Graph-Clustering

Autor und Herausgeber

Alexander Jung*, Department of Computer Science, Aalto University, Espoo, Finland (ORCID 0000-0001-7538-0990)
Editors: Konstantina Olioumtsevits (ORCID 0000-0003-2697-8659), Ekkehard Schnoor (ORCID 0000-0001-8252-1499), Aalto University, Espoo, Finland
*Corresponding author: alex.jung@aalto.fi

Zusammenfassung

Das Graph-Clustering zerlegt die Knoten eines gewichteter Graphen, ein Knoten pro Datenpunkt, in Cluster: Teilmengen von Knoten mit einem großen gesamten Kantengewicht im Inneren und einem kleinen gesamten Kantengewicht zum Rest. Jedes Kantengewicht drückt die Ähnlichkeit zweier Datenpunkte aus. Der Graph ist entweder ein aus Merkmalsvektoren aufgebauter Ähnlichkeitsgraph, wie etwa ein k-Nächste-Nachbarn-Graph, oder durch die Anwendung gegeben, wie etwa durch Zitate verbundene Fachartikel oder durch Freundschaften verbundene Nutzer. Der Schnitt eines Clusters ist das gesamte Kantengewicht der ihn verlassenden Kanten, und der Ratio-Cut eines Clusterings ist die Summe der Schnitte seiner Cluster, jeweils geteilt durch die Anzahl der Knoten im Cluster, was verhindert, dass ein einzelner Knoten abgetrennt wird. Die Partition mit dem kleinsten Ratio-Cut zu finden, ist rechnerisch schwierig, weshalb das Spektrales Clustering sie durch Eigenvektoren der Laplace-Matrix approximiert, das flussbasiertes Clustering durch Netzwerkflüsse und das konvexes Clustering durch einen mit den Kantengewichte gewichteten Strafterm. Zu den Anwendungen gehören die Bildsegmentierung, mit Pixeln als Knoten, sowie die Erkennung von Gemeinschaften in sozialen Netzwerken.

Definition

Betrachtet sei jeder 18 Tag des Jahres 2024 in Krems an der Donau (Österreich), insgesamt 21 Tage, jeder ein Datenpunkt mit einem Merkmalsvektor $\mathbf{x} \in \mathbb{R}^2$, der die Lufttemperatur um 06:00 und um 15:00 (UTC) enthält. Anstatt zu messen, wie weit zwei Tage voneinander entfernt sind, wird erfasst, welche Tage einander ähnlich sind. Jeder Tag ist ein Knoten eines Graphen, und zwei Tage sind durch eine Kante verbunden, wenn einer zu den sechs nächsten Nachbarn des anderen gehört. Jede Kante trägt das Kantengewicht 1 und jedes andere Paar das Kantengewicht 0, sodass der Graph ausschließlich erfasst, welche Tage einander ähnlich sind, und ein Schnitt Kanten zählt (Abb. 1(a)). Das Graph-Clustering zerlegt die Knoten eines solchen Graphen in Cluster: Teilmengen von Knoten mit einem großen gesamten Kantengewicht im Inneren und einem kleinen gesamten Kantengewicht zum Rest des Graphen. Für die 21 Tage bestehen die beiden Cluster aus 10 kalten Tagen und 11 warmen Tagen, die durch sechs Kanten verbunden sind.

Der Graph wird aus Merkmalsvektoren aufgebaut, als ein Ähnlichkeitsgraph wie etwa der k-Nächste-Nachbarn-Graph mit gaußschen Kantengewichte (Hastie et al., 2009, Abschn. 14.5.3; von Luxburg, 2007, Abschn. 2), oder er ist durch die Anwendung gegeben: wissenschaftliche Arbeiten, die durch Zitate verbunden sind, Nutzer eines sozialen Netzwerks, die durch Freundschaften verbunden sind, oder Webseiten, die durch Links verbunden sind. Im zweiten Fall benötigen die Datenpunkte überhaupt keine Merkmalsvektoren, was der Hauptgrund dafür ist, einen Graphen statt eines Merkmalsraums zu clustern.

Formal sind die Datenpunkte die Knoten $i = 1, \ldots, m$ eines gewichteter Graphen $\mathcal{G}$, der ein ungerichteter Graph mit den Kantengewichte $A_{i,i'} = A_{i',i} \geq 0$ ist, wobei $A_{i,i'} = 0$ bedeutet, dass die Knoten i und i' nicht durch eine Kante verbunden sind. Der Schnitt eines Clusters $\mathcal{C} \subseteq \{1, \ldots, m\}$ ist das gesamte Kantengewicht der Kanten, die ihn verlassen,

$$\operatorname{cut}(\mathcal{C}) := \sum_{i \in \mathcal{C}} \sum_{i' \notin \mathcal{C}} A_{i,i'}. \tag{1}$$

Der Schnitt (1) eines einzelnen Knotens ist sein Knotengrad $d^{(i)} = \sum_{i'=1}^{m} A_{i,i'}$. Für einen Vektor $\mathbf{s} \in \{-1, 1\}^m$ mit $s_i = 1$ für $i \in \mathcal{C}$ und $s_i = -1$ andernfalls ist der Schnitt die quadratische Form $\operatorname{cut}(\mathcal{C}) = (1/4)\, \mathbf{s}^\top \mathbf{L}^{(\mathcal{G})} \mathbf{s}$ der Laplace-Matrix $\mathbf{L}^{(\mathcal{G})}$ (von Luxburg, 2007, Abschn. 5.1). Ein Clustering in k Cluster $\mathcal{C}^{(1)}, \ldots, \mathcal{C}^{(k)}$ weist jeden Knoten genau einem davon zu, und die Summe ihrer Schnitte zählt jede Kante zwischen zwei Cluster doppelt. Werden die Cluster allein zur Minimierung dieser Summe gewählt, trennt dies oft einen einzelnen Knoten ab (von Luxburg, 2007, Abschn. 5). Für die 21 Tage beträgt der kleinste Schnitt über alle 2^{20} Zweiteilungen 6 Kanten, und er trennt den einzelnen Tag 20 Januar ab, den Tag mit dem kleinsten Knotengrad.

Der Ratio-Cut teilt den Schnitt jedes Clusters durch seine Knotenanzahl (Hagen and Kahng, 1992),

$$\sum_{c=1}^{k} \frac{\operatorname{cut}(\mathcal{C}^{(c)})}{|\mathcal{C}^{(c)}|}. \tag{2}$$

Einem kleinen Cluster wird ein großer Anteil seines Schnitts angerechnet, sodass der Ratio-Cut (2) Cluster ähnlicher Größe bevorzugt. Der Normalized-Cut teilt stattdessen durch den gesamten Knotengrad des Clusters (Shi and Malik, 2000), sodass ein Cluster anhand seiner Kantengewichte statt anhand seiner Knotenanzahl gemessen wird (von Luxburg, 2007, Abschn. 5). Für die 21 Tage beträgt der kleinste Ratio-Cut über alle Zweiteilungen 1.145, für die kalten und die warmen Tage. Er durchtrennt ebenfalls sechs Kanten, jedoch sechs andere: Elf Partitionen erreichen den kleinsten Schnitt von sechs Kanten, und der Ratio-Cut behält die ausgewogene unter ihnen bei, während der Schnitt allein sie nicht unterscheiden kann. Die sechs von ihm durchtrennten Kanten verbinden Spätwinter- und Frühlingstage mit Herbsttagen, während die sechs Kanten des kleinsten Schnitts alle am 20 Januar zusammentreffen. Das Abtrennen des 20 Januars hat den Ratio-Cut 6.30, das Fünfeinhalbfache. Geordnet nach dem Schnitt ist diese Aufteilung die beste aller Partitionen; geordnet nach

dem Ratio-Cut steht sie an 71983 Stelle. Eine zufällige Partition in 11 und 10 Tage durchtrennt 40 der 74 Kanten, mit dem Schnitt 40 und dem Ratio-Cut 7.64. Über alle 184756 Partitionen dieser Größen beträgt der Ratio-Cut im Mittel 7.39. Eine Partition mit dem kleinsten Ratio-Cut oder Normalized-Cut zu finden, ist rechnerisch schwierig (von Luxburg, 2007, Abschn. 5). Für 21 Knoten lässt sich jede Partition aufzählen, für die 12288 Bildausschnitte eines Fotos nicht.

Verfahren des Graph-Clusterings unterscheiden sich darin, wie sie einen Minimierer approximieren. Das Spektrales Clustering lockert die Nebenbedingung, dass jeder Knoten einem Cluster zugewiesen wird, und löst das relaxierte Problem über Eigenvektoren der Laplace-Matrix (von Luxburg, 2007, Abschn. 5; Ng et al., 2001). Für den Ratio-Cut (2) eines zusammenhängenden Graphen und $k = 2$ ist der relaxierte Minimierer der Fiedler-Vektor (Fiedler, 1973), der Eigenvektor zum zweitkleinsten Eigenwert, der algebraische Konnektivität. Die Knoten werden dann anhand der Vorzeichen seiner Einträge oder am besten der $m-1$ Schwellenwerte entlang seiner sortierten Einträge aufgeteilt. Für die 21 Tage, deren Graph zusammenhängend ist, liefert der beste der 20 Schwellenwerte entlang des Fiedler-Vektors den exakten Minimierer des Ratio-Cuts (Abb. 1(b)). Das Flussbasiertes Clustering bildet für jeden Knoten einen Merkmalsvektor aus Netzwerkflüssen zwischen ausgewählten Knoten und wendet k-Means auf diese Merkmalsvektoren an (SarcheshmehPour et al., 2021). Das Konvexes Clustering nutzt sowohl den Graphen als auch die Merkmalsvektoren der Knoten. Es gibt jedem Knoten seinen eigenen Clusterzentrum und fügt einen mit den Kantengewichte gewichteten Strafterm hinzu, der die Clusterzentren verbundener Knoten zusammenzieht, bis sie zusammenfallen (Hocking et al., 2011). Die Knoten, die sich einen Clusterzentrum teilen, bilden einen Cluster.

Graph-Clustering ist unüberwachtes Lernen: Die Knoten tragen keine Labels, und die gelernte Hypothese ist die Zuweisung jedes Knotens zu einem Cluster-Index. Der Schnitt (1) ist eine Summe über die Kanten statt über die Datenpunkte. Eine Kante kostet ihr Kantengewicht, wenn ihre beiden Knoten verschiedenen Cluster zugewiesen sind, und andernfalls nichts, sodass der Schnitt die verallgemeinerte totale Variation (GTV) der Zuweisung für diejenige Diskrepanz ist, die 1 für zwei verschiedene Indizes und 0 für gleiche beträgt. Ohne den Nenner des Ratio-Cuts (2) wäre es die günstigste Wahl, alle Knoten bis auf einen demselben Cluster zuzuweisen, ganz so, wie ein Trainingsfehler von null durch eine Hypothese erreicht wird, die anderswo nutzlos ist.

Ein Foto wird mittels Graph-Clustering in Regionen unterteilt, mit den Pixeln als Knoten und Kantengewichte, die für nahe beieinander liegende Pixel ähnlicher Farbe groß sind; dies ist das Verfahren zur Bildsegmentierung, mit dem der Normalized-Cut eingeführt wurde (Shi and Malik, 2000). In einem sozialen Netzwerk findet das Graph-Clustering mit den Freundschaften als Kanten Gemeinschaften, also Gruppen von Nutzern mit vielen Freundschaften untereinander und wenigen zum Rest (Fortunato, 2010).

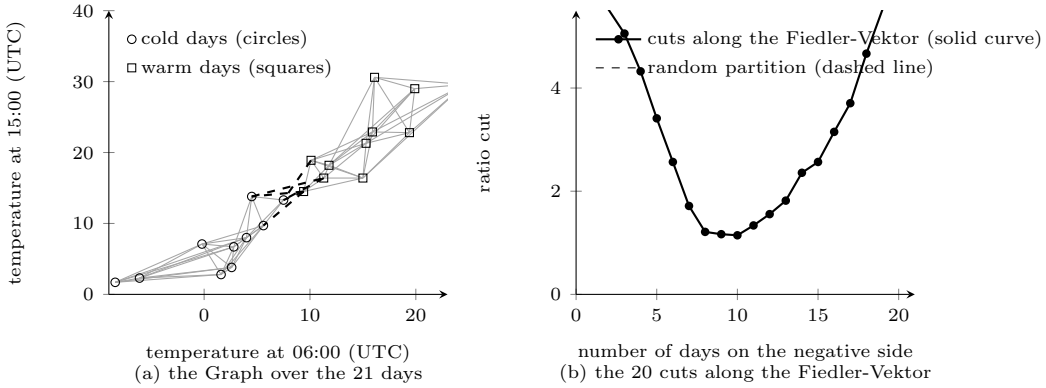


Abbildung 1: (a) Jeder 18 der 365 Tage des Jahres 2024 in Krems, 21 Tage, als die Knoten eines Graphen: Zwei Tage sind verbunden, wenn einer zu den sechs nächsten Nachbarn des anderen gehört, und jede Kante trägt das Kantengewicht 1, sodass ein Schnitt Kanten zählt. Die Partition mit dem kleinsten Ratio-Cut (2), 1.145, besteht aus 10 kalten Tagen (Kreise) und 11 warmen Tagen (Quadrate); 68 der 74 Kanten verlaufen innerhalb eines Clusters (dünne graue Strecken) und sechs werden durchtrennt (gestrichelte Strecke). (b) Der Ratio-Cut der 20 Partitionen, welche die Tage mit den kleinsten Einträgen des Fiedler-Vektors auf eine Seite legen (durchgezogene Kurve): Das Minimum, bei 10 Tagen, ist die Partition in (a). Die gestrichelte Linie ist der Ratio-Cut 7.64 einer zufälligen Partition in 11 und 10 Tage. Daten erzeugt mit `pythondemos/graphclustering.py`

Siehe auch: Clustering, Cluster, Graph, gewichteter Graph, ungerichteter Graph, Kantengewicht, Ähnlichkeitsgraph, k-Nächste-Nachbarn-Graph, Knotengrad, Laplace-Matrix, Fiedler-Vektor, algebraische Konnektivität, Eigenvektor, Spektrales Clustering, flussbasiertes Clustering, konvexes Clustering, k-Means, GTV, Bildsegmentierung, unüberwachtes Lernen.

Querverweise

Literatur

1. Fiedler (1973). Algebraic connectivity of graphs. Czechoslovak Mathematical Journal, 23(2), pp. 298–305.

2. SarcheshmehPour et al. (2021). Flow-Based Clustering and Spectral Clustering: A Comparison. In: 2021 55th Asilomar Conf. Signals, Syst., Comput., pp. 1292–1296.

3. Fortunato (2010). Community detection in graphs. Physics Reports, 486(3–5), pp. 75–174.

4. Hagen and Kahng (1992). New Spectral Methods for Ratio Cut Partitioning and Clustering. IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst., 11(9), pp. 1074–1085.

5. Hocking et al. (2011). Clusterpath: An Algorithm for Clustering using Convex Fusion Penalties. In: Proc. of the 28th International Conference on Machine Learning,.

6. von Luxburg (2007). A tutorial on spectral clustering. Statist. Comput., 17(4), pp. 395–416.

7. Ng et al. (2001). On Spectral Clustering: Analysis and an Algorithm. In: Adv. Neural Inf. Process. Syst., pp. 849–856.

8. Shi and Malik (2000). Normalized Cuts and Image Segmentation. IEEE Trans. Pattern Anal. Mach. Intell., 22(8), pp. 888–905.

9. Hastie et al. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed., Springer Science+Business Media, New York, NY, USA.