Comparação de Técnicas Estatísticas para Detecção de Anomalias em Tráfego de IoT

Miqueas Galdino dos Santos, Paulo Ribeiro Lins Júnior

Grupo de Pesquisa em Sistemas de Comunicações – GCOM
Instituto Federal de Educação, Ciência e Tecnologia da Paraíba
Campus Campina Grande
Email: miqueasgaldino@gmail.com, paulo.lins@ifpb.edu.br

***Resumo*—Neste trabalho, são avaliadas cinco técnicas estatísticas para a detecção de anomalias em redes de sensores sem fio, infra estrutura básica para o paradigma da Internet das Coisas. As técnicas são comparadas com relação a sua acurácia e ao desempenho computacional, considerando o tempo de execução como métrica, considerando os cenários de redes *indoor* e *outdoor*, para os casos de medidas em configurações *singlehop* e *multihop*.**

***Index Terms*—IoT, Detecção de Anomalias, Redes de Sensores sem Fio**

I. Introdução

Dados representam um valioso ativo dentro do padarigma de Internet das Coisas (IoT – *Internet of Things*). A qualidade dos dados obtidos do sensoriamento das redes IoT, mais que um parâmetro, é uma necessidade que consumidores e gerentes desse tipo de rede tem, exigindo isso do próprio sistema.

Uma questão crítica no contexto de IoT é o surgimento de anomalias (ou anormalidades) na medições feitas pelos nós [1].

Dados corrompidos por ruído ou que não representam medidas reais podem ser oriundas da faltas, como a completa falha do *hardware* do nó sensor, de falhas intermitentes no funcionamento do sensor, da transmissão contínua de sinais de níveis muito altos ou muito baixos ou de interferência de outros equipamentos [2].

Nesse trabalho, são avaliadas cinco técnicas estatísticas de detecção de anomalias em redes de sensores sem fio, infraestrutura básica de IoT, selecionadas principalmente pela sua simplicidade de execução, tendo em vista a necessidade de uso na detecção em tempo real em tráfego de redes de sensores sem fio. Em comum, todos os métodos trabalham com a definição de intervalos de verificação, chamados aqui de Iv, em que dados que estejam fora deles são considerados anômalos. Essa abordagem é simples, pois não necessita de comparação com padrões de tráfego ou de técnicas mais complexas de verificação de distribuição de probabilidade dos dados medidos. São considerados dois cenários distintos, e as técnicas são comparadas com relação a acurácia e a eficiência computacional. O restante do artigo se apresenta da seguinte forma: na Seção II são apresentadas as técnicas de detecção de anomalia estudadas, na Seção III são apresentados os resultados e as discussões das avaliações e na Seção IV, as considerações finais.

II. Técnicas de Detecção de Anomalias

A. Desvio Padrão (SD)

Nesse método, considera-se como anomalia todos os dados fora do intervalo definido por

$$Iv = \bar{x} \pm 2SD, \tag{1}$$

em que $\bar{x}$ é media da amostra de dados e $2SD$, o dobro do desvio padrão.

B. Desvio Absoluto da Mediana (MAD)

O intervalo baseado no desvio absoluto da mediana é dado por

$$Iv = \tilde{x} \pm 2.96(Md|x_i - \tilde{x}|), \tag{2}$$

em que $\tilde{x}$ é a mediana do grupo de dados, e Md, a mediana dos desvios absolutos calculados para todos os dados do grupo.

C. Z-score Modificado (ZS)

O intervalo dado pelo Z-score modificado é [5]

$$Iv = \left| \frac{0.6475(x - \tilde{x})}{Md|x_i - \tilde{x}|} \right| > 3.5, \tag{3}$$

em que dados acima desse intervalo são possíveis anomalias.

D. Intervalo Interquartil (IQR)

Aqui, o intervalo é dado por

$$Iv = [Q_1 - 1.5(Q_3 - Q_1)] \cup [Q_3 + 1.5(Q_3 - Q_1)], \tag{4}$$

em que Q_1 e Q_3 representam, respectivamente, o primeiro e o terceiro quartis. No caso, valores abaixo do limite inferior ou acima do superior, são anômalos.

E. Regra da Mediana (MR)

Variação do método anterior, no qual o intervalo é dado por

$$Iv = \tilde{x} \pm 2.3(Q_3 - Q_1), \tag{5}$$

sendo $\tilde{x}$ a mediana dos dados analisados.

III. Resultados e Discussões

Nesse trabalho, afim de avaliar as técnicas consideradas, usou-se dois *datasets* rotulados. Os dados, disponíveis em [6], foram coletados a partir de uma rede de sensores sem fios *multi-hop* e *single-hop* implantadas utilizando módulos TelosB.

Duas métricas para comparação das técnicas são consideradas: acurácia e tempo de execução.

A acurácia é definida como

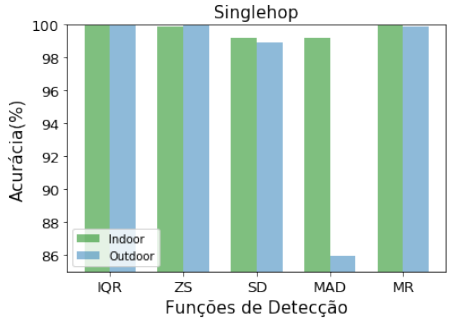
$$Acuracia = \frac{(VP + VN)}{(VP + VN + FP + FN)}, \tag{6}$$

em que VP e VN são os verdadeiros positivo e negativo, e FP e FN, os falsos positivo e negativo.

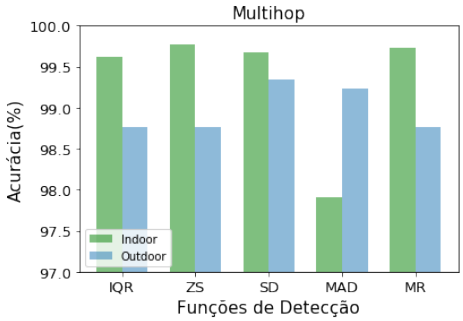
A acurácia foi medida usando uma matriz de confusão, implementada na biblioteca Scikit-Learn do Python [7].

Os gráficos da Fig1.(a) e Fig1.(b) representam os valores de acurácia das técnicas de detecção para medições feitas em *singlehop* e *multihop*, considerando ambientes *indoor* e outdoor. Dos gráficos é

possível perceber que a consideração de um número maior de saltos na medição de tráfego diminui a acurácia da medição, independente do método usado, o que é esperado, devido o número maior de dados coletados nesse caso. É possível ver também, principalmente no contexto *multihop*, que dados oriundo de medições *outdoor* apresentam maior dificuldade de detecção de anomalias, possivelmente pela maior interferência sofrida pelos sensores.



(a) Gráficos de acurácia para dados Singlehop.



(b) Gráficos de acurácia para dados Multihop

Figura 1: Acurácia dos métodos em ambientes distintos.

Os gráficos da Fig2(a) e Fig2(b) mostram o tempo de execução, aferido em um desktop com um processador Core 2 Duo, 500 GB de HD, e 4 GB de memória RAM. Nota-se, dos gráficos, não haver significativas diferenças entre o desempenho computacional entre as técnicas avaliadas, destacando-se, apenas, e como é esperado, um tempo um pouco maior de execução para o cenário *multihop*.

IV. Conclusões

Os resultados apontam que a acurácia tem dependência direta da forma de medição dos dados, sendo menor para medições em múltiplos enlaces, e do local de medição, sendo maior para ambientes *indoor*. Nesse contexto, perceu-se que o desempenho de técnicas mais robustas, relacionando medidas relativas de medianas e/ou intervalos interquartis, como IQR, ZS e MR, tem um desempenho melhor na detecção de anomalias. Percebeu-se também que as técnicas estudadas possuem desempenho computacional praticamente similar, sendo essa uma métrica pouco relevante nesse caso.

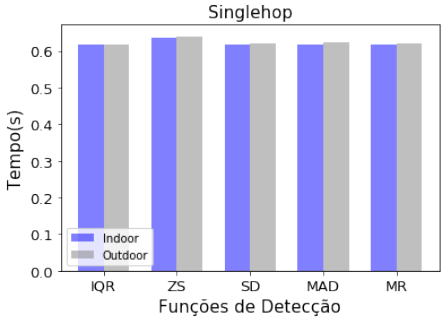
Agradecimentos

Os autores agradecem ao IFPB, *Campus* Campina Grande pelo apoio institucional e estrutural e ao CNPq, pelo financiamento do projeto.

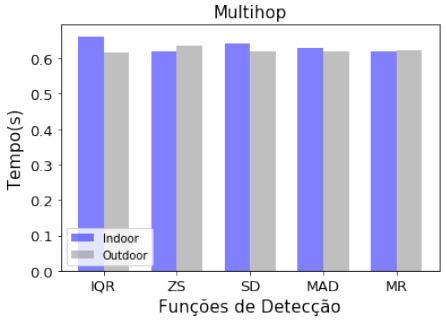
Referências

[1] SBC, “Grandes Desafios da Pesquisa em Computação no Brasil – 2006-2016,” tech. rep., Sociedade Brasileira de Computação, 2006.

[2] H. Sagha, J. d. R. Mill, R. Chavarriaga et al., “Detecting and Rectifying Anomalies in Body Sensor Networks,” in 2011 International Conference on Body Sensor Networks. IEEE, 2011, pp. 162–167.



(a) Gráficos de tempo de execução para dados Singlehop.



(b) Gráficos de acurácia para dados Multihop.

Figura 2: Tempo de execução dos métodos em ambientes distintos.

[3] R. G. d. S. Ramos, P. Ribeiro, and J. V. d. M. Cardoso, “Anomalies Detection in Wireless Sensor Networks Using Bayesian Changepoints,” in Mobile Ad Hoc and Sensor Systems (MASS), 2016 IEEE 13th International Conference on, pp. 384–385, IEEE, 2016. 1, 12, 17

[4] S.Seo,"A Review and Comparison of Methods for Detecting Outliers in Univariate Data Sets", in the Submitted to the Graduate Faculty of Graduate School of Public Health in partial fulfillment of the requirements for the degree of Master of Science on, pp. 9-13, 2002.

[5] Iglewicz, B., Hoaglin, D. How to detect and handle outliers. ASQC Quality Press, 1993

[6] S. Suthaharan, M. Alzahrani, S. Rajasegarar, C. Leckie, and M. Palaniswami, “Labelled data collection for anomaly detection in wireless sensor networks,” in Intelligent sensors, sensor networks and information processing (ISSNIP), 2010 sixth international conference on, pp. 269–274, IEEE, 2010. vi, 5, 19, 30

[7] "Sklearn metrics confusion matrix". http://scikit-learn.org/stable/modules/generated/sklearn.metrics.confusion_matrix. Visitado em: 28 julho 2017