DBSCAN (DBSCAN)

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

Dichtebasiertes räumliches Clustering von Anwendungen mit Rauschen (DBSCAN) ist eine Clustering-Methode für Datenpunkte mit numerischen Merkmalsvektoren, die Cluster beliebiger Form findet und deren Anzahl selbst bestimmt. Gegeben ein Radius ε und eine Anzahl m^*, ist ein Datenpunkt ein Kernpunkt, wenn mindestens m^* Datenpunkte innerhalb der euklidischer Abstand ε von ihm liegen, und ein Cluster ist eine maximale Menge von Datenpunkte, die durch Kernpunkte verbunden sind, die höchstens ε voneinander entfernt sind. Datenpunkte, die innerhalb von ε von keinem Kernpunkt liegen, sind Rauschen und gehören zu keinem Cluster. Die Cluster sind die zusammenhängend Komponenten des ε-Nachbarschafts-Graph auf den Kernpunkten, berechnet in einem einzigen Durchlauf, sodass DBSCAN keinen Operator iteriert und kein empirisches Risiko minimiert. Im Gegensatz zu k-Means und einem Gaußsches Mischmodell (GMM), deren Cluster konvex Regionen und Ellipsoide sind, findet DBSCAN längliche und gekrümmte Cluster, wobei der Radius ε und die Anzahl m^* seine Hyperparameter sind.

Definition

Ein Luftbild der Wachau mit Weißenkirchen, Rossatz und Dürnstein entlang einer Donauschleife wird in quadratische Parzellen von 76 m geschnitten, und die 1000 Parzellen, die größtenteils auf Weinbergen liegen, sind Datenpunkte, wobei jede durch einen Merkmalsvektor $\mathbf{x} \in \mathbb{R}^2$ gekennzeichnet ist, der ihre Position enthält: Ost- und Nordwert in km (Abb. 1). Die Weinanbaugebiete, die diese Parzellen bilden, folgen den Flussufern und den Hängen. Sie sind daher länglich und gekrümmt, und ihre Anzahl ist nicht im Voraus bekannt. DBSCAN ist eine Clustering-Methode für Datenpunkte mit numerischen Merkmalsvektoren, die Cluster beliebiger Form findet und deren Anzahl selbst bestimmt: Ein Cluster ist eine Region, in der Datenpunkte dicht liegen, und Datenpunkte, die in keiner dichten Region liegen, sind Rauschen (Ester et al., 1996; Han et al., 2011). DBSCAN clustert nach Dichte, nicht nach Form, und lässt isolierte Datenpunkte als Rauschen zurück. Es ist eine hartes Clustering-Methode in dem Sinne, dass jeder Datenpunkt zu höchstens einem Cluster gehört, aber im Gegensatz zu einer Partition des Datensatz lässt es die Rauschpunkte außerhalb jedes Cluster.

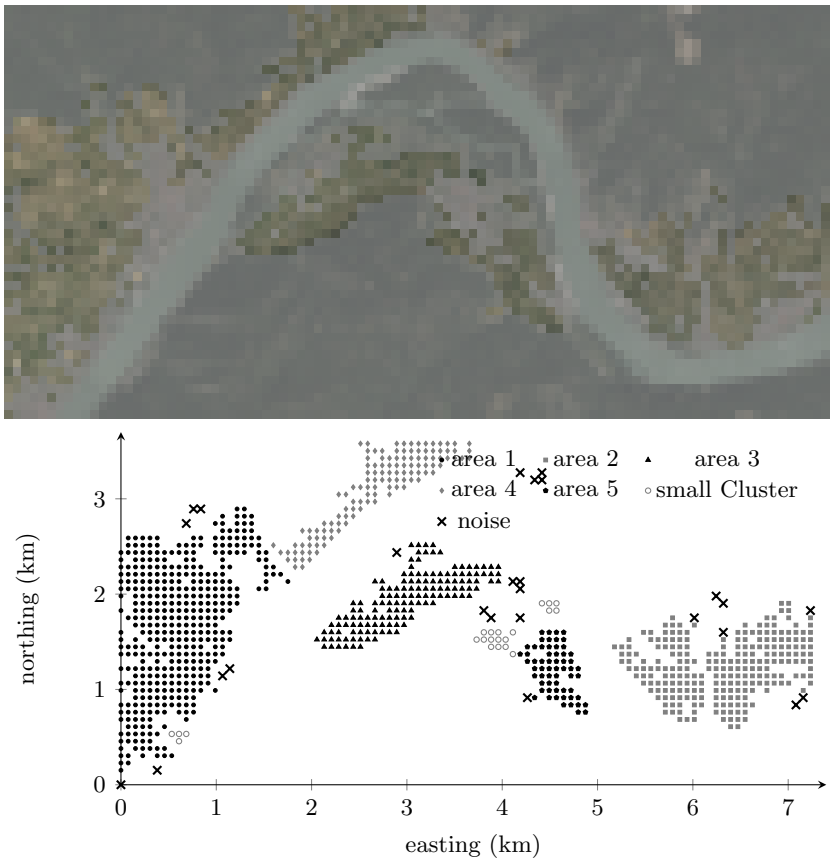


Abbildung 1: Oben: das Luftbild der Wachau in Parzellenauflösung (7.31×3.65 km, 96×48 Parzellen von 76 m); die 1000 Parzellen, die größtenteils auf Weinbergen liegen, behalten ihre Farbe, alle anderen sind abgedunkelt. Unten: dieselben Parzellen als Datenpunkte, jede an ihrer Position gezeichnet und markiert durch das Cluster, dem DBSCAN mit $\varepsilon = 1.5$ Parzellenabständen und $m^* = 4$ sie zuweist: fünf Weinanbaugebiete, drei kleine Cluster und 26 Rauschparzellen. Orthofoto: basemap.at (CC BY 4.0); Parzellen: INVEKOS 2025, AgrarMarkt Austria (CC BY 3.0 AT). Daten generiert von `pythondemos/dbscan.py`

DBSCAN hat zwei Hyperparameter: einen Radius $\varepsilon > 0$ und eine Anzahl m^*. Die ε-Umgebung eines Datenpunkt ist die Menge der Datenpunkte innerhalb der euklidischer Abstand ε von ihm, der Datenpunkt selbst eingeschlossen. Ein Datenpunkt ist ein Kernpunkt, wenn seine ε-Nachbarschaft mindestens m^* Datenpunkte enthält. Es ist ein Randpunkt, wenn es kein Kernpunkt ist, aber in der ε-Nachbarschaft eines solchen liegt, und es ist sonst Rauschen (Abb. 2). Ein Cluster ist eine maximale Menge von Datenpunkte, die durch Kernpunkte zusammenhängend sind: Zwei Kernpunkte, die höchstens ε voneinander entfernt sind, gehören zum selben Cluster, und ein Randpunkt gehört zum Cluster eines

Kernpunkts in seiner ε-Nachbarschaft (Ester et al., 1996). Die Cluster sind daher die zusammenhängend Komponenten des ε-Nachbarschafts-Graph auf den Kernpunkten, des Graph, dessen Kanten Kernpunkte verbinden, die höchstens ε voneinander entfernt sind, wobei die Randpunkte angehängt sind. DBSCAN ist eine Graph-basierte Methode.

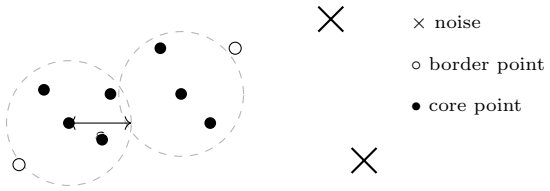


Abbildung 2: Kernpunkte, Randpunkte und Rauschen für $m^* = 4$. Die beiden gestrichelten Kreise sind die ε-Nachbarschaften zweier Kernpunkte; jeder enthält mindestens vier Datenpunkte. Die beiden Randpunkte liegen innerhalb von ε eines Kernpunkts, haben aber weniger als vier Datenpunkte in ihren eigenen Nachbarschaften, und die beiden Rauschpunkte liegen innerhalb von ε von keinem Kernpunkt. Alle gefüllten und offenen Punkte bilden ein Cluster

DBSCAN berechnet die Cluster in einem einzigen Durchlauf über die Datenpunkte. Es wählt einen Kernpunkt aus, der noch keinem Cluster zugewiesen ist, sammelt jeden Datenpunkt, der von ihm aus über Kernpunkte, die höchstens ε voneinander entfernt sind, erreichbar ist, erklärt die Sammlung zu einem Cluster und geht zum nächsten noch nicht zugewiesenen Kernpunkt über. Algorithmus 1 beschreibt den Durchlauf vollständig.

Algorithm 1 DBSCAN (Ester et al., 1996)

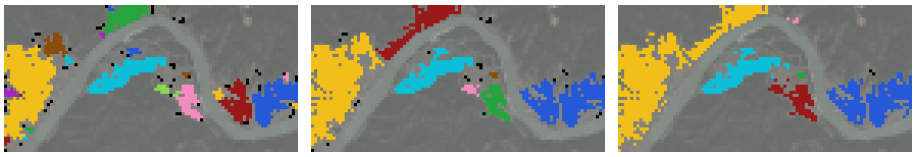
Input: Merkmalsvektoren of the Datenpunkte, radius $\varepsilon > 0$, count m^*
1: mark a Datenpunkt as a core point if at least m^* Datenpunkte, itself included, lie within euklidischer Abstand ε of it
2: mark every Datenpunkt as unassigned
3: **for** each unassigned core point **do**
4: open a new Cluster and assign the core point to it
5: **while** the Cluster holds a core point whose ε-Umgebung has not been collected **do**
6: assign every unassigned Datenpunkt of that ε-Umgebung to the Cluster
7: **end while**
8: **end for**
Output: the Cluster; every Datenpunkt still unassigned is noise

Es wird kein Operator iteriert, sodass die Fixpunktiteration-Ansicht, die auf den Lloyd-Algorithmus und den EM-Algorithmus (EM) Algorithmus zutrifft, hier nicht anwendbar ist. Ein zweiter Durchlauf über das Ergebnis ändert nichts,

und das Ergebnis hängt von keiner Initialisierung ab, sodass DBSCAN keine lokalen Optima hat, in denen es sich verfangen könnte. DBSCAN minimiert kein empirisches Risiko; es ist allein durch seinen Cluster-Begriff definiert.

Auf den Weinbergparzellen mit $\varepsilon = 1.5$ Parzellenabständen und $m^* = 4$ findet DBSCAN 934 Kernpunkte, 40 Randpunkte und 26 Rauschparzellen, gruppiert in acht Cluster, von denen fünf mindestens 20 Parzellen enthalten: die Weinanbaugebiete des Streifens (Abb. 1). Der Fluss, breiter als ε, trennt die Gebiete an seinen beiden Ufern.

Der Radius ε entscheidet, welche Strukturen als Cluster zählen (Abb. 3). Mit $\varepsilon = 1.1$ Parzellenabständen, sodass nur die vier achsenparallelen Nachbarn einer Parzelle erreichbar sind, fallen die Parzellen in 24 Cluster mit 6% Rauschen, und mehrere Gebiete zerbrechen entlang ihrer dünnen Abschnitte. Mit $\varepsilon = 3$ Parzellenabständen bleiben sechs Cluster übrig und keine Parzelle ist Rauschen. Allerdings werden zwei Gebiete, die $\varepsilon = 1.5$ getrennt hält, zusammengeführt. Die Anzahl der Cluster und der Rauschanteil sinken beide, wenn ε wächst. Eine gängige Methode zur Wahl von ε ist, es dort zu platzieren, wo die sortierten euklidische Abstände der Datenpunkte zu ihrem m^*-ten nächsten Nachbar steil anzusteigen beginnen (Ester et al., 1996).



(a) (b) (c)

Abbildung 3: Die Weinbergparzellen von Abb. 1, nach Cluster eingefärbt für $m^* = 4$ und (a) $\varepsilon = 1.1$, (b) $\varepsilon = 1.5$ und (c) $\varepsilon = 3$ Parzellenabstände; Rauschparzellen sind schwarz. Eine Farbe identifiziert ein Cluster nur innerhalb eines Panels, und die Farben unterscheiden sich in der Helligkeit, sodass die Cluster in einem Graustufendruck getrennt bleiben. Orthofoto: basemap.at (CC BY 4.0). Daten generiert von `pythondemos/dbscan.py`

DBSCAN unterscheidet sich in dreierlei Hinsicht von k-Means und von einem Gaußsches Mischmodell (GMM). Erstens ist ein Cluster von k-Means eine konvex Region des Merkmalsraum und ein Cluster eines GMM ist ein Ellipsoid, während ein DBSCAN-Cluster jede Form annehmen kann, die eine Kette benachbarter Datenpunkte nachzeichnen kann. Auf den Weinbergparzellen teilt k-Means mit fünf Cluster ein Gebiet auf, und eines seiner Cluster verbindet zwei Gebiete, da ein Gebiet, das sich entlang des Flusses krümmt, nicht konvex ist. Zweitens ist die Anzahl der Cluster ein Input von k-Means und von einem GMM, aber ein Output von DBSCAN, zum Preis der beiden Hyperparameter ε und m^*. Drittens weisen k-Means und ein GMM jeden Datenpunkt einem Cluster zu, während DBSCAN Ausreißer als Rauschen belässt. DBSCAN teilt seine graphenbasierte Sichtweise eines Cluster mit Spektrales Clustering. Diese Methode erstellt einen Ähnlichkeitsgraph, einen Graph, dessen Kanten ähnliche

Datenpunkte verbinden, und clustert seine Knoten anhand der Eigenvektoren der Laplace-Matrix, einer Matrix, die den Graph kodiert, anstatt anhand der zusammenhängend Komponenten des ε-Nachbarschafts-Graph.

Siehe auch: Clustering, Cluster, hartes Clustering, k-Means, GMM, Spektrales Clustering, Graph, Umgebung, euklidischer Abstand, Ausreißer, Hyperparameter.

Querverweise

Literatur

1. Ester et al. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, pp. 226–231.

2. Han et al. (2011). Data Mining: Concepts and Techniques. 3rd ed., Morgan Kaufmann, Waltham, MA, USA.