Estudo de técnicas de detecção de anomalias e suas aplicações

Helena Almeida Victoretti e Luciana de Melo e Abud
Orientador: João Eduardo Ferreira
Co-orientador: Pedro Losco Takecian

Instituto de Matemática e Estatística - Universidade de São Paulo

Objetivo

- Estudo de caracterização e delimitação de algumas das principais técnicas de detecção de anomalias
- Escolha de uma técnica de detecção de anomalias para ser estudada e implementada em um módulo para o núcleo de um sistema de segurança transfusional de sangue

Introdução

- O que são anomalias?
  - Instâncias que não seguem um comportamento padrão esperado

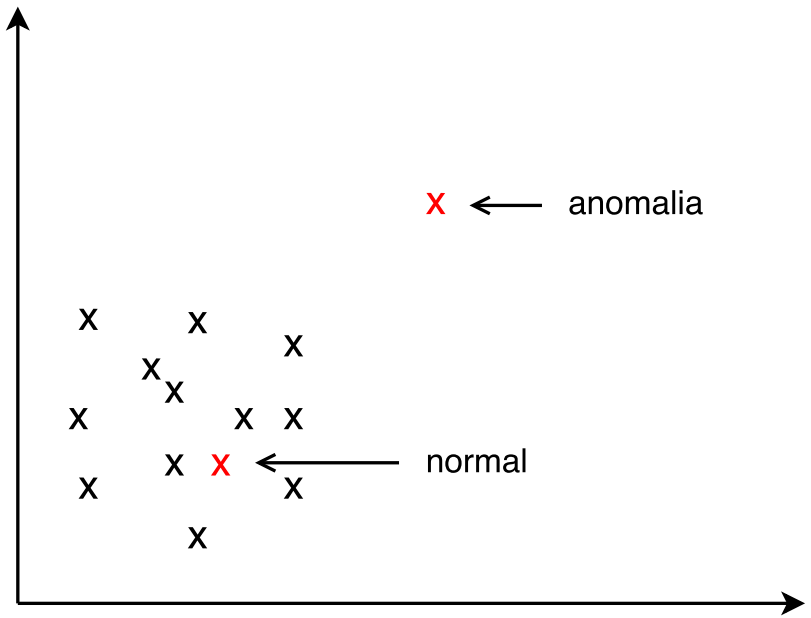


Figura 1: Exemplo de instâncias normais e de uma anomalia

Técnicas de Detecção de Anomalias

- São utilizadas em sistemas com diferentes finalidades:
  - Detecção de invasão em redes de computadores
  - Verificação de fraude bancária
  - Detecção de doença por análise de imagens
- Utilizam-se métodos de diferentes áreas, como aprendizado de máquina
- Escolha da técnica a ser utilizada depende basicamente do domínio do problema e dos dados de entrada

Técnicas de Detecção de Anomalias

Principais fases

- Fase de treino
  - Fase inicial, em que se define os parâmetros utilizados pelo detector
- Fase de teste
  - Determina se uma instância é normal ou anomalia

Classificação das técnicas

- Supervisionadas
- Semi-supervisionadas
- Não supervisionadas

Principais Técnicas de Detecção de Anomalias

- Baseadas em classificadores
- Baseadas em distância
- Baseadas em densidade
- Baseadas em *clustering*
- Baseadas em métodos estatísticos

Principais Técnicas de Detecção de Anomalias

Baseadas em Classificadores

- Classificam os dados recebidos entre anômalos ou normais
- Principais subcategorias:
  - Redes Neurais
  - Redes Bayesianas
  - Máquinas de Vetores Suporte (SVM)
  - Regras
- Vantagens:
  - Existência de muitos algoritmos poderosos de classificação
  - Rapidez do processo da fase de teste
- Desvantagens:
  - Necessidade de os dados de treino serem rotulados
  - Falta de uma atribuição de pontuação de anomalia associada às instâncias

Principais Técnicas de Detecção de Anomalias

Baseadas em Distância

- Classificam os dados entre anômalos e normais considerando a distância entre uma instância e seus vizinhos
- Dois principais métodos:
  - Método da vizinhança local
    - Uma instância O é uma anomalia se pelo menos uma fração de instâncias se encontra a uma distância maior que um determinado d de O
  - Método do k-ésimo vizinho mais próximo (kNN)
    - Uma instância O é uma anomalia se no máximo $n - 1$ outras instâncias O' possuem $D^k(O') > D^k(O)$
- Vantagem: Não são feitas hipóteses sobre a disposição dos dados
- Desvantagem: Ineficiência quando dados possuem alta dimensionalidade

Principais Técnicas de Detecção de Anomalias

Baseadas em Densidade

- Estimam a densidade da vizinhança de cada instância
- Uma instância que pertence a uma vizinhança pouco densa é considerada anomalia
- Principais métodos:
  - *Local Outlier Factor* (LOF)
  - *Connectivity-based Outlier Factor* (COF)
  - *Outlier Detection using In-degree Number* (ODIN)
  - *Influenced Outlierness* (INFLO)
  - *Multi-granularity Deviation Factor* (MDEF)
- Vantagens: Eficácia e a possibilidade de serem utilizadas em dados não rotulados
- Desvantagem: Custo computacional

Principais Técnicas de Detecção de Anomalias

Baseadas em *Clustering*

- Agrupam dados similares
- Principais subcategorias:
  - Regra 1: instâncias normais pertencem a algum *cluster* e instâncias anômalas a nenhum.
  - Regra 2: instâncias normais estão próximas do centro do *cluster* mais próximo e instâncias anômalas estão distantes do centro do *cluster* mais próximo.
  - Regra 3: instâncias normais pertencem a *clusters* grandes e densos, enquanto instâncias anômalas pertencem a *clusters* pequenos e esparsos.
- Vantagens: Possibilidade de serem utilizadas com dados não rotulados e a rapidez no processo da fase de teste
- Desvantagem: Dependência dos algoritmos de clustering

Principais Técnicas de Detecção de Anomalias

Baseadas em Métodos Estatísticos

- Baseiam-se na hipótese de que existe uma distribuição estatística que modela o conjunto de dados
- Uma anomalia é uma instância que não é modelada por essa distribuição
- Vantagens: Justificativa matemática para a classificação da instância entre normal ou anomalia e eficiência quando se obtém uma boa função de densidade da distribuição dos dados
- Desvantagens: Necessidade de existir uma distribuição estatística que modele os dados e performance limitada quando há poucas instâncias de treino

Validador Estatístico de Dados

Sistema de validação estatístico

- Dados provenientes de hemocentros brasileiros
- Estrutura de lote de dados
- Detecção de erros não capturados por validação sintática ou semântica utilizando detecção de anomalias
- Classificação de lotes em normais ou anomalias
- Acoplamento de técnicas de detecção de anomalias em um núcleo
  - Atualmente: método gaussiano

Processo de validação de instâncias

- Proporções dos valores do atributo como medida estatística

Fator Rh sanguíneo	Proporção no lote
Positivo	0.46
Negativo	0.45
Indeterminado	0.09

Tabela 1: Proporções das categorias do atributo fator Rh sanguíneo em um lote exemplo

Processo de validação de instâncias

- Proporções dos valores do atributo como medida estatística

Fator Rh sanguíneo	Proporção no lote
Positivo	0.46
Negativo	0.45
Indeterminado	0.09

Tabela 1: Proporções das categorias do atributo fator Rh sanguíneo em um lote exemplo

$$\text{Instância } x = (x_1, x_2, x_3) = (0.46, 0.45, 0.09)$$

Processo de validação de instâncias

- Proporções dos valores do atributo como medida estatística

Fator Rh sanguíneo	Proporção no lote
Positivo	0.46
Negativo	0.45
Indeterminado	0.09

Tabela 1: Proporções das categorias do atributo fator Rh sanguíneo em um lote exemplo

$$\text{Instância } x = (x_1, x_2, x_3) = (0.46, 0.45, 0.09)$$

- Eliminação de uma dimensão

Problema do método gaussiano

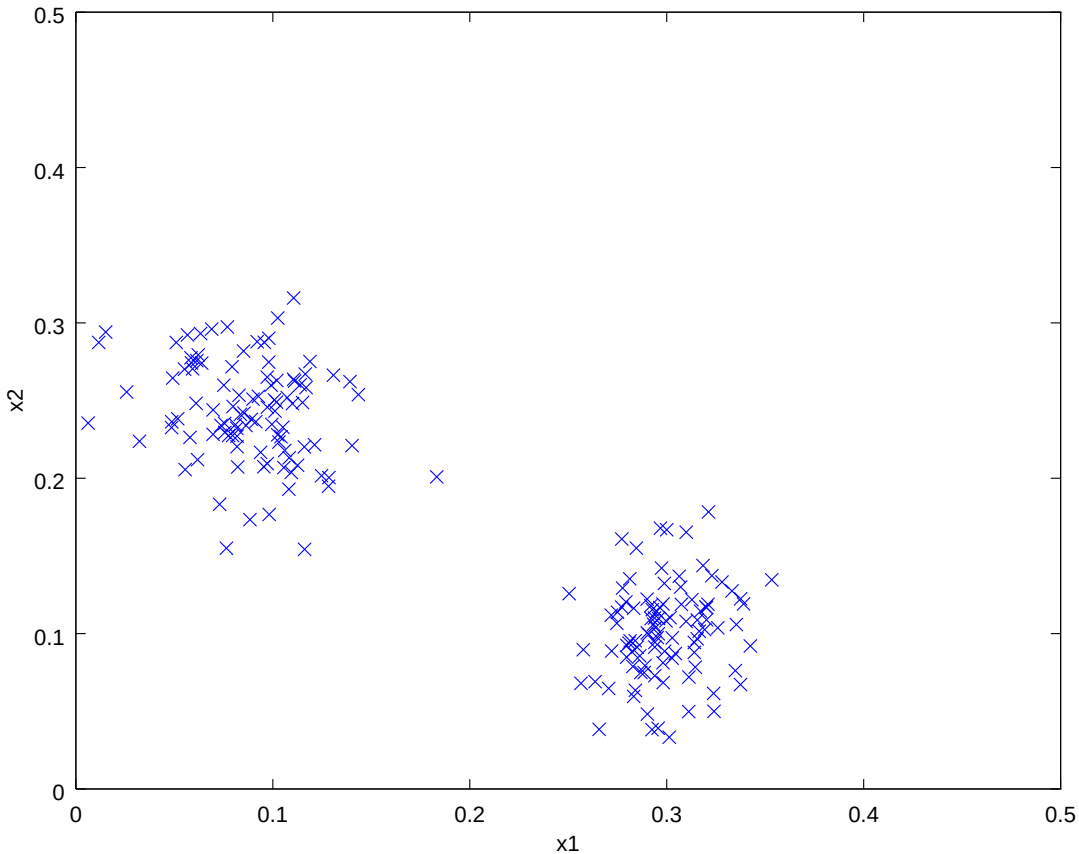


Figura 2: Exemplo de instâncias que não são adequadamente representadas pelo modelo gaussiano. Fonte: Pedro L. Takecian. Diretrizes metodológicas e validação estatística de dados para a construção de *data warehouses*, 2014 [4]

Detecção de Anomalias Baseada em Distância

Escolha da técnica baseada em distância

- Não são feitas hipóteses sobre a disposição dos dados
- Generalização de conceitos de métodos que assumem que os dados seguem uma certa distribuição
- Complexidade computacional menor do que técnicas baseadas em densidade

Detecção de Anomalias Baseada em Distância

Métodos

- Método da vizinhança local
  - "Uma instância O é uma anomalia se pelo menos uma fração de instâncias se encontra a uma distância maior que um determinado d de O"
  - Dependência do parâmetro D, difícil de ser encontrado
  - Não são atribuídas pontuações às anomalias
- Método do k-ésimo vizinho mais próximo (kNN)
  - "Uma instância O é uma anomalia se no máximo $n-1$ outras instâncias O' possuem $D^k(O') > D^k(O)$"
  - Dependência do parâmetro n

Método do k-ésimo vizinho mais próximo (kNN)

Principais algoritmos

- Algoritmo *simple nested-loop*
  - $O(N^2)$
- Algoritmo *index-based*
  - $O(N^2)$ no pior caso
- Algoritmo *partition-based*
  - $O(N^2)$ no pior caso
- Algoritmo *nested-loop ANNS*
  - Linear na prática
  - $O(N^2)$ no pior caso

Implementação do Módulo de Detecção de Anomalias

- Implementação em Python
  - NumPy, SciPy
- Algoritmo *nested-loop ANNS*
  - Treinamento para encontrar k e c (*threshold* para classificar instância em normal ou anomalia)
- Distância de Manhattan

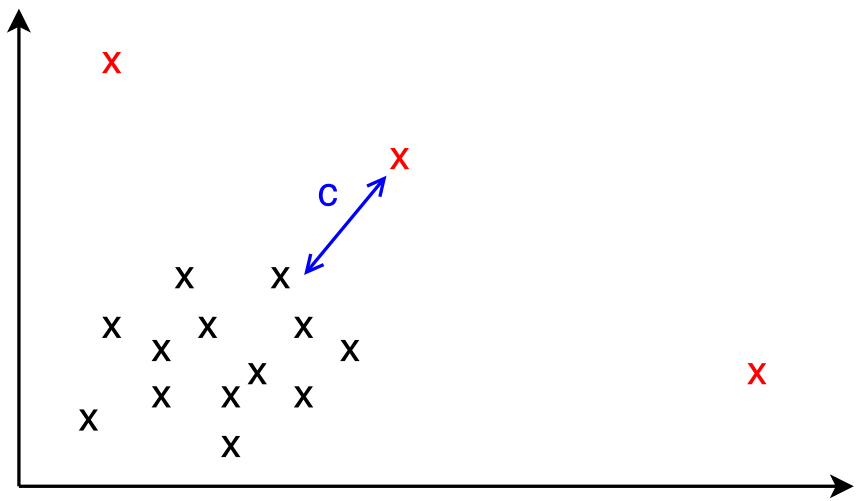
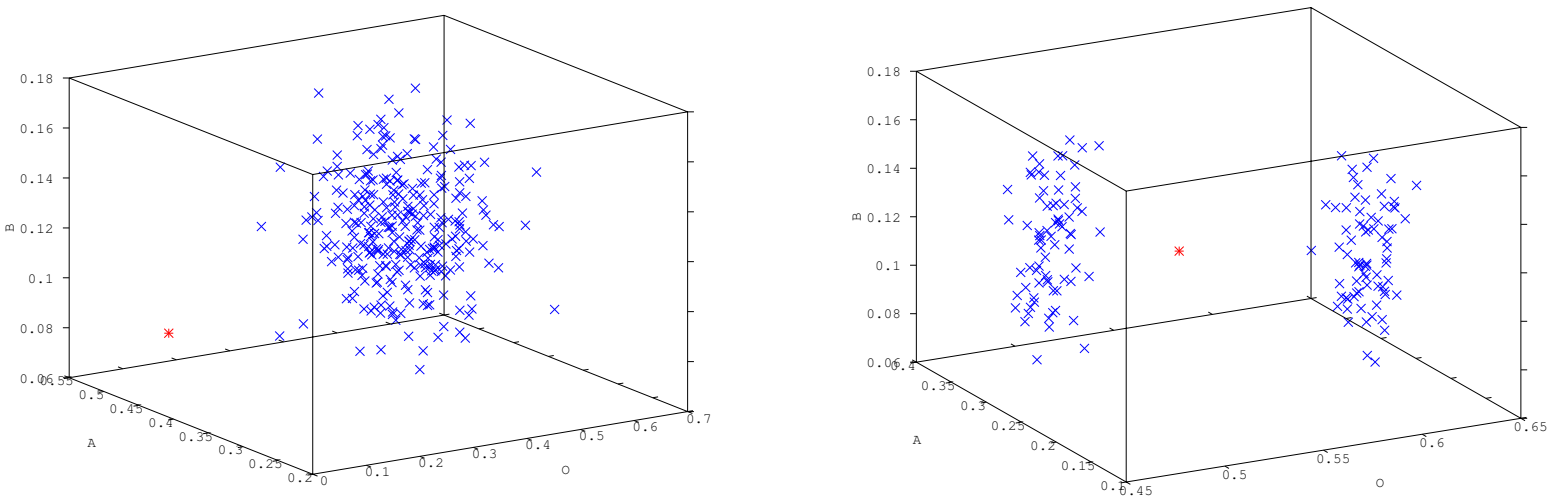


Figura 3: Escolha do *threshold* c para $k = 1$

Estudo de Caso

- Análise comparativa entre os métodos gaussiano e kNN



(a) Distribuição gaussiana

(b) Distribuição em clusters

Figura 4: Instâncias geradas seguindo diferentes distribuições. Cada eixo corresponde às proporções de cada categoria (O, A e B) em cada lote representado por um "x".

Conclusão

- Escolha da técnica de detecção de anomalias depende do domínio dos dados
- Apesar de funcionar bem em muitos casos, a técnica gaussiana não detecta anomalias quando os dados se encontram em uma disposição de *clusters*
- Técnicas de distância em geral possuem uma boa performance, entretanto são computacionalmente mais complexas

Referências

[1] Stephen D. Bay and Mark Schwabacher.
Mining distance-based outliers in near linear time with randomization and a simple pruning rule, 2003.

[2] Varun Chandola, Arindam Banerjee, and Vipin Kumar.
Anomaly detection: A survey.
ACM Comput. Surv., 41(3):15:1 – 15:58, 2009.

[3] Eduardo Dias Filho.
Implementação de um sistema de validação estatística configurável de dados, 2014.

[4] Pedro L. Takecian.
Diretrizes metodológicas e validação estatística de dados para a construção de *data warehouses*, 2014.

Estudo de técnicas de detecção de anomalias e suas aplicações

Helena Almeida Victoretti e Luciana de Melo e Abud
Orientador: João Eduardo Ferreira
Co-orientador: Pedro Losco Takecian

Instituto de Matemática e Estatística - Universidade de São Paulo