Algorithmische Methoden der Netzwerkanalyse

Marco Gaertler

Algorithmic Group
Faculty of Informatics
Universität Karlsruhe (TH)
Research University · founded 1825

20. Januar, 2009

≫ Clustering

Definition

Definition

Sei $G = (V, E)$ ein ungerichteter, einfacher Graph. Ein *Clusterung* $\mathcal{C} = (C_1, \dots, C_k)$ ist eine Partition der Knotenmenge in nicht-leere, disjunkte Teilmengen, d.h.:

$$V = \biguplus_{i=1}^{k} C_i \quad \text{und} \quad C_i \neq \emptyset \text{ für } 1 \leq i \leq k$$

Eine Kante $e = \{u, v\} \in E$ heisst *Intra-Clusterkante*, falls die beide Knoten u, v in der gleichen Teilmenge C_j liegen und sonst *Inter-Clusterkante*.

Notation

- $\vec{E}(V', V'')$ Menge aller Kanten $(u, v) \in E$ mit $u \in V'$ und $v \in V''$
- ungerichteter Fall: $E(V', V'') = \vec{E}(V', V'') \cup \vec{E}(V'', V')$
- $E(\mathcal{C}) = E(C_1, C_1) \uplus \cdots \uplus E(C_k, C_k)$ Menge aller Intra-Clusterkanten
- $\overline{E(\mathcal{C})} = E \setminus E(\mathcal{C})$ Menge aller Inter-Clusterkanten
- $G[C_i] := (C_i, E(C_i, C_i))$ von C_i-induzierte Teilgraph

Falls $k = 1$, dann heißt $\mathcal{C}$ *1-Clusterung*.
Falls $k = n$, dann heißt $\mathcal{C}$ *singeltons*.
Falls $k = 2$, dann wird $\mathcal{C}$ auch *Schnitt* genannt.

Alternative Darstellung als Äquivalenzrelation

Lemma

Sei $G = (V, E)$ ein beliebiger Graph mit $n \geq 1$. Jede Clusterung $\mathcal{C}$ kann eindeutig durch die Äquivalenzrelation

$$u \sim_{\mathcal{C}} v :\iff \exists\, C \in \mathcal{C} : u, v \in C$$

beschrieben werden und die Cluster C_i entsprechen genau den Äquivalenzklassen. Ebenso kann jede Äquivalenzrelation über V kann als Clusterung aufgefasst werden.

Lemma

Lemma

Sei $G = (V, E)$ ein Graph, dann ist die Relation, die durch E repräsentiert wird, genau dann eine Äquivalenzrelation, wenn G die disjunkte Vereinigung von vollständigen Teilgraphen ist.

Definition: Graphen, deren Kantenmenge eine Äquivalenzrelation darstellt, nennt man auch *Clustergraphen*.

Editierdistanz

Definition

Sei $G = (V, E)$ ein Graph. Eine Menge $E'' \subseteq \binom{V}{2}$ von Knotenpaaren wird *Kantenkorrektur* genannt, falls $G' = (V, E')$ mit

$$E' := E \Delta E'' := (E \cup E'') \setminus (E \cap E'')$$

ein Clustergraph ist. Die Kardinalität von E'' wird *Korrekturgröße* genannt. Die *Editierdistanz* von G ist die kleinste mögliche Korrekturgröße.

Editierdistanz: Eigenschaften

- Die Editierdistanz kann als minimale Anzahl von Kanteneinfüge- und Kantenlösungsoperationen angesehen werden, um den Graphen in einen Clustergraphen umzuwandeln.
- Die maximale Korrekturgröße ist $\binom{n}{2}$.
- Die obere Schranke für die Editierdistanz ist $\min(|E| - 1, \binom{n}{2} - |E|)$.

Varianten

Definition

Sei $G = (V, E)$ ein Graph. Eine Kantenkorrektur $E' \subseteq E$ wird als *Kantenreduktionsmenge* und eine Kantenkorrektur $E'' \cap E = \emptyset$ wird als *Kantenerweiterungsmenge* bezeichnet.

Die *Lösch-Editierdistanz* ist die minimale Korrekturgröße von Kantenreduktionsmengen und die *Hinzufüge-Editierdistanz* ist die minimale Korrekturgröße von Kantenerweiterungsmengen.

Komplexität

Satz

Das Problem zu einem gegebenen Graphen $G = (V, E)$ zu entscheiden, ob es eine Kantenerweiterungsmenge der Größe höchstens K gibt, ist polynomiell lösbar.

Erweitert man das Problem, um die Nebenbedingung, dass höchstens k Clusters vorkommen dürfen, so ist es für festes k immer noch polynomiell lösbar.

Komplexität II

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.

Komplexität III

Satz

Das Problem zu einem gegebenen Graphen $G = (V, E)$ zu entscheiden, ob es eine Kantenkorrekturmenge gibt, die höchstens K Elemente enthält, ist $\mathcal{NP}$-vollständig.

Beweis: Reduktion auf *3-Exact 3-Cover*

3-Exact 3-Cover

Problem

Gegeben sei eine Menge C von 3-Tupeln über einer Menge $U = \{u_1, \ldots, u_{3n}\}$ von Elementen, so dass jedes Elemente von U in höchstens 3 der 3-Tupeln auftritt. Bestimme, falls möglich, eine Teilmenge $I \subseteq C$ der Größe n, die U überdeckt.

3-Exact 3-Cover ist $\mathcal{NP}$-vollständig. (Garey&Johnson, 1979, SP2)