MINERAÇÃO DE DADOS PARA DETECÇÃO DE SPAMs EM REDES DE COMPUTADORES

Kelton Costa; Patricia Ribeiro; Atair Camargo; Victor Rossi; Henrique Martins; Miguel Neves; Ricardo Fontes.

kelton.costa@gmail.com; patriciabellin@yahoo.com.br; junior.camargo@hotmail.co.uk; vic.v.rossi@gmail.com; henmartins@gmail.com; miguel.neves01@fatec.sp.gov.br; ricardoffabris@gmail.com

Faculdade de Tecnologia de Bauru, LASIC – Laboratório de Sistemas Inteligentes e Computação

Resumo

Nas últimas décadas houve um aumento na quantidade de diferentes anomalias em redes de computadores, levando a preocupação com a identificação destes ataques. Um método proposto neste estudo, é a utilização da Mineração de Dados como apoio na tentativa de uma correta identificação destas anomalias. A utilização da ferramenta Weka, permite através de uma base de dados rotulada, uma identificação e análise das anomalias em um ambiente de redes de computadores, servindo como base para melhorias neste mesmo ambiente.

Palavras-chave*: Anomalias, Redes de computadores, Mineração de dados, Redes Neurais Artificiais.*

Abstract

Anomalies in computer networks has increased in the last decades and raised concern to create techniques to identify these unusual traffic patterns. This research aims to use data mining techniques in order to correctly identify these anomalies. Weka is a collection of machine learning algorithms for data mining tasks – was used to identify and analyse anomalies of a data set called SPAMBASE in order to improve this environment.

Keywords*: Anomalies, Computer networks, Data Mining, Artificial Neural Networks.*

1. Introdução

Pesquisas na área da computação estão sendo desenvolvidas, destacando neste contexto a utilização de alguns métodos como as Redes Neurais Artificiais e Processos de Mineração de Dados utilizando por exemplo árvore de decisão, com o objetivo de minimizar os efeitos danosos de SPAMs.

Mineração de dados é parte do processo em KDD (*Knowledge Discovery in Databases*) que objetiva a seleção das técnicas a serem utilizadas para localização de padrões nos dados, tendo como finalidade a busca dos referidos padrões relacionados a

um interesse exclusivo (FAYYAD, PIATESKY-SHAPIRO e SMYTH, 1996; NARENDRAN, 2009).

As etapas do processo de descoberta de conhecimento em KDD (FAYYAD, PIATESKY-SHAPIRO e SMYTH, 1996; NARENDRAN, 2009) podem se apresentar de forma cognitiva, interativa e exploratória, compreendendo nos seguintes passos: definição do tipo de conhecimento a se buscar, definição de um conjunto ou subconjunto de dados a se pesquisar, pré-processamento, redução dos dados, mineração dos dados, interpretação dos padrões minerados e implantação do conhecimento descoberto.

Neste trabalho, tendo uma base de dados rotulada, ou seja, uma base que contempla os SPAMs identificados, propõe-se na base de dados denominada SPAMBASE (MARK et al., 2013) que contém 4.601 SPAMs, sendo, 1.813 classificados SPAMs normais e 2.232 como anormais, a aplicação do processo de mineração de dados, através da ferramenta Weka, com intuito de analisar e quantificar os tipos de SPAMs presente na base de dados estudada, auxiliando no processo da gerência de redes. Neste sentido, o presente estudo aborda um processo de mineração de dados extraída de uma base de dados rotulada, SPAMBASE.

2. Metodologia

Neste trabalho foi investigado o potencial de dois classificadores, sendo eles, árvore de decisão (J48) e a rede neural *Multi-Layer Perceptron* (MLP) para a mineração de dados utilizando uma base de dados pública de spams chamada SPAMBASE, e a partir dos dados desta base, identificar possíveis anomalias.

2.1 Base de Dados

Os experimentos realizados utilizaram a base de dados pública SPAMBASE que contém cinquenta e sete atributos de dados e um atributo de classificação a fim de determinar a normalidade no conteúdo. Esta base de dados foi criada com o objetivo de prover melhoria nos softwares desenvolvidos para a segurança em redes de computadores, visto que ataques através de SPAMs com anomalias pode gerar muito prejuízo tais como, gastos desnecessário de tempo, aumento de custos, perda de produtividade, conteúdo impróprio ou ofensivo e prejuízos financeiros causados por fraude (FOGEL e SHLIVKO; 2010).

O presente estudo utiliza a base de dados SPAMBASE convertida em formato de valores separados em virgulas (CSV – *Comma Separated Values*) e estão no formato ARFF, compatível com o programa minerador de dados Weka. Mesmo a base SPAMBASE possuindo cinquenta e sete atributos, como cada atributo representa uma palavra e a frequência em que esta aparece no e-mail, pode representar um SPAMs normal ou anormal. Foram utilizados todos os atributos uma vez que todas estas características são relevantes para os testes.

2.2 *Software* Weka

Para análise e quantificação dos tipos de anomalias presente na base de dados SPAMBASE foi utilizado o processo de mineração de dados, realizado através do *software* Weka. O Weka é um *software* desenvolvido na Universidade de Waikato na

Nova Zelândia, escrito em linguagem Java, possui chave de licença pública e código aberto. Os dados das anomalias podem ser carregados no Weka utilizando o formato de Arquivo de Atributo Relação (ARFF). Nesse arquivo são definidos cada coluna com um tipo de dado, por exemplo, numérico ou caractere, em cada linha são fornecidos os dados, com seus respectivos tipos de dados, delimitados por vírgulas.

2.3 Árvore de Decisão e MLP

Para a realização da mineração de dados foi utilizado o filtro Discretize para a normalização dos dados de entrada em um intervalo numérico (Exemplo: entre 0 e 1). Após o filtro, foram selecionadas duas técnicas, o algoritmo árvore de decisão (J48) e a Rede Neural Artificial (RNA) MLP, muito utilizada tanto para mineração como classificação (RIBEIRO, SCHIABEL e ROMERO, 2010; 2011).

O algoritmo J48 constrói um modelo de árvore de decisão baseado num conjunto de dados de treinamento, sendo que esse modelo é utilizado para classificar as instâncias de um conjunto de teste. Já a rede neural MLP possui aprendizado supervisionado e tem o objetivo calcular o erro para a camada de saída e propagar este no sentido saída-entrada (*backpropagation*), ajustando os pesos de todas as camadas, através da retro propagação do erro (HAYKIN, 2009; SILVA; SPATTI e FLAUZINO, 2010).

2.4 Avaliação de Desempenho dos Classificadores

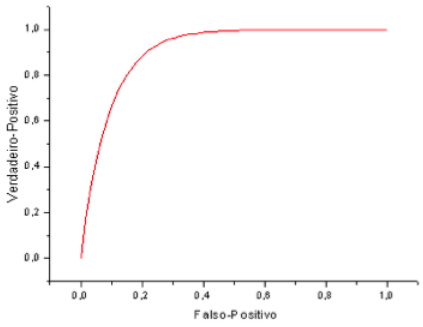
Para a avaliação dos algoritmos, foi escolhida uma ferramenta padrão da estatística, conhecida como *cross-validation* (HAYKIN, 2009), na qual o conjunto disponível de dados, com *n* elementos, é dividido aleatoriamente em um conjunto de treinamento, com 75% do conjunto de dados, e em um conjunto de teste, com 25% do mesmo conjunto de dados. Os exemplos são divididos em 10 partições mutuamente exclusivas, com 75% do conjunto de dados, para treinar o algoritmo. E os 25% de elementos restantes, de cada partição, são utilizados para testar o classificador. Repete-se esse procedimento para todas as partições. A avaliação do grau de generalização do classificador ficará desta forma, garantida com este método. Cabe ressaltar que foi utilizada a curva ROC *(Receiver Operating Characteristics)* que é uma técnica para análise do desempenho dos classificadores.

3. Resultados

Após o a mineração de dados utilizando os dois algoritmos foi possível obter excelentes taxas de acerto, com base em um total de 4.601 amostras e utilizando todos os atributos da base de dados SPAMBASE como entrada. Cabe ressaltar que foi utilizada a técnica *cross-validation* para a validação dos testes.

Utilizando o algoritmo J48 foi possível obter, uma taxa de acerto médio de 92,76%, sendo 89,79% de acerto para classificações do tipo SPAMs normais e 93,34% de SPAMs anormais obtendo curva ROC (EVANS , 1981) com Az igual a 0,941, conforme Figura 1.

Figura 1 – Curva ROC dos resultados obtidos com o algoritmo J48.

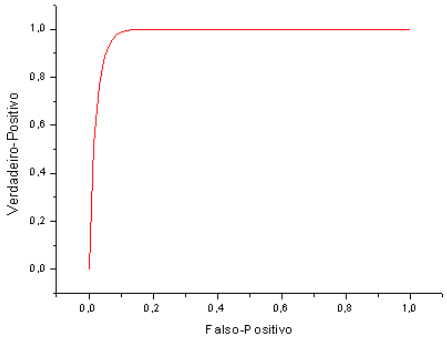


Fonte: Elaborado pelos autores.

Com a rede neural MLP utilizando a arquitetura composta por 57 atributos de entrada, uma camada intermediária com topologia contendo 68 neurônios, camada neural de saída contendo um neurônio, taxa de aprendizagem igual a 0,3, momentum igual a 0,2.

Após o a mineração de dados utilizando a rede MLP foi possível obter, com base em um total 4.601 amostras, uma taxa de acerto médio de 93,89%, sendo 93,93% de acerto para classificações do tipo SPAMs normais e 93,87% de SPAMs anormais obtendo curva ROC com Az igual a 0,98, conforme Figura 2.

Figura 2 – Curva ROC dos resultados obtidos com a rede neural MLP.



Fonte: Elaborado pelos autores.

4. Conclusão

Atualmente muitas pesquisas na área da computação, mais precisamente em redes de computadores, estão sendo desenvolvidas utilizando métodos de mineração de dados para identificar o comportamento da rede. A base de dados denominada SPAMBASE, que contempla amostras de SPAMs, foi utilizada neste estudo. Neste trabalho propomos a utilização de todos os atributos da base citada. Após a utilização do filtro Discretize e a utilização de duas técnicas para a mineração de dados, a árvore de decisão (J48) e a rede MLP, foi possível obter excelentes taxas de acertos sendo 92,76% de acerto com o algoritmo J48 e 93,89% de acerto com a rede neural MLP. Demonstrando assim, a vantagem do uso de técnicas inteligentes em redes de computadores para detecção de SPAMs. Cabe ressaltar que a rede neural MLP apresentou, durante os testes, taxa de

acerto superior à árvore de decisão (J48) devido a sua característica de generalização, sendo a melhor opção até o momento para a mineração de dados de SPAMs.

Referências

EVANS, A.L. The Evaluation of Medical Images. Adam Hilger Ltda, Bristol, Great Britain, 1981.

FAYYAD, U. M.; PIATESKY-SHAPIRO, G.; SMYTH, P. From Data Mining to Knowledge Discovery: An Overview. In: Advances in Knowledge Discovery and Data Mining, AAAI Press, 1996.

FOGEL, J.; SHLIVKO, S. Weight Problems and Spam E-mail for Weight Loss Products Southern Medical Journal, v. 103, ed. 1, pp 31-36, 2010.

HAYKIN, S. Neural Networks and Learning Machines. Editora Prentice Hall, 3a. Edição, p. 936, 2009.

MARK, H.; REEBER, E.; FORMAN, G.; SUERMONDT, J. SPAMBASE, Disponível em: <http://www.ics.uci.edu/~mlearn/databases/spambase/>. Acesso em: 28 maio 2013.

NARENDRAN, C. R., Data Mining - Classification Algorithm – Evaluation, May 8th, 2009.

RIBEIRO, P. B.; SCHIABEL, H.; ROMERO, R. A F. Comparativo entre Classificadores de Nódulos Mamários In: XXII Congresso Brasileiro de Engenharia Biomédica (CBEB, 2010), Tiradentes, MG. Anais do XXII CBEB - 2010 (ISSN: 2179-3220), 2010. pp.165 – 168, 2010.

RIBEIRO, P. B.; SCHIABEL, H.; ROMERO, R. A F. Artificial Neural Networks versus Systems Fuzzy in Breast Masses Classification Schemes In: Society for Imaging Informatics in Medicine (SIIM) - Junho 2-5/2011, 2011, Washington, DC.Society for Imaging Informatics in Medicine (SIIM). , 2011.

SILVA, I. N. da; SPATTI, D.H.; FLAUZINO, R. A. Redes Neurais Artificiais: para engenharia e ciências aplicada. Ed. Artliber, p. 399, 2010.