Implementação de um sistema de validação estatística configurável de dados

Eduardo Dias Filho
Orientadores: João Eduardo Ferreira, Pedro Losco Takecian

Introdução

Definindo um lote como um volume de dados com muitas instâncias de uma ou mais estruturas de dados. Supondo que um lote tenha sofrido um erro em sua coleta de dados de forma que a estrutura do lote e os domínios de seus atributos ainda sejam respeitados mas os valores não condizem com a realidade. Assim, tal lote, mesmo inválido, seria aceito por validações sintáticas e semânticas.
Para detectar este tipo de erro é aplicada a validação estatística, que consiste em decidir se um novo lote é válido ou inválido através da técnica de detecção de anomalias, que utiliza as informações do novo lote e dos lotes que já foram classificados como válidos ou inválidos para tomar sua decisão.
Este trabalho visa implementar a validação estatística configurável de lotes, ou seja, implementar um algoritmo geral de detecção de anomalias e uma interface de configuração e gerenciamento do processo de validação estatística.

Detecção de Anomalias

Detecção de anomalias refere-se ao problema de encontrar exemplares em dados que não seguem o comportamento esperado em uma amostra. Tais exemplares são geralmente chamados de anomalias [1].

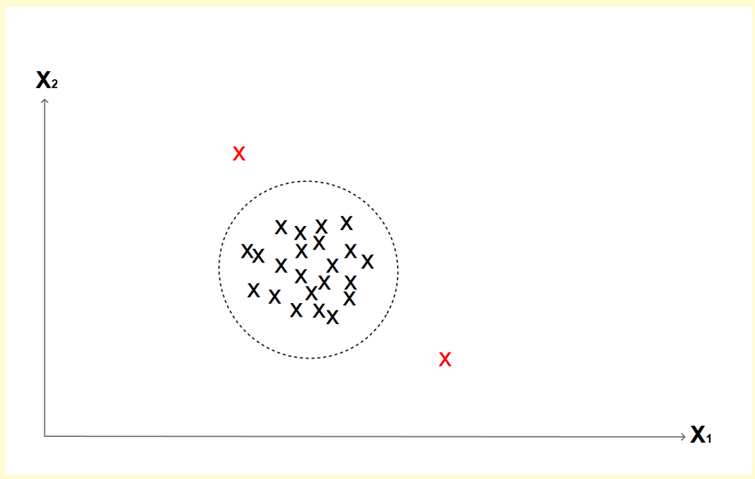

Figura 1: Exemplo de anomalias marcadas de vermelho em um conjunto de dados bidimensionais.

Há diversas técnicas para realizar a detecção, usaremos neste trabalho a técnica da detecção gaussiana multivariada, que é uma variante da técnica da detecção gaussiana.

Detecção Gaussiana

O método da detecção gaussiana assume que as características X_i de $X = (X_1, ..., X_n)$ seguem uma função de distribuição de probabilidade normal e são independentes entre si.
Portanto, sejam μ_i e σ_i respectivamente a média e a variância dos valores da característica i, calculados a partir de um conjunto de instâncias já classificadas como não-anômalas, então X_i segue a função de distribuição dada por:

$$P(X_i) = \frac{1}{\sigma_i\sqrt{2\pi}} e^{-(X_i-\mu_i)^2/2\sigma_i^2}$$

E como o método assume independência entre características:

$$P(X) = \prod_{i=1}^{n} P(X_i)$$

Detecção Gaussiana

Assim, para uma instância X:

Se $P(X) < \varepsilon$, então X é marcada como anomalia.
Se $P(X) >= \varepsilon$, então X é marcada como não-anômala.

Para decidir qual valor ε é o mais adequado para a decisão é usada a validação cruzada, na qual um conjunto de instâncias já classificadas que não foram usadas para calcular μ e σ é submetido ao algoritmo, que reclassifica as instâncias usando um determinado valor de ε.
Assim, para medir o desempenho do algoritmo usando um valor de ε, é usada a métrica F1-Score.

	Anomalia	Normal
Anomalia	TP	FP
Normal	FN	TN

Tabela 1: Cada linha representa uma possível resposta do algoritmo usando ε e cada coluna a classificação anterior da instância

Sejam $P = \frac{TP}{TP+FP}$ e $C = \frac{TP}{TP+FN}$, então:

$$\text{F1-Score} = \frac{2*P*C}{P+C}$$

Deste modo, é preciso encontrar um valor de ε que maximiza o F1-Score.

Detecção Gaussiana Multivariada

A detecção gaussiana multivariada assume que há dependência entre as características, portanto a função de probabilidade $P(X)$ deve ser calculada de forma diferente.

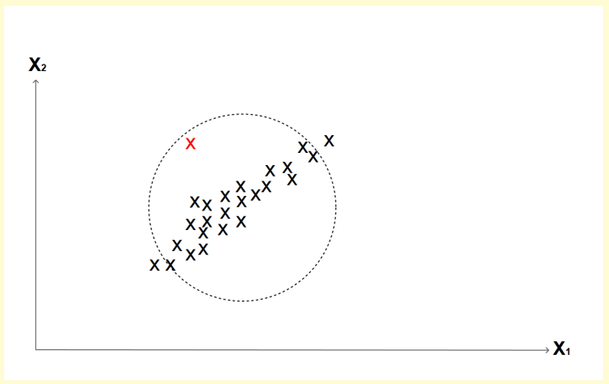

Figura 2: Exemplo que mostra a inadequação da detecção gaussiana quando há dependência entre características.

Na detecção gausiana multivariada, a função usada para comparar com ε é dada por:

$$P(X) = \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}} e^{(-\frac{1}{2}(X-\mu)^T\Sigma^{-1}(X-\mu))}.$$

Onde Σ é a matriz de covariância das características na amostra.

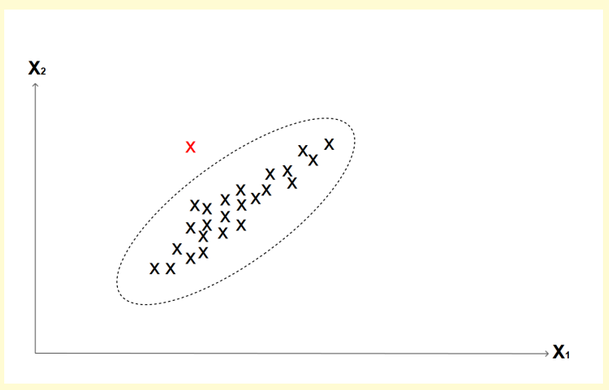

Figura 3: Exemplo da aplicação da detecção gaussiana multivariada.

Validação Estatística de Lotes

O processo de validação estatística de lotes consiste na aplicação da detecção de anomalias sobre um lote a ser classificado, usando um conjunto de treinamento composto de outros lotes já rotulados por um especialista de domínio ou pelo próprio algoritmo.
Cada detector de anomalia verifica os valores de um só atributo do lote, o detector recebe como instância uma tupla na qual as características representam as porcentagens de ocorrência de cada valor de tal atributo.

Interface de Gerenciamento

A interface de gerenciamento permite ao usuário configurar o processo de validação estatística, administrar o treinamento dos detectores de anomalia e executar a validação para novos lotes.
A configuração permite definir a estrutura dos lotes e quais os critérios serão levados em conta na detecção em tais lotes.

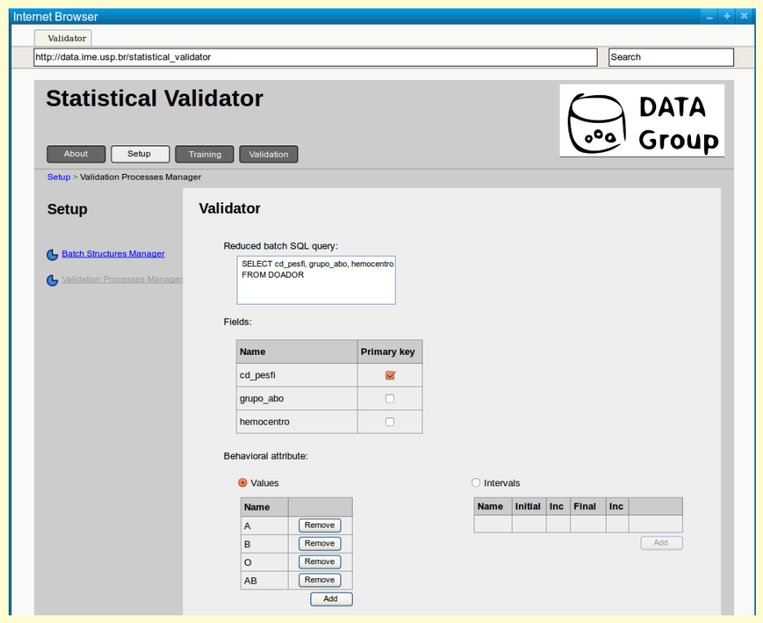

Figura 4: Exemplo de tela de configuração.

A administração do treinamento permite enviar instâncias rotuladas para melhorar os detectores de anomalia e decidir quais instâncias devem ser usadas ou não em um detector específico.

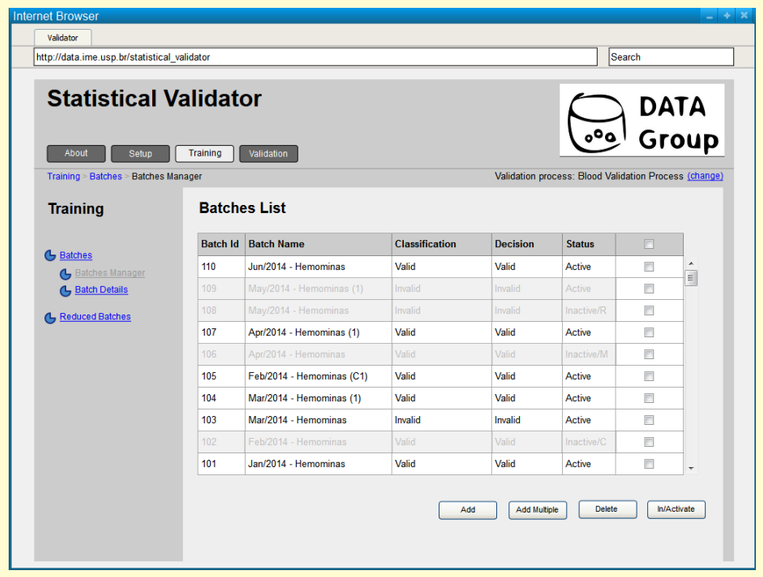

Figura 5: Exemplo de tela de administração de treinamento.

Referências

[1] V. Chandola; A. Banerjee and V. Kumar. Anomaly detection: A survey. *ACM Computing Surveys*, 41(3), July 2009.

[2] Andrew Ng. Machine learning, xv. anomaly detection (week 9), 2014. https://class.coursera.org/ml-005/lecture/preview.