Algorithmische Methoden der Netzwerkanalyse

Marco Gaertler

27. Januar, 2009

Komplexität II – Korrektur

Satz

Das Problem zu einem gegebenen Graphen $G = (V, E)$ zu entscheiden, ob es eine Kantenreduktionsmenge gibt, die höchstens K Elemente enthält (und den Graphen in mindestens drei Cluster aufspaltet), ist $\mathcal{NP}$-vollständig.

Anmerkung: Es gibt eine Konstante $\varepsilon > 0$, so dass es $\mathcal{NP}$-schwer ist das obige, allgemeine Entscheidungsproblem mit einem Faktor von $1 + \varepsilon$ zu approximieren.

Berechnung mittels ILP

Satz

Das folgende ILP berechnet die Editierdistanz eines beliebigen Graphen $G = (V, E)$, wobei n^2 binäre Variablen $X_{u,v}$ und $n + n^2 + n^3$ viele Nebenbedingungen benötigt werden.

$$\text{obj} : \min \sum_{u \in V} \sum_{v \in V} \Big((1 - E_{u,v}) \cdot X_{u,v} + E_{u,v} \cdot (1 - X_{u,v}) \Big)$$

$$\begin{aligned} \forall\, v \in V &: \quad X_{v,v} = 1 \\ \forall\, u, v \in V &: \quad X_{u,v} - X_{v,u} = 0 \\ \forall\, u, v, w \in V &: \quad X_{u,v} + X_{v,w} - 2 \cdot X_{u,w} \leq 1 \end{aligned}$$

Bewertungsfunktion

Definition

Eine Bewertungsfunktion index ordnet jeder Clusterung eines Graphens eine nicht-negative reelle Zahl zu (typischer in $[0, 1]$), die die Qualität der Clusterung bezüglich eines Paradigmas beschreibt.

Verfeinerung

Eine Bewertungsfunktion index für das Paradigma *intra-cluster density vs. inter-cluster sparsity* hat oft die Form:

$$\mathrm{index}(\mathcal{C}) = \frac{f(\mathcal{C}) + g(\mathcal{C})}{\max_{\mathcal{C}'}(f(\mathcal{C}') + g(\mathcal{C}'))} \quad ,$$

wobei f die Dichte der Cluster und g die Dünnheit der Verbindungen mißt.

Interpretation

- Bewertungsfunktionen stellen eine Formalisierung der umgangsprachlich Beschreibung eines Paradigmas da
- je größer der Wert einer Bewertungsfunktion desto besser ist die Clusterung bzgl. des Paradigmas
- die Bestimmung des Nenners, also die Berechnung der maximalen Qualität, ist oft schwer
- $\max_{\mathcal{C}'}(f(\mathcal{C}') + g(\mathcal{C}'))$ wird oft durch eine obere Schranke M ersetzt

Distanz/Ähnlichkeit zum Clustergraphen

Idee: Eine Clusterung ist dann gut, wenn es weniger Kantenkorrekturen bedarf, um den Clustergraphen zu erzeugen, der die Cluster als Zusammenhangskomponenten hat.

Distanz/Ähnlichkeit zum Clustergraphen

Idee: Eine Clusterung ist dann gut, wenn es weniger Kantenkorrekturen bedarf, um den Clustergraphen zu erzeugen, der die Cluster als Zusammenhangskomponenten hat.

$$
\begin{aligned}
f(\mathcal{C}) &= \sum_{C \in \mathcal{C}} |E(C)| \\
g(\mathcal{C}) &= \frac{1}{2} \sum_{C \in \mathcal{C}} \sum_{\substack{u \in C, \\ v \in V \setminus C}} [\{u, v\} \notin E] \\
M &= \binom{n}{2}
\end{aligned}
$$

Der zugehörige Index wird als *Performance* $\mathrm{perf}(\mathcal{C})$ bezeichnet.

Einschränkung auf die Dichte

Einschränkung: Messe nur die Dichte der Cluster und nicht die Dünnheit dazwischen.

$$f(\mathcal{C}) = \sum_{C \in \mathcal{C}} |E(C)| \qquad g(\mathcal{C}) = 0 \qquad \max = |E|$$

Der zugehörige Index wird als *Coverage* $\mathrm{cov}(\mathcal{C})$ bezeichnet.

Einschränkung auf die Dichte

Einschränkung: Messe nur die Dichte der Cluster und nicht die Dünnheit dazwischen.

$$f(\mathcal{C}) = \sum_{C \in \mathcal{C}} |E(C)| \qquad g(\mathcal{C}) = 0 \qquad \max = |E|$$

Der zugehörige Index wird als *Coverage* $\mathrm{cov}(\mathcal{C})$ bezeichnet.

Lemma

Eine Clusterung $\mathcal{C}$ eines Graphens G hat genau dann eine Coverage von 1, wenn die Cluster aus der disjunkten Vereinigung von Zusammenhangskomponenten bestehen.

Schnitte als Maß für Dichte

Definition

Sei $G = (V, E)$ ein Graph und $\mathcal{C}' = (C_1, C_2)$ ein beliebiger Schnitt. Die *Conductance* (*Leitfähigkeit*) $\varphi(\mathcal{C}')$ von $\mathcal{C}'$ ist definiert als:

$$\varphi(\mathcal{C}') := \begin{cases} 1 & , \text{falls } C_1 \in \{\emptyset, V\} \\ 0 & , \text{falls } C_1 \notin \{\emptyset, V\}, \overline{E(\mathcal{C}')} = 0 \\ \frac{\left|\overline{E(\mathcal{C}')}\right|}{\left|\overline{E(\mathcal{C}')}\right| + \min(|E(C_1)|, |E(C_2)|)} & , \text{sonst} \end{cases} .$$

Die *Conductance* $\varphi(G)$ von G ist definiert als

$$\varphi(G) := \min_{C \subseteq V} \varphi((C, V \setminus C)) .$$

Conductance-Qualitätsmaß

Definition

Sei $G = (V, E)$ ein Graph und $\mathcal{C} = (C_1, \ldots, C_k)$ eine Clusterung von G. Der Index *Intra-Cluster Conductance* ist definiert durch:

$$f(\mathcal{C}) = \min_{1 \leq i \leq k} \varphi(G[C_i]) \quad g \equiv 0 \quad M = 1 \ .$$

und der Index *Inter-Cluster Conductance* ist gegeben durch:

$$M = 1 \quad f \equiv 0 \quad g = \begin{cases} 1 & , \text{falls } \mathcal{C} = \{V\} \\ 1 - \max_{1 \leq i \leq k} \varphi((C_i, V \setminus C_i)) & , \text{sonst} \end{cases} .$$

Modularity / Significance

Definition

Sei $G = (V, E)$ ein Graph und $\mathcal{C} = (C_1, \ldots, C_k)$ eine Clusterung von G. Das Maß *Modularity* ist definiert durch:

$$\mathrm{mod}(\mathcal{C}) := \sum_{i=1}^{k} \left(\frac{|E(C_i)|}{|E|} - \frac{1}{4\,|E|^2} \left(\sum_{v \in C_i} \deg(v) \right)^2 \right)$$

Modularity / Significance

Definition

Sei $G = (V, E)$ ein Graph und $\mathcal{C} = (C_1, \ldots, C_k)$ eine Clusterung von G. Das Maß *Modularity* ist definiert durch:

$$\mathrm{mod}(\mathcal{C}) := \sum_{i=1}^{k} \left(\frac{|E(C_i)|}{|E|} - \frac{1}{4\,|E|^2} \left(\sum_{v \in C_i} \deg(v) \right)^2 \right)$$

Es gilt: $\mathrm{mod}(\mathcal{C}) \in [-0.5, 1]$.