Detecção de Anomalias

ACH5504 – Mineração de Dados

Notas de aulas baseadas no livro

"*Introduction to Data Mining*"

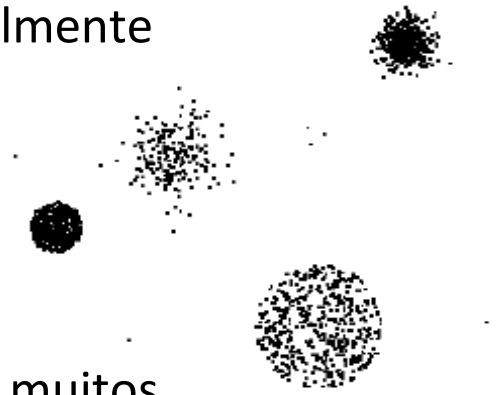
Tan, Steinbach, Karpatne, Kumar

Resumo

- Importância e causas de anomalias ou outliers
- Métodos de detecção de anomalias

Detecção de Anomalia/Outlier

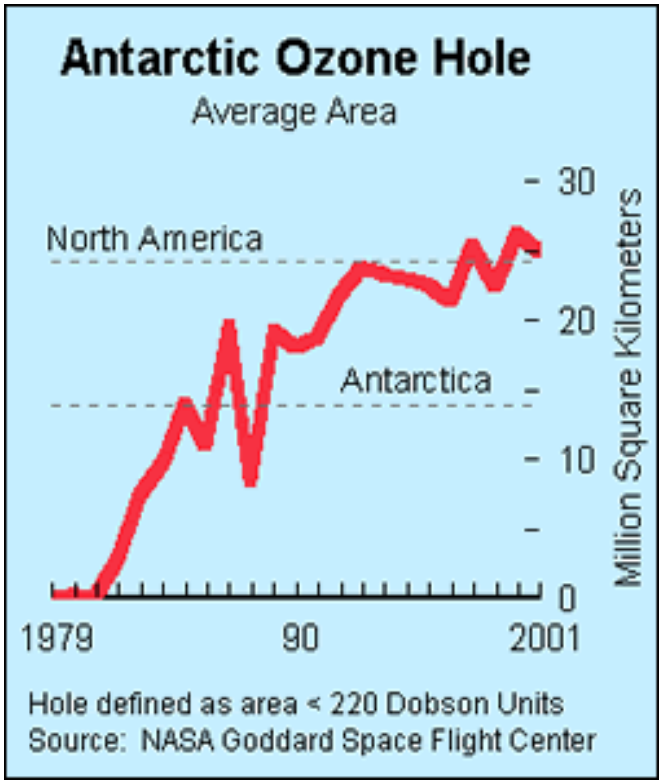
- O que são anomalias/outliers?
  - O conjunto de pontos de dados que são consideravelmente diferentes do restante dos dados
- A implicação natural é que as anomalias são relativamente raras
  - Um em cada mil ocorre com frequência se você tem muitos dados
  - O contexto é importante, por exemplo, as temperaturas congelantes em julho
- Pode ser importante ou não
  - 2 metros de altura com 2 anos de idade
  - Pressão arterial excepcionalmente alta



Importância da detecção de anomalias

História do esgotamento do ozônio

- Em 1985, três pesquisadores (Farman, Gardinar e Shanklin) ficaram intrigados com os dados coletados pelo British Antarctic Survey mostrando que os níveis de ozônio para a Antártida caíram 10% abaixo dos níveis normais
- Por que o satélite Nimbus 7, que tinha instrumentos a bordo para registrar os níveis de ozônio, não registrou concentrações de ozônio igualmente baixas?
- As concentrações de ozônio registradas pelo satélite eram tão baixas que estavam sendo tratadas como outliers por um programa de computador e descartadas!
-



Fontes:
http://exploringdata.cqu.edu.au/ozone.html
http://www.epa.gov/ozone/science/hole/size.html

Causas de anomalias

- Dados de diferentes classes
  - Medindo os pesos das laranjas, mas algumas toranjas são misturadas junto
- Variação natural
  - Pessoas excepcionalmente altas
- Erros de dados
  - Criança de 2 anos com 100 kg de peso

Diferenças entre ruído e anomalias

- O ruído é errôneo, talvez aleatório, valores ou objetos contaminadores
  - Peso registrado incorretamente
  - Toranjas misturadadas com as laranjas

- O ruído não produz necessariamente valores ou objetos incomuns
- O ruído não é interessante
- Anomalias podem ser interessantes se não forem resultado do ruído
- O ruído e as anomalias são conceitos relacionados mas distintos

Questões gerais: Número de atributos

- Muitas anomalias são definidas em termos de um único atributo
  - Altura
  - Forma
  - Cor
- Pode ser difícil encontrar uma anomalia usando todos os atributos
  - Atributos ruidosos ou irrelevantes
  - Objeto é apenas anômalo em relação a alguns atributos
- No entanto, um objeto pode ser não anômalo em alguns atributos

Questões gerais: Pontuação de anomalias

- Muitas técnicas de detecção de anomalias fornecem apenas uma categorização binária
  - Um objeto é uma anomalia ou não é
  - Isto é especialmente verdadeiro para abordagens baseadas em classificação
- Outras abordagens atribuem uma pontuação a todos os pontos
  - Esta pontuação mede o grau em que um objeto é uma anomalia
  - Isso permite que os objetos sejam classificados
- No final, muitas vezes você precisa de uma decisão binária
  - Esta transação de cartão de crédito deve ser sinalizada?
  - Ainda útil para ter uma pontuação
- Quantas anomalias existem nos dados?

Outras questões para a detecção de anomalias

- Encontre todas as anomalias de uma só vez ou uma de cada vez
  - Afogando
  - Mascarando
- Avaliação
  - Como medir o desempenho?
  - Situações supervisionadas versus sem supervisão
- Eficiência
- Contexto

Tipos de problemas de detecção de anomalias

- Dado um conjunto de dados D, encontrar todos os pontos de dados **x** $\in$ D com pontuações de anomalia maior do que um valor t

- Dado um conjunto de dados D, encontrar todos os pontos de dados **x** $\in$ D que tem n maiores pontuações de anomalia

- Dado um conjunto de dados D, contendo principalmente pontos de dados normais (mas não rotulados) e um ponto de teste **x**, calcule a pontuação de anomalia de **x** em relação a D

Detecção de anomalias baseadas em modelos

- Crie um modelo para os dados e veja
  - no caso de analise não supervisionada
    - Se as anomalias são aqueles pontos que não se encaixam bem
    - Se as anomalias são aqueles pontos que distorcem o modelo
    - Exemplos:
      - Distribuição estatística
      - Grupos ou clusters
      - Regressão
      - Geometria
      - Gráfico
  - no caso de analise supervisionada
    - Se anomalias são consideradas como uma classe rara
    - Se precisar ter mais dados de treinamento para classificação

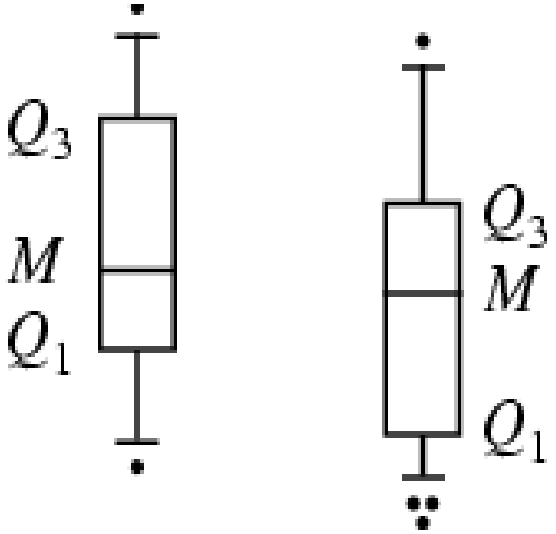
Técnicas adicionais de detecção de anomalias

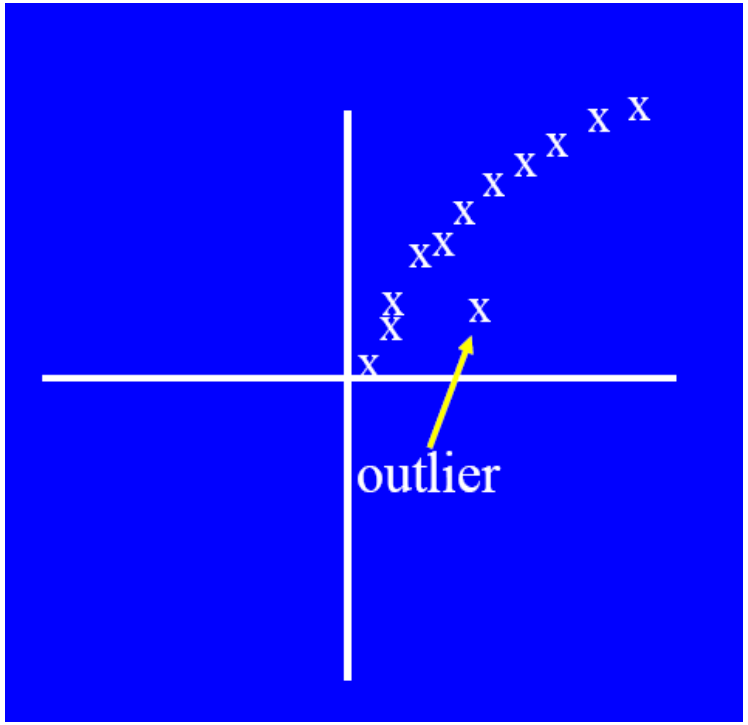
- Baseadas em proximidade
  - Anomalias são pontos distantes de outros pontos
  - Pode detectar isso graficamente em alguns casos
- Baseadas em densidade
  - Outliers são pontos de baixa densidade
- Combinação de padrões
  - Crie perfis ou modelos de eventos ou objetos atípicos, mas importantes
  - Algoritmos para detectar esses padrões geralmente são simples e eficientes

Abordagens visuais

- Boxplots ou gráficos de disperssão

- Limitações
  - Não automática
  - Subjetiva



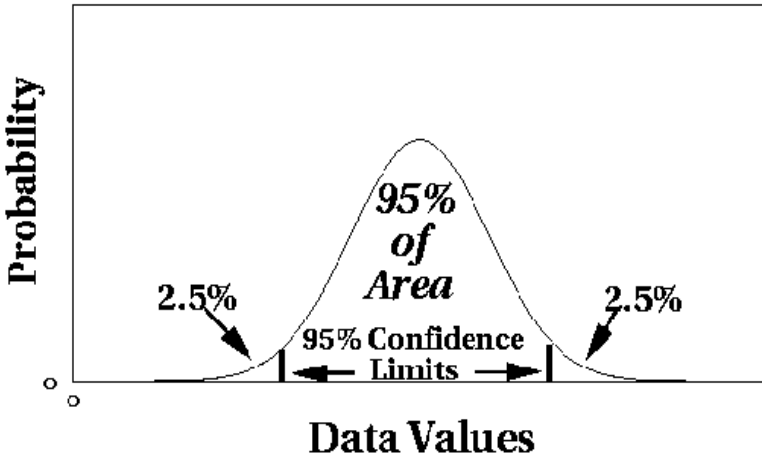


Abordagens estatísticas

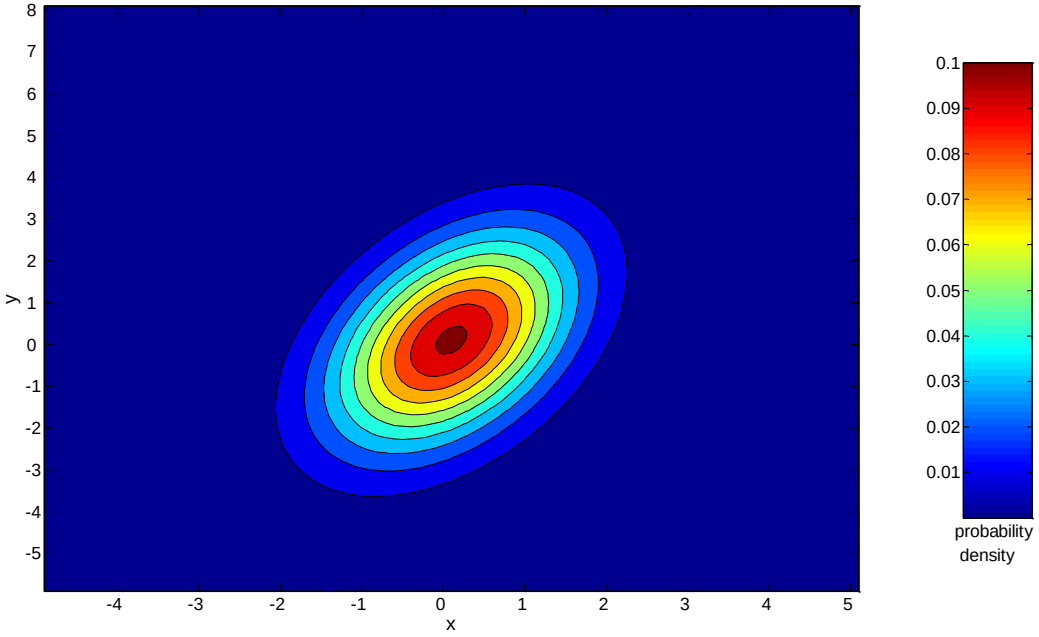
Definição probabilística de um outlier: Um outlier é um objeto que tem uma baixa probabilidade em relação a um modelo de distribuição de probabilidade dos dados.

- Normalmente assume um modelo paramétrico descrevendo a distribuição dos dados (por exemplo, distribuição normal)
- Podemos aplicar um teste estatístico que depende
  - Distribuição de dados
  - Parâmetros de distribuição (por exemplo, média, variação)
  - Número de outliers esperados (limite de confiança)
- Problemas em
  - Identificação da distribuição de um conjunto de dados
    - Distribuição de cauda pesada
  - Número de atributos
  - Os dados são uma mistura de distribuições?

Distribuições normais



Gaussianaa unidimensional



Gaussiana bidimensional

Teste de Grubbs

- Detecte outliers em dados univariados
- Suponha que os dados vêm da distribuição normal
- Detecta um outlier de cada vez, remova o outlier, e repete
  - H_0: Não há nenhum outlier nos dados
  - H_A: Há pelo menos um outlier
- Estatística de teste de Grubbs:

$$G = \frac{\max\left|X - \overline{X}\right|}{s}$$

- Rejeite H_0 se:

$$G > \frac{(N-1)}{\sqrt{N}} \sqrt{\frac{t^2_{(\alpha/N,N-2)}}{N-2+t^2_{(\alpha/N,N-2)}}}$$

É o limite superior da distribuição t de Student para N-2 graus de liberdade e nível de confiança α/2N

Abordagem baseada em probabilistica

- Suponha que o conjunto de dados D contenha amostras de uma mistura de duas distribuições de probabilidade:
  - M (distribuição de maioria)
  - A (distribuição anômala)
- Abordagem geral:
  - Inicialmente, suponha que todos os dados pertencem a M
  - Deixe $L_t(D)$ ser a log de probabilidade de D no momento t
  - Para cada ponto x_t que pertence a M, movê-lo para A
    - Deixe $L_{t+1}(D)$ ser a log de probabilidade nova.
    - Calcule a diferença, $\Delta = L_t(D) - L_{t+1}(D)$
    - Se $\Delta > c$ (um valor limite), então x_t é declarado como anomalia e transferido permanentemente de M para A

Abordagem baseada em probabilistica

- Distribuição de dados, D = (1 – λ) M + λ A
- M é uma distribuição de probabilidade estimada a partir de dados
  - Pode ser baseado em qualquer método de modelagem (Baías ingênuas, entropia máxima, etc)
- A é inicialmente assumido como uma distribuição uniforme
- Probabilidade no tempo t:

$$L_t(D) = \prod_{i=1}^{N} P_D(x_i) = \left((1-\lambda)^{|M_t|} \prod_{x_i \in M_t} P_{M_t}(x_i) \right) \left(\lambda^{|A_t|} \prod_{x_i \in A_t} P_{A_t}(x_i) \right)$$

$$LL_t(D) = |M_t| \log(1-\lambda) + \sum_{x_i \in M_t} \log P_{M_t}(x_i) + |A_t| \log \lambda + \sum_{x_i \in A_t} \log P_{A_t}(x_i)$$

Pontos fortes e fracos das abordagens estatísticas

- Fundação matemática firme
- Pode ser muito eficiente
- Bons resultados se a distribuição for conhecida
- Em muitos casos, a distribuição de dados pode não ser conhecida
-
- Para dados de alta dimensão, pode ser difícil estimar a verdadeira distribuição
- Anomalias podem distorcer os parâmetros da distribuição
-

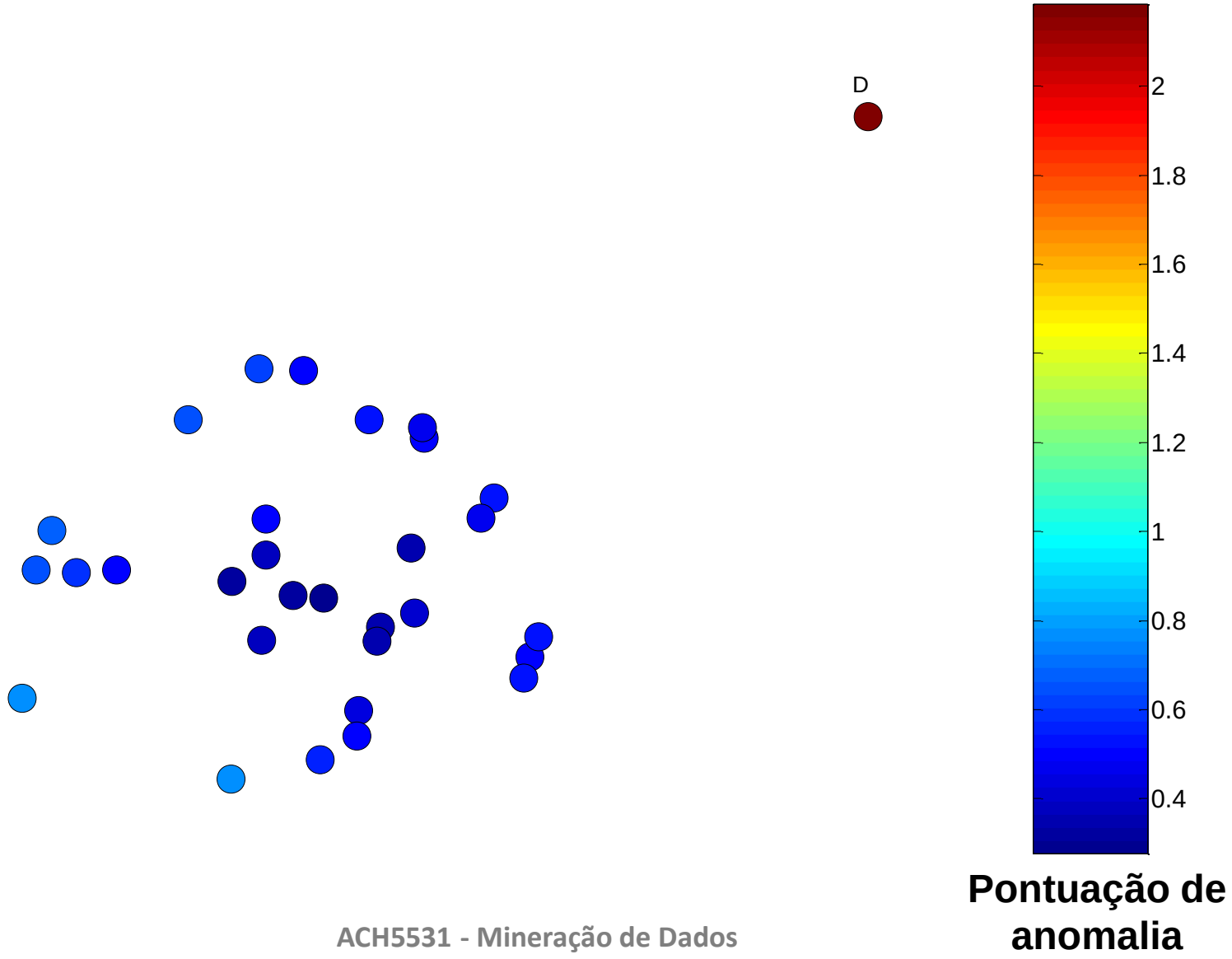
Abordagens baseadas na distância

- Várias técnicas diferentes

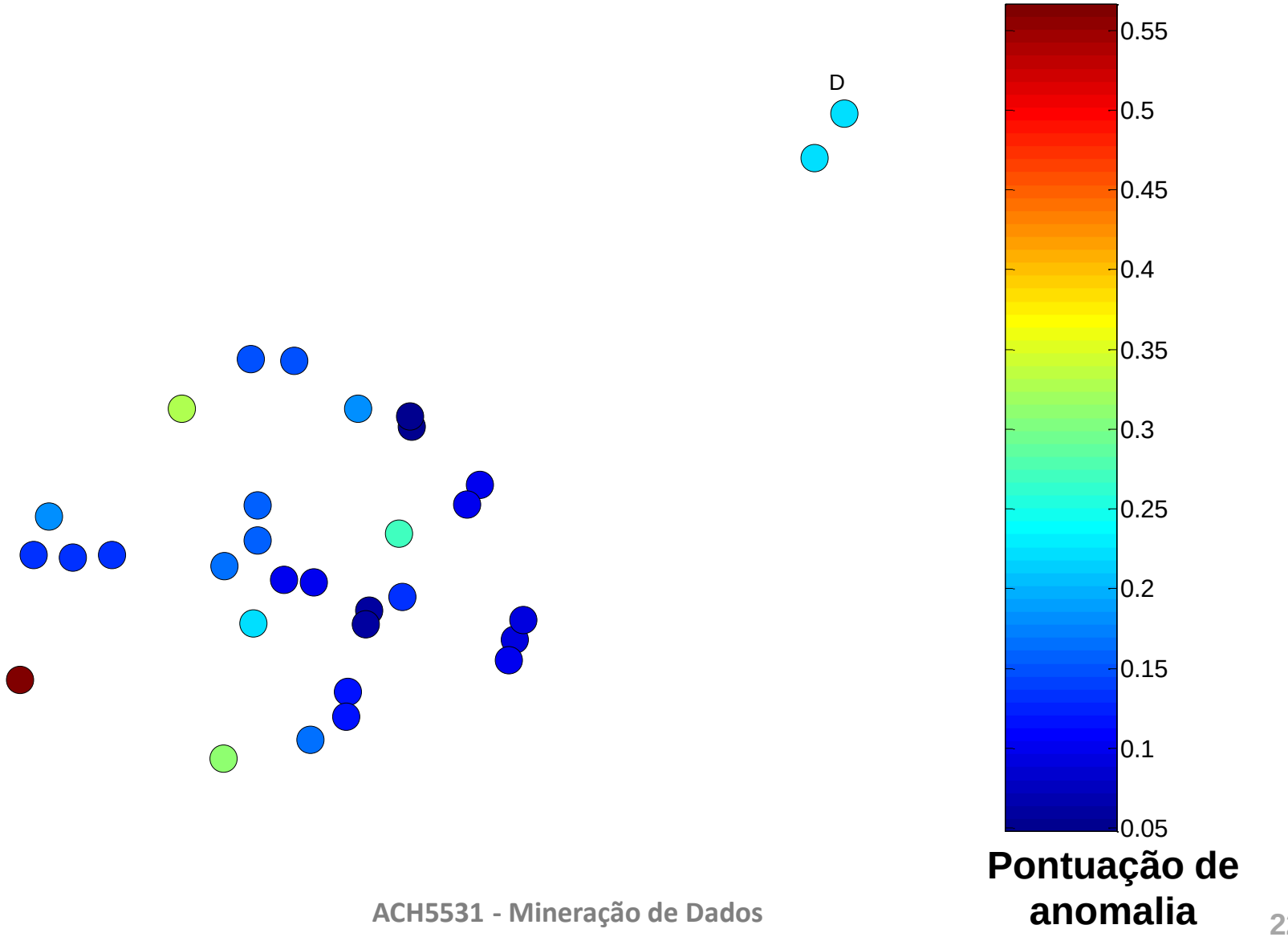
- Um objeto é um outlier se uma fração especificada dos objetos é maior do que uma distância especificada (Knorr, Ng 1998)
  - Algumas definições estatísticas são casos especiais desta definição

- A pontuação de outlier de um objeto é a distância para o seu k-ésimo vizinho mais próximo

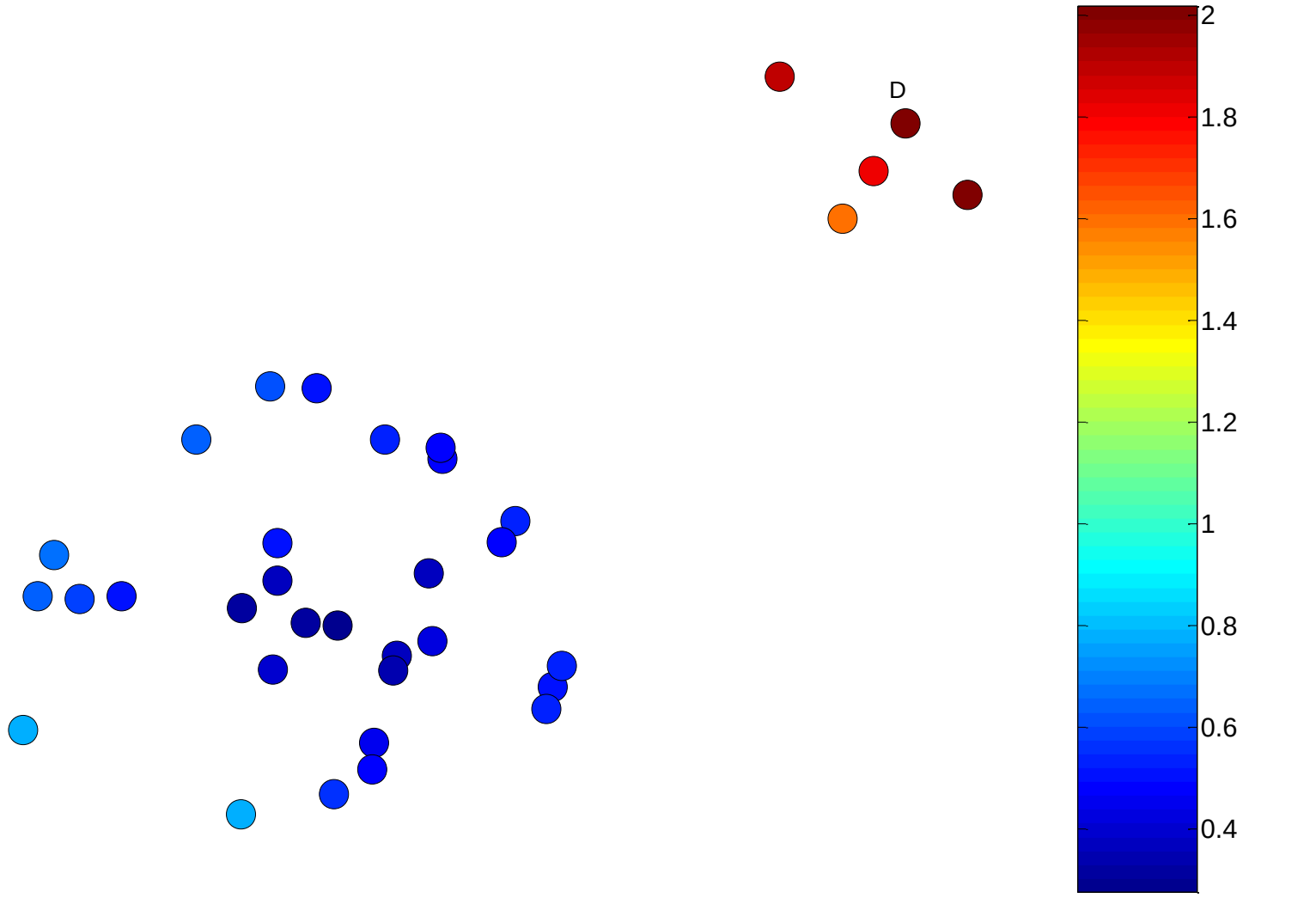
Um vizinho mais próximo - um outlier



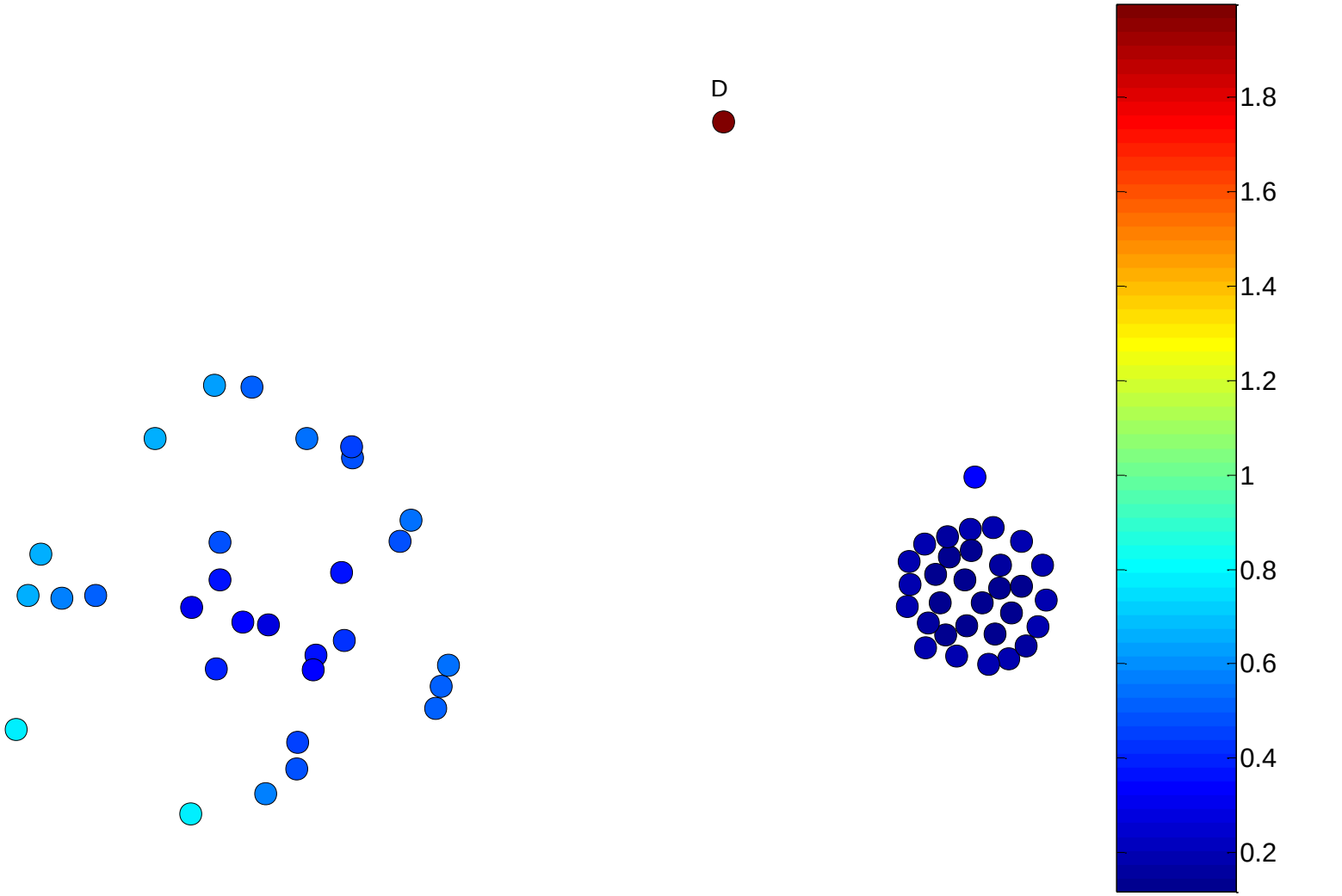
Um vizinho mais próximo - Dois outliers



Cinco vizinhos mais próximos - Pequeno cluster



Cinco vizinhos mais próximos - densidade diferente



Pontos fortes e fracos de abordagens baseadas na distância

- Simples
- Caras – $O(n^2)$
- Sensível aos parâmetros
- Sensível às variações na densidade
- A distância torna-se menos significativa no espaço de alta dimensão

Abordagens baseadas em densidade

- **Anomalia baseada em densidade:** A pontuação de anomalia de um objeto é o inverso da densidade em torno do objeto.
  - Pode ser definido em termos dos k vizinhos mais próximos
  - Uma definição: Inverso de distância ao k-ésimo vizinho
  - Outra definição: Inverso da distância média para k vizinhos
  - Definição DBSCAN

- Se existem regiões de diferentes densidades, esta abordagem pode ter problemas

Densidade relativa

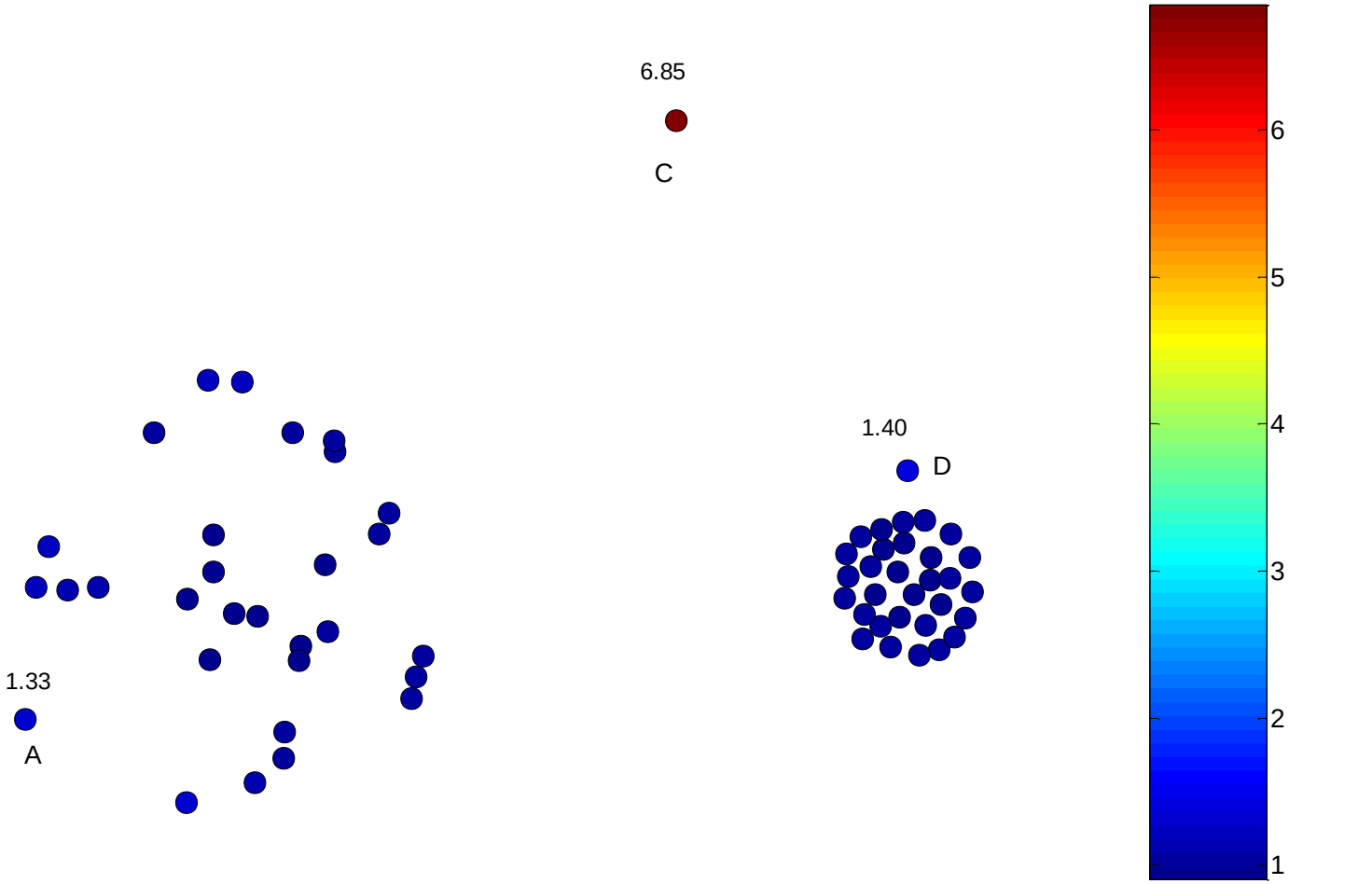
- Considere a densidade de um ponto em relação ao seus k vizinhos mais próximos
-

$$average\ relative\ density(\mathbf{x}, k) = \frac{density(\mathbf{x}, k)}{\sum_{\mathbf{y} \in N(\mathbf{x},k)} density(\mathbf{y}, k)/|N(\mathbf{x}, k)|}. \quad (10.7)$$

Algorithm 10.2 Relative density outlier score algorithm.

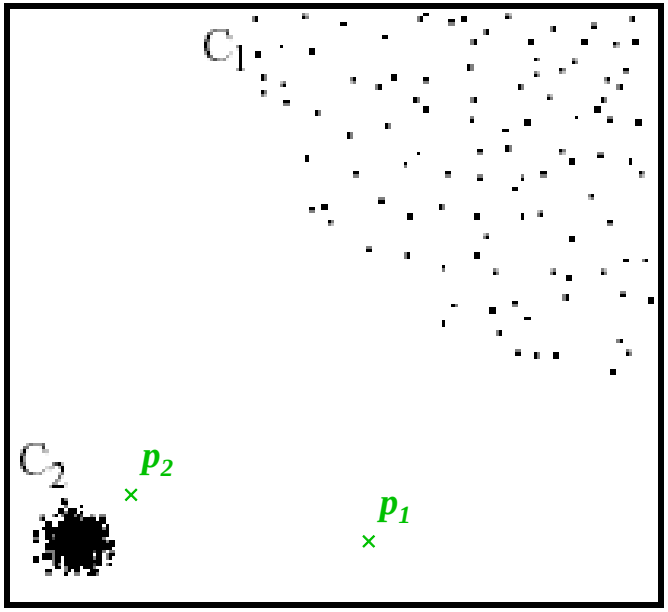
1: $\{k$ is the number of nearest neighbors$\}$
2: **for all** objects $\mathbf{x}$ **do**
3: Determine $N(\mathbf{x}, k)$, the k-nearest neighbors of $\mathbf{x}$.
4: Determine $density(\mathbf{x}, k)$, the density of $\mathbf{x}$, using its nearest neighbors, i.e., the objects in $N(\mathbf{x}, k)$.
5: **end for**
6: **for all** objects $\mathbf{x}$ **do**
7: Set the $outlier\ score(\mathbf{x}, k) = average\ relative\ density(\mathbf{x}, k)$ from Equation 10.7.
8: **end for**

Pontuação de anomalia com densidade relativa



Baseado em densidade: abordagem LOF

- Para cada ponto, calcula a densidade local
- Calcule o fator local de outlier (LOF) de uma amostra p como a média das proporções da densidade da amostra p e a densidade de seus vizinhos mais próximos
- Outliers são pontos com maior valor LOF



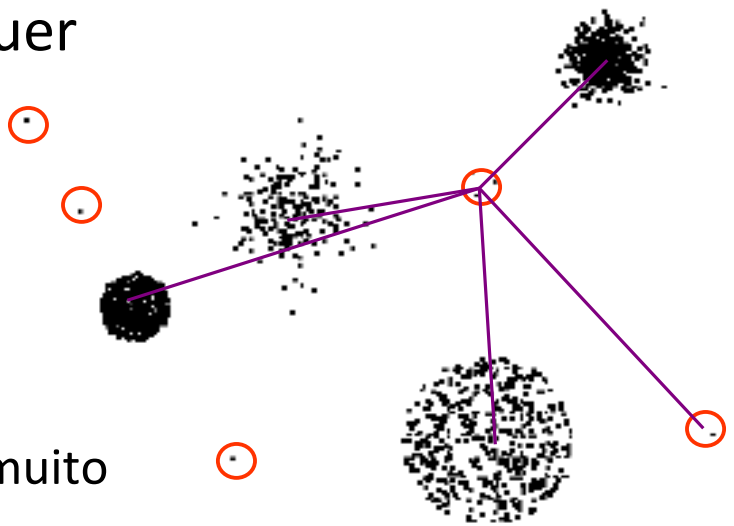
Na abordagem de vizinho próximo, p_2 não é considerado como anomalia, enquanto a abordagem LOF acha ambos p_1 e p_2 como anomalias

Pontos fortes e fracos de abordagens baseadas em densidade

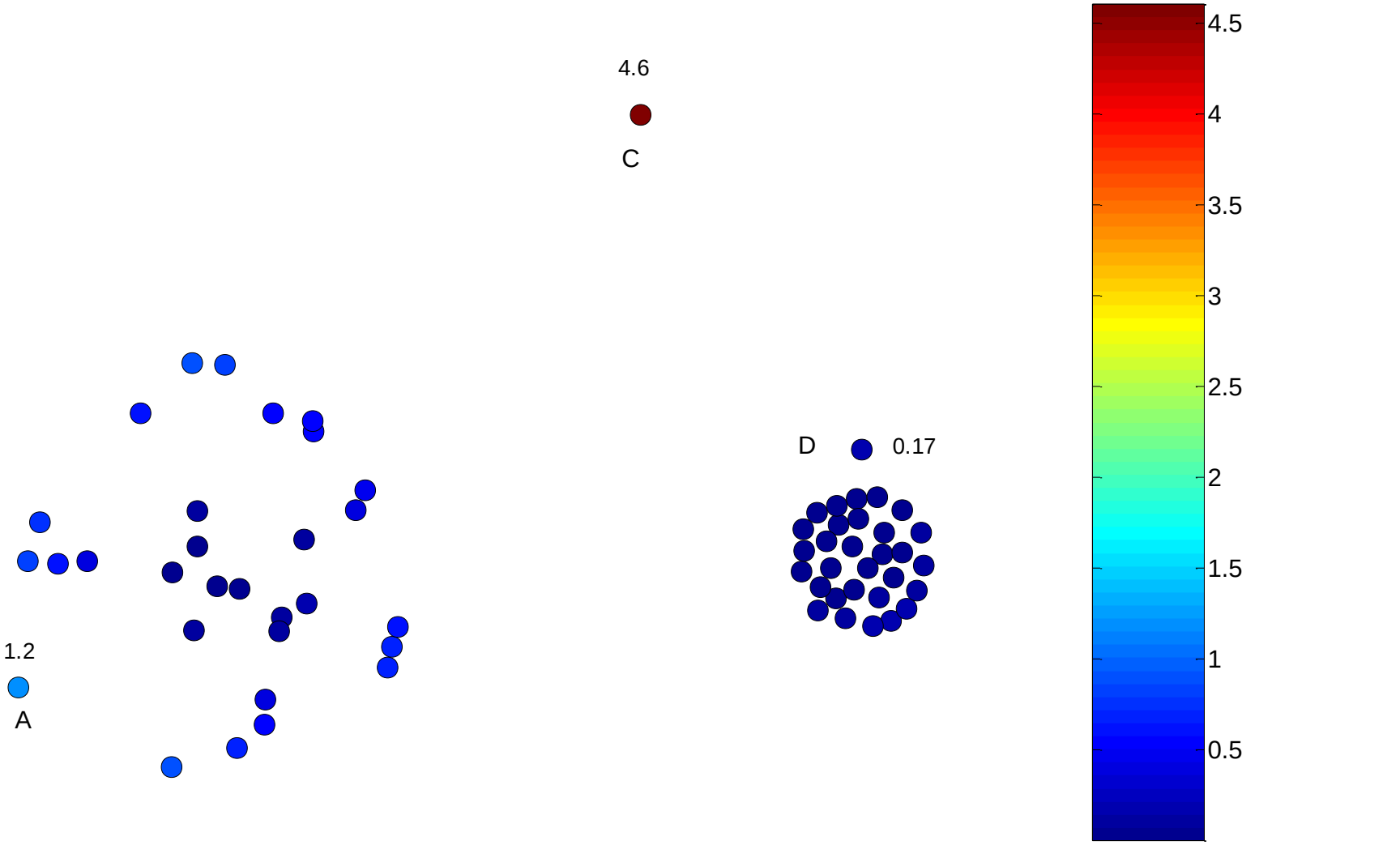
- Simples
- Caras – $O(n^2)$
- Sensível aos parâmetros
- A densidade torna-se menos significativa no espaço de alta dimensão

Abordagens baseadas em agrupamento

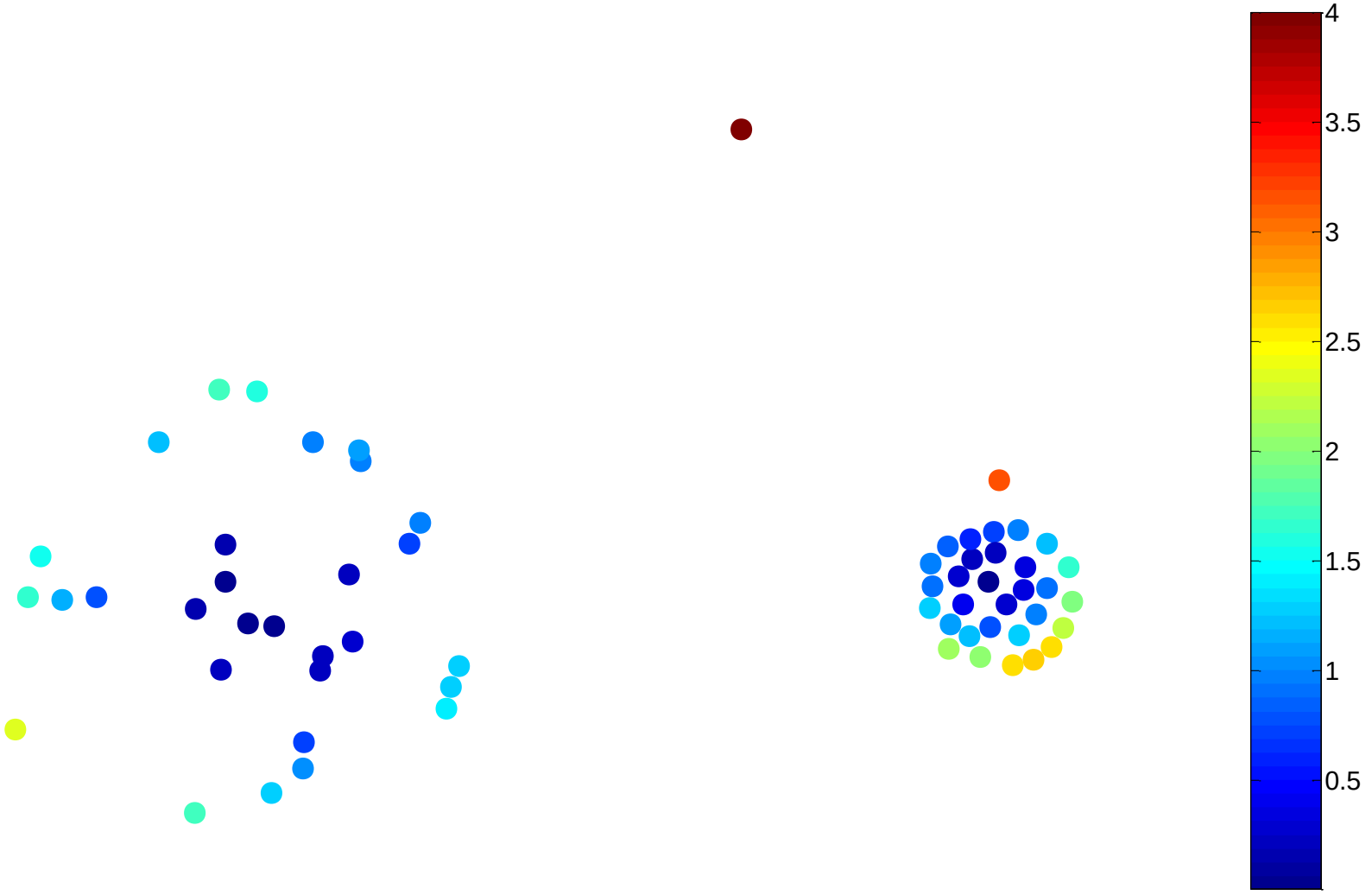
- **Anomalia baseada no agrupamento:** Um objeto é um outlier baseado em cluster se ele não pertencer fortemente a qualquer cluster
  - Para clusters baseados em protótipos, um objeto é um outlier se não estiver perto suficiente de um centro de um cluster
  - Para clusters baseados em densidade, um objeto é um outlier se sua densidade for muito baixa
  - Para clusters baseados em gráfico, um objeto é um outlier se não estiver bem conectado
- Outras questões incluem o impacto dos outliers nos clusters e o número de clusters



Distância dos pontos ao centroides mais próximos



Distância relativa dos pontos ao centroide mais próximo



Pontos fortes e fracos de abordagens baseadas na agrupamento

- Simples
- Muitas técnicas de agrupamento podem ser usadas
- Pode ser difícil decidir sobre uma técnica de agrupamento
- Pode ser difícil decidir sobre o número de clusters
- Outliers podem distorcer os clusters