Mineração de Dados

PUC-SP

Prof. Dr. Daniel Rodrigues da Silva

Mineração de Dados

DBSCAN

PUC-SP

Bibliografia básica:

Introdução à mineração de dados : conceitos básicos, algoritmos e aplicações. Leandro Nunes de Castro e Daniel Gomes Ferrari. – São Paulo : Saraiva, 2016.

Inteligência Artificial - Uma Abordagem de Aprendizado de Máquina . André Carlos Ponce de Leon Ferreira et al. 2 Ed. LTC, 2024.

DBSCAN (Agrupamento Espacial de Aplicativos com Ruído Baseado em Densidade), é um método de agrupamento eficiente que agrupa pontos que estão próximos uns dos outros no espaço de dados.

Ao contrário de outros métodos de agrupamento, o DBSCAN não pede que você determine previamente o número de agrupamentos (k).

O método funciona definindo clusters como regiões densas separadas por regiões de menor densidade. Dessa forma, DBSCAN descubre grupos de formato arbitrário e identifica exceções como ruído.

PUC-SP

DBSCAN funciona sobre três pilares:

Pontos principais: São pontos que têm um número mínimo de outros pontos (MinPts) dentro de uma distância especificada (ϵ).

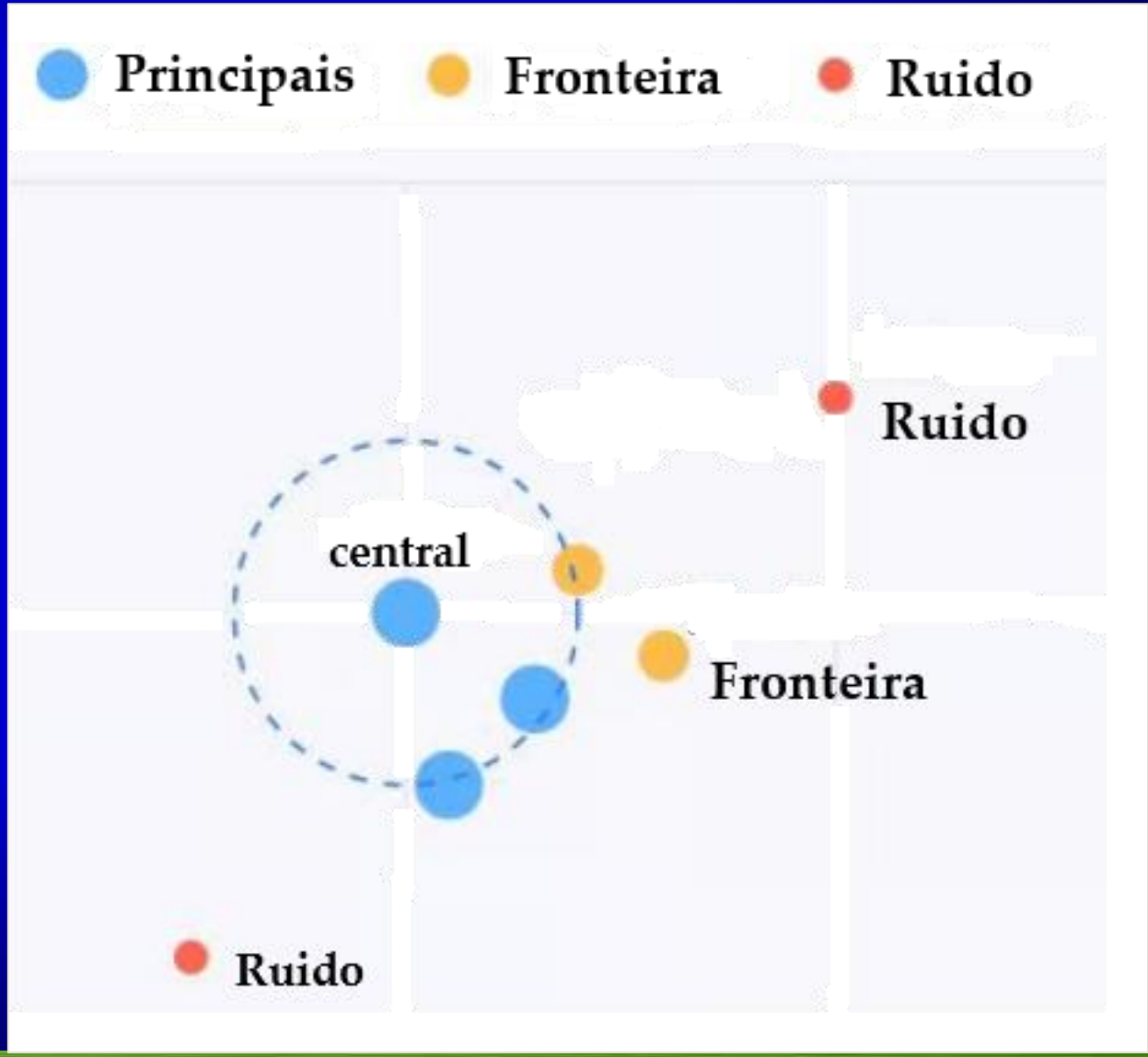
Pontos de fronteira: São pontos que estão dentro da distância ϵ de um ponto central, mas não têm MinPts vizinhos.

Pontos de ruído: Esses são pontos que não são nem pontos principais nem pontos de fronteira. Eles não estão próximos o suficiente de nenhum cluster para serem incluídos.

Os pontos centrais (azul) formam o coração dos clusters, os pontos de borda (laranja) estão na borda dos clusters e os pontos de ruído (vermelho) estão isolados.

O DBSCAN usa dois parâmetros principais:
ϵ (epsilon): A distância máxima entre dois pontos para que eles sejam considerados vizinhos.

MinPts: O número mínimo de pontos necessários para formar uma região densa.



PUC-SP

Ajustando esses parâmetros, dá para controlar como o método define os clusters, permitindo que ele se adapte a diferentes tipos de conjuntos de dados e requisitos de agrupamentos.

PUC-SP

Funcionamento do DBSCAN:

O DBSCAN examina a vizinhança de cada ponto no dataset. O método faz o processo passo a passo para identificar clusters com base na densidade dos pontos de dados:

Seleção de parâmetros

- Escolha ϵ: Distância máxima entre dois pontos para que sejam considerados vizinhos.
- Selecione MinPts: Número mínimo de pontos para formar uma região densa.

PUC-SP

- **Selecione um ponto inicial:** O método inicia com um ponto arbitrário não visitado do dataset.

- **Examine a vizinhança:** O método recupera todos os pontos dentro da distância ϵ do ponto inicial indicado.
  Se o número de pontos vizinhos for menor que MinPts, o ponto será rotulado como ruído (por enquanto).
  Se houver pelo menos MinPts pontos a uma distância ε, o ponto será marcado como um ponto central e um novo cluster será formado.

- **Expandir o cluster:** Todos os vizinhos do ponto central são adicionados ao cluster.
  Para cada um desses vizinhos:
  - Se for um ponto central, seus vizinhos serão adicionados ao cluster recursivamente.
    Se não for um ponto central, ele será marcado como um ponto de borda e a expansão será interrompida.

- **Repetir o processo**: O método passa para o próximo ponto não visitado do dataset.
  As etapas 3 e 4 são repetidas até que todos os pontos sejam visitados.

PUC-SP

- **Finalizar clusters**: Após o processamento de todos os pontos, o algoritmo identifica todos os clusters.
  Os pontos inicialmente rotulados como ruído podem agora ser pontos de borda se estiverem a uma distância ε de um ponto central.

- **Ruído de manuseio**: Os pontos que não pertencem a nenhum cluster ficam classificados como ruído.

Este processo faz com que o DBSCAN forme grupos de formas arbitrárias e identifique outliers de forma bem eficaz.

A capacidade do método em encontrar clusters sem exigir a priori o número de grupos é um de seus pontos fortes.

É importante observar que a escolha de ε e MinPts pode afetar significativamente os resultados do agrupamento.

Na próxima seção, discutiremos como escolher esses parâmetros de forma eficaz e apresentaremos métodos como o gráfico de distância k para a seleção de parâmetros.

Como já dito, a eficiência do DBSCAN depende bastante da escolha de seus parâmetros principais: ϵ e MinPts.

Escolha de ϵ:

1. **Use o conhecimento do domínio**: O conhecimento prévio do domínio pode ser crucial para uma boa escolha.

2. **Gráfico de distância K**: Abordagem sistemática:
   Calcule a distância até o k-ésimo vizinho mais próximo de cada ponto (onde k = MinPts).
   Trace essas distâncias k em ordem crescente.
   Procure um "cotovelo" no gráfico - um ponto em que a curva começa a se nivelar.

PUC-SP

Escolha de MinPts:

1. **Regra geral**: Uma boa prática é definir MinPts = 2 * num_features. Em que num_features é o número de dimensões em seu dataset.

2. Consideração de ruído: Se seu dataset tiver ruído ou se você quiser detectar clusters menores, pode ser melhor diminuir o MinPts.

3. Tamanho do dataset: Para dataset maiores, talvez seja necessário aumentar o MinPts para evitar a criação de vários clusters de pequeno tamanho.

PUC-SP

PS. Como a escolha dos parâmetros pode afetar de forma significativa o desempenho do método, em geral, é interessante fazer simulações com alguns valores diferentes e avaliar os resultados obtidos visando à melhor opção para seu dataset específico.