Implementação de um sistema de validação estatística configurável de dados

Eduardo Dias Filho
Supervisores: João Eduardo Ferreira e Pedro Losco Takecian

16 de novembro de 2014

Table of Contents

Validação estatística de lotes

- Lotes: Conjuntos de instâncias de uma ou mais estrutura de dados.
- Problema: Quando o lote sofre um erro na coleta de dados que não é detectável por validações sintáticas ou semânticas.
- Solução: Validação estatística
- Através da detecção de anomalias, determinar se as informações do lote são provavelmente corretas ou provavelmente incorretas.

Interface de gerenciamento

- Implementado com o framework Django em Python, usando HTML, CSS e JavaScript. Usando Postgres como sistema gerenciador de banco de dados.
- Permitirá a configuração, administração e uso do processo de validação estatística.

Table of Contents

Detecção de anomalias

- Identificação de exemplares de dados que não seguem um comportamento esperado dentro de uma amostra de dados.
- Métodos usados: detecção gaussiana, detecção gaussiana multivariada.

Detecção gaussiana

Seja A uma amostra com dados já classificados em anômalos ou não-anômalos:

- Encontrar anomalias baseando-se no restante da amostra.
- Conjunto de treinamento: por convenção, cerca de 60% da amostra, apenas dados não-anômalos.
- Assume que cada característica X_i de X segue uma distribuição normal $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$ com média μ_i e variância σ_i^2 calculados usando o conjunto de treinamento.

Detecção gaussiana

- Assume também que as características são independentes entre si, portanto $P(X) = \prod P(X_i)$.
- Se $P(X) < \varepsilon$, então X é anomalia.
- Caso contrário, X é não anômalo.

Detecção gaussiana

Como escolher ε?

- Partir de um conjunto de possíveis valores de ε.
- Conjunto de validação cruzada: Os dados já classificados da amostra que não foram usados no conjunto de treinamento.
- Calcular para cada X no conjunto de validação cruzada $P(X)$ e comparar, para todo ε, a classificação do algoritmo com a classificação anterior.
- Medir o quão bem o algoritmo se saiu usando ε através da métrica F1-Score.

Detecção gaussiana

F1-Score

- VP = o número de verdadeiros positivos, ou seja, instâncias corretamente classificadas como anomalias.
- FP = o número de falsos positivos, ou seja, instâncias não-anômalas classificadas como anômalas.
- FN o número de falsos negativos, ou seja, anomalias classificadas como não-anômalas.
- P é a precisão dada por $P = \frac{VP}{VP+FP}$.
- C é a cobertura dada por $C = \frac{VP}{VP+FN}$.
- F1-Score $= \frac{2*P*C}{P+C}$

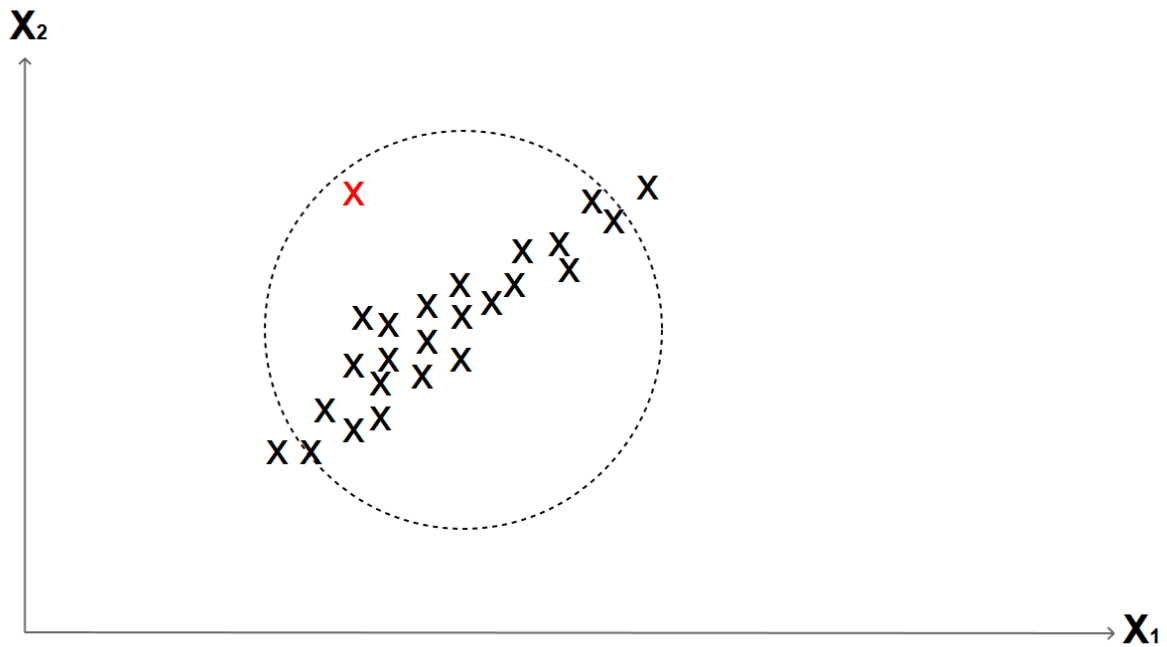
Encontrar um valor de ε que maximiza o F1-Score.

Detecção gaussiana multivariada

A detecção gaussiana multivariada assume que há dependência entre as características de X. Portanto P(X) deve ser calculado de outra forma.

- Seja Σ a matriz de covariância das características na amostra, onde $\Sigma_{ij} = COV(X_i, X_j)$.
- Seja μ o vetor de médias onde μ_i representa a média da característica X_i.
- Então X segue uma distribuição normal multivariada $X \sim \mathcal{N}(\mu, \Sigma^2)$ e P(X) é usada para se comparar a ε.

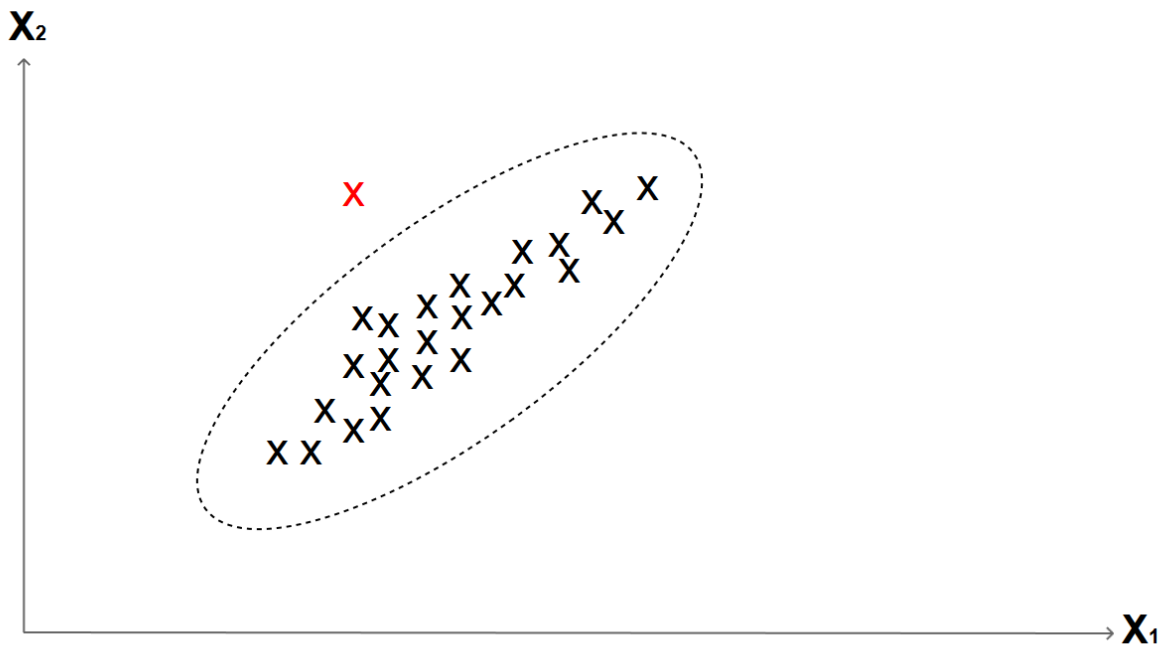
Detecção gaussiana multivariada



[1]

[1] Exemplo bidimensional que mostra a inadequação da detecção gaussiana quando há dependência entre características.

Detecção gaussiana multivariada



[2]Exemplo bidimensional da aplicação da detecção gaussiana multivariada.

Como se aplica na validação estatística de lotes

- Cada detecção de anomalias é aplicada sobre um atributo dos lotes, chamado de atributo comportamental.
- Seja uma categoria de um atributo como um possível valor ou intervalo de valores.
- Cada característica representa a porcentagem de ocorrência de uma categoria do atributo comportamental no lote.
- A amostra é composta de lotes já classificados em anômalos ou não-anômalos seja por um especialista de domínio quanto pelo próprio algoritmo. (aprendizado semi-supervisionado)

Table of Contents

Configuração

A configuração permite definir a estrutura dos lotes e quais os critérios serão levados em conta na detecção em tais lotes.

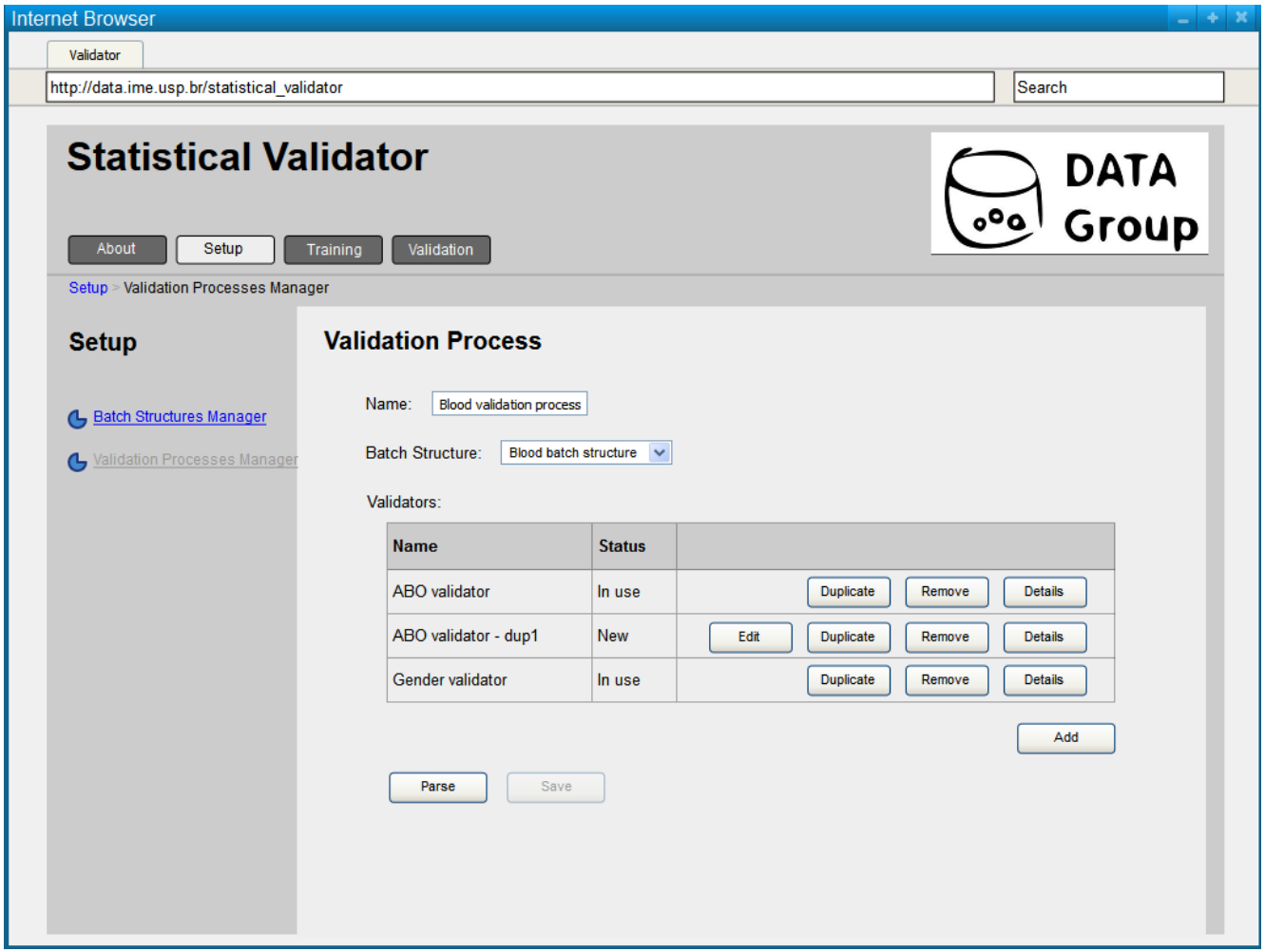
- Processo de validação: representa o processo de validação como um todo, é composto por uma estrutura de lotes e por um ou mais validadores.
- Estrutura de lote: representa como estão dispostos os dados em um lote. Quais arquivos fazem parte do lote e qual a organização e domínio dos atributos em um lote.

Configuração

Caso as distribuições das categorias do atributo comportamental se alterem dependendo dos valores de um ou mais atributos, chamamos estes de atributos contextuais.

- Contexto: Conjunto de valores de um ou mais atributos contextuais.
- Validadores físicos: Representam uma aplicação do algoritmo detector de anomalias sob determinado contexto.
- Validadores lógicos: Representam o nível mais alto dos validadores, podendo representar mais de uma aplicação do detector de anomalias, uma para cada contexto. Contém qual é o atributo comportamental, quais são os atributos contextuais e os contextos que definem seus validadores físicos.

Configuração



[3]

[3] Protótipo da tela de edição de um processo de validação.

Treinamento

Os dados de treinamento dos detectores são obtidos através do envio de lotes rotulados pelo especialista do domínio ou através do envio de lotes para a validação, classificados pelo próprio algoritmo.

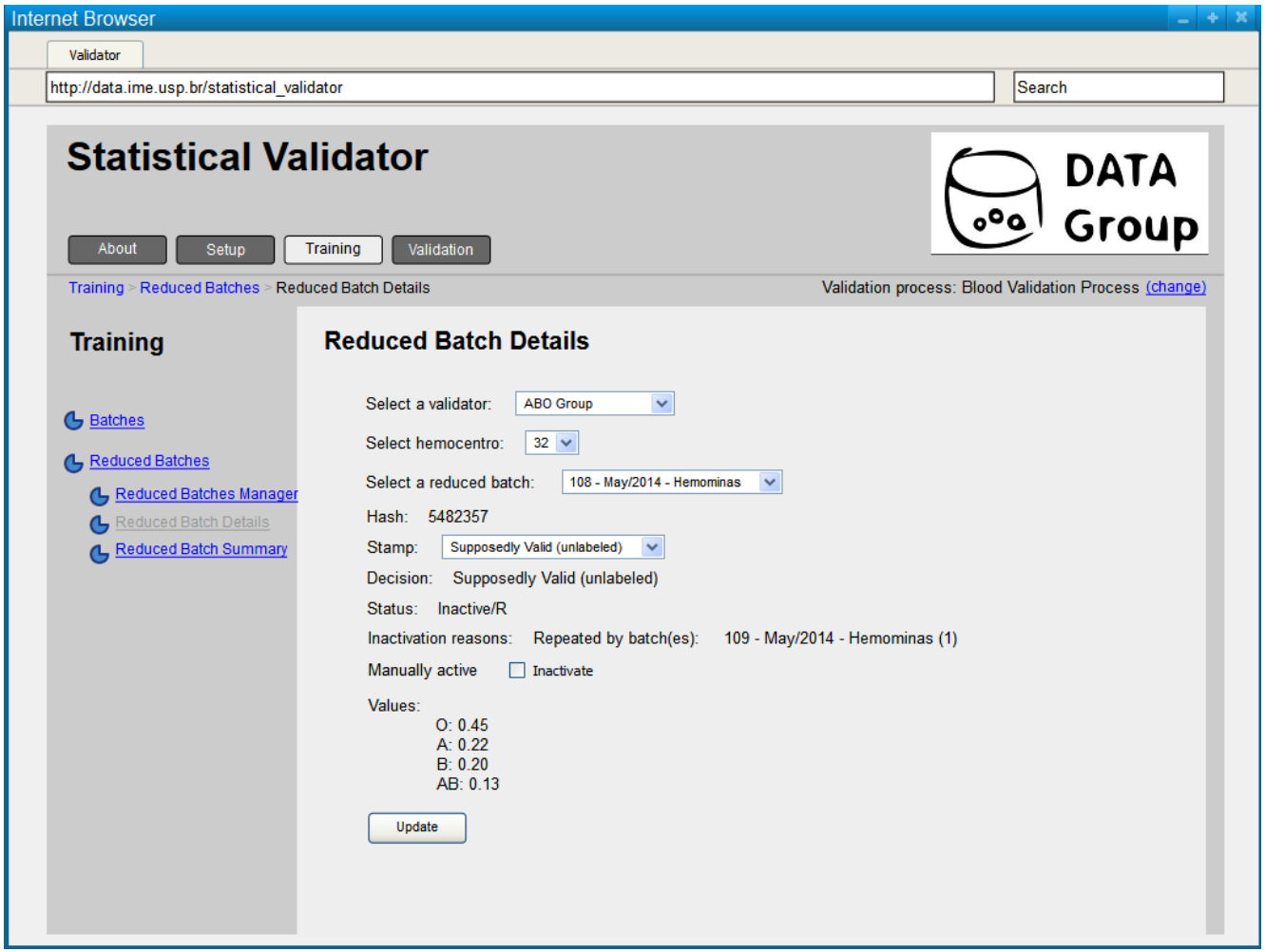
- Lotes reduzidos: A parte do lote usada em um detector de anomalias, ou seja, apenas as linhas do contexto levado em conta e apenas os valores do atributo comportamental. Um detector de anomalias usa lotes reduzidos como instâncias do conjunto de treinamento.
- Possíveis marcas para lotes reduzidos: Válido (rotulado), Supostamente válido (não rotulado), Desconhecido (não rotulado), Inválido (rotulado), Supostamente inválido (não rotulado).

Treinamento

Status: ativo ou inativo

- Manual: a interface permite mudar o status de lotes e lotes reduzidos.
- Sistêmica por repetição: o sistema inativa lotes e lotes reduzidos repetidos.
- Sistêmica por correção: o sistema inativa lotes que foram corrigidos por outros lotes.

Treinamento



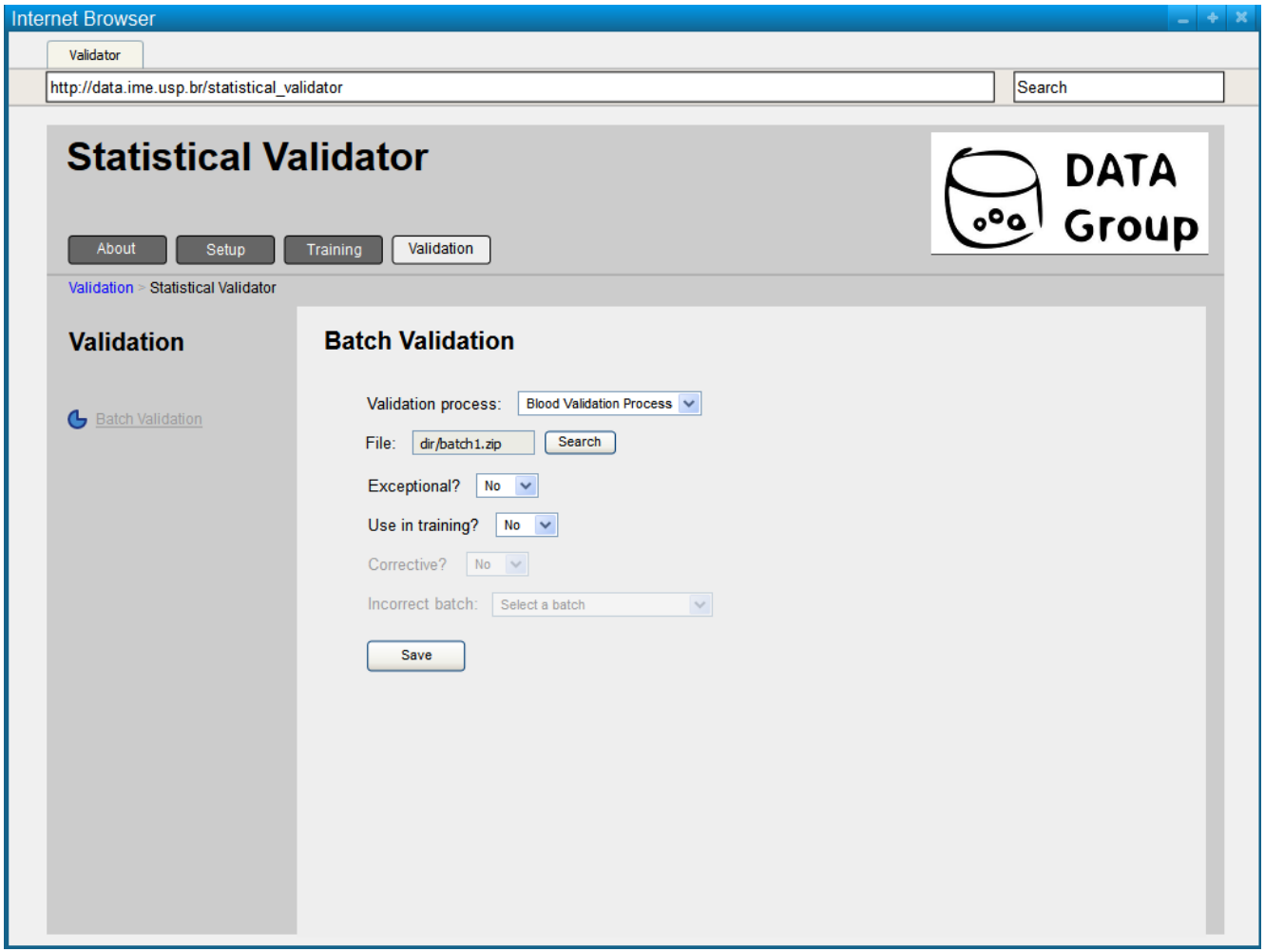
[4]

[4] Protótipo da tela de edição de um lote reduzido de treinamento.

Validação estatística

A interface permite ao usuário escolher um processo de validação pronto para uso e enviar lotes para a validação estatística.

Validação estatística



[5]

[5] Protótipo da tela de envio de lotes para a validação.