Aula 14 – Mineração de Dados
Detecção de Anomalia

Profa. Elaine Faria

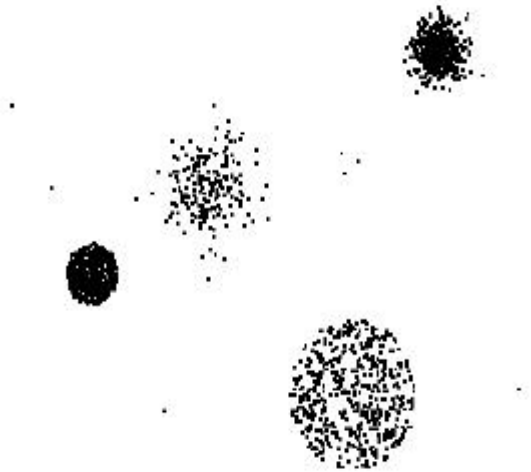
UFU

Material

- Este material foi construído por meio de traduções dos slides do prof. Tan, disponíveis em:
  - https://www-users.cse.umn.edu/~kumar001/dmbook/index.php
  - Todas as figuras usadas foram retiradas dos slides do prof. Tan

- O material é baseado no livro
  - Tan P., SteinBack M. e Kumar V. Introduction to Data Mining, Pearson, 2006.
  - Faceli, K., Lorena, A. C., Gama, J., Carvalho, A. C. P. L. F., Inteligência Artificial: Uma abordagem de Aprendizado de Máquina, LTC, 2011.

Detecção de Anomalias/*Outliers*



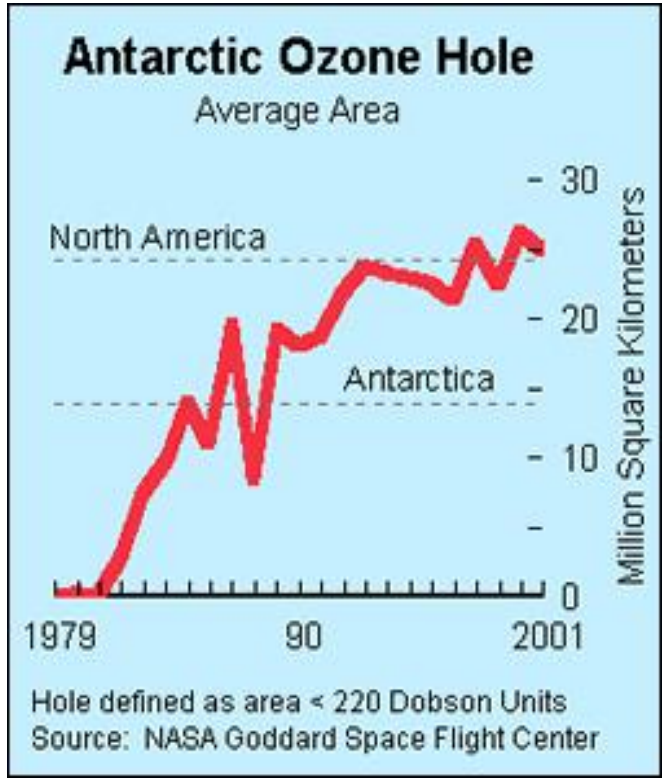
- O que são anomalias/outliers?
  - Conjunto de pontos que são consideravelemente diferentes dos demais pontos da base de dados
- As anomalias são relativamente raras
  - Não significa que não ocorrem frequentemente em termos absolutos
    - Um em mil dados -> pode ocorrer milhões de vezes quando temos bilhões de dados
- Pode ser importante ou um incômodo
  - 50 kgs, 2 anos

Aplicações para a detecção de anomalias

- Detecção de Fraude
  - Comportamento de compra no cartão de crédito
- Detecção de Intrusão
  - Ataques em redes de computadores
- Distúrbios no ecosistema
  - Eventos atipicos que possam afetar os seres humanos
- Saúde pública
  - Ex: ocorrência de uma doença em um grupo vacinado
- Medicina
  - Sintomas não usuais pode indicar um problema de saúde

Detecção de Anomalias/*Outliers*

- Em 1985, três pesquisadores (Farman, Gardinar e Shanklin) ficaram intrigados com os dados coletados pela British Antarctic Survey mostrando que os níveis de ozônio na Antártica 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 tão baixas?
- As concentrações de ozônio registradas pelo satélite eram tão baixas que estavam sendo tratadas como *outliers* por um programa de computador e descartadas!



Fonte:
http://www.epa.gov/ozone/science/hole/size.html

Causa das Anomalias

- Dados de diferentes classes
  - Medir o peso de laranjas, mas algumas outras frutas estão misturadas
  - Fraude em cartão de crédito
- Variação não usual
  - Pessoa com uma altura não usual
- Erro nos dados
  - 200 kg e 2 anos
  - Remover anomalias é o foco do pré-processamento (limpeza dos dados)

Problemas gerais: Pontuação de anomalias

- Muitas técnicas de detecção de anomalias fornecem apenas uma categoriazação binária
  - Um objeto é uma anomalia ou não
  - Isso é especialmente verdadeiro para abordagens baseadas em classificação

- Outras abordagens associam uma pontuação (*score*) a todos os objetos
  - Essa pontuação mede o grau de um objeto ser uma anomalia
  - Isso permite ranquear objetos

- No fim, sempre é necessária uma decisão binária
  - Essa transação de cartão de crédito deve ser sinalizada?
  - Ainda é útil ter uma pontuação

Variantes do problema de detecção de anomalias

- Dado um conjunto D, contendo em sua maioria objetos normais (mas não rotulados), e um objeto de teste **x**, calcule a pontuação de anomalia de **x** com relação a D

- Dado um conjunto D, encontre todos os pontos **x** $\in$ D com pontuação de anomalia maior que um limiar **t**

- Dado um conjunto D, encontre todos os objetos **x** $\in$ D tendo as *top-n* maiores pontuações de anomalia

Visão geral das Propostas para detecção de anomalias

- Técnicas baseada em Modelos
- Técnicas baseadas em Proximidade
- Técnicas baseadas em Densidade

Detecção de Anomalia baseada em Modelos

- Constrõem um modelo dos dados; as anomalias são objetos que não se encaixam muito bem no modelo
- Não-supervisionado
    - Anomalias são aqueles pontos que não se encaixam bem em nenhum dos clusters
    - Anomalias são os pontos que distorcem o modelo
- Supervisionado
    - As anomalias são consideradas uma classe rara
    - Precisa ter dados de treinamento

Detecção de Anomalia baseada em Proximidade

- Técnicas baseadas em uma medida de proximidade entre os objetos
- Objetos anômalos são aqueles distantes dos outros objetos
- Muitas dessas técnicas são baseadas em distância
  - Técnicas de detecção de *outliers* baseadas em distância
- Quando os objetos podem ser plotados em uma *scatter plot* (gráfico de dispersão) de 2 ou 3 dimensões, os objetos *outliers* podem ser detectados visualmente

Detecção de Anomalia baseada em Densidade

- O cálculo da densidade de objetos são relativamente fáceis de calcular se a medida de proximidade entre eles está disponível
- Objetos que estão em regiões de baixa densidade estão relativamente distantes de seus vizinhos e podem ser considerados anômalos
  - Propostas sofisticadas consideram que os *datasets* tem diferentes densidades
  - Classificam um objeto como *outlier* somente se ele tem uma densidade local significantemente menor que a maioria dos seus vizinhos

Questões importantes a serem tratadas na detecção de anomalias

- Número de atributos a ser usado para definir uma anomalia
  - Um objeto pode ter um valor anômalo para algum atributo, mas valores normais para outros atributos
  - Um objeto pode ser anômalo mesmo se nenhum dos valores dos seus atributos forem anômalos individualmente
- Perspectiva global versus local
  - Um objeto pode ser não usual com relação a todos os objetos, mas não com respeito a sua vizinhança local
  - Ex: altura de um jogador
- Limiar para considerar um objeto como uma anomalia
  - Decisão binária nem sempre reflete a realidade
  - Alguns objetos são anomalias mais extremas que outras --> uso de *score*

Questões importantes a serem tratadas na detecção de anomalias

- Detectar uma anomalia por vez versus muitas anomalias de uma vez
  - Algumas técnicas removem uma anomalia por vez (o objeto mais anômalo)
  - Algumas técnicas detectam uma coleção de anomalias
- Avaliação
  - Se os rótulos estão disponíveis para identificar os objetos anômalos e os normais --> medidas de classificação podem ser usadas
  - Se os rótulos não estiverem disponíveis --> a avaliação torna-se mais difícil
    - Ex: Qual a melhora no modelo quando o *outlier* foi eliminado?
- Eficiência
  - Há diferenças significativas no custo computacional de diferentes técnicas de detecção de anomalia

Técnicas de detecção de anomalias

- Propostas estatísticas
- Baseadas em proximidade
  - As anomalias são pontos distantes de outros pontos
- Baseadas em agrupamento
  - Os pontos distantes dos centros do cluster são *outliers*
- Baseadas em reconstrução

Propostas estatísticas

- São propostas baseadas em modelos
  - Um modelo é criado e os objetos são avaliados com relação a quão bem eles se ajustam ao modelo

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 que descreve a distribuição dos dados (por exemplo, distribuição normal)
- Aplicam um teste estatístico que depende
  - Distribuição de dados
  - Parâmetros de distribuição (por exemplo, média, variância)
  - Número de *outliers* esperados (limite de confiança)

Propostas estatísticas

- Problemas
  - Identificar a distribuição de um conjunto de dados
    - A base de dados pode ser modelada por uma distribuição gaussiana?
  - Número de atributos
    - Algumas técnicas se baseiam em apenas um atributo, mas os dados são multidimensionais
  - Mistura de distribuições
    - Dados modelados por uma mistura de distribuições
      - Modelos complexos

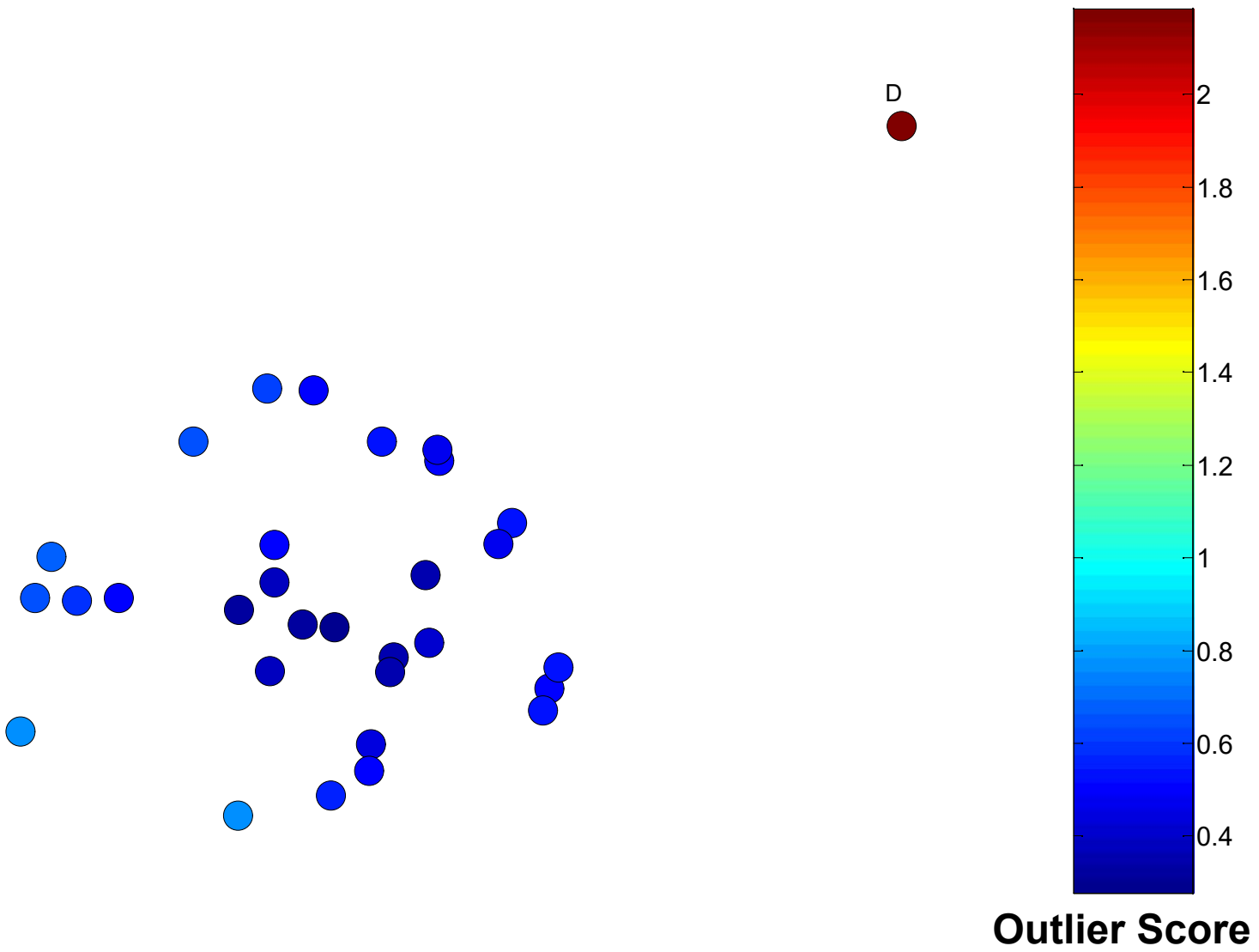
Pontos fortes / fracos das abordagens estatísticas

- Base 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 dimensionais elevados, pode ser difícil estimar a verdadeira distribuição
- As anomalias podem distorcer os parâmetros da distribuição

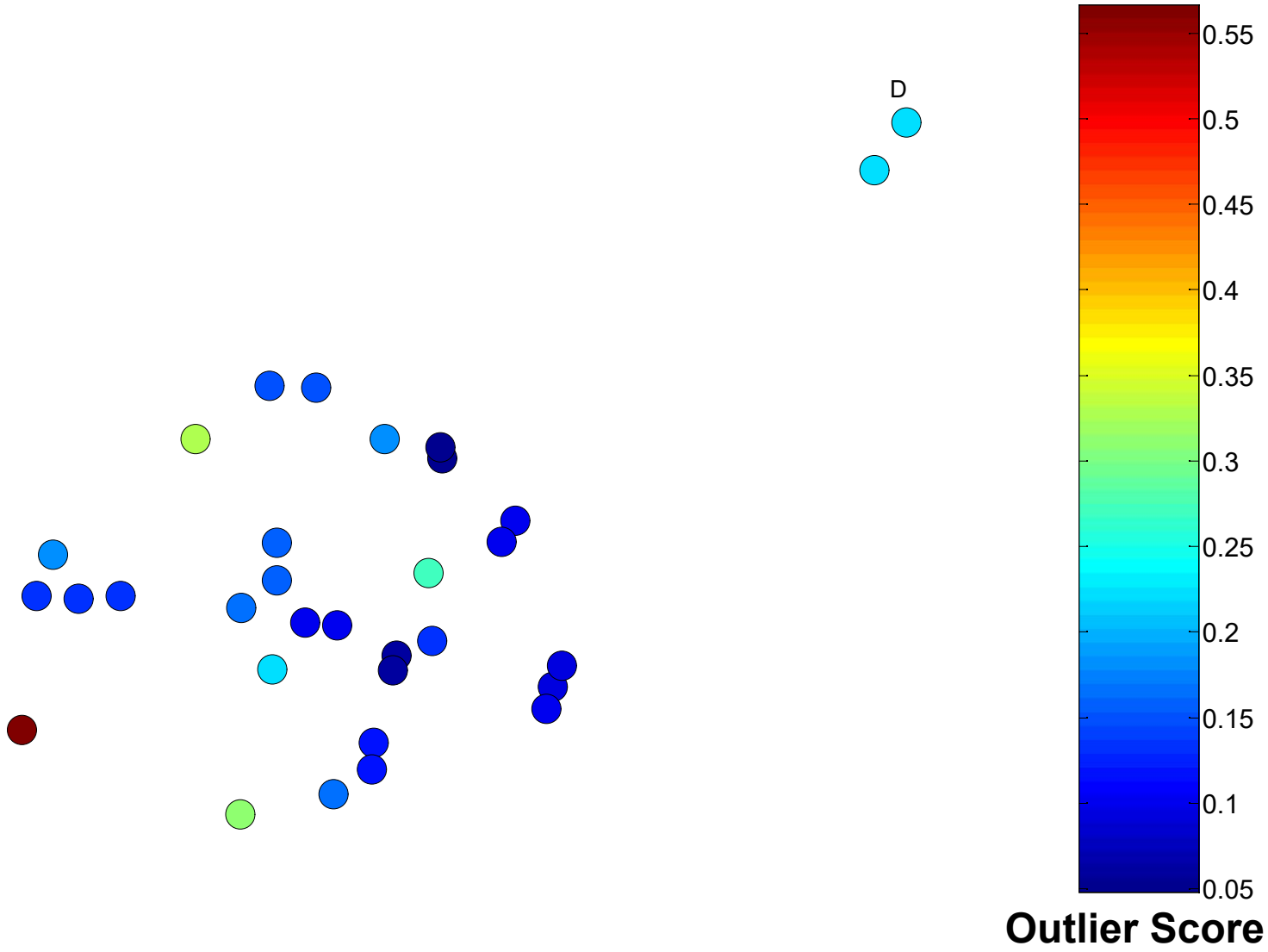
Abordagens baseadas em proximidade

- Um objeto é anômalo se está distante da maioria dos outros objetos
  - Proposta mais geral e mais fácil de ser aplicada do que as estatísticas
  - É mais fácil determinar uma medida de proximidade significativa do que uma distribuição estatística
- Distância aos k-vizinhos mais próximos
  - Uma das formas de medir se um objeto está distante da maioria dos objetos
  - A pontuação *outlier* de um objeto é a distância até seu k-ésimo vizinho mais próximo
    - O *score* pode ser sensível ao valor de k

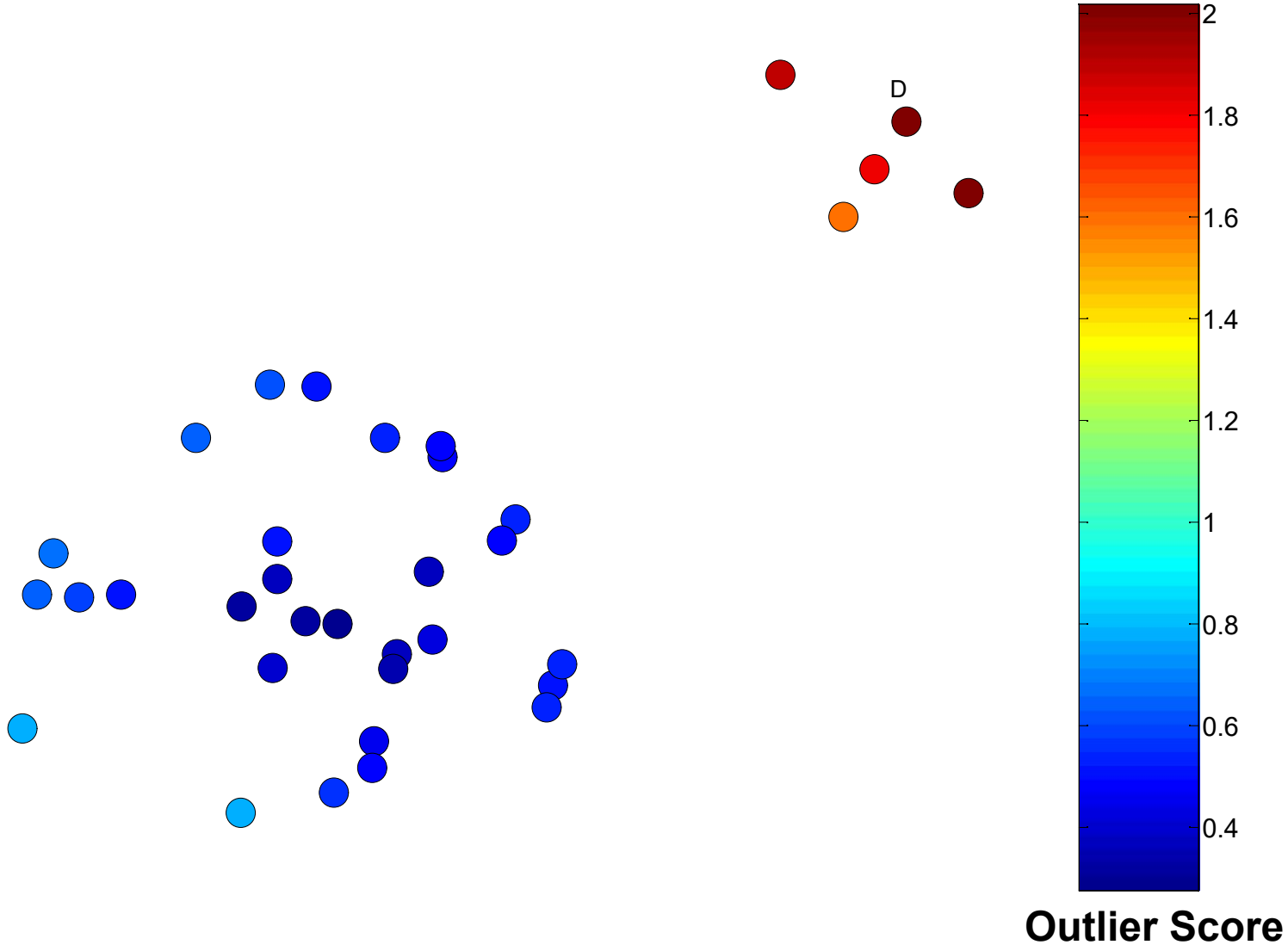
Um vizinho mais próximo - Um *outlier*



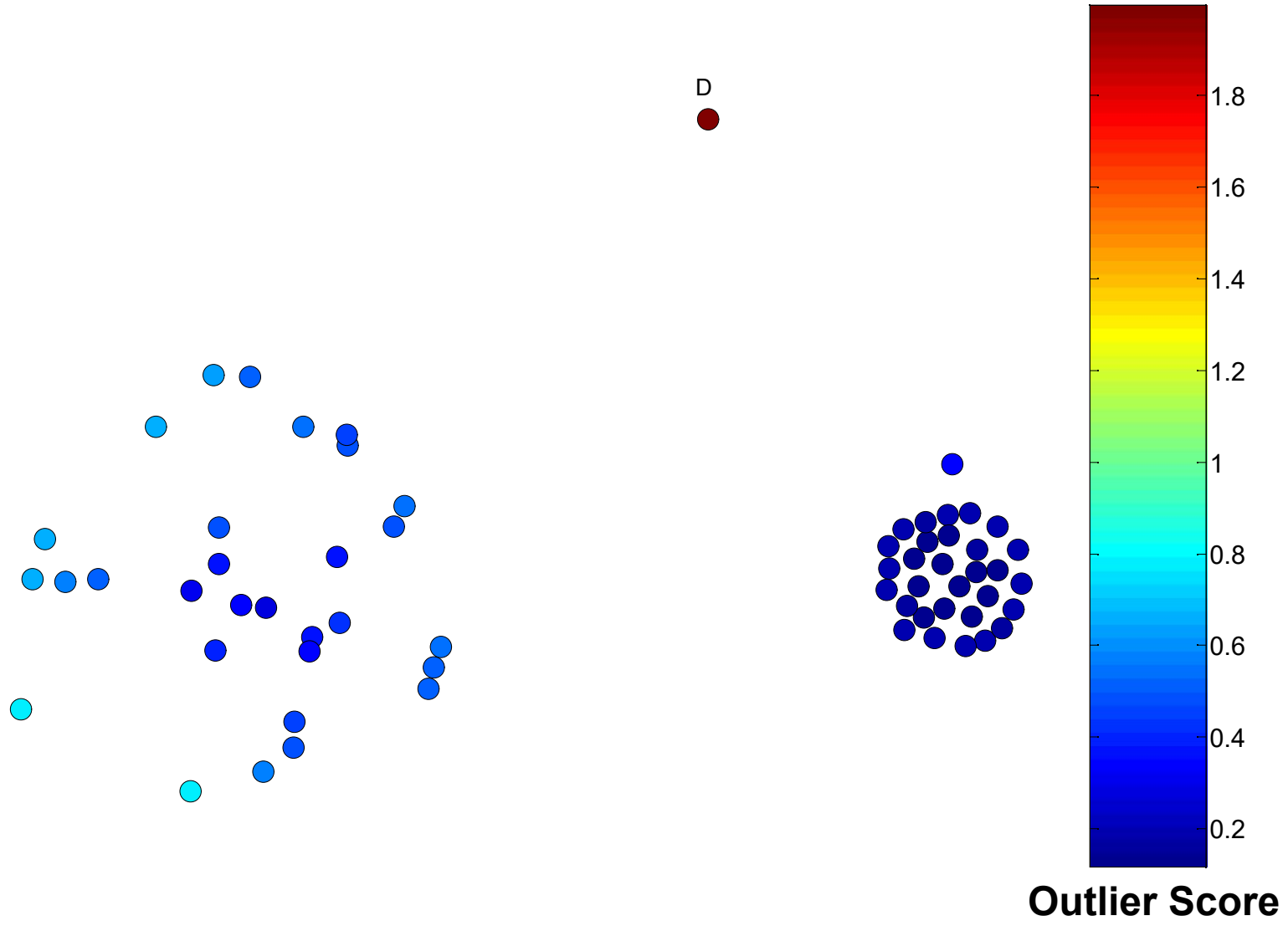
Um vizinho mais próximo - Dois *outliers*



Cinco vizinhos mais próximos - Pequeno grupo



Cinco vizinhos mais próximos - Diferente densidades



Pontos fortes / fracos das abordagens baseadas em distância

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

Propostas baseadas em densidade

- *Outliers* são objetos que estão em região de baixa densidade
- *Outlier* baseado em densidade: a pontuação de *outlier* de um objeto é o inverso da densidade ao redor do objeto
  - Pode ser definido em termos dos k vizinhos mais próximos
  - Uma definição: inverso da distância até o k-ésimo vizinho
  - Outra definição: inverso da distância média para k vizinhos
  - Definição DBSCAN
    - a densidade de um objeto é igual ao nro de objetos que estão dentro de uma distância especificada *d* do objeto
- Se houver regiões de densidade diferente, esta abordagem pode ter problemas

Pontos fortes / fracos das abordagens baseadas em densidade

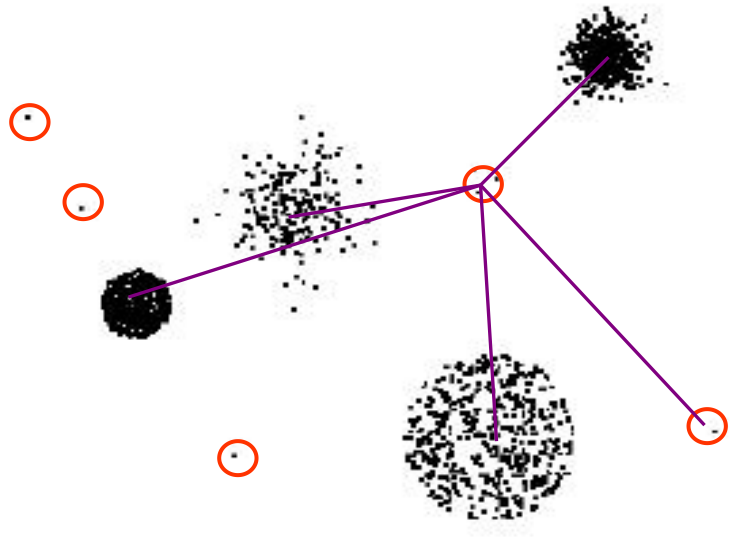
- Simples
- Caro - $O(n^2)$
- Seleção de parâmetros pode ser difícil
- A densidade torna-se menos significativa no espaço de alta dimensão

Propostas baseadas em agrupamento

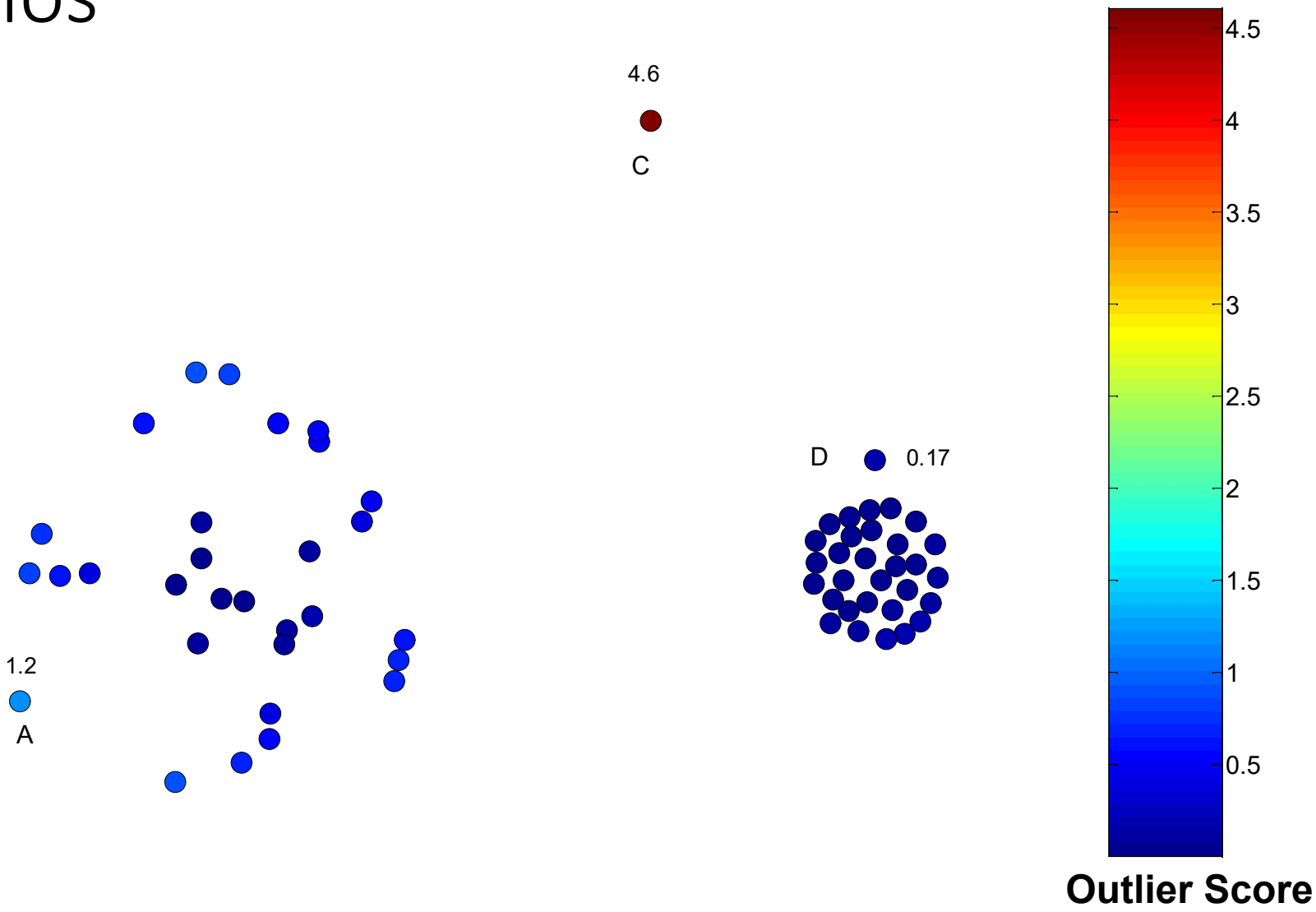
Outlier baseado em cluster: um objeto é um *outlier* baseado em *cluster* se não pertencer fortemente a nenhum cluster

- Para clusters baseados em protótipo, um objeto é um *outlier* se não estiver perto o suficiente de um centro de cluster
- Para clusters baseados em densidade, um objeto é um *outlier* se sua densidade for muito baixa
- Para clusters baseados em grafo, um objeto é um *outlier* se não estiver fortemente conectado

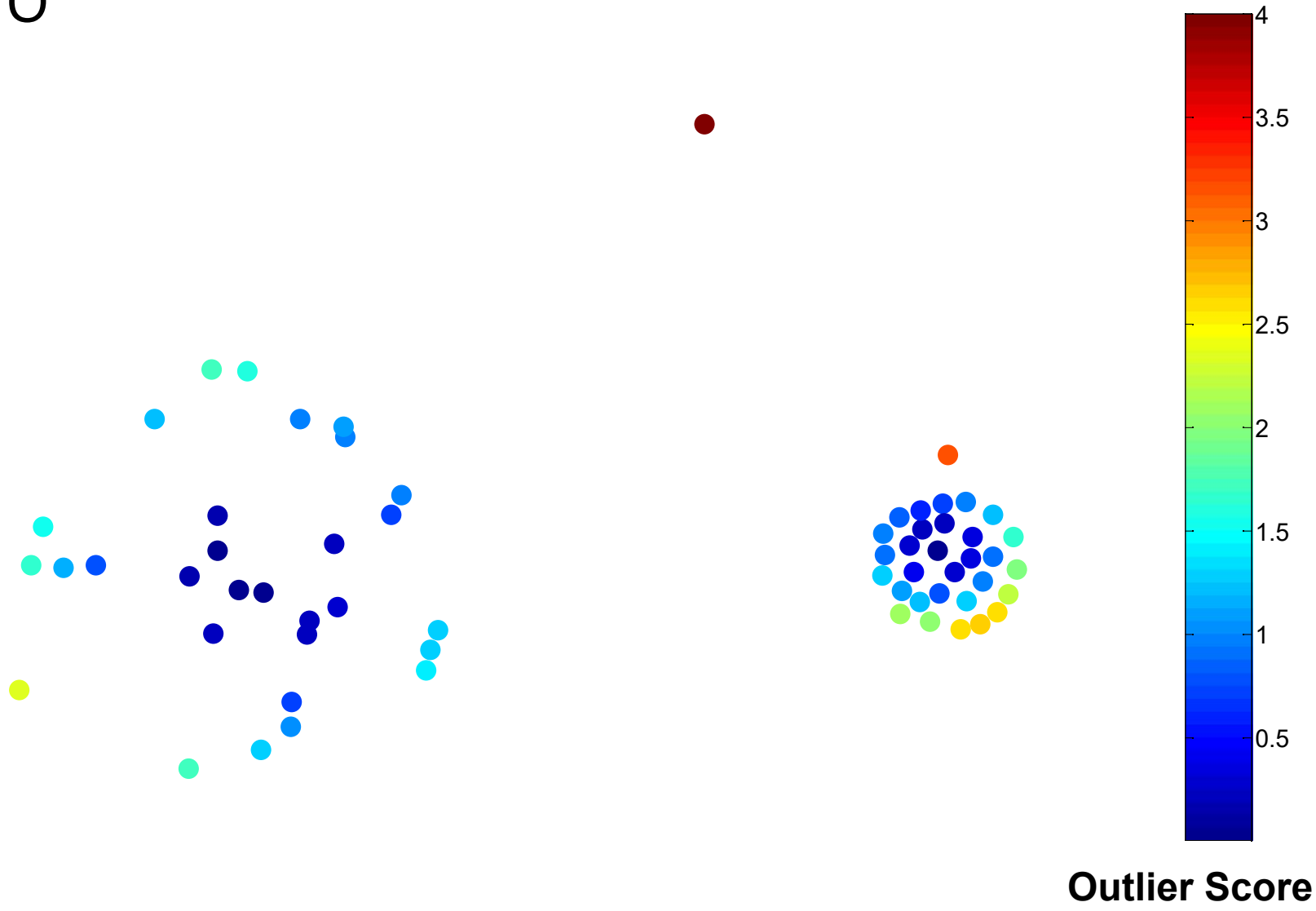
- Outros problemas incluem o impacto de *outliers* nos agrupamento inicial e o número de clusters



Distância dos pontos aos centróides mais próximos



Distância Relativa dos Pontos ao Centróide Mais Próximo



Pontos fortes / fracos das abordagens baseadas em cluster

- 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

Abordagens baseadas na reconstrução

- Com base em suposições, existem padrões na distribuição da classe normal que podem ser capturados usando representações de dimensões inferiores
- Reduza os dados para dados dimensionais inferiores
  - Pode usar Análise de Componentes Principais (PCA) ou outras técnicas de redução de dimensionalidade
  - Também pode usar redes neurais
- Meça o erro de reconstrução para cada objeto
  - A diferença entre a versão original e a de dimensionalidade reduzida

Erro de reconstrução

- Seja x o objeto de dados original
- Encontre a representação do objeto em um espaço dimensional inferior
- Projete o objeto de volta ao espaço original
- Chame este objeto de $\hat{x}$
- Erro de reconstrução (x) = $\|x - \hat{x}\|$
- Objetos com grandes erros de reconstrução são anomalias

Forças e fraquezas

- Não requer suposições sobre a distribuição da classe normal
- Pode usar muitas abordagens de redução de dimensionalidade
- O erro de reconstrução é calculado no espaço original
- Isso pode ser um problema se a dimensionalidade for alta

Abordagens teóricas da informação

- A ideia principal é medir a quantidade de informações que diminui quando você exclui uma observação

$$Gain(x) = Info(D) - Info(D \setminus x)$$

- As anomalias devem mostrar maior ganho

- Os pontos normais devem ter menos ganho

Forças e fraquezas

- Base teórica sólida
- Teoricamente aplicável a todos os tipos de dados
- Difícil e computacionalmente caro de implementar na prática

Avaliação da detecção de anomalias

- Se rótulos de classe estiverem presentes, use abordagens de avaliação padrão para classes raras, como precisão, *recall* ou taxa de falsos positivos
  - FPR também é conhecido como taxa de falso alarme

- Para detecção de anomalia não supervisionada, use medidas fornecidas pelo método de anomalia
  - Erro de reconstrução ou ganho

- Também pode ver histogramas de pontuações de anomalias.

Leitura Recomendada

- Leitura do
  - Capítulo 10 do livro Tan et al, 2006.