# Técnicas de detecção de outlier

## Introdução

**O que é um outlier?**

**Definição de Hawkins [Hawkins 1980]:**

"Um outlier é uma observação que se desvia tanto das outras observações a ponto de levantar suspeitas de que foi gerado por um mecanismo diferente"

**Intuição baseada em estatísticas**

         Objetos de dados normais seguem um "mecanismo gerador", por exemplo, algum processo estatístico dado

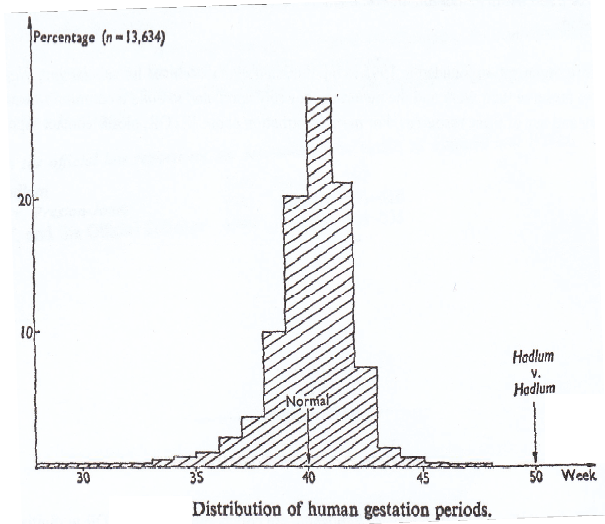
         Objetos anormais desviam-se desse mecanismo gerador

Exemplo: Hadlum vs. Hadlum (1949) [Barnett 1978]

• O nascimento de um filho da Sra. Hadlum aconteceu 349 dias depois que o Sr. Hadlum partiu para o serviço militar.

• O período médio de gestação humana é de 280 dias (40 semanas).

• Estatisticamente, 349 dias é um ponto fora da curva.



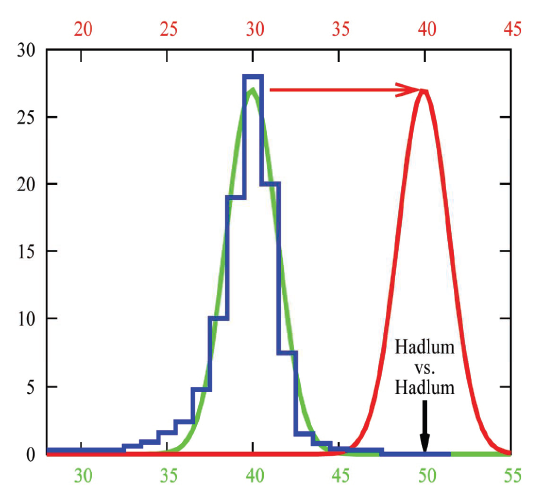
         azul: base estatística (13634 observações dos períodos de gestação)

         verde: suposta subjacente ao processo gaussiano

o Baixíssima probabilidade do nascimento da Sra. Hadlums ser gerada por esse processo

         vermelho: assunção do Sr. Hadlum (outro processo gaussiano responsável pelo parto observado, onde o período de gestação responsável)

o Sob esse pressuposto, o período de gestação tem duração média e maior probabilidade possível



 Aplicações de exemplo de detecção de outliers

o Detecção de fraudes

  O comportamento de compra do titular de um cartão de crédito geralmente muda quando o cartão é roubado

  Padrões de compra anormais podem caracterizar abuso de cartão de crédito

o Medicina

  Sintomas incomuns ou resultados de testes podem indicar possíveis problemas de saúde de um paciente

  Se um determinado resultado de teste é anormal pode depender de outras características dos pacientes (por exemplo, sexo, idade, ...)

o Saúde pública

  A ocorrência de uma determinada doença, por exemplo, tétano, espalhada por vários hospitais de uma cidade indica problemas com o programa de vacinação correspondente naquela cidade

  Se uma ocorrência é anormal depende de diferentes aspectos como frequência, correlação espacial, etc.

o Estatísticas esportivas

  Em muitos esportes, vários parâmetros são registrados para os jogadores, a fim de avaliar o desempenho dos jogadores

  Jogadores proeminentes (tanto no sentido positivo quanto negativo) podem ser identificados como tendo valores de parâmetros anormais

  Às vezes, os jogadores mostram valores anormais apenas em um subconjunto ou uma combinação especial dos parâmetros gravados

o Detecção de erros de medição

  Os dados derivados de sensores (por exemplo, numa determinada experiência científica) podem conter erros de medição

  Valores anormais podem fornecer uma indicação de um erro de medição

  A remoção desses erros pode ser importante em outras tarefas de mineração e análise de dados

  "O barulho de uma pessoa pode ser o sinal de outra."

         Discussão da intuição básica baseada em Hawkins

• Os dados são geralmente multivariados, ou seja, multidimensionais => modelo básico é univariado, ou seja, 1-dimensional

• Geralmente há mais de um mecanismo gerador/processo estatístico subjacente aos dados => modelo básico pressupõe apenas um mecanismo gerador "normal"

• Anomalias podem representar uma classe diferente (mecanismo gerador) de objetos, então pode haver uma grande classe de objetos semelhantes que são os outliers => modelo básico assume que outliers são observações raras

         Consequência: muitos modelos e abordagens evoluíram nos últimos anos para superar essas premissas e não é fácil acompanhar essa evolução.

         Novos modelos geralmente envolvem suposições e restrições típicas, novas, embora geralmente ocultas.

         Cenários gerais de aplicativos

o Cenário supervisionado

  Em alguns aplicativos, dados de treinamento com objetos de dados normais e anormais são fornecidos

  Pode haver múltiplas classes normais e/ou anormais

  Muitas vezes, o problema de classificação é altamente desequilibrado

o Cenário Semi-Supervisionado

  Em algumas aplicações, apenas os dados de treinamento para a(s) classe(s) normal(ais) (ou apenas a(s) classe(s) anormal(ais)) são fornecidos

o Cenário não supervisionado

  Na maioria dos aplicativos, não há dados de treinamento disponíveis

         Neste tutorial, nos concentramos no cenário não supervisionado

         Os outliers são apenas um produto secundário de alguns algoritmos de clustering?

• Muitos algoritmos de clustering não atribuem todos os pontos aos clusters, mas contabilizam objetos de ruído

• Procure por outliers aplicando um desses algoritmos e recupere o conjunto de ruído

• Problema:

  Os algoritmos de clustering são otimizados para encontrar clusters em vez de outliers

  A precisão da detecção de outliers depende de quão bom o algoritmo de agrupamento captura a estrutura dos clusters

  Um conjunto de muitos objetos de dados anormais que são semelhantes entre si seria reconhecido como um cluster em vez de como ruído/outliers

         Focalizaremos três abordagens de classificação diferentes

• Detecção de outlier global versus local (Considera o conjunto de objetos de referência em relação aos quais o "outlierness" de cada ponto é julgado)

• Rotulação versus pontuação de outliers (Considera a saída de um algoritmo)

• Propriedades de modelagem (Considera os conceitos com base nos quais o "outlierness" é modelado)

NOTA: nos concentramos em modelos e métodos para dados euclidianos, mas muitos deles também podem ser usados para outros tipos de dados (porque exigem apenas uma medida de distância)

         Abordagens globais versus locais

• Considera a resolução do conjunto de referência w.r.t. que o "outlierness" de um determinado objeto de dados é determinado

• Abordagens globais

  O conjunto de referência contém todos os outros objetos de dados

  Pressuposto básico: existe apenas um mecanismo normal

  Problema básico: outros outliers também estão no conjunto de referência e podem falsificar os resultados

• Abordagens locais

  A referência contém um (pequeno) subconjunto de objetos de dados

  Nenhuma suposição sobre o número de mecanismos normais

  Problema básico: como escolher um conjunto de referência adequado

• NOTA: Algumas abordagens são um pouco intermediárias

  A resolução do conjunto de referência é variada, por exemplo, de apenas um único objeto (usuário local) para todo o banco de dados (global) automaticamente ou por um parâmetro de entrada definido pelo usuário

         Rotulagem versus pontuação

• Considera a saída de um algoritmo de detecção de outlier

• Abordagens de rotulagem

  Saída binária

  Os objetos de dados são rotulados como normais ou discrepantes

• Abordagens de pontuação

  Saída contínua

  Para cada objeto, uma pontuação atípica é calculada (por exemplo, a probabilidade de ser um outlier)

  Os objetos de dados podem ser classificados de acordo com suas pontuações

• Notas

  Muitas abordagens de pontuação se concentram em determinar os *outliers*top-n (o parâmetro *n*é geralmente dado pelo usuário)

  As abordagens de pontuação geralmente também podem produzir saída binária, se necessário (por exemplo, definindo um limiar adequado nos valores de pontuação)

 • Abordagens classificadas pelas propriedades da abordagem de modelagem subjacente

  Abordagens baseadas em modelos

         Racional

o Aplicar um modelo para representar pontos de dados normais

o Outliers são pontos que não se encaixam nesse modelo

         Exemplos de abordagens

o Testes probabilísticos baseados em modelos estatísticos

o Abordagens baseadas em profundidade

o Abordagens baseadas em desvios

o Algumas abordagens de detecção de outlier de subespaço

  Abordagens baseadas na proximidade

         Racional

o Examinar a proximidade espacial de cada objeto no espaço de dados

o Se a proximidade de um objeto se desvia consideravelmente da proximidade de outros objetos, é considerado um outlier

         Exemplos de abordagens

o Abordagens baseadas na distância

o Abordagens baseadas em densidade

o Algumas abordagens de detecção de outlier de subespaço

  Abordagens baseadas em ângulos

         Racional

o Examinar o espectro de ângulos pares entre um determinado ponto e todos os outros pontos

o Outliers são pontos que possuem um espectro caracterizado por alta flutuação

## Testes estatísticos

  Ideia geral

o Dado um certo tipo de distribuição estatística (por exemplo, gaussiana)

o Calcular os parâmetros assumindo que todos os pontos de dados foram gerados por tal distribuição estatística (por exemplo, média e desvio padrão)

o Outliers são pontos que têm uma baixa probabilidade de serem gerados pela distribuição geral (por exemplo, desviar mais de 3 vezes o desvio padrão da média)

  Pressuposto básico

o Os objetos de dados normais seguem uma distribuição (conhecida) e ocorrem em uma região de alta probabilidade deste modelo

o Outliers se desviam fortemente dessa distribuição

  Um grande número de testes diferentes estão disponíveis diferindo em

o Tipo de distribuição de dados (por exemplo, gaussiana)

o Número de variáveis, ou seja, dimensões dos objetos de dados (univariada/multivariada)

o Número de distribuições (modelos de mistura)

o Paramétrico versus não paramétrico (por exemplo, baseado em histograma)

  Exemplo nos slides a seguir

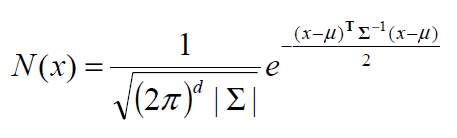
o Distribuição gaussiana

o Multivariada

o 1 modelo

o Paramétrico

  Função densidade de probabilidade de uma distribuição normal multivariada



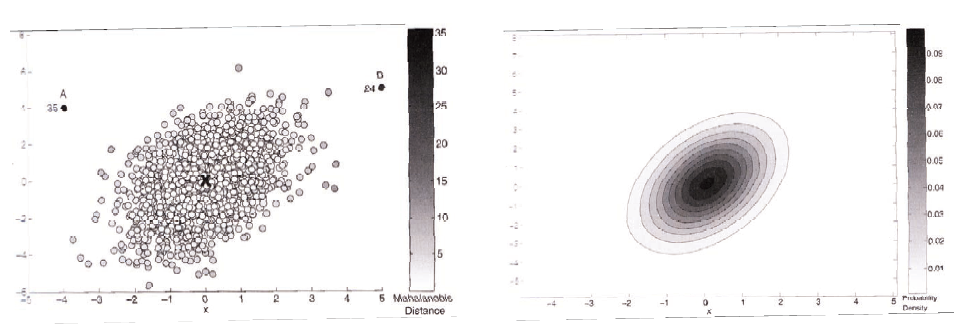
o μ é o valor médio de todos os pontos (geralmente os dados são normalizados de tal forma que μ=0)

o Σ é a matriz de covariância da média

o é a distância de Mahalanobis do ponto x ao μ

o MDist segue uma distribuição χ2 com d graus de liberdade (d = dimensionalidade dos dados)

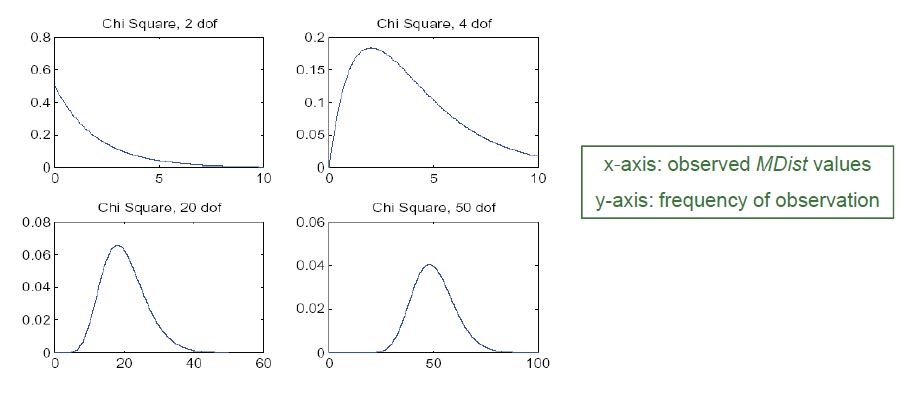
o Todos os pontos x, com MDist(x,μ) > χ2(0,975) [≈ 3.σ]

Visualização (2D) [Tan et al. 2006]

  Problemas

o Maldição da dimensionalidade

  Quanto maior o grau de liberdade, mais semelhantes são os valores *do MDist*para todos os pontos



o Robustez

  Média e desvio padrão são muito sensíveis a outliers

  Esses valores são calculados para o conjunto de dados completo (incluindo possíveis outliers)

  O MDist é usado para determinar outliers, embora os valores de *MDist*sejam influenciados por esses outliers => Mínimo Covariância Determinante [Rousseeuw e Leroy 1987] minimiza a influência de outliers na distância de Mahalanobis

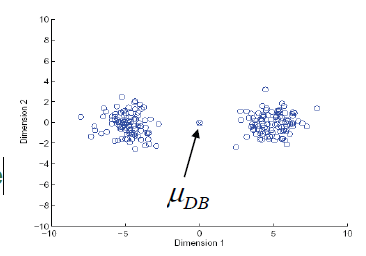
o Discussão

  A distribuição de dados é fixa

  Baixa flexibilidade (sem modelo de mistura)

  Método global

  Gera um rótulo, mas também pode gerar uma pontuação



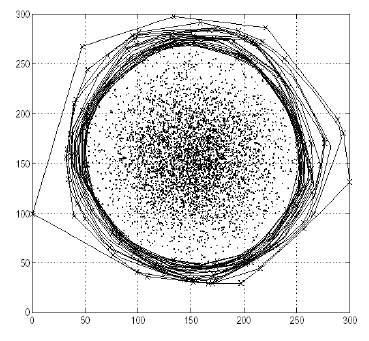
## Abordagens baseadas em profundidade

  Ideia geral

o Busca de outliers na borda do espaço de dados, mas independente de distribuições estatísticas

o Organizar objetos de dados em camadas de casco convexas

o Outliers são objetos em camadas externas



  Pressuposto básico

o Outliers estão localizados na borda do espaço de dados

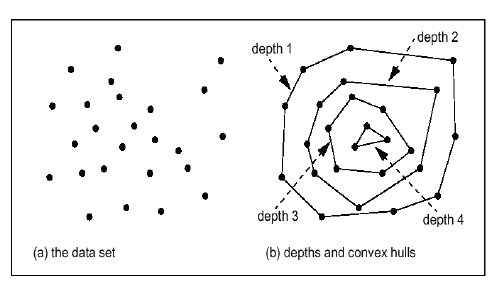
o Objetos normais estão no centro do espaço de dados

  Modelo [Tukey 1977]

o Pontos no casco convexo do espaço de dados completo têm profundidade = 1

o Pontos no casco convexo do conjunto de dados após a remoção de todos os pontos com profundidade = 1 têm profundidade = 2

o Pontos com profundidade ≤ k são relatados como outliers



  Algoritmos de exemplo

o ISOPROFUNDIDADE [Ruts e Rousseeuw 1996]

o FDC [Johnson et al. 1998]

  Discussão

o Ideia semelhante às abordagens estatísticas clássicas (k = 1 distribuições), mas independente do tipo de distribuição escolhido

o A computação do casco convexo geralmente só é eficiente em espaços 2D / 3D

o Originalmente produz um rótulo, mas pode ser estendido para pontuar facilmente (tomar profundidade como valor de pontuação)

o Usa um conjunto de referência global para detecção de outliers

## Abordagens baseadas em desvios

  Ideia geral

o Dado um conjunto de pontos de dados (grupo local ou conjunto global)

o Outliers são pontos que não se ajustam às características gerais daquele conjunto, ou seja, a variância do conjunto é minimizada ao remover os outliers

  Pressuposto básico

o Outliers são os pontos mais externos do conjunto de dados

  Modelo [Arning et al. 1996]

o Dado um fator de suavização SF(I) que calcula para cada I ⊆ DB quanto a variância do DB é diminuída quando eu é removido do DB

o Com igual diminuição na variância, um conjunto de exceções menor é melhor

o Os outliers são os elementos do conjunto de exceções E ⊆ DB para os quais se aplica o seguinte: 

  Discussão:

o Ideia semelhante às abordagens estatísticas clássicas (k = 1 distribuições), mas independente do tipo de distribuição escolhido

o Solução ingênua está em O(2n) para n objetos de dados

o Heurísticas como amostragem aleatória ou melhor primeira pesquisa são aplicadas

o Aplicável a qualquer tipo de dados (depende da definição de SF)

o Originalmente concebido como um método global

o Gera uma etiquetagem

## Abordagens baseadas em distância

  Ideia Geral

o Julgar um ponto com base na(s) distância(s) de seus vizinhos

o Várias variantes propostas

  Pressuposto Básico

o Objetos de dados normais têm uma vizinhança densa

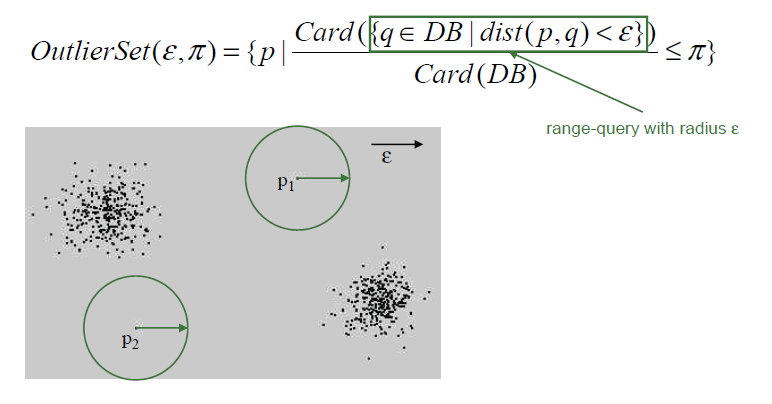
o Outliers estão distantes de seus vizinhos, ou seja, têm um Bairro menos denso

  DB(ε,π)-Outliers

o Modelo básico [Knorr e Ng 1997]

  Dado um raio ε e uma porcentagem π

  Um ponto p é considerado um outlier se, no máximo, π% de todos os outros pontos tiverem uma distância a p menor que ε



  Algoritmos

o Baseado em índices [Knorr e Ng 1998]

  Calcular a junção do intervalo de distâncias usando a estrutura do índice espacial

  Exclua o ponto de consideração adicional se sua vizinhança ε contiver mais de Card(DB) . π pontos

o • Baseado em loop aninhado [Knorr e Ng 1998]

  Divida o buffer em duas partes

  Use a segunda parte para digitalizar/comparar todos os pontos com os pontos da primeira parte

o • Baseado em grade [Knorr e Ng 1998]

  Construa uma grade de forma que quaisquer dois pontos da mesma célula de grade tenham uma distância de no máximo ε um do outro

  Os pontos só precisam ser comparados com pontos de células vizinhas

  Derivando conhecimento intensional [Knorr e Ng 1999]

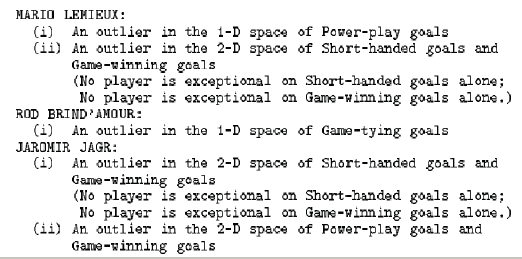
o Baseia-se no modelo DB(ε,π)-outlier

o Encontre o(s) subconjunto(s) mínimo(s) de atributos que explica o "outlierness" de um ponto, ou seja, no qual o ponto ainda é um outlier

o Exemplo

  Valores atípicos identificados



  Conhecimento intensional derivado (esboço)

  Pontuação de outlier com base em distâncias kNN

o Modelos gerais

  Tome a distância kNN de um ponto como sua pontuação atípica [Ramaswamy et al 2000]

  Agregar as distâncias de um ponto a todos os seus 1NN, 2NN, ..., kNN como um escore outlier [Angiulli e Pizzuti 2002]

o Algoritmos

  Abordagens gerais

         Loop aninhado

o Abordagem ingênua: Para cada objeto: compute kNNs com uma varredura sequencial

o Aprimoramento: use estruturas de índice para consultas kNN

         Baseado em partição

o Particionar dados em micro clusters

o Informações agregadas para cada partição (por exemplo, retângulos delimitadores mínimos)

o Permite podar micro clusters que não podem se qualificar ao procurar os kNNs de um determinado ponto

o Algoritmos de amostra (computação top-n outliers)

  Loop aninhado [Ramaswamy et al 2000]

         Algoritmo NL simples com suporte de índice para consultas kNN

         Algoritmo baseado em partição (baseado em um algoritmo de clustering que tem complexidade de tempo linear)

         Algoritmo para o modelo simples kNN-distância

  Linearização [Angiulli e Pizzuti 2002]

         Linearização de um conjunto de dados multidimensionais usando curvas de preenchimento de espaço

         A representação 1D é particionada em micro clusters

         Algoritmo para o modelo kNN-distância média

  ORCA [Baía e Schwabacher 2003]

         Algoritmo NL com randomização e poda simples

         Poda: se um ponto tiver uma pontuação maior do que o outlier top-n até agora (cut-off), remova esse ponto de consideração adicional

o não-outliers são podados

o funciona bem em dados aleatórios (pode ser feito em tempo linear)

o pior caso: algoritmo NL ingênuo

         Algoritmo para ambos os modelos kNN-distância e o modelo DB(ε,π)-outlier

  RBRP [Ghoting et al. 2006],

         Ideia: tentar aumentar o corte o máximo possível => aumentar o poder de poda

         Calcule kNNs aproximados para cada ponto para obter um melhor ponto de corte

         Para pesquisa aproximada de kNN, os pontos de dados são particionados em microclusters e kNNs são pesquisados apenas dentro de cada micro cluster

         Algoritmo para ambos os modelos kNN-distância

  Outras abordagens

         Também aplicar algoritmos baseados em particionamento usando micro clusters [McCallum et al 2000], [Tao et al. 2006]

         Solução aproximada baseada em pontos de referência [Pei et al. 2006]

o Discussão

  A saída pode ser uma pontuação (modelos kNN-distância) ou uma etiquetagem (modelos kNNdistância e o modelo DB(ε,π)-outlier)

  As abordagens são locais (a resolução pode ser ajustada pelo usuário via ε ou k)

  Variante

o Detecção de Outlier usando o número de grau [Hautamaki et al. 2004]

  Idéia

         Construir o gráfico kNN para um conjunto de dados

o Vértices: pontos de dados

o Borda: se q∈kNN(p) então há uma aresta direcionada de p para q

         Um vértice que tem um grau menor que igual a T (limite definido pelo usuário) é um outlier

  Discussão

         O grau de um vértice no gráfico kNN é igual ao número de kNNs reversos (RkNN) do ponto correspondente

         Os RkNNs de um ponto p são aqueles objetos de dados que têm p entre seus kNNs

         Intuição do modelo: outliers são

o pontos que estão entre os kNNs de menos de T outros pontos

o têm menos de T RkNNs

         Gera um rótulo atípico

         É uma abordagem local (dependendo do parâmetro k definido pelo usuário)

  Fator de outlier baseado em resolução (ROF) [Fan et al. 2006]

o Modelo

  Dependendo da resolução dos limiares de distância aplicados, os pontos são discrepantes ou dentro de um cluster

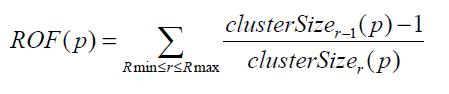
  Com a resolução máxima Rmax (limiar de distância mínima) todos os pontos são outliers

  Com a resolução mínima Rmin (limite de distância máxima) todos os pontos estão dentro de um cluster

  Altere a resolução de Rmax para Rmin para que, em cada etapa, pelo menos um ponto mude de outlier para membro de um cluster

  Cluster é definido similarmente como em DBSCAN [Ester et al 1996] como um fechamento transitivo de r-bairros (onde are é a resolução atual)

  Valor ROF



o Discussão

  Gera uma pontuação (o valor de ROF)

  A resolução varia automaticamente de local para global

## Abordagens baseadas em densidade

  Ideia geral

o Compare a densidade em torno de um ponto com a densidade em torno de seus vizinhos locais

o A densidade relativa de um ponto em comparação com seus vizinhos é calculada como uma pontuação atípica

o As abordagens também diferem na forma de estimar a densidade

  Pressuposto básico

o A densidade em torno de um objeto de dados normal é semelhante à densidade em torno de seus vizinhos

o A densidade em torno de um outlier é consideravelmente diferente da densidade em torno de seus vizinhos

  Local Outlier Factor (LOF) [Breunig et al. 1999], [Breunig et al. 2000]

o Motivação:

  Modelos de detecção de outliers baseados em distância têm problemas com densidades diferentes

  Como comparar a vizinhança de pontos de áreas de diferentes densidades?

  Exemplo

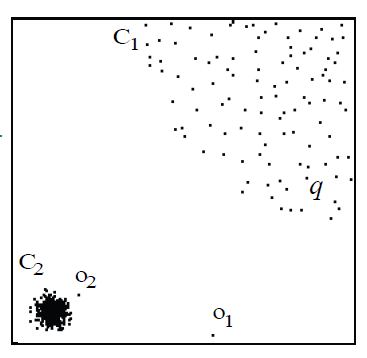
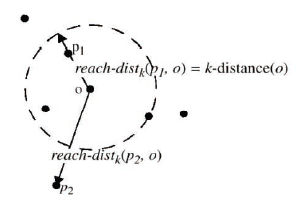
         Modelo DB(ε,π)-outlier

o Os parâmetros ε e π não podem ser escolhidos para que o2 seja um outlier, mas nenhum dos pontos no cluster C1 (por exemplo, q) é um outlier

         Outliers baseados em kNN-distância

o kNN-distâncias de objetos em C1 (por exemplo, q) são maiores do que a kNN-distância de o2

o Solução: considere a densidade relativa

o Modelo

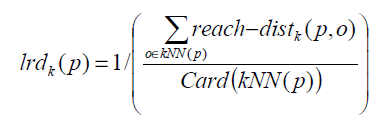
  Distância de acessibilidade

         Introduz um fator de suavização



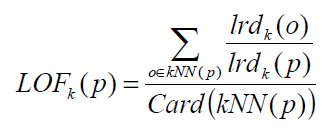
  Distância de acessibilidade local (lrd) do ponto p

         Inverso dos dists de alcance médios dos kNNs de p



  Fator de outlier local (LOF) do ponto p

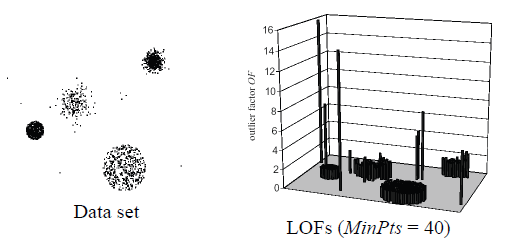
         Razão média de LRD de vizinhos de p e de lrd de p



o Propriedades

  LOF ≈ 1: ponto está em um aglomerado (região com densidade homogênea ao redor do ponto e seus vizinhos)

  LOF >> 1: ponto é um outlier



o Discussão

  A escolha de k (MinPts no artigo original) especifica o conjunto de referência

  Originalmente implementa uma abordagem local (a resolução depende da escolha do usuário para k)

  Gera uma pontuação (atribui um valor LOF a cada ponto)

  Variantes de LOF

o Mineração top-n outliers locais [Jin et al. 2001]

  Idéia:

         Normalmente, um usuário só está interessado nos outliers top-n

         Não calcular a LOF para todos os objetos de dados => salvar tempo de execução

  Método

         Comprimir pontos de dados em micro clusters usando os FCs de BIRCH [Zhang et al. 1996]

         Derivar limites superior e inferior das distâncias de acessibilidade, valores lrd e valores LOF para pontos dentro de um micro clusters

         Calcular limites superior e inferior de valores LOF para microclusters e classificar resultados w.r.t. limite inferior ascendente

         Prune micro clusters que não podem acomodar pontos entre os top-n outliers (n maiores valores de LOF)

         Refine iterativamente os micro cachos restantes e os pontos de poda de acordo

  Variantes da LOF (cont.)

o Fator outlier baseado em conectividade (COF) [Tang et al. 2002]

  Motivação

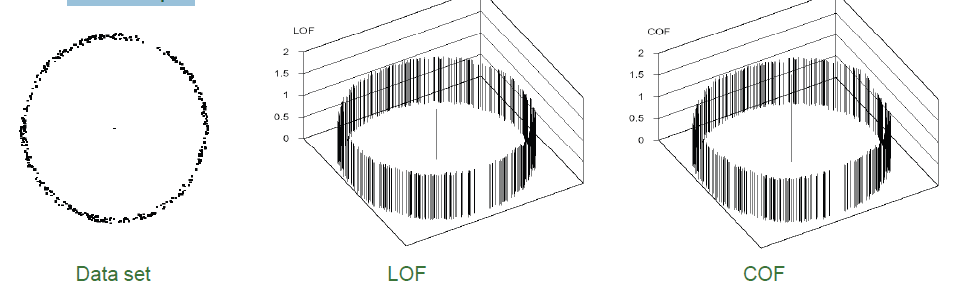
         Em regiões de baixa densidade, pode ser difícil detectar outliers

         Escolher um valor baixo para k muitas vezes não é apropriado

  Solução

         Tratar a "baixa densidade" e o "isolamento" de forma diferente

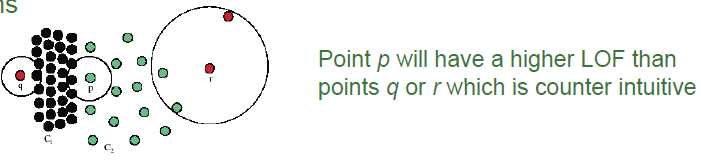
  Exemplo



  Outlierness Influenciado (INFLO) [Jin et al. 2006]

o Motivação

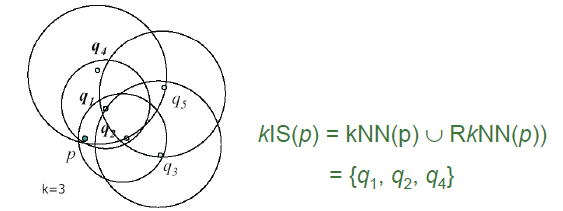
  Se aglomerados de diferentes densidades não forem claramente separados, a LOF terá problemas



o Ideia

  Leve em conta a relação simétrica de vizinhança

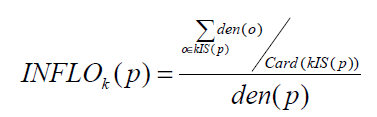
  O espaço de influência (kIS(p)) de um ponto p inclui seus kNNs (kNN(p)) e seus kNNs reversos (RkNN(p))



o Modelo

  A densidade é medida simplesmente pelo inverso da distância kNN, ou seja, den(p) = 1/k-distance(p)

  Influenciou o outlierness de um ponto p



  INFLO toma a razão entre a densidade média de objetos na vizinhança de um ponto p (ou seja, em kNN(p) ∪ RkNN(p)) e a densidade de p

o Algoritmos propostos para mineração de outliers top-n

  Baseado em índice

  Abordagem bidirecional

  Abordagem baseada em micro clusters

o Propriedades

  Semelhante ao LOF

  INFLO ≈ 1: ponto está em um cluster

  INFLO >> 1: ponto é um outlier

o Discussão

  Gera uma pontuação atípica

  Originalmente proposto como uma abordagem local (a resolução do conjunto de referência kIS pode ser ajustada pelo parâmetro de configuração do usuário k)

  Integral de correlação local atípica (LOCI) [Papadimitriou et al. 2003]

o Ideia é semelhante ao LOF e variantes

o Diferenças para LOF

  Tome o ε-bairro em vez de kNNs como conjunto de referência

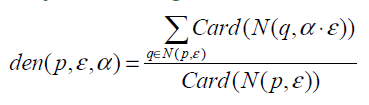
  Teste várias resoluções (aqui chamadas de "granularidades") do conjunto de referência para se livrar de qualquer parâmetro de entrada

o Modelo

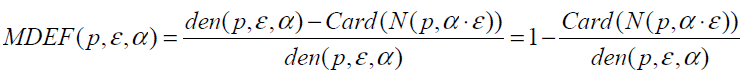
  ε-vizinhança de um ponto p: N(p,ε) = {q | dist(p,q) ≤ ε}

  Densidade local de um objeto p: número de objetos em N(p,ε)

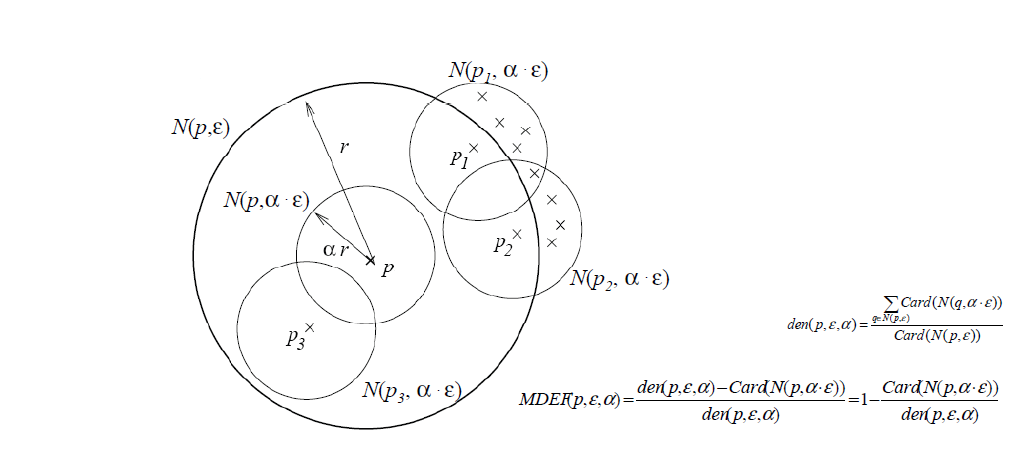
  Densidade média do bairro



  Fator de desvio de multigranularidade (MDEF)



* + Intuição



o σMDEF(p,ε,α) é o desvio padrão normalizado das densidades de todos os pontos de N(p,ε)

o Propriedades

  MDEF = 0 pontos para dentro de um cluster

  MDEF > 0 para outliers ou MDEF > 3.σMDEF => outlier

o Características

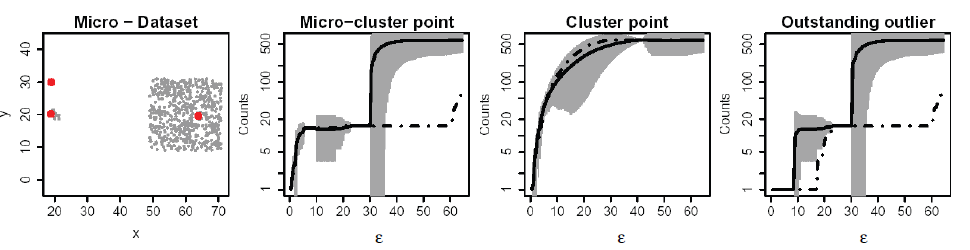
  Os parâmetros ε e α são determinados automaticamente

  Na verdade, todos os valores possíveis para ε são testados

  O gráfico LOCI exibe para um determinado ponto p os seguintes valores w.r.t. ε

         Cartão(N(p, α.ε))

         den(p, ε, α) com uma borda de +- 3.σden(p, ε, α)



* + Algoritmos

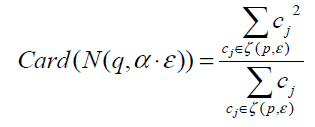
  A solução exata é bastante cara (calcular valores MDEF para todos os valores ε possíveis)

  aLOCI: solução rápida e aproximada

  Discretize o espaço de dados usando uma grade com comprimento lateral 2αε

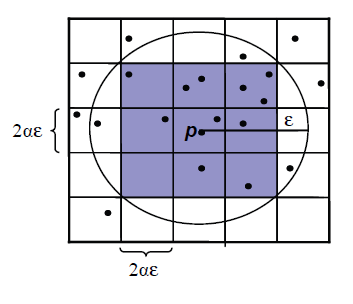
  Consultas de intervalo aproximado através de células de grade

  ε - vizinhança do ponto p: ζ(p,ε) todas as células que estão completamente cobertas por ε-esfera aroun p

Em seguida, onde cj é o objeto conte a célula correspondente

  Uma vez que diferentes valores de ε são necessários, diferentes grades são construídas com resolução variável

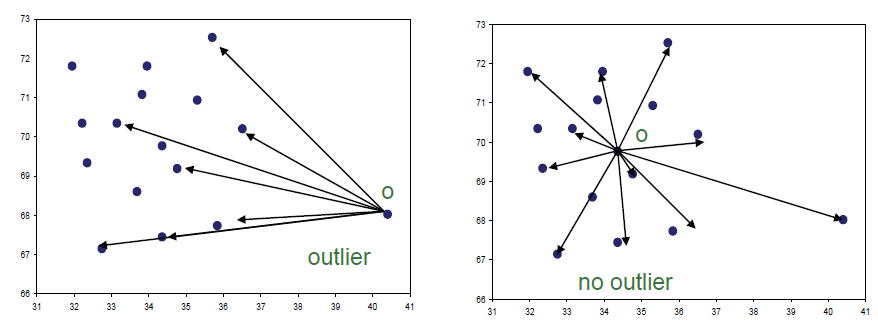
  These different grids can be managed efficiently using a Quad-tree



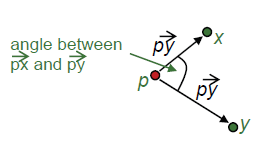
* + Discussão
    - Tempo de execução exponencial w.r.t. dimensionalidade de dados
    - Saída:
      * Rótulo: se MDEF de um ponto > 3.σMDEF então este ponto é marcado como outlier
      * Gráfico LOCI
        + Em qual resolução é um ponto um outlier (se houver)
        + Informações adicionais, como diâmetro de clusters, distâncias para clusters, etc.
    - Todas as resoluções interessantes, ou seja, possíveis valores para ε, (de local para global) são testadas

## Abordagens de alta dimensão

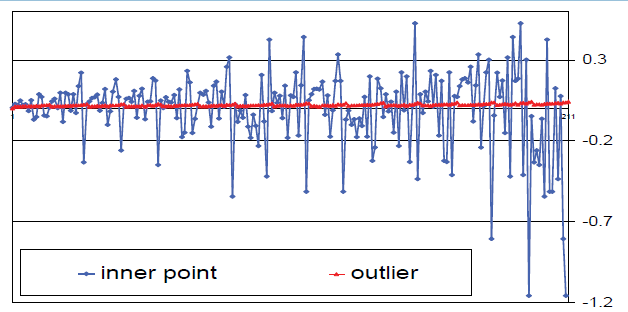
* o Desafios
  +  Maldição da dimensionalidade
    - • O contraste relativo entre as distâncias diminui com o aumento da dimensionalidade
    - • Os dados são muito esparsos, quase todos os pontos são discrepantes
    - • O conceito de bairro torna-se sem sentido
  +  Soluções
    - • Use funções de distância mais robustas e encontre outliers de dimensão total
    - • Encontre outliers em projeções (subespaços) do espaço de recurso original
* o ABOD – grau outlier baseado em ângulo [Kriegel et al. 2008]
  +  Racional
    -  Os ângulos são mais estáveis ​​do que as distâncias em espaços dimensionais altos (por exemplo, a popularidade de medidas de similaridade baseadas em cosseno para dados de texto)
    -  O objeto o é um outlier se a maioria dos outros objetos estiver localizada em direções semelhantes
    -  Objeto o não é atípico se muitos outros objetos estiverem localizados em direções variadas



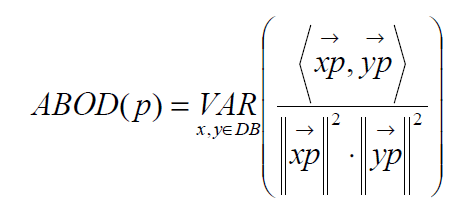
* Suposição básica
  + Outliers estão na borda da distribuição de dados
  + Os pontos normais estão no centro da distribuição de dados
* Modelo



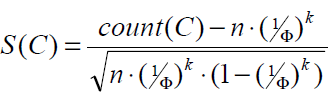
* + Considere para um determinado ponto p o ângulo entre px e py para quaisquer dois x, y do banco de dados
  + Considere o espectro de todos esses ângulos
  + A amplitude deste espectro é uma pontuação para a exceção de um ponto



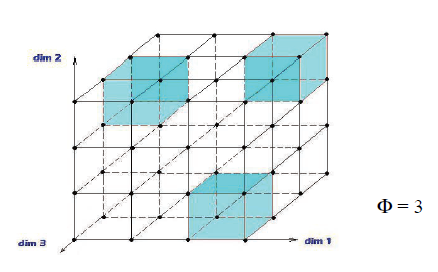
* + Medir a variância do espectro angular
  + Ponderado pelas distâncias correspondentes (para conjuntos de dados dimensionais inferiores, onde os ângulos são menos confiáveis)



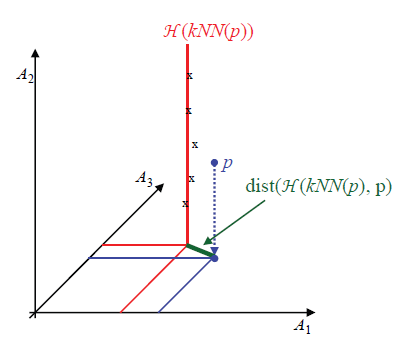
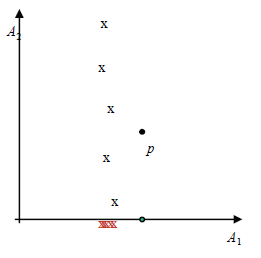
* Propriedades
  + ABOD pequeno => outlier
  + Alto ABOD => nenhum outlier
* Algoritmos
  + Algoritmo ingênuo está em O(n3)
  + Algoritmo aproximado baseado em amostragem aleatória para minerar top-n outliers
    - Não considere todos os pares de outros pontos x,y no banco de dados para calcular os ângulos
    - Calcular ABOD com base em amostras => limite inferior do ABOD real
    - Filtre os pontos que têm um limite inferior alto
    - Refinar (calcular o valor ABOD exato) apenas para um pequeno número de pontos
  + Discussão
    - Abordagem global para detecção de outliers
    - Emite uma pontuação atípica
* Detecção de subespaço baseado em grade [Aggarwal e Yu 2000]
  + Modelo
    - Espaço de dados de partição por uma grade de profundidade equivalente (Φ = número de células em cada dimensão)
    - Coeficiente de esparsidade S(C) para uma célula de grade k-dimensional C

, onde count(C) é o número de objetos de dados em C

* + - S(C) < 0 => count(C) é menor que o esperado
    - Outliers são aqueles objetos que estão localizados em células de menor dimensão com coeficiente de esparsidade negativo



* Algoritmo
  + Encontre as m células da grade (projeções) com os menores coeficientes de esparsidade
  + O algoritmo de força bruta está em O(Φd)
  + Algoritmo Evolutivo (entrada: m e a dimensionalidade das células)
* Discussão
  + Os resultados não precisam ser os pontos das células ideais
  + Modelo muito grosseiro (todos os objetos que estão na célula com menos pontos do que o esperado)
  + A qualidade depende da resolução da grade e da posição da grade
  + Emite uma rotulagem
  + Implementa uma abordagem global (critério chave: número globalmente esperado de pontos dentro de uma célula)
* SOD – grau outlier subespacial [Kriegel et al. 2009]
  + Motivação: Outliers podem ser visíveis apenas em subespaços dos dados originais
  + Modelo
    - Calcular o subespaço no qual os kNNs de um ponto p minimizam a variância
    - Calcule o hiperplano p)) que é ortogonal a esse subespaço
    - Tome a distância de p ao hiperplano como medida para sua dist(H (kNN(p) “outlierness”



* Discussão
  + Assume que kNNs de outliers têm uma projeção de dimensão inferior com pequena variância
  + A resolução é local (pode ser ajustada pelo usuário através do parâmetro k)
  + A saída é uma pontuação (valor SOD)

## Resumo

* Diferentes modelos são baseados em diferentes suposições para modelar outliers
* Diferentes modelos fornecem diferentes tipos de saída (rotulagem/pontuação)
* Diferentes modelos consideram outlier em diferentes resoluções (global/local)
* Assim, modelos diferentes produzirão resultados diferentes
* Ainda falta uma comparação completa e abrangente entre diferentes modelos e abordagens

Outlook

* Avaliação experimental de diferentes abordagens para entender e comparar diferenças e propriedades comuns
* Um primeiro passo para a unificação das diversas abordagens: fornecer pontuações discrepantes baseadas em densidade como valores de probabilidade [Kriegel et al. 2009a]: julgar o desvio da pontuação atípica em relação ao valor esperado
* Visualização [Achtert et al. 2010]
* Novos modelos
* Problemas de desempenho
* Tipos de dados complexos
* Dados de alta dimensão