ESTUDO DE TÉCNICAS DE AGRUPAMENTO EM DETECÇÃO DE ANOMALIAS EM REDES DE COMPUTADORES.

CLUSTERING TECHNIQUES FOR DETECTING ANOMALIES IN COMPUTER NETWORKS

Bruno Prece da Silva[1]
bprece@gmail.com
Mario Henrique Akihiko da Costa Adaniya[2]
mario.adaniya@unifil.br

Resumo

A detecção de anomalias em redes é uma área de pesquisa muito dinâmica, que envolve diversas técnicas para detectar anomalias conhecidas e desconhecidas. Diversos sistemas e algoritmos são propostos e há muitas discussões referentes à eficácia de cada método na literatura. Este artigo apresenta alguns aspectos da detecção de anomalias, e discute em particular as técnicas de agrupamento. Um estudo de caso utilizando a base de dados do KDD99 e o algoritmo K-means é apresentado.

Palavras-chave: Detecção de Anomalias; Clustering; K-means.

Abstract

The detection of anomalies in networks is an area of very dynamic research that involves various techniques to detect known and unknown anomalies. Several systems and algorithms are proposed and there are many discussions concerning the effectiveness of each method in the literature. This article presents some aspects of anomaly detection, and discusses in particular the clustering techniques. A case study using the KDD99 database and K-means algorithm is presented.

Keywords: Anomaly Detection; Clustering; K-means.

INTRODUÇÃO

A constante expansão na utilização da Internet tornou a rede um serviço essencial para a interconexão global, e especialmente para algumas empresas onde seu negócio gira em torno da conectividade. A cada dia há um crescimento na demanda por novos serviços, que necessitam de políticas de segurança que mantenham a integridade e a privacidade dos dados. Isso torna os sistemas mais complexos e os dados cada vez mais heterogêneos, impossibilitando que a gerência da rede seja realizada por um operador humano. Devido à estas dificuldades, surgem comportamentos anômalos no tráfego.

[1]Centro Universitário Filadélfia de Londrina - UniFil
[2]Centro Universitário Filadélfia de Londrina - UniFil

DETECÇÃO DE ANOMALIAS

No tráfego de rede, as anomalias representam ações que se diferem do comportamento normal da rede. Mesmo quando uma anomalia não impacta profundamente a rede, ela atinge a qualidade dos serviços que são entregues aos usuários finais (LAKHINA; CROVELLA; DIOT, 2004). O comportamento anômalo pode ser gerado por eventos de origem maliciosa, como, por exemplo, ataques de negação de serviço, varredura de portas e ataques de penetração. Além de eventos não maliciosos, como, aumentos repentinos no volume do tráfego, tempestades de *broadcast* e congestionamentos.

As principais vertentes de pesquisas dividem as técnicas de detecção de anomalias em duas categorias, a detecção baseada em assinaturas de anomalias que se destaca em relação à detecção de anomalias já conhecidas, e a detecção baseada na caracterização do comportamento normal da rede, que possui um melhor desempenho na detecção de anomalias desconhecidas. Os sistemas de detecção são classificados de acordo com as assinaturas utilizadas na detecção. Sistemas supervisionados utilizam as duas classes de assinaturas, diferente dos sistemas semi-supervisionados, que fazem uso apenas da assinatura do perfil normal do tráfego e os sistemas não supervisionados que não utilizam assinaturas, sendo mais fáceis de serem aplicados na rede, porém, com maior dificuldade de implementação (CHANDOLA; BANERJEE; KUMAR, 2009).

Vários métodos e algoritmos são utilizados nas pesquisas e nos sistemas para detecção de anomalias. Dentre estas técnicas, a clusterização de dados é muito explorada e possui bons resultados. A clusterização consiste no particionamento e agrupamento de um conjunto de dados, de tal forma que os objetos que compartilhem características comuns e que se diferem de outros cluster sejam alocados em um mesmo grupo, assim, as técnicas de clustering precisam encontrar automaticamente a assinatura dos dados, ou seja, as características mais relevantes de cada grupo (REHMAN; REHMAN; KHAN, 2009).

METODOLOGIA EXPERIMENTAL

Para a elaboração deste trabalho foi utilizado o *K-means*, um algoritmo simples e muito poderoso para clusterização de dados. A técnica divide um conjunto de

n elementos em um número *k* de clusters, de um modo que a similaridade entre os elementos de um mesmo grupo seja alta e a similaridade entre os grupos seja baixa. O algoritmo consiste primeiramente na escolha de *k* objetos centrais definidos por parâmetro. Assim, cada ponto de dado é atribuído ao seu objeto central mais próximo, formando os grupos. Os centróides de cada cluster são atualizados de acordo com a média de similaridade dos objetos dos grupos, atualizando a média dos clusters a cada iteração. Assim, o algoritmo cria grupos de dados normais e grupos de anomalias.

A base de dados KDDcup99 é um dos conjuntos de dados mais utilizados para avaliação de sistemas de detecção de anomalia e detecção de intrusões. A base contém aproximadamente cinco milhões de vetores de conexões, que são compostos por 41 atributos referentes aos recursos utilizados em cada conexão.

De maneira padrão, o algoritmo *K-means* utilizado primeiramente aplica o algoritmo *K-means++* (ARTHUR; VASSILVITSKII, 2007) para a escolha dos centros iniciais. Porém, este método se mostrou ineficaz neste caso, devido à proximidade entre os centros que se concentravam apenas no primeiro terço dos vetores de conexões, assim, o algoritmo não identificava os vários grupos de anomalias. Após a utilização de centróides iniciais que fossem distribuídos por toda base, o algoritmo passou a subdividir o grupo de anomalias em vários grupos de acordo com a característica de cada anomalia, evidenciando a eficácia do mesmo para detecção de anomalias.

CONCLUSÃO

Neste artigo, foram apresentados aspectos gerais da área de detecção de anomalias e de algoritmos de agrupamento, que se mostraram capazes de realizar o agrupamento de extensas bases de dados de redes, realizando a divisão entre o tráfego normal e o tráfego anômalo.

Esta metodologia mostra-se simples e eficiente, contendo diversas possibilidades de extensões e combinações com outras técnicas que podem refinar o agrupamento, gerando dados altamente precisos e com custo computacional baixo.

REFERÊNCIAS

ARTHUR, D.; VASSILVITSKII, S. k-means++: The advantages of careful seeding. In: SOCIETY FOR INDUSTRIAL AND APPLIED MATHEMATICS. *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*. [S.l.], 2007. p. 1027–1035. 3

CHANDOLA, V.; BANERJEE, A.; KUMAR, V. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, ACM, v. 41, n. 3, p. 15, 2009. 2

LAKHINA, A.; CROVELLA, M.; DIOT, C. Diagnosing network-wide traffic anomalies. In: *Proceedings of the 2004 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications*. New York, NY, USA: ACM, 2004. (SIGCOMM '04), p. 219–230. ISBN 1-58113-862-8. Disponível em: <http://doi.acm.org/10.1145/1015467.1015492>. 2

REHMAN, Z.; REHMAN, S. A.; KHAN, L. Survey reports on four selected research papers on data mining based intrusion detection system. *University of the West Indies at Mona*, 2009. 2