$ sind logisch identisch
  (ignorieren möglicher zusätzlicher Informationen)
- für jede eindeutige Zuordnung existieren k! Repräsentationen
  (k -> Anzahl der Gruppen)

Wie können wir damit arbeiten?

Beobachtungen

$$\lambda^{(1)} = (1;1;1;2;2;3;3)^T \ \lambda^{(2)} = (2;2;2;3;3;1;1)^T$$
$$\lambda^{(3)} = (1;1;2;2;3;3;3)^T \ \lambda^{(4)} = (1;2;?;1;2;?;?)^T$$

- $\lambda^{(1)}$ und $\lambda^{(2)}$ sind logisch identisch
  (ignorieren möglicher zusätzlicher Informationen)
- für jede eindeutige Zuordnung existieren k! Repräsentationen
  (k -> Anzahl der Gruppen)
- Vereinheitlichung der Darstellung

Wie können wir damit arbeiten?

Darstellung als Hypergraph

	$\lambda^{(1)}$	$\lambda^{(2)}$	$\lambda^{(3)}$	$\lambda^{(4)}$
x_1	1	2	1	1
x_2	1	2	1	2
x_3	1	2	2	?
x_4	2	3	2	1
x_5	2	3	3	2
x_6	3	1	3	?
x_7	3	1	3	?

$\Leftrightarrow$

	$\mathbf{H}^{(1)}$			$\mathbf{H}^{(2)}$			$\mathbf{H}^{(3)}$			$\mathbf{H}^{(4)}$	
	$\mathbf{h}_1$	$\mathbf{h}_2$	$\mathbf{h}_3$	$\mathbf{h}_4$	$\mathbf{h}_5$	$\mathbf{h}_6$	$\mathbf{h}_7$	$\mathbf{h}_8$	$\mathbf{h}_9$	$\mathbf{h}_{10}$	$\mathbf{h}_{11}$
v_1	1	0	0	0	1	0	1	0	0	1	0
v_2	1	0	0	0	1	0	1	0	0	0	1
v_3	1	0	0	0	1	0	0	1	0	0	0
v_4	0	1	0	0	0	1	0	1	0	1	0
v_5	0	1	0	0	0	1	0	0	1	0	1
v_6	0	0	1	1	0	0	0	0	1	0	0
v_7	0	0	1	1	0	0	0	0	1	0	0

- Zugehörigkeit zur Gruppe (1), sonst (0)
- Pro Spalte eine Hyperkante

Wie können wir damit arbeiten?

Darstellung als Hypergraph

	$\lambda^{(1)}$	$\lambda^{(2)}$	$\lambda^{(3)}$	$\lambda^{(4)}$
x_1	1	2	1	1
x_2	1	2	1	2
x_3	1	2	2	?
x_4	2	3	2	1
x_5	2	3	3	2
x_6	3	1	3	?
x_7	3	1	3	?

$\Leftrightarrow$

	$\mathbf{H}^{(1)}$			$\mathbf{H}^{(2)}$			$\mathbf{H}^{(3)}$			$\mathbf{H}^{(4)}$	
	$\mathbf{h}_1$	$\mathbf{h}_2$	$\mathbf{h}_3$	$\mathbf{h}_4$	$\mathbf{h}_5$	$\mathbf{h}_6$	$\mathbf{h}_7$	$\mathbf{h}_8$	$\mathbf{h}_9$	$\mathbf{h}_{10}$	$\mathbf{h}_{11}$
v_1	1	0	0	0	1	0	1	0	0	1	0
v_2	1	0	0	0	1	0	1	0	0	0	1
v_3	1	0	0	0	1	0	0	1	0	0	0
v_4	0	1	0	0	0	1	0	1	0	1	0
v_5	0	1	0	0	0	1	0	0	1	0	1
v_6	0	0	1	1	0	0	0	0	1	0	0
v_7	0	0	1	1	0	0	0	0	1	0	0

- Erweiterung einer Spalte in k(q)Spalten
  (k := Anzahl der Gruppen)
  $\longrightarrow$ für jede Gruppe eine Spalte

Cluster-based Similarity Partitioning Algorithm

Verfahren

Cluster-based Similarity Partitioning Algorithm

Verfahren

- Ermittlung der paarweisen Ähnlichkeit der Elemente

Cluster-based Similarity Partitioning Algorithm

Verfahren

- Ermittlung der paarweisen Ähnlichkeit der Elemente
- Matrizenmutiplikation als Basis
  $S = \frac{1}{r} H H^T$

Cluster-based Similarity Partitioning Algorithm

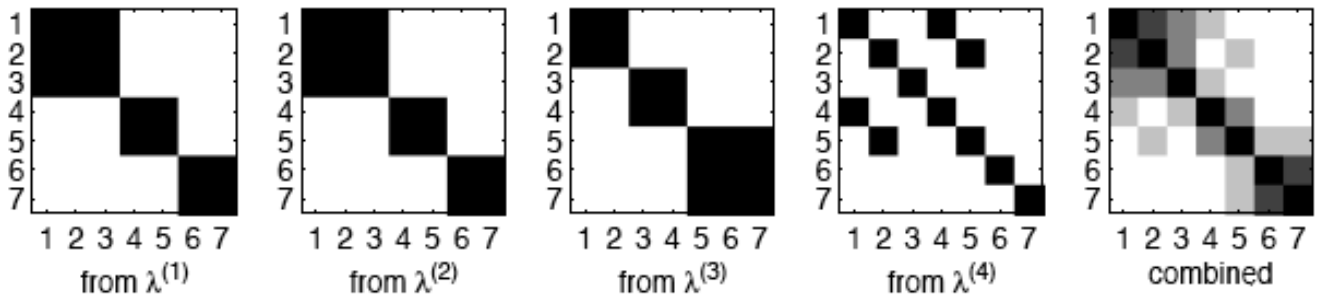
Verfahren

- Ermittlung der paarweisen Ähnlichkeit der Elemente
- Matrizenmutiplikation als Basis
  $S = \frac{1}{r} H H^T$
- Umgruppierung der Elemente mit Hilfe der errechneten Matrix durch einen ähnlichkeitsbasierten Zuordnungsalgorithmus (z. B. METIS)

Cluster-based Similarity Partitioning Algorithm

Verfahren

	$H^{(1)}$			$H^{(2)}$			$H^{(3)}$			$H^{(4)}$	
	h_1	h_2	h_3	h_4	h_5	h_6	h_7	h_8	h_9	h_{10}	h_{11}
v_1	1	0	0	0	1	0	1	0	0	1	0
v_2	1	0	0	0	1	0	1	0	0	0	1
v_3	1	0	0	0	1	0	0	1	0	0	0
v_4	0	1	0	0	0	1	0	1	0	1	0
v_5	0	1	0	0	0	1	0	0	1	0	1
v_6	0	0	1	1	0	0	0	0	1	0	0
v_7	0	0	1	1	0	0	0	0	1	0	0



HyperGraph-Partitioning Algorithm

Verfahren

HyperGraph-Partitioning Algorithm

Verfahren

- Hypergraph als Basis

HyperGraph-Partitioning Algorithm

Verfahren

- Hypergraph als Basis
- Alle Hyperkanten und Knoten besitzen die gleiche Gewichtung

HyperGraph-Partitioning Algorithm

Verfahren

- Hypergraph als Basis
- Alle Hyperkanten und Knoten besitzen die gleiche Gewichtung
- Ziel: Entfernung möglichst weniger Hyperkanten zu k unverbundenen Komponente (ungefähr gleicher Größe)

HyperGraph-Partitioning Algorithm

Verfahren

- Hypergraph als Basis
- Alle Hyperkanten und Knoten besitzen die gleiche Gewichtung
- Ziel: Entfernung möglichst weniger Hyperkanten zu k unverbundenen Komponente (ungefähr gleicher Größe)
- Vorteil: Die Elemente werden nicht nur paarweise betrachtet

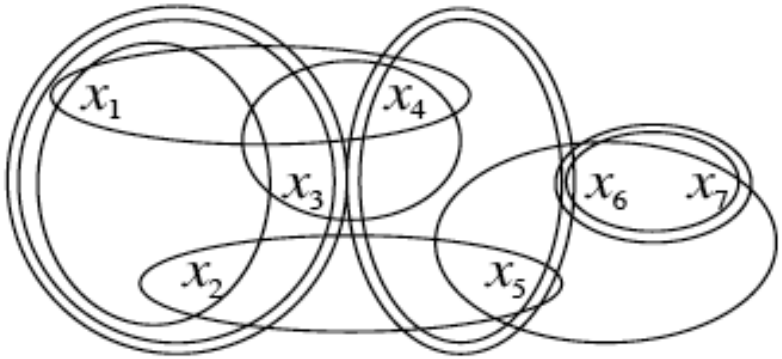
HyperGraph-Partitioning Algorithm

Transformation

	$\lambda^{(1)}$	$\lambda^{(2)}$	$\lambda^{(3)}$	$\lambda^{(4)}$
x_1	1	2	1	1
x_2	1	2	1	2
x_3	1	2	2	?
x_4	2	3	2	1
x_5	2	3	3	2
x_6	3	1	3	?
x_7	3	1	3	?

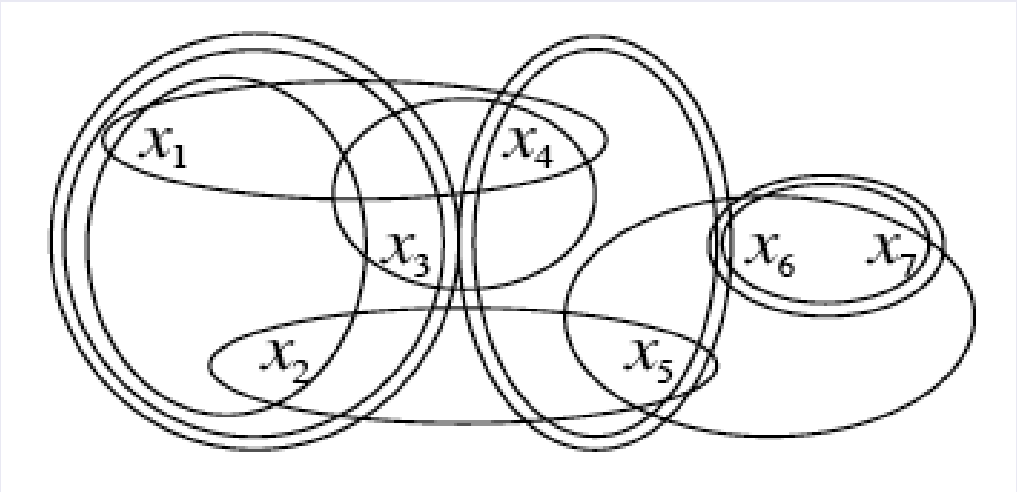
$\Leftrightarrow$

	$H^{(1)}$			$H^{(2)}$			$H^{(3)}$			$H^{(4)}$	
	h_1	h_2	h_3	h_4	h_5	h_6	h_7	h_8	h_9	h_{10}	h_{11}
v_1	1	0	0	0	1	0	1	0	0	1	0
v_2	1	0	0	0	1	0	1	0	0	0	1
v_3	1	0	0	0	1	0	0	1	0	0	0
v_4	0	1	0	0	0	1	0	1	0	1	0
v_5	0	1	0	0	0	1	0	0	1	0	1
v_6	0	0	1	1	0	0	0	0	1	0	0
v_7	0	0	1	1	0	0	0	0	1	0	0



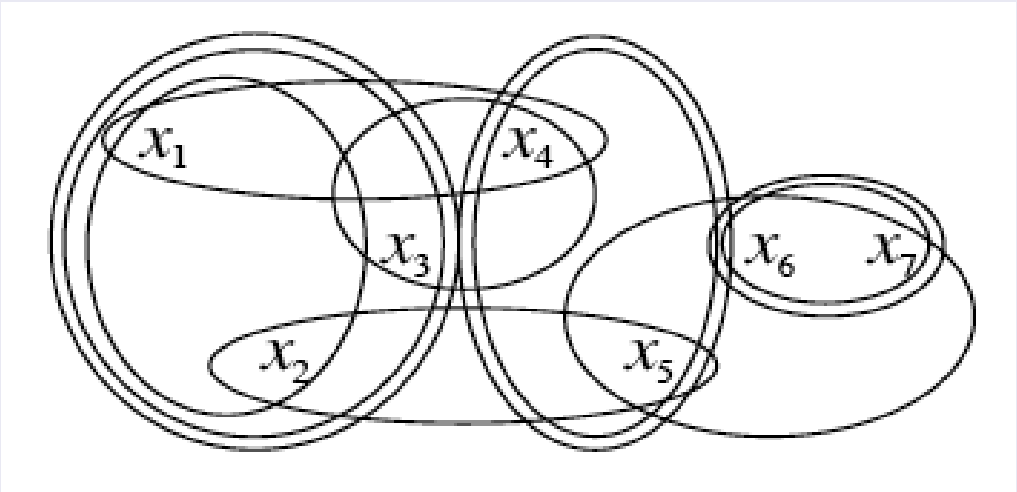
HyperGraph-Partitioning Algorithm

Graphische Darstellung



HyperGraph-Partitioning Algorithm

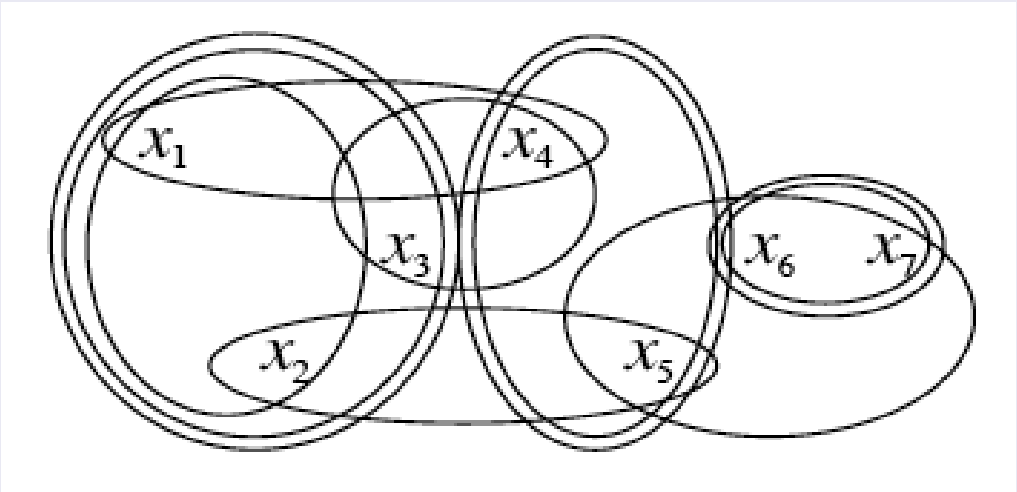
Graphische Darstellung



- Optimal: $\{\{x_1, x_2, x_3\}, \{x_4, x_5\}, \{x_6, x_7\}\}$

HyperGraph-Partitioning Algorithm

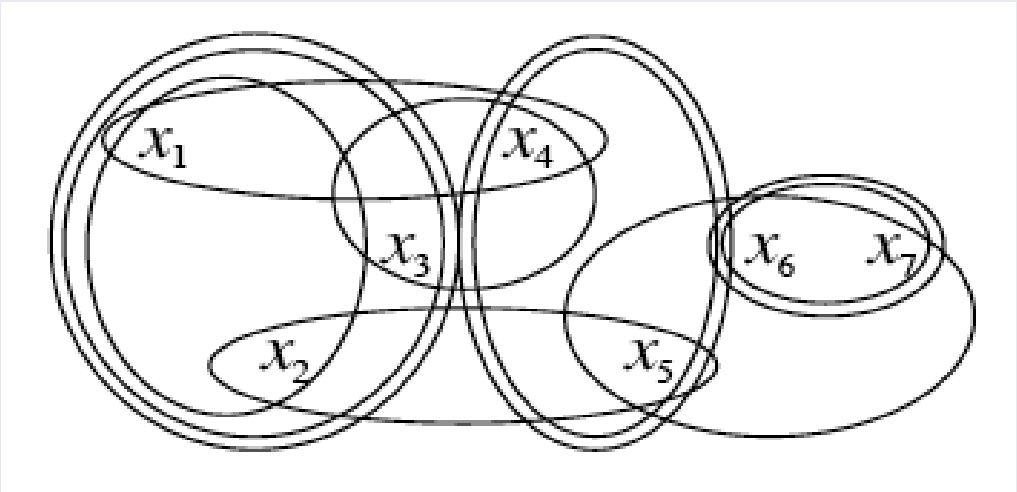
Graphische Darstellung



- Optimal: $\{\{x_1, x_2, x_3\}, \{x_4, x_5\}, \{x_6, x_7\}\}$
- $\{\{x_1, x_2, x_7\}, \{x_3, x_4\}, \{x_5, x_6\}\}$ und
  $\{\{x_1, x_7\}, \{x_3, x_4\}, \{x_2, x_5, x_6\}\}$ brechen alle Kanten auf

HyperGraph-Partitioning Algorithm

Graphische Darstellung

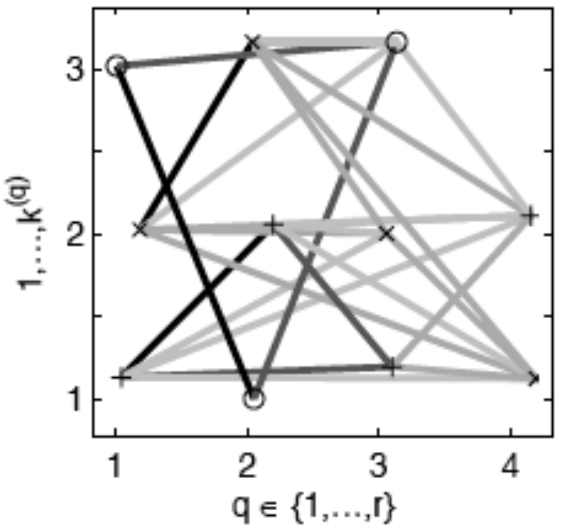


- Optimal: $\{\{x_1, x_2, x_3\}, \{x_4, x_5\}, \{x_6, x_7\}\}$
- $\{\{x_1, x_2, x_7\}, \{x_3, x_4\}, \{x_5, x_6\}\}$ und $\{\{x_1, x_7\}, \{x_3, x_4\}, \{x_2, x_5, x_6\}\}$ brechen alle Kanten auf , allerdings bricht die 1. Kante $\{x_1, x_2, x_3\}$ nur einmal (keine Berücksichtigung)

Meta-Clustering Algorithm

Verfahren

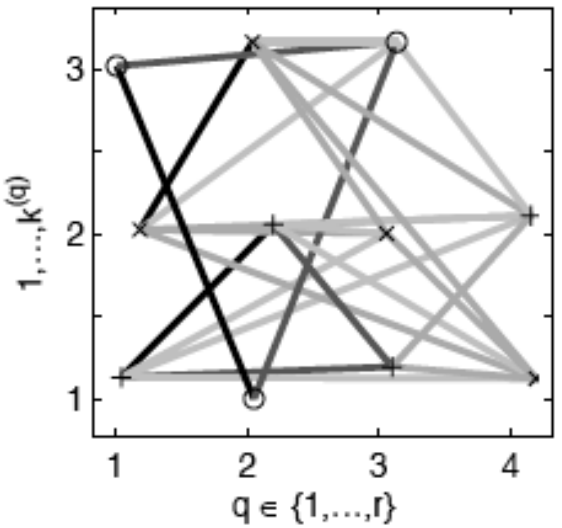
	$H^{(1)}$			$H^{(2)}$			$H^{(3)}$			$H^{(4)}$	
	h_1	h_2	h_3	h_4	h_5	h_6	h_7	h_8	h_9	h_{10}	h_{11}
v_1	1	0	0	0	1	0	1	0	0	1	0
v_2	1	0	0	0	1	0	1	0	0	0	1
v_3	1	0	0	0	1	0	0	1	0	0	0
v_4	0	1	0	0	0	1	0	1	0	1	0
v_5	0	1	0	0	0	1	0	0	1	0	1
v_6	0	0	1	1	0	0	0	0	1	0	0
v_7	0	0	1	1	0	0	0	0	1	0	0



Meta-Clustering Algorithm

Verfahren

	$H^{(1)}$			$H^{(2)}$			$H^{(3)}$			$H^{(4)}$	
	h_1	h_2	h_3	h_4	h_5	h_6	h_7	h_8	h_9	h_{10}	h_{11}
v_1	1	0	0	0	1	0	1	0	0	1	0
v_2	1	0	0	0	1	0	1	0	0	0	1
v_3	1	0	0	0	1	0	0	1	0	0	0
v_4	0	1	0	0	0	1	0	1	0	1	0
v_5	0	1	0	0	0	1	0	0	1	0	1
v_6	0	0	1	1	0	0	0	0	1	0	0
v_7	0	0	1	1	0	0	0	0	1	0	0

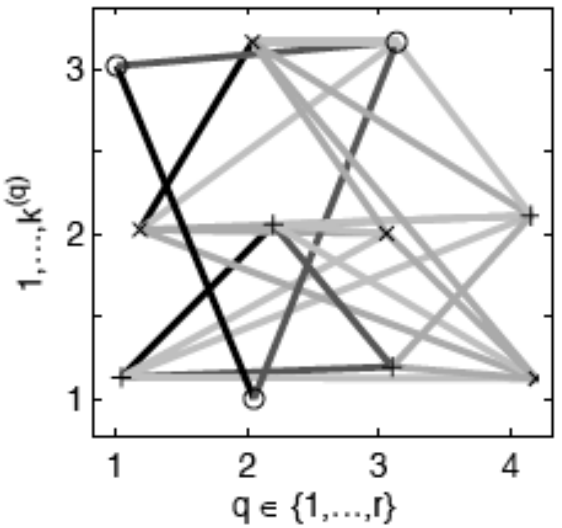


- Jedes Cluster h_x erhält einen Knoten

Meta-Clustering Algorithm

Verfahren

	$H^{(1)}$			$H^{(2)}$			$H^{(3)}$			$H^{(4)}$	
	h_1	h_2	h_3	h_4	h_5	h_6	h_7	h_8	h_9	h_{10}	h_{11}
v_1	1	0	0	0	1	0	1	0	0	1	0
v_2	1	0	0	0	1	0	1	0	0	0	1
v_3	1	0	0	0	1	0	0	1	0	0	0
v_4	0	1	0	0	0	1	0	1	0	1	0
v_5	0	1	0	0	0	1	0	0	1	0	1
v_6	0	0	1	1	0	0	0	0	1	0	0
v_7	0	0	1	1	0	0	0	0	1	0	0

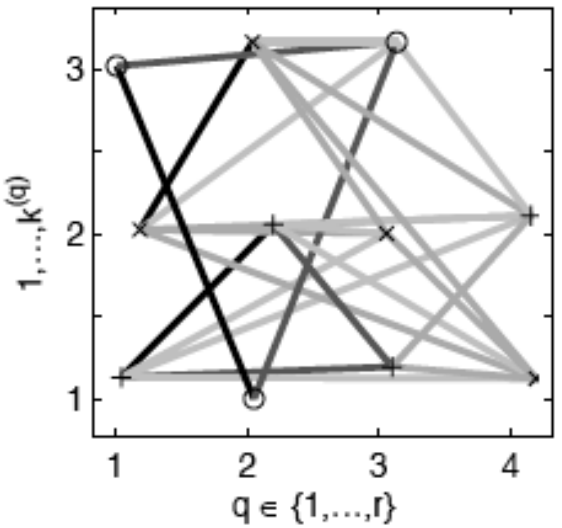


- Jedes Cluster h_x erhält einen Knoten
- Kantengewicht = Ähnlichkeit der Cluster (Jaccard measure)

Meta-Clustering Algorithm

Verfahren

	$H^{(1)}$			$H^{(2)}$			$H^{(3)}$			$H^{(4)}$	
	h_1	h_2	h_3	h_4	h_5	h_6	h_7	h_8	h_9	h_{10}	h_{11}
v_1	1	0	0	0	1	0	1	0	0	1	0
v_2	1	0	0	0	1	0	1	0	0	0	1
v_3	1	0	0	0	1	0	0	1	0	0	0
v_4	0	1	0	0	0	1	0	1	0	1	0
v_5	0	1	0	0	0	1	0	0	1	0	1
v_6	0	0	1	1	0	0	0	0	1	0	0
v_7	0	0	1	1	0	0	0	0	1	0	0

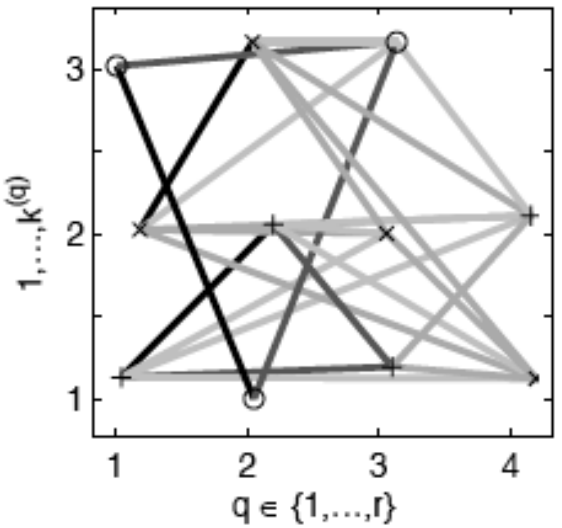


- Ähnlichkeitsbasierter Gruppierungsalgorithmus (METIS)

Meta-Clustering Algorithm

Verfahren

	$H^{(1)}$			$H^{(2)}$			$H^{(3)}$			$H^{(4)}$	
	h_1	h_2	h_3	h_4	h_5	h_6	h_7	h_8	h_9	h_{10}	h_{11}
v_1	1	0	0	0	1	0	1	0	0	1	0
v_2	1	0	0	0	1	0	1	0	0	0	1
v_3	1	0	0	0	1	0	0	1	0	0	0
v_4	0	1	0	0	0	1	0	1	0	1	0
v_5	0	1	0	0	0	1	0	0	1	0	1
v_6	0	0	1	1	0	0	0	0	1	0	0
v_7	0	0	1	1	0	0	0	0	1	0	0

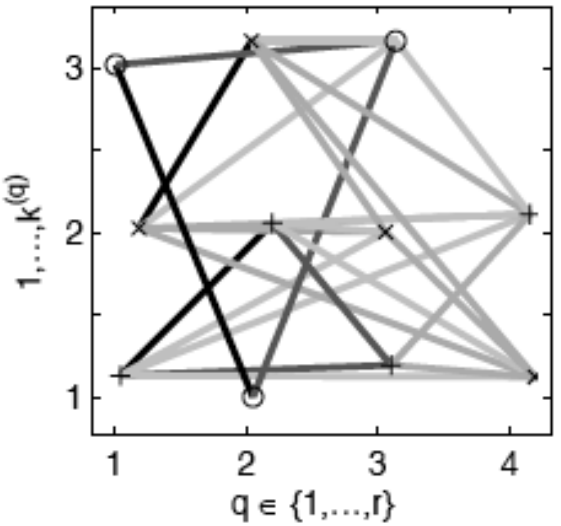


- Ähnlichkeitsbasierter Gruppierungsalgorithmus (METIS)
- Zuordnen der Elemente zu einem Cluster

Meta-Clustering Algorithm

Verfahren

	$H^{(1)}$			$H^{(2)}$			$H^{(3)}$			$H^{(4)}$	
	h_1	h_2	h_3	h_4	h_5	h_6	h_7	h_8	h_9	h_{10}	h_{11}
v_1	1	0	0	0	1	0	1	0	0	1	0
v_2	1	0	0	0	1	0	1	0	0	0	1
v_3	1	0	0	0	1	0	0	1	0	0	0
v_4	0	1	0	0	0	1	0	1	0	1	0
v_5	0	1	0	0	0	1	0	0	1	0	1
v_6	0	0	1	1	0	0	0	0	1	0	0
v_7	0	0	1	1	0	0	0	0	1	0	0

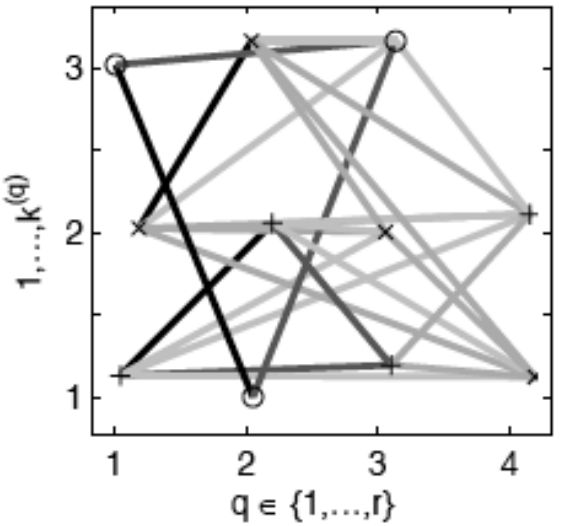


- $\{h_3, h_4, h_9\}, \{h_2, h_6, h_8, h_{10}\}, \{h_1, h_4, h_7, h_{11}\}$

Meta-Clustering Algorithm

Verfahren

	$H^{(1)}$			$H^{(2)}$			$H^{(3)}$			$H^{(4)}$	
	h_1	h_2	h_3	h_4	h_5	h_6	h_7	h_8	h_9	h_{10}	h_{11}
v_1	1	0	0	0	1	0	1	0	0	1	0
v_2	1	0	0	0	1	0	1	0	0	0	1
v_3	1	0	0	0	1	0	0	1	0	0	0
v_4	0	1	0	0	0	1	0	1	0	1	0
v_5	0	1	0	0	0	1	0	0	1	0	1
v_6	0	0	1	1	0	0	0	0	1	0	0
v_7	0	0	1	1	0	0	0	0	1	0	0



- $\{h_3, h_4, h_9\}, \{h_2, h_6, h_8, h_{10}\}, \{h_1, h_4, h_7, h_{11}\}$
- $\{h_3, h_4, h_9\} \rightarrow (0, 0, 0, 0, \frac{1}{3}, 1, 1)$

Estimation-Maximization-Algorithmus

Verfahren

Estimation-Maximization-Algorithmus

Verfahren

- Basiert auf k-means Algorithmus

Estimation-Maximization-Algorithmus

Verfahren

- Basiert auf k-means Algorithmus
  - Initialisierung: Auswahl von k Clusterzentren

Estimation-Maximization-Algorithmus

Verfahren

- Basiert auf k-means Algorithmus
  - Initialisierung: Auswahl von k Clusterzentren
  - Jedes Objekt wird dem nächsten Clusterzentrum zugeordnet

Estimation-Maximization-Algorithmus

Verfahren

- Basiert auf k-means Algorithmus
  - Initialisierung: Auswahl von k Clusterzentren
  - Jedes Objekt wird dem nächsten Clusterzentrum zugeordnet
  - Neuberechnung der Clusterzentren

Estimation-Maximization-Algorithmus

Verfahren

- Basiert auf k-means Algorithmus
  - Initialisierung: Auswahl von k Clusterzentren
  - Jedes Objekt wird dem nächsten Clusterzentrum zugeordnet
  - Neuberechnung der Clusterzentren
  - Wiederholung der Zuordnung

Estimation-Maximization-Algorithmus

Verfahren

- Basiert auf k-means Algorithmus
  - Initialisierung: Auswahl von k Clusterzentren
  - Jedes Objekt wird dem nächsten Clusterzentrum zugeordnet
  - Neuberechnung der Clusterzentren
  - Wiederholung der Zuordnung
- Estimation: Wahrscheinlichkeitsverteilung (häufig Normalverteilung) für Zuordnung

Estimation-Maximization-Algorithmus

Verfahren

- Basiert auf k-means Algorithmus
  - Initialisierung: Auswahl von k Clusterzentren
  - Jedes Objekt wird dem nächsten Clusterzentrum zugeordnet
  - Neuberechnung der Clusterzentren
  - Wiederholung der Zuordnung
- Estimation: Wahrscheinlichkeitsverteilung (häufig Normalverteilung) für Zuordnung
- Maximization: Neubestimmung der Parameter, die Mittelpunkt bestimmen

Wie schnell sind die Algorithmen?

Laufzeitabschätzung

Wie schnell sind die Algorithmen?

Laufzeitabschätzung

- CSPA: $O(n^2kr)$

Wie schnell sind die Algorithmen?

Laufzeitabschätzung

- CSPA: $O(n^2kr)$
- HGPA: $O(nkr)$

Wie schnell sind die Algorithmen?

Laufzeitabschätzung

- CSPA: $O(n^2kr)$
- HGPA: $O(nkr)$
- MCLA: $O(nk^2r^2)$

Wie schnell sind die Algorithmen?

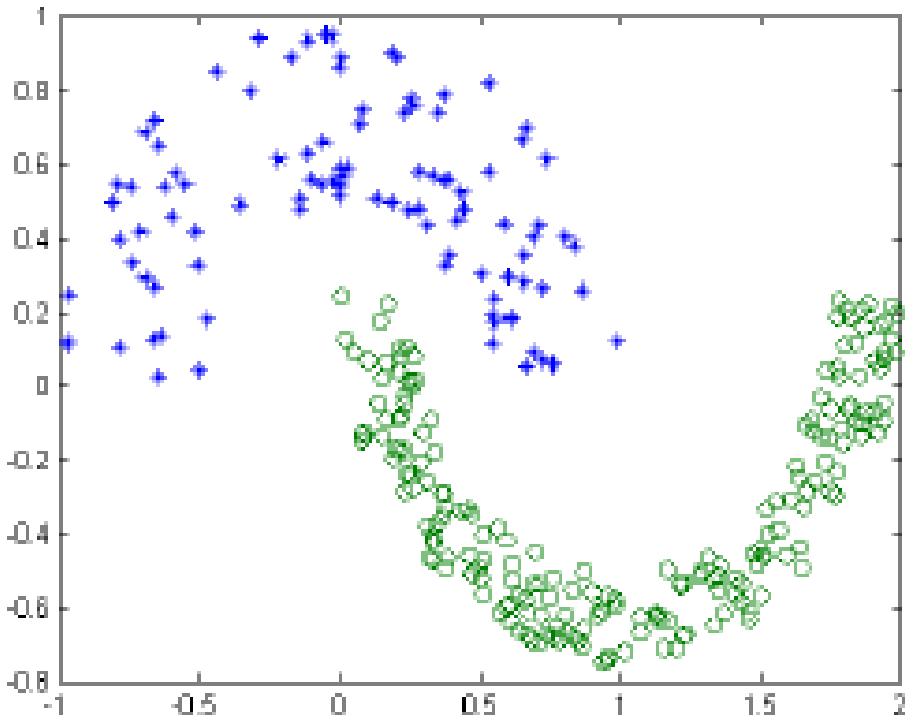
Laufzeitabschätzung

- CSPA: $O(n^2kr)$
- HGPA: $O(nkr)$
- MCLA: $O(nk^2r^2)$
- EM Algorthmus: $O(kNH)$

Wie gut sind die Ergebnisse?

Qualitative Analyse

Mean error rate (%) for the Half-rings dataset.



H	k	EM	CSPA	HGPA	MCLA
5	2	25.4	25.5	50.0	25.4
5	3	24.0	26.2	48.8	25.1
10	2	26.7	28.6	50.0	23.7
10	3	33.5	24.9	26.0	24.2
30	2	26.9	26.2	50.0	26.0
30	3	29.3	26.2	27.5	26.2
50	2	27.2	29.5	50.0	21.1
50	3	28.8	25.0	24.8	24.6

Ende

Vielen Dank für die Aufmerksamkeit