Proposta

Eduardo Dias Filho
7278100

29 de abril de 2014

1 Supervisores

João Eduardo Ferreira
Pedro Losco Takecian

2 Tema

Detecção de anomalias e análise de componentes principais

3 Objetivo

O trabalho visa criar um serviço que receba lotes de dados e realize, usando as ferramentas de aprendizagem computacional, como detecção de anomalias e análise de componentes principais, a validação estatística de dados sobre tais lotes. Ou seja, a partir de lotes já classificados e dos atributos de um novo lote, classifica o novo lote em válido ou anômalo.

4 Contexto e justificativa

O sistema Bloddis (`https://data.ime.usp.br/bloddis`) é alimentado com o envio de lotes de dados providos pelos principais hemocentros do Brasil.

Quando o sistema recebe um lote, dois tipos de validação são feitos pelo sistema. Primeiro a validação sintática, na qual verificamos se os tipos dos campos recebidos são os esperados. A próxima etapa é a validação semântica, na qual usamos um conjunto de regras para buscar por absurdos. Caso qualquer uma das validações encontre um erro, o sistema exibe uma mensagem de erro e não realiza a carga dos dados.

Outro tipo de validação se mostrou necessária neste sistema, a validação estatística dos lotes. Por se tratarem de um conjunto de dados, os lotes estão sucetíveis a erros que vão além da percepção das validações até aqui descritas. Erros como repetições ou remoções de linhas ou de valores de atributos ao longo das linhas.

Para detectar este tipo de erro, é necessário fazer a validação estatística, que usa métodos de aprendizagem de máquina para, a partir de lotes já classificados como válidos ou anômalos no banco de dados, classificar um novo lote.

Exemplificando, supondo um lote com erro no qual a quantidade total de coletas de sangue é 5.000 e temos 4.000 instâncias de coletas com grupo sanguíneo AB, ou seja, 80% das instâncias como AB. Supondo também que,

dentre os lotes já classificados, temos, nos exemplos válidos, de 2% a 5% de ocorrências de coletas do grupo AB e em um dos lotes já classificados como anomalia temos 65% de AB. Usando os dados dos lotes já classificados, um algoritmo de detecção de anomalias é capaz de detectar o erro e classificar corretamente o lote como anômalo.

A validação estatística submete o lote a diversos classificadores previamente configurados por um outro serviço, que permitirá a parametrização do classificador como, por exemplo, escolher os atributos considerados e o algoritmo de classificação.

Um arquivo de lote tem muitas instâncias e o serviço de validação estatística precisa processá-lo e ainda aplicar os diversos classificadores para certos atributos do lote, portanto esta validação não acontece imediatamente.

Com o envio de um lote que passou pelas duas primeiras etapas de validação, o serviço será executado em background e, no término da validação estatística, enviará uma resposta ao usuário via email.

5 Principais atividades previstas

- Planejar e implementar um serviço de configuração dos validadores
- Planejar e implementar o processamento e armazenamento dos dados
- Implementar os classificadores
- Estudar e aplicar mais algoritmos de classificação
- Integrar o serviço e implementar respostas

6 Cronograma

Atividade	Mai	Jun	Jul	Ago	Set	Out	Nov
Serviço de configuração dos validadores	X	X	X				
Processamento e armazenamento de dados	X	X	X	X			
Implementação dos classificadores		X	X	X	X		
Integração do serviço				X	X	X	
Monografia		X	X	X	X	X	X
Pôster						X	X
Apresentação						X	X