DBSCAN: Density-Based Spatial Clustering of Applications with Noise

Presented by Wondong Lee

Written by M.Ester, H.P.Kriegel, J.Sander and Xu.
Why Density-Based Clustering?

- Results of k-means algorithm for $k = 4$

→ The result is not satisfiable!!
DBSCAN

- Relies on a density-based notion of cluster
- Discovers clusters of arbitrary shape in spatial databases with noise

Basic Idea
- Group together points in high-density
- Mark as outliers ➔ points that lie alone in low-density regions
DBSCAN

- Local point density at a point \( p \) defined by two parameters

\begin{align*}
(1) \quad & \varepsilon \rightarrow \text{radius for the neighborhood of point } p:
\varepsilon\text{-Neighborhood: all points within a radius of } \varepsilon \text{ from the point } p \\
& N_\varepsilon(p) := \{q \in \text{data set } D \mid \text{dist}(p, q) \leq \varepsilon\}
\end{align*}

\begin{align*}
(2) \quad & \text{MinPts } \rightarrow \text{minimum number of points in the given neighborhood } N(p)
\end{align*}
High Density?

- $\epsilon$-Neighborhood of an point contains at least $\text{MinPts}$

$\epsilon$-Neighborhood of $p$
$\epsilon$-Neighborhood of $q$

Q. When $\text{MinPts} = 4$?

*Density of $p$ is “high”*

*Density of $q$ is “low”*
Core, Border & Outlier

- Three category for each point
  - Core point: if its density is high
  - Border point: density is low (but in the neighborhood of a core point)
  - Noise point: any point that is not a core point nor a border point

MinPts = 5
Density-Reachability

• Directly density-reachable
  • A point $q$ is **directly density-reachable** from a point $p$:
    - If $p$ is a core point and $q$ is in $p$’s $\varepsilon$-neighborhood

Q. $p$ is directly density-reachable from $q$?
No, why?

Q. Density-reachability is **asymmetric**

Minpts = 4
Density-Reachability

- Density-reachable
  - A point $p$ is **density-reachable** from a point $q$ if there is a chain of points $p_1, \ldots, p_n$, with $p_1 = q$, $p_n = p$ such that $p_{i+1}$ is directly density-reachable from $p_i$

  **Q.** $q$ is density-reachable from $p$?  
  **No, why?**

$\text{MinPts} = 7$
Density-Connectivity

- Density-connected
  - A pair of points p and q are density-connected
    - If they are commonly density-reachable from a point o

Q. o is density-reachable from p? Yes, why?

Q. Density-connectivity is symmetric

MinPts = 7
Formal Description of Cluster

- Given a data set $D$, parameter $\varepsilon$ and $MinPts$,

- A cluster $C$ is a subset of $D$ satisfying two criteria:
  
  - **Maximality**
    - $\forall p, q$ if $p \in C$ and if $q$ is density-reachable from $p$, then also $q \in C$
  
  - **Connectivity**
    - $\forall p, q \in C$, $p$ and $q$ are density-connected

- **Note**: cluster contains core points as well as border points
Parameter
  • $\varepsilon = 2$, $MinPts = 3$

```latex
\textbf{if} p \textbf{is not classified} \textbf{then} \\
\quad \textbf{if} p \textbf{is a core-point} \textbf{then} \\
\quad \quad \text{collect all points density-reachable from } p \\
\quad \quad \text{and assign them to a new cluster.} \\
\textbf{else} \\
\quad \text{assign } p \text{ to NOISE} \\
\textbf{else} \\
\quad \text{assign } p \text{ to NOISE}
```

Parameter

- $\varepsilon = 2$, $MinPts = 3$

\[
\begin{align*}
\forall p \in D \text{ do} \\
\quad \text{if } p \text{ is not yet classified then} \\
\quad \quad \text{if } p \text{ is a core-point then} \\
\quad \quad \quad \text{collect all points density-reachable from } p \\
\quad \quad \quad \text{and assign them to a new cluster.} \\
\quad \text{else} \\
\quad \quad \text{assign } p \text{ to NOISE}
\end{align*}
\]
Parameter

- $\varepsilon = 2$, $MinPts = 3$

\[ \forall p \in D \text{ do}
\begin{align*}
&\text{if } p \text{ is not yet classified then } \\
&\quad \text{if } p \text{ is a core-point then} \\
&\quad \quad \text{collect all points density-reachable from } p \\
&\quad \quad \text{and assign them to a new cluster.} \\
&\text{else} \\
&\quad \text{assign } p \text{ to NOISE}
\end{align*} \]
Example

Original Points

Point types: core, border and outliers

$\varepsilon = 10$, $\text{MinPts} = 4$
When DBSCAN Works Well

- Resistant to Noise
- Can handle clusters of different shapes and sizes
When DBSCAN Does Not Work Well

- Cannot handle varying densities

Original Points

\((\varepsilon = 9.92, \text{MinPts}=4)\)
When DBSCAN Does Not Work Well

- Sensitive to parameters
K-means VS DBSCAN

(1) When k = 3
MinPts = 4

(2) When k = 2
MinPts = 3

Winner is DBSCAN
K-means VS DBSCAN

(1) When $k = 2$
MinPts = 3

Winner is K-means
Thank you for attention

Any Questions?
Reference

- **Comparing Clustering Algorithm**
  - http://www.cise.ufl.edu/~jmishra/clustering/DataMiningPresentation.ppt

- **Density-Based Clustering**
  - http://www.cse.buffalo.edu/faculty/azhang/cse601/density-based.ppt