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
Universidade de São Paulo - Instituto de Matemática e Estatística

IME-USP

Introdução

Anomalias são instâncias de um conjunto de dados que não seguem um comportamento padrão esperado. Essas instâncias não possuem uma característica específica e são de grande interesse para os especialistas do domínio em questão, pois indicam eventos importantes e requerem maior atenção. Detecção de anomalias refere-se ao problema de encontrar tais instâncias [2].

Existem diferentes técnicas de detecção de anomalias que podem ser aplicadas em um conjunto de dados, e cada uma possui suas características. A escolha da técnica a ser utilizada depende basicamente do domínio do problema e dos dados de entrada fornecidos.

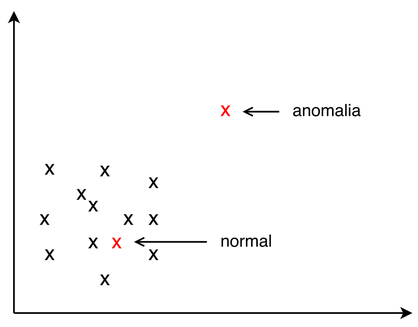

Figura 1: Exemplo de anomalia

Um sistema de validação estatística de dados que utiliza detecção de anomalias foi idealizado na tese de doutorado de Pedro Takecian [4] e desenvolvido por Eduardo Dias Filho [3]. O intuito do sistema é manter a consistência dos dados referentes a doações de sangue recebidos de diferentes hemocentros brasileiros, utilizando detecção de anomalias para encontrar lotes inválidos. Nesse contexto um lote é um conjunto de arquivos de dados referentes a um período de funcionamento do sistema transacional.

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

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

Os lotes são analisados pelo sistema através das proporções das categorias dos atributos de interesse, em que cada instância de um detector de anomalia verifica um atributo. A tabela 1 ao lado mostra um exemplo das proporções de um atributo.

Esse sistema possui um núcleo que tem como objetivo acoplar módulos com diferentes técnicas de detecção de anomalias.

Atualmente o sistema possui um módulo com um detector de anomalias baseado no modelo de detecção gaussiana.

Principais Técnicas de Detecção de Anomalias

Baseadas em Classificadores

Um classificador tem por objetivo categorizar os dados em classes. No contexto de detecção de anomalias essas classes correspondem às classes dos dados normais e dos dados anômalos.

Um exemplo dessas técnicas são algoritmos baseados em Máquinas de Vetores Suporte (*Support Vector Machines* – SVM), em que a principal ideia é encontrar uma fronteira de separação entre duas classes - a de instâncias normais e a de anomalias. Outras subcategorias dessas técnicas são as baseadas em redes neurais, redes bayesianas e em regras.

Baseadas em Distância

As técnicas dessa categoria levam em consideração a distância entre uma instância do conjunto de dados a ser analisada e seus vizinhos. Existem dois principais métodos que diferem entre si na definição de anomalia:

Método da vizinhança local: "Uma instância O em um conjunto T de dados é um *DB(p, D)-outlier* se pelo menos uma fração p ($0 < p < 1$) de instâncias em T se encontra a uma distância maior que D de O."

Método do k-ésimo vizinho mais próximo (*kNN – k nearest neighbors*): "Dado um conjunto de entrada com N instâncias e parâmetros n e k, uma instância p é um D_n^k-*outlier* se não mais que $n-1$ outras instâncias p' possuem $D^k(p') > D^k(p)$."

Baseadas em Densidade

As técnicas dessa categoria estimam a densidade da vizinhança de cada instância. Uma instância que pertence a uma vizinhança pouco densa é considerada anômala. A diferença entre as técnicas que pertencem a essa categoria são os métodos utilizados para atribuir a pontuação de anomalia às instâncias. Os principais métodos são: *LOF (Local Outlier Factor)*, *ODIN (Outlier Detection using In-degree Number)*, *INFLO (Influenced Outlierness)* e *MDEF (Multi-granularity Deviation Factor)*.

Baseadas em *Clustering*

Clustering é um método utilizado para agrupar dados similares. Os métodos utilizados nessa técnica diferem em relação à regra utilizada para determinar uma anomalia:

Regra 1: Assume-se que instâncias normais pertencem a algum *cluster* e instâncias anômalas a nenhum.

Regra 2: Assume-se que uma instância normal está próxima do centro do *cluster* mais próximo e uma instância anômala está distante do centro do *cluster* mais próximo.

Regra 3: Assume-se que instâncias normais pertencem a *clusters* grandes e densos, enquanto instâncias anômalas pertencem a *clusters* pequenos e esparsos.

Baseadas em Métodos Estatísticos

As técnicas dessa categoria se baseiam na hipótese de que existe uma distribuição estatística que modela o conjunto de dados. As instâncias anômalas são as que não são modeladas por essa distribuição. Os métodos utilizados nessa técnica podem ser divididos em dois grupos:

Paramétricos: Assume explicitamente um modelo estatístico para o conjunto de dados.

Não Paramétricos: Usa um modelo estatístico não paramétrico, tal que a estrutura da distribuição dos dados não é definida previamente, mas obtida a partir do conjunto de dados.

Aplicações de detecção de anomalias

Aplicação	Técnicas
Detectar fraude de cartão de crédito	Baseadas em Classificadores Baseadas em *Clustering*
Detectar ataques a rede de computadores	Baseadas em Classificadores Baseadas em *Clustering* Baseadas em Métodos Estatísticos Baseadas em Densidade
Detectar fraudes em redes de comunicação	Baseadas em Classificadores
Detectar comportamento anômalo através de imagens de vídeo	Baseadas em Distância
Testes de motores de jatos	Baseadas em Classificadores
Detecção precoce de surtos de doenças	Baseadas em Classificadores

Tabela 2: Aplicações de detecção de anomalias

Detecção de Anomalias Baseada em Distância

O módulo gaussiano desenvolvido para o núcleo do sistema de validação modela adequadamente muitas situações, mas em alguns casos outro método é necessário para representar melhor o problema. A figura abaixo exemplifica um caso desses:

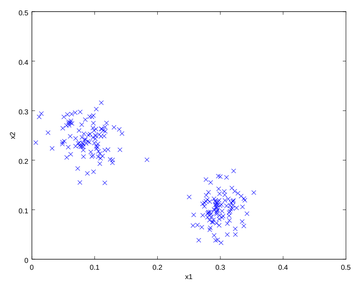

Tabela 3: Exemplo de instâncias que não são adequadamente representadas pelo modelo gaussiano. Fonte: Pedro L. Takecian. Diretrizes metodológicas e validação estatística de dados para a construção de *data warehouses*, 2014.

Escolhemos para estudo as técnicas baseadas em distância, pois se adequaria a situação acima, logo seria uma melhora para o sistema de validação. Além disso, as técnicas baseadas em densidade possuem um alto custo computacional, as baseadas em *clustering* dependem dos algoritmos de *clustering* que não são aptos a encontrarem anomalias, mas a formar *clusters*.

Entre os dois principais tipos de técnicas baseadas em distância optamos por escolher o método kNN. Existem quatro principais algoritmos que utilizam o método kNN: *simple nested-loop*, *index-based partition-based* e *nested-loop ANNS*.

O algoritmo *nested-loop ANNS* [1] se mostrou o melhor em termos de eficiência e simplicidade. Ele utiliza a ideia de *loops* encadeados e uma técnica de *prunning* simples: *Approximate Nearest Neighbour Search – ANNS*, em que a principal ideia é encontrar as n melhores anomalias.

Algorithm 1 Algoritmo *nested-loop ANNS*

entrada: k (número de vizinhos mais próximos a ser considerado), n (número de anomalias a serem encontradas), conjunto D de dados de entrada
saída: O (conjunto de anomalias encontradas)

```
Inicialize c = 0
Inicialize O como vazio
para cada b em D faça
    Inicialize Vizinhos(b) como vazio
    para cada d em D, d ≠ b faça
        se a quantidade de elementos em Vizinhos(b) for menor que k ou a distância entre b e d for
        menor que a distância entre b e os elementos de Vizinhos(b) então
            Atualize os valores de Vizinhos(b) com as k instâncias do conjunto Vizinhos(b) ∪d que
            estão mais próximas de b
            se D^k(b) < c então
                break
            fim se
        fim se
    fim para
    Atualize o conjunto O obtendo as n melhores anomalias entre as que estão no conjunto O ∪ b
    Atualize c com o valor da menor distância para os k-ésimos vizinhos mais próximos dos elementos que estão em O
fim para
```

Conclusão

Realizamos uma análise comparativa entre a técnica de detecção de anomalias baseada no método gaussiano e a baseada em distância. A figura ao lado mostra um conjunto de dados simulados para teste em que a técnica baseada em distância detecta a anomalia, porém a baseada no método gaussiano não. Podemos perceber que a distribuição dos dados interfere na qualidade da técnica de detecção de anomalias. Apesar disso, as técnicas baseadas em distância possuem uma maior complexidade computacional que a técnica de detecção gaussiana.

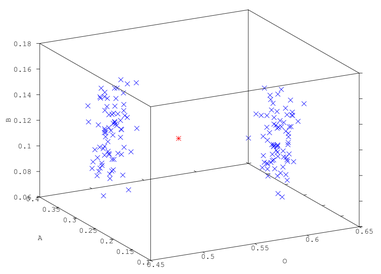

Figura 2: Conjunto de dados em que cada eixo corresponde a uma categoria de sangue (A, O ou B). A anomalia está em vermelho.

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.