Departamento de Ciência da Computação
Instituto de Matemática e Estatística
Universidade de São Paulo

Trabalho de Formatura Supervisionado

Implementação de um sistema de detecção de anomalias

Eduardo Dias Filho

Orientadores:
Prof. Dr. João Eduardo Ferreira
Prof. Dr. Pedro Losco Takecian

Resumo

Sumário

1 Introdução

1.1 Motivação

Com o aumento do número de sistemas que recebem grandes volumes de dados de fontes diversas, mostrou-se necessário garantir a integração e corretude da informação recebida. Neste trabalho, será chamado de lote um volume de dados que contém muitas instâncias de uma ou mais estrutura de dados.

Para alcançar a integração e corretude dos lotes, os dados recebidos devem estar dispostos em um formato preestabelecido e cada atributo deve respeitar um determinado domínio, ou seja, o sistema deve realizar uma validação sintática para checar estas condições. Também é necessário que os dados estejam de acordo com um conjunto de regras lógicas predeterminadas, isto é, o sistema deve realizar uma validação semântica. Porém as medidas citadas são insuficientes para garantir a corretude, pois embora o lote siga as regras sistáticas e semânticas, ainda pode não representar a realidade por conta de algum erro de coleta antes do envio.

Pelo fato de considerarmos que um lote possui muitas instâncias, um lote que contenha um erro deste tipo pode ser detectado caso, devido ao erro, as porcentagens de ocorrência dos valores de um atributo diferem o suficiente das porcentagens de outros lotes considerados corretos a ponto de levantar suspeitas quanto seu processo de coleta. Ou seja, a partir de um conjunto de treinamento, ou seja, uma amostra de lotes previamente recebidos e classificados em válidos e inválidos por um especialista de domínio, é possível usar um algoritmo para decidir se um novo lote recebido é supostamente válido ou supostamente inválido. Tal técnica é conhecida como detecção de anomalias e é utilizada para realizar a validação estatística do lote.

1.2 Contexto

Inicialmente, o objetivo deste trabalho era implementar a validação estatística para o sistema de integração de dados Bloddis (Blood donation data integration system), que recebe lotes de diversos hemocentros. A validação estatística para os lotes do Bloddis consiste em uma série de detecções de anomalias sobre os atributos dos lotes selecionados para validação. Devido à complexidade da validação estatística, surgiu a necessidade de uma interface de configuração e treinamento para facilitar futuras alterações no processo de validação e realizar o controle dos conjuntos de treinamento.

Portanto, este trabalho visa não só implementar a detecção de anomalias como desenvolver uma interface para configurar o processo de validação estatística e gerenciar instâncias de treinamento. Deste modo, através da especificação da estrutura dos lotes enviados e de quais atributos serão usados para decidir se o lote é anômalo, a interface permite ao usuário adaptar a validação estatística a qualquer problema de validação estatística de lotes desejado.

2 Fundamentos

2.1 Detecção de anomalias

Sendo $A = \{A_1, A_2, ..., A_n\}$ uma amostra de dados, uma instância A_i é considerada uma anomalia quando se desvia acentuadamente do restante das outras instâncias de A. Por serem tão diferentes do padrão, anomalias podem causar falhas críticas quando ocorrem em sistemas, gerando a necessidade de detectá-las previamente.

Em geral, as técnicas de detecção de anomalias visam rotular uma instância A_i como anômala ou não-anômala a partir de uma amostra A de instâncias já rotuladas.

2.2 Método da distribuição gaussiana

Considerando $A_i = (X_1, ..., X_m)$ e assumindo que toda característica X_K segue uma distribuição gaussiana e é independente das outras características de A_i, é possível usar as instâncias restantes de A, que chamaremos de instâncias de treinamento, para descobrir a média μ_k e variância σ_k^2 de cada característica X_k para usar a função de distribuição de probabilidade normal dada por $P(X_k) = \frac{1}{\sigma_k\sqrt{2\pi}} e^{-(X_k-\mu_k)^2/2\sigma_k^2}$, e como estamos assumindo independência entre características, seja $P(X_k^{(i)})$ o equivalente a $P(X_k)$ calculado para a característica X_k de A_i, podemos calcular $P(A_i) = \prod_{k=1}^{m} P(X_k^{(i)})$ para decidir se uma nova instância A_i é considerada uma anomalia na amostra.

Pela definição de anomalia, nas amostras de treinamento usadas na detecção de anomalias o número de instâncias rotuladas como não-anômalas ser bem maior do que o número de instâncias rotuladas como anômalas. Deste modo, é possível usar uma parte das instâncias rotuladas como não-anômalas para inferir os parâmetros, por convenção, usa-se 60% das instâncias de treinamento não-anômalas para tal tarefa.

Com os parâmetros da normal calculados, o algoritmo deve partir de um parâmetro ε pré-definido e para cada uma das instâncias A_j restantes de treinamento, ou seja, os outros 40% das instâncias rotuladas não-anômalas e as rotuladas anômalas, calcular $P(A_j)$ e caso $P(A_j) < \varepsilon$, a instância de treinamento é classificada como anômala, caso contrário, é classificada como não-anômala. Esta etapa do treinamento visa achar o ε que maximiza o acerto do algoritmo, comparando a classificação dada pelo algoritmo para as instâncias de treinamento com os rótulos originais de tais instâncias.

Para medir o quão preciso é o algoritmo para um certo ε geralmente é usada uma medida chamada F1-Score, tal que, seja 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 e FN o número de falsos negativos, ou seja, anomalias classificadas como não-anômalas, então F1-Score = $\frac{2*P*C}{P+C}$, onde P é a precisão dada por $P = \frac{VP}{VP+FP}$ e C é a cobertura dada por $C = \frac{VP}{VP+FN}$. Deste modo, o maior valor de F1-Score é 1 e o menor é 0, o objetivo desta etapa é encontrar um valor de ε para o qual F1-Score = 1.

Encontrado o parâmetro ε, o treinamento se encerra e a próxima etapa é classificar uma nova instância A_i ainda não rotulada. Portanto, calcula-se $P(A_i)$ para a instância A_i e caso $P(A_i) \geq \varepsilon$, então A_i é rotulada como uma instância não-anômala, caso $P(A_i) < \varepsilon$, A_i é considerada uma anomalia na amostra A.

2.3 Método da distribuição gaussiana multivariada

Assumindo agora que há dependência entre as características da instância A_i, devido à dependência das características as funções de probabilidade $P(X_k)$ não podem ser obtidas, portanto devemos usar outro método para a detecção de anomalias, a distribuição gaussiana multivariada. Este método é similar ao anterior, mas leva em conta a dependência entre variáveis para calcular $P(A_i)$.

Seja $\mu \in \mathbb{R}^m$ o vetor no qual μ_i é a média da variável i e Σ a matriz de covariância, calculada por $\Sigma_{i,j} = COV(X_i, X_j)$, onde $COV(X_i, X_j)$ é a covariância entre as variáveis X_i e X_j na amostra A, sejam X_i e X_j características da instância A_k simbolizadas por $X_i^{(k)}$ e $X_j^{(k)}$, então $COV(X_i, X_j) = \frac{1}{n}\sum_{k=1}^{n}(X_i^{(k)} - \mu_i)(X_j^{(k)} - \mu_j)$.

A detecção multivariada usa $P(A_i) = \frac{1}{(2\pi)^{\frac{m}{2}}|\Sigma|^{\frac{1}{2}}} e^{(-\frac{1}{2}(A_i-\mu)^T\Sigma^{-1}(A_i-\mu))}$ para comparar com ε e decidir se a instância é ou não anômala, onde $|\Sigma|$ é o determinante da matriz de covariância.

3 Tecnologias

3.1 Interface de configuração e treinamento

3.1.1 Django

3.1.2 Postgres

3.2 Algoritmos de detecção de anomalias

3.2.1 Numpy e Scipy

4 Implementação

4.1 Processo de validação estatística

O processo de validação estatística de lotes consiste basicamente na aplicação de detectores de anomalias sobre um lote a ser classificado, usando um conjunto de treinamento composto de outros lotes rotulados como válidos ou inválidos por um especialista de domínio.

Cada detector de anomalia verifica os valores de um só atributo do lote, chamado de atributo comportamental. O detector recebe como instância uma tupla $(X_1, ..., X_m)$ na qual as características representam cada valor do atributo comportamental do lote a ser validado e suas respectivas porcentagens de ocorrência no lote. Por exemplo, um detector de anomalias que age sobre um lote de doação de sangue onde o atributo comportamental é o grupo ABO, pode receber como instância a tupla $(O = 0.48, A = 0.36, B = 0.12, AB = 0.04)$. Caso o atributo comportamental apresente valores numéricos, as categorias representam intervalos de valores.

Portanto, para obter as características da instância, o algoritmo de detecção de anomalias deve saber como estão organizados os dados no lote. Do mesmo modo, cada detector também precisa saber qual o atributo comportamental e outros detalhes sobre a validação. As informações sobre os dados pode ser obtida através da estrutura de lote e as informações de validação através de validadores, ambas as entidades, validadores e estruturas de lote podem ser configuradas pelo usuário e fazem parte de um processo de validação.

4.1.1 Estruturas de lote

Como descrito na seção 1.1, um lote é composto de instâncias de estruturas de dados, tais estruturas de dados são representadas por um conjunto de atributos e para cada estrutura há um arquivo de valores separados por vírgulas (CSV - Comma Separated Values) no qual as linhas representam as instâncias que compõe o lote com os valores dos atributos especificados. Uma estrutura de arquivo consiste no conjunto de atributos do arquivo e no nome do arquivo dentro do lote.

Chamamos de estrutura de lote o conjunto de uma ou mais estruturas de arquivos. Todo processo de validação deve utilizar uma única estrutura de lote, mas a mesma estrutura de lote pode ser usada por mais de um processo de validação.

Uma estrutura de lote também tem um campo de estado que indica de a estrutura está em uso por algum processo de validação ou se a estrutura é nova e não está em uso. Se a estrutura estiver com o estado "nova"ela pode ser alterada e removida através da interface com o usuário, porém se o estado for "em uso", uma alteração ou remoção causaria erros em todos os processos de validação que usam a estrutura de lote, portanto, neste caso não são permitidas alterações ou remoções. Independentemente do estado de uma estrutura de lote, a interface permite ao usuário realizar uma duplicação da estrutura de lote, originando uma nova estrutura com a mesma especificação da estrutura duplicada, mas com o estado "nova", ou seja, uma cópia que não está em uso.

Configurando a estrutura de lote, o usuário pode adaptar o lote ao modelo desejado e usar a estrutura de lote em um processo de validação estatística.

4.1.2 Validadores

Supondo que em um caso de validação estatística, as porcentagens de ocorrência dos valores do atributo comportamental dependam de alguma forma dos valores de um ou mais atributos do lote, que chamaremos aqui de atributos contextuais. Assim, para cada diferente

valor dos atributos contextuais pode existir um diferente padrão de distribuição dos valores do atributo comportamental. O que dificulta a detecção de anomalia pois uma ocorrência que seria considerada anômala em um dos padrões, pode ser considerada normal quando consideramos todos os padrões.

Chamando de contexto uma tupla de valores dos atributos contextuais. Para resolver o problema de dependência entre atributos a detecção de anomalia é fragmentada, para cada contexto há uma detecção de anomalias separada. Assim, os validadores, que são estruturas usadas na representação da detecção, são segmentados em validadores lógicos e validadores físicos.

Validadores lógicos representam de forma geral todos os possíveis contextos de um atributo comportamental

Validadores físicos representam as informações de cada uma destas detecções separadas por contexto.

Dependendo do problema, pode ser desejável utilizar outro tipo de algoritmo para detectar anomalias. A interface permite configurar o algoritmo usado pelo validador. Como um determinado contexto pode exigir um algoritmo diferente, no validador lógico é definido apenas o padrão, ou seja, se não for especificado um algoritmo no validador físico, a detecção de tal contexto deve usar o algoritmo padrão definido no validador lógico.

Durante a configuração dos validadores, a interface também permite agrupar categorias do atributo comportamental, assim ao invés de usar as porcentagens de ocorrência das categorias que fazem parte de um grupo, o validador usa a soma de tais porcentagem como se fosse uma só categoria. Também é possível agrupar contextos, deste modo os dados de mais de um contexto que estão no mesmo grupo são classificados pelo mesmo detector de anomalias e portanto, pelo mesmo validador físico.

4.1.3 Gerenciamento de processos de validação

4.2 Treinamento

O treinamento dos detectores de anomalias podem ser feitos de duas maneiras. A primeira é pelo envio de lotes já rotulado por um especialista de domínio. A segunda é através do envio de lotes para validação, que após serem classificados são usados no conjunto de treinamento de validações futuras. Como para realizar a validação de um lote é necessário ter um conjunto de treinamento antes, então é sempre necessário que o treinamento se inicie com o primeiro método.

Quando um processo de validação recebe um lote de treinamento, para cada um de seus validadores físicos são geradas uma estrutura chamada de lote reduzido que representa a parte importante do lote para o detector de anomalia representado pelo validador físico. Um lote reduzido é composto pela coluna que representa o atributo comportamental de todas as linhas do lote recebido que estão dentro do contexto de seu respectivo validador físico.

Para que sejam usados no treinamento, os lotes reduzidos precisam ser rotulados, por isso possuem um campo “marca”. A marca do lote reduzido pode indicar os seguintes valores:

Válido (rotulado) representa que o lote reduzido foi confirmado como válido por um especialista de domínio.

Supostamente válido (não rotulado) representa que o lote reduzido passou pelo processo de validação e foi considerado válido pelo algoritmo, portanto ele é provavelmente válido, mas não confirmado por especialistas de domínio.

Desconhecido (não rotulado) representa que o lote reduzido foi enviado para treinamento sem passar pelo processo de validação e ainda não recebeu rótulo pelo especialista de domínio.

Inválido (rotulado) representa que o lote reduzido foi confirmado como inválido por um especialista de domínio.

Supostamente inválido (não rotulado) representa que o lote reduzido passou pelo processo de validação e foi considerado inválido pelo algoritmo, portanto ele é provavelmente inválido, mas não confirmado por especialistas de domínio.

Para o detector de anomalias, lotes reduzidos com as marcas "Válido (rotulado)", "Supostamente válido (não rotulado)"ou "Desconhecido (não rotulado)"são considerados não anômalos, ou seja, válidos (1). Já nos lotes reduzidos com as marcas "Inválido (rotulado)"ou "Supostamente inválido (não rotulado)"são considerados como anomalias, ou seja, inválidos (0).

Os lotes enviados ao processo de validação também possuem marcas, que são dadas pela conjunção lógica das marcas de seus respectivos lotes reduzidos. Assim, caso todas as marcas de seus lotes reduzidos forem consideradas válidas (1), a marca do lote terá valor "válido", caso uma das marcas de seus lotes reduzidos seja considerada inválida (0), então a marca do lote será "inválido".

Um lote também tem um estado que pode ter valor "ativo"ou "inativo". Um lote inativo deixa de participar do conjunto de treinamento de seus detectores. Os lotes reduzidos também tem seus próprios estados, mas quando um lote está inativo, todos seus lotes reduzidos também estão inativos. Quando um lote está ativo, seus lotes reduzidos podem estar ativos ou inativos. A inativação de lotes ou lotes reduzidos pode acontecer manualmente ou automaticamente.

Para lotes, a inativação automática pode ocorrer por dois motivos. O primeiro caso é quando ocorre a submissão de um lote corretivo, ou seja, um lote que corrige um lote anteriormente enviado, o que pode viesar o treinamento dos validadores que usam o lote, portanto, sempre que um lote corretivo é enviado para treinamento, o lote corrigido é inativado pelo sistema. O segundo caso também pode ocorrer para lotes reduzidos, é quando ocorre o envio de um lote ou lote reduzido repetido, para evitar o viés causado pela repetição no conjunto de treinamento, todos os lotes ou lotes reduzidos que são repetidos por um lote ou lote reduzido mais atual são inativados pelo sistema.

A interface permite que o usuário mude o status de lotes ou lotes reduzidos manualmente em alguns casos. A inativação manual de lotes e lotes reduzidos é permitida pela interface, mas a ativação manual só é possível caso o lote ou lote reduzido está apenas manualmente inativo e não automaticamente inativo pelo sistema.

4.3 Detectores de anomalias

5 Conclusão