Estudo de técnicas de detecção de anomalias e suas aplicações

Helena Almeida Victoretti
Luciana de Melo e Abud

Proposta do Trabalho de Conclusão de Curso
Instituto de Matemática e Estatística
da
Universidade de São Paulo

Orientador: Prof. Dr. João Eduardo Ferreira
Coorientador: Dr. Pedro Losco Takecian

São Paulo, 2015

1 Motivação

Detecção de anomalia refere-se ao problema de encontrar instâncias que apresentam um grande afastamento das demais, ou que são inconsistentes com elas. Tais instâncias são denominadas anomalias. Nas palavras de D. Hawkins [1], anomalia é "uma observação que se desvia tanto das outras observações a ponto de levantar suspeita de que ela foi gerada por um mecanismo diferente". Por frequentemente indicarem eventos importantes, anomalias requerem uma maior análise para que se possa chegar a uma conclusão sobre suas causas.

Técnicas de detecção de anomalias são úteis em áreas como mineração de dados, limpeza de dados e *data warehousing* [2]. Elas são utilizadas em sistemas com diferentes finalidades. Alguns exemplos são:

- Detecção de invasão em redes de computadores [3]: os padrões de comportamento da rede são analisados e são gerados mapeamentos de transmissão de seus pacotes. Uma anomalia detectada nesses mapeamentos pode indicar uma invasão na rede.
- Verificação de fraude bancária [4]: a detecção de uma anomalia neste contexto pode significar uma fraude no uso de cartão de crédito. As fraudes são refletidas em registros de transações e correspondem a pagamentos de alto valor, compra de itens que nunca foram adquiridos anteriormente pelo cliente, taxa elevada de compras, entre outros.
- Detecção de doença por análise de imagens [5]: a anomalia é detectada no processamento de imagens de ressonância magnética de mama, e pode indicar presença de tumor.

Existem diferentes técnicas de detecção de anomalias que podem ser aplicadas em um conjunto de dados, e cada uma possui suas características. Elas utilizam métodos de diferentes áreas, como aprendizado de máquina, e podem ser baseadas em classificadores, em métodos estatísticos, em distâncias, em densidade, entre outros [6].

Técnicas que utilizam aprendizado de máquina podem ser classificadas entre supervisionadas (dados de entrada rotulados como anômalos ou normais), semissupervisionadas (alguns dados de entrada rotulados e outros não) e não supervisionadas (dados de entrada sem nenhum rótulo) [7].

A escolha da técnica a ser utilizada depende basicamente do domínio do problema e dos dados de entrada fornecidos [8]. Por exemplo, uma técnica capaz de tratar instâncias quantitativas contínuas com mais de uma característica (multivariadas) é o método de distribuição gaussiana [9], que é um método de detecção estatístico semissupervisionado, baseado na teoria das densidades probabilísticas das curvas gaussianas. Ele recebe um conjunto X = $\{x^{(1)}, x^{(2)}, ..., x^{(m)}\}$ com m instâncias de treinamento não rotuladas, e um conjunto V = $\{v^{(1)}, v^{(2)}, ..., v^{(r)}\}$ com r instâncias de treinamento rotuladas.

Cada instância $x^{(i)}$ é um vetor de n características contínuas, para i = 1, 2, ..., m. Então:

$$x^{(i)} \in \mathbb{R}^n \text{ , e } x^{(i)} = [x_1^{(i)}, x_2^{(i)}, ..., x_n^{(i)}], i = 1, 2, ..., m.$$

Cada instância $v^{(i)}$ também é um vetor de n características contínuas, mas a cada uma está associado um rótulo $y^{(i)}$ que contém a real classificação da instância, representada por 1 (no caso de anomalia) e 0 (no caso de normalidade). O par é representado por ($v^{(i)}$, $y^{(i)}$), para i = 1, 2, ..., r.

Supõe-se que cada uma das características x_i segue uma distribuição normal com média μ_i e variância σ_i^2. Os cálculos da média e da variância são dados por:

$$\mu_i = \frac{1}{m} \sum_{j=1}^{m} x_i^{(j)}$$

$$\sigma_i^2 = \frac{1}{m} \sum_{j=1}^{m} (x_i^{(j)} - \mu_i)^2$$

Assim, temos:

$$\forall i \in \mathbb{Z}, 1 \leq i \leq n,\ x_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$$

E então a função de probabilidade de cada característica x_i é dada por:

$$p(x_i, \mu_i, \sigma_i^2) = \frac{1}{\sigma_i(\sqrt{2\pi})} e^{-\frac{1}{2}(\frac{x_i - \mu_i}{\sigma_i})^2}$$

A função densidade de probabilidade conjunta de uma instância $x = [x_1,\ x_2,\ ...,\ x_n]$ é descrita por:

$$\mathrm{p}(x) = \prod_{j=1}^{n} p(x_j, \mu_j, \sigma_j^2)$$

Dessa forma, dado um limiar ε, uma instância x com $\mathrm{p}(x) < \varepsilon$ será classificada como anômala, enquanto que uma instância x com $\mathrm{p}(x) \geq \varepsilon$ será classificada como normal. Esse limiar é escolhido de forma que seja o que maximiza a eficiência do algoritmo. Para medir a eficiência, é escolhida uma métrica que quantifique a eficiência do algoritmo quando executado sobre o conjunto V de instâncias rotuladas e um limiar ε .

Outras técnicas utilizam maneiras diferentes para detectar anomalias, partindo de diferentes premissas e tratando de outros domínios.

Além de saber aplicar cada técnica, é importante entender a característica e o funcionamento de cada uma.

2 Objetivo

Este projeto tem como objetivo estabelecer um estudo de caracterização e delimitação de algumas das diferentes técnicas de detecção de anomalias.

Para atingir esse objetivo, o primeiro passo será estudar o sistema de validação de lotes de dados que foi desenvolvido para o projeto REDS-II *(Retrovirus Epidemiology Donor Study - II)*, uma iniciativa cujo intuito é desenvolver projetos de pesquisa com enfoque em segurança transfusional de sangue, e cujo banco de dados armazena dados provenientes de diversos anos de doações de seus hemocentros participantes.

Esse validador utiliza métodos de detecção de anomalias para a validação de dados, a fim de manter a consistência e a integridade do banco de dados, uma vez que inserções de novas entradas podem levar o banco a um estado inconsistente ou até mesmo inutilizá-lo. Esses métodos são implementados seguindo uma certa interface e obedecendo a certas regras, e depositados no núcleo do sistema como módulos.

Também é de interesse desse trabalho estudar os módulos já implementados para o núcleo do sistema de validação, além de escolher uma técnica de detecção de anomalias que seja adequada ao domínio do problema para ser implementada em um módulo para esse núcleo. Após implementada, a técnica será utilizada em um conjunto de dados a ser escolhido, de modo a testar a implementação e os resultados das análises esperados.

3 Plano de Estudo

Neste plano de estudo, incluímos as seguintes atividades:

1. Levantamento das principais técnicas e algoritmos para detecção de anomalia.
2. Estudo aprofundado de pelos menos três técnicas, levantadas no item 1.
3. Descrição e caracterização das técnicas estudadas.
4. Escolha de uma técnica para implementação de um módulo do sistema de validação no ambiente de segurança transfusional de sangue.
5. Teste da técnica implementada usando estudo de caso de segurança transfusional de sangue.
6. Elaboração de um documento que descreve a implementação, bem como o estudo de caso dos itens anteriores.

4 Cronograma de Atividades

As atividades do projeto serão desenvolvidas de acordo com o cronograma da tabela abaixo, em que os tópicos são os descritos na seção 3.

Tópico	Mês							
	Abril	Maio	Junho	Julho	Agosto	Setembro	Outubro	Novembro
1	×							
2		×	×					
3			×	×				
4					×	×		
5						×	×	
6			×	×	×	×	×	×

Tabela 1: *Cronograma de atividades do projeto*

Referências

[1] D.M. Hawkins. *Identification of Outliers*. Monographs on applied probability and statistics. Chapman and Hall, 1980. 1

[2] H. Jair Escalante. A comparison of outlier detection algorithms for machine learning, 2005. 1

[3] Paulo M. Mafra, Joni da Silva Fraga, Vinicius Moll, and Altair Olivo Santin. POLVO-IIDS: Um sistema de detecção de intrusão inteligente baseado em anomalias, 2008. 1

[4] Prakash N. Kalavadekar Amruta D. Pawar and Swapnali N. Tambe. A survey on outlier detection techniques for credit card fraud detection. *IOSR Journal of Computer Engineering*, 16(2):44–48, 2014. 1

[5] Clay Spence, Lucas Parra, and Paul Sajda. Detection, synthesis and compression in mammographic image analysis with a hierarchical image probability model, 2001. 1

[6] Ji Zhang. Advancements of outlier detection: A survey. *EAI Endorsed Trans. Scalable Information Systems*, 1:e2, 2013. 1

[7] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM Comput. Surv.*, 41(3):15:1 – 15:58, 2009. 1

[8] Pedro L. Takecian. Diretrizes metodológicas e validação estatística de dados para a construção de *data warehouses*, 2014. 1

[9] Machine learning: Lecture xv. anomaly detection. `https://class.coursera.org/ml-005/lecture/preview`. Universidade de Stanford, Plataforma Coursera. 1