Classificação online dos tráfegos TCP e UDP baseada em sub-fluxos

Raimir Holanda Filho[1], Victor Pasknel de Alencar Ribeiro[1], José Everardo Bessa Maia[2]

[1]Mestrado em Informática Aplicada – Universidade de Fortaleza (UNIFOR)

[2]Departamento de Estatística e Computação – Universidade Estadual do Ceara (UECE)

raimir@unifor.br, pasknel@hotmail.com, jmaia@uece.br

Abstract. *Classification based on sub-flows is a step towards meeting the requirements for online classification of network traffic. This work describes a classification strategy based on sub-flows, applied to TCP and UDP traffic, which opens a perspective to classify the entire traffic of a link. The proposed work uses the One-Against-All (OAA) decomposition applied to statistical features of sub-flows obtained only from the information of packet headers. The OAA decomposition explores the characteristic to build highly specialized binary classifiers for each application class. The results show a mean accuracy of 98% for the selected classes.*

Resumo. *A classificação baseada em sub-fluxos é um passo no sentido de cumprir os requisitos para a classificação online de tráfego de rede. Este trabalho descreve uma estratégia de classificação baseada em sub-fluxos, aplicado à tráfego TCP e UDP, que abre uma perspectiva para classificar todo o tráfego de um link. O trabalho proposto utiliza a decomposição de um-contra-todos (1ct) aplicada a variáveis estatísticas dos sub-fluxos, as quais são obtidas somente a partir da informação dos cabeçalhos dos pacotes. A decomposição 1ct explora a característica de construir classificadores binários altamente especializados para cada classe de aplicação. Os resultados mostram uma precisão média de 98% para as classes selecionadas.*

1. Introdução

Classificação online de tráfego de rede é uma atividade essencial para diversos sistemas, tais como: detecção de intrusão, gerenciamento de rede e QoS (*Quality of Service*). As técnicas clássicas de classificação de tráfego envolvem normalmente a inspeção do conteúdo de pacotes e a análise das portas TCP/UDP utilizadas. Entretanto, a eficácia de tais técnicas tem se mostrado bastante limitada em decorrência da utilização de criptografia e do uso de portas não convencionais. Por outro lado, a classificação de tráfego baseada em métodos estatísticos e técnicas de aprendizado de máquina têm atraído um grande interesse, em virtude a capacidade de utilizar informações retiradas somente dos cabeçalhos dos pacotes.

Em se tratando de classificação online, surgem um conjunto de restrições relacionadas ao funcionamento de um classificador de tráfego, principalmente relacionadas à compatibilização entre o tempo de classificação e a duração do fluxo. Trabalhos recentes apresentam abordagens de classificação de tráfego IP baseadas nos

primeiros N pacotes de fluxos bidirecionais [Bernaille 2006], [Li 2006]. Entretanto, em um ambiente real, classificadores de tráfego podem não ter acesso aos primeiros pacotes de um fluxo e/ou não serem capazes de identificar a direção dos pacotes (do cliente para o servidor ou o inverso).

Este trabalho propõe uma arquitetura para classificação online de tráfego de rede através de um algoritmo um-contra-todos (1ct) supervisionado. Esta abordagem tem a vantagem de permitir a utilização de classificadores binários, que são altamente especializados. Em nossa variante da abordagem 1ct, o problema de classificar um fluxo em uma das M classes é reduzido em M problemas de classificação binária, cada qual com uma regra de decisão associada.

Neste trabalho, a abordagem 1ct é aplicada em duas estratégias de classificação online baseadas em variáveis estatísticas para sub-fluxos TCP e UDP:

- Sub-fluxos constituídos dos N primeiros pacotes de um fluxo TCP.
- Sub-fluxos compostos por N pacotes tomados a partir de uma posição aleatória de um fluxo UDP.

Estas duas estratégias são comparadas com os resultados obtidos a partir de um classificador multi-classe. Foi investigado o efeito da variação no tamanho de N sobre os resultados da classificação e da utilização de um conjunto mínimo de variáveis diferentes em cada um dos problemas acima.

Os testes foram realizados com a utilização de traços reais, contendo um total de 7 classes de aplicação, e os resultados obtidos apresentaram uma precisão de 98,43% para sub-fluxos iniciais TCP e 98,75% para sub-fluxos UDP. Na estratégia apresentada, o número de variáveis e tamanhos de sub-fluxo utilizados variam respectivamente de 4 a 18 e de 5 a 10.

Este trabalho é organizado da seguinte maneira: a seção 2 apresenta uma breve revisão sobre trabalhos relacionados ao tema deste artigo; uma introdução sobre classificadores bayesianos e sua aplicação em classificação de tráfego IP é demonstrada na seção 3; o procedimento para coleta e rotulação de fluxos utilizados neste trabalho é apresentado na seção 4; a seção 5 ilustra a abordagem proposta através de um experimento, assim como apresenta uma análise dos resultados obtidos; a conclusão é apresentada na seção 6 com as principais considerações e os direcionamentos para trabalhos futuros.

2. Trabalhos Relacionados

Publicações mais recentes têm abordado o tema de classificação do tráfego de rede sob uma perspectiva de classificação online [Bernaille 2006], [Este 2009]. A classificação de sub-fluxos, quando comparada com uma abordagem baseada em fluxos completos, reduz o processamento e não necessita esperar até o final de um fluxo (latência). Para a atividade de classificação, dois aspectos importantes devem ser verificados: o método de seleção de variáveis e o algoritmo de classificação.

Algoritmos para seleção de variáveis são divididos em três categorias gerais: o modelo de filtro, o modelo *wrapper* e o modelo híbrido [Erman 2006]. Algoritmos de

aprendizado de máquina têm sido utilizados supondo que uma classe de tráfego pode ser identificada através da análise estatística das suas características de tráfego.

Relacionado ao tema de seleção de variáveis para uma classificação on-line, o trabalho de [Zhang 2009] explora uma análise comparativa de dois algoritmos diferentes para identificar o subconjunto de variáveis apropriado para um algoritmo de *cluster* como uma questão crítica na classificação de tráfego online.

Em [Wang 2009], os autores realizam uma avaliação da eficácia de métodos estatísticos para o problema de classificação online de tráfego. Este trabalho avalia três conjuntos diferentes de variáveis, os quais são utilizados para capturar as propriedades distintas de diferentes aplicações, sendo dois deles constituídos por variáveis geradas a partir de fluxos completos, enquanto o terceiro conjunto é composto por variáveis estatísticas retiradas de sub-fluxos formados pelos primeiros pacotes de cada fluxo.

No trabalho proposto em [Li 2006], o algoritmo *Naïve Bayes* é utilizado para a classificação de tráfego IP baseada em propriedades estatísticas dos fluxos, tais como o tamanho médio do segmento, a variância do tamanho de carga e tamanho da janela inicial. Um total de 10 variáveis foram selecionadas e uma taxa de acerto de 96% foi obtida na classificação do tráfego de 10 classes de aplicação. No entanto, neste trabalho, foram apenas avaliados fluxos TCP.

Em [Nguyen and Armitage 2006], é proposto um classificador baseado em propriedades estatísticas retiradas de N pacotes, os quais são tomados a partir de qualquer ponto arbitrário de um fluxo. O classificador proposto foi treinado usando variáveis estatísticas calculadas sobre vários sub-fluxos extraídos de fluxos completos. Este trabalho, no entanto, é aplicado para identificar apenas uma classe de jogo online.

Os mesmos autores de [Nguyen and Armitage 2006] estendem seu trabalho anterior, sobre treinamento com múltiplos sub-fluxos, para incluir a ideia de usar algoritmos de aprendizagem de máquina não supervisionados para a seleção automatizada de sub-fluxos [Nguyen and Armitage 2008]. Entretanto, como no trabalho anterior, a abordagem proposta é limitada a classificar uma única classe de aplicação.

O trabalho aqui proposto difere dos trabalhos anteriores por utilizar em conjunto as seguintes premissas: caracterização de um tamanho ótimo de sub-fluxo para cada classe de aplicação; explorar uma minimização no número de variáveis, para cada classe, sem comprometimento do desempenho e utilização de técnicas de baixa complexidade estatística.

3. Metodologia

Nesta seção é apresentada uma introdução sobre classificadores bayesianos e suas aplicações para classificação de tráfego IP, assim como o processo de seleção de variáveis e tamanhos de sub-fluxos para o treinamento do classificador proposto.

3.1 Classificação bayesiana e Naïve Bayes

O classificador *Naïve Bayes* foi utilizado neste trabalho [Duda and Hart 2000]. Considere uma coleção de fluxos $x = (x_1, \ldots, x_n)$, aonde cada fluxo x_i é descrito por m discriminadores $\{d_1(i), \ldots, d_m(i)\}$, os quais podem ser valores numéricos ou discretos.

Dentro do contexto de classificação de tráfego, $d_j(i)$ é um discriminador do fluxo x_i, por exemplo: média de bytes do fluxo x_i. Neste trabalho, cada fluxo x_i pertence exclusivamente a uma única classe. O classificador supervisionado bayesiano trata de construir um modelo estatístico, para descrever cada classe analisada, através de uma fase de treinamento, na qual cada fluxo *y* recebe uma probabilidade de ser classificado como uma única classe de acordo com a regra de Bayes:

$$p(c_j|y) = \frac{p(c_j)f(y|c_j)}{\sum_{c_j} p(c_j)f(y|c_j)} \quad (1)$$

na qual $p(c_j)$ representa a probabilidade de obter a classe c_j independente dos dados observados, $f(y \mid c_j)$ é a função de distribuição (ou probabilidade de *y* dado c_j) e o denominador funciona como uma constante de normalização. A técnica de *Naïve Bayes* considerada neste artigo, assume a independência dos discriminadores $d_1, \ldots, d_m$, assim como, um comportamento Gaussiano simples destes.

A regra de classificação consiste na escolha de uma classe com a probabilidade máxima de adesão, de acordo com a equação 2:

$$c = c_k : k = \arg\max_j p(c_j|y) \quad (2)$$

3.2 Seleção de variáveis e tamanho de sub-fluxo

Devido a grande quantidade de variáveis que podem ser utilizadas para a classificação de um fluxo, um classificador pode ter que tratar variáveis que contêm características irrelevantes e redundantes, causando assim um processo de classificação mais lento, com maior consumo de variáveis, bem como uma precisão de classificação ruim. Portanto, a seleção de variáveis tem um papel fundamental na optimização do desempenho. Como encontrar um sub-conjunto ótimo de variáveis para sub-fluxos, ainda é uma questão crítica. Métodos de seleção de variáveis têm sido aplicados com sucesso para classificação, mas raramente aplicados à algoritmos de *cluster* devido à indisponibilidade de informações do rótulo da classe.

O avaliador *Wrapper* [Hall 1998] foi utilizado neste trabalho para a seleção de variáveis. *Wrapper* avalia variáveis usando estimativas de precisão, produzidas por um algoritmo de aprendizagem, as quais serão utilizadas na classificação. Uma abordagem de seleção foi realizada para cada classificador binário, os quais produzem um conjunto específico de variáveis para cada classe.

A implementação em *Java* do avaliador *Wrapper* encontrado em Weka [Weka 2010] foi utilizada para a seleção de variáveis de cada modelo de classes criado. O classificador *Naive Bayes* foi utilizado como o algoritmo de aprendizado e *Best First* foi escolhida como método de pesquisa para o avaliador *Wrapper*.

Os passos seguintes foram aplicados para as duas estratégias de classificação (pacotes iniciais e aleatórios), para selecionar um número reduzido de variáveis e tamanhos de sub-fluxos:

1. Um conjunto de dados 1ct é criado para cada classe e tamanho de sub-fluxo analisado. Nesta pesquisa, foi estudado o efeito da variação do tamanho de sub-fluxos de 5 até 10 pacotes.

2. O avaliador *Wrapper* é executado em cada conjunto de dados criado no passo 1. As variáveis e tamanhos de sub-fluxo de cada classe são selecionados com base no maior resultado obtido a partir do avaliador *Wrapper*. Em caso de igualdade entre os resultados mais elevados, o modelo com o menor tamanho de sub-fluxo é selecionado.

O número ideal de atributos e tamanho de sub-fluxo obtidos para cada classe TCP são apresentados na Tabela 1. Os resultados obtidos pelo avaliador *Wrapper*, ao analisar os N pacotes iniciais de sub-fluxos TCP, são apresentados na Figura. 1. A descrição de todos as variáveis selecionadas, para cada classe TCP, pode ser vista na Tabela 10 (Anexos).

Tabela 1. Modelos 1ct (TCP)

Classe	Número de Atributos	Tamanho do Sub-fluxo
CHAT	4	5
MAIL	6	10
P2P	18	10
SSL	11	6
WWW	8	6

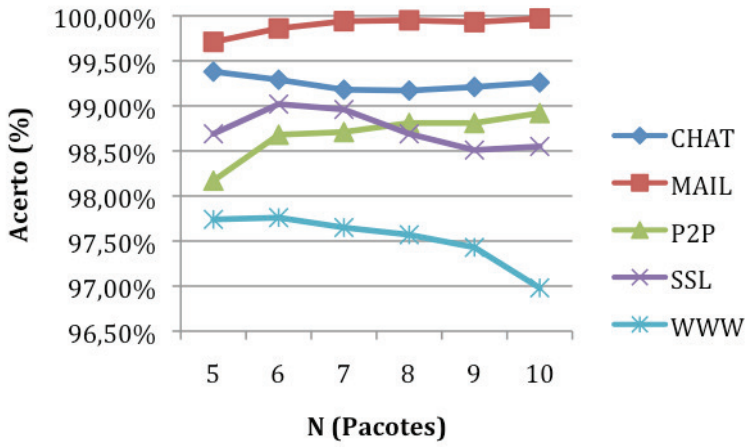

Figura 1. Evolução da taxa de acerto para sub-fluxos TCP iniciais

Observe que para algumas classes (ex: WWW) a taxa de acerto cai, enquanto cresce o tamanho do sub-fluxo. Esse fenômeno tem sido observado em outros estudos [Nguyen and Armitage 2006] e é devido ao fato de que as variáveis utilizadas pelos pacotes iniciais, quando isolados, podem diferenciar de forma mais adequada essa classe, ao invés de quando usadas com um número maior de pacotes.

O número ideal de atributos e tamanhos de sub-fluxos obtidos para cada classe UDP são apresentados na Tabela 2. Os resultados obtidos durante a análise do avaliador *Wrapper*, utilizando N pacotes retirados a partir de uma posição aleatória de sub-fluxos

UDP, são apresentados na Figura 2. A descrição de todas as variáveis selecionadas para cada classe pode ser vista na Tabela 11 (Anexos).

Tabela 2. Modelos 1ct (UDP)

Classe	Número de Atributos	Tamanho de Sub-fluxo
DNS	4	8
VOIP	3	9

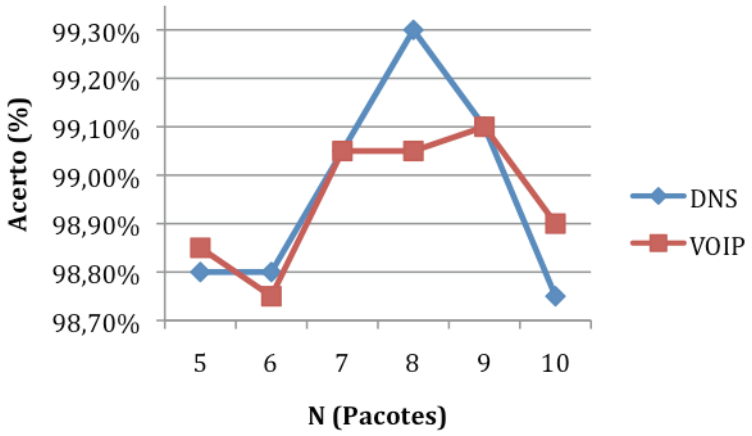

Figura 2. Evolução da taxa de acerto para sub-fluxos UDP aleatórios

A partir da Figura 2, é possivel constatar as altas taxas de acerto obtidas pelas classes UDP. Entre todos os tamanhos de sub-fluxos analisados (5 a 10 pacotes), os resultados apresentaram uma taxa de acerto mínima de 98,75% e uma taxa de acerto máxima de 99,3 %.

4. Dados e Medidas

Esta seção apresenta as definições básicas utilizadas durante este trabalho, assim como o processo de coleta de dados e rotulação de fluxos.

4.1. Definição de fluxo

Nossa proposta é baseada na análise de fluxos de tráfego IP. Um fluxo consiste em uma sequência de pacotes que são transmitidos entre um par de *hosts* [Moore 2005]. O fluxo também pode ser definido como uma 5-tupla: endereços IP (origem e destino), números de portas (origem e destino) e um protocolo (TCP ou UDP).

Fluxos devem ter pelo menos um pacote em cada sentido de tráfego para serem considerados válidos. Fluxos TCP são iniciados com uma apresentação de três vias e são considerados finalizados se um flag *FIN* e/ou *RST* são vistos no cabeçalho TCP. Fluxos UDP são considerados expirados se nenhum pacote for transmitido entre os *hosts* durante um intervalo de 60 segundos.

4.2. Aquisição de Dados

Para verificar a validade da abordagem, a metodologia proposta foi aplicada com dados de tráfego de rede. Uma série de passos deve ser realizada a fim de obter todos os dados necessários: capturar pacotes em estado bruto, reconstruir fluxos e rotular as classes.

A primeira etapa consiste na captura de pacotes da rede usando uma interface de rede em modo promíscuo. Os pacotes capturados são armazenados temporariamente e, em seguida, os fluxos são reconstruídos. Durante a etapa final, cada fluxo deve ser rotulado com uma única classe. O processo de rotulação de cada fluxo foi realizado de forma semi-automatizada através do uso da ferramenta de inspeção de pacotes OpenDPI [OpenDPI 2010] e a biblioteca Jpcap [Jpcap 2010].

O classificador online proposto foi treinado e validado através da utilização dos conjuntos de dados de tráfego recolhidos a partir do *gateway* de rede, durante o período de 22 a 30 de Abril de 2010, da Universidade de Fortaleza. A Tabela 3 apresenta as classes, aplicações, protocolos e o total dos fluxos encontrados nos conjuntos de dados.

Tabela 3. Descrição dos dados

Classe	Número de Fluxos	Aplicações	Protocolo
CHAT	12141	MSN Messenger	TCP
DNS	31643	DNS	UDP
MAIL	19490	SMTP, IMAP e POP3	TCP
P2P	43675	Bittorrent, Gnutella e eDonkey	TCP
SSL	25851	SSL e TLS	TCP
VOIP	1115	IAX, RTP e SIP	UDP
WWW	571692	HTTP	TCP

Um total de 25.000 fluxos TCP (5.000 por classe) e 2.000 fluxos UDP (1.000 por classe) foram selecionados de forma aleatória para a fase de treinamento do classificador proposto (descrito na seção 3.2).

4.3. Sub-fluxos e variáveis

As variáveis utilizadas para a classificação de fluxos nesta pesquisa foram calculadas com base apenas nas informações obtidas dos cabeçalhos dos pacotes. Nenhuma inspeção foi realizada no payload e os números de porta TCP/UDP não foram utilizados para calcular este grupo de variáveis. As variáveis foram calculadas para ambos os sentidos de um fluxo UDP e para cada direção de um fluxo TCP: cliente para servidor (CS) e servidor para cliente (SC).

Os sub-fluxos utilizados nesta pesquisa consistem em grupos de N pacotes, retirados de fluxos completos. Os tamanhos de sub-fluxos analisados variam de 5 a 10 pacotes. Durante este trabalho, duas estratégias de sub-fluxos foram utilizadas na abordagem proposta: a primeira contém os N pacotes iniciais de fluxos TCP, enquanto os N pacotes tomados a partir de uma posição aleatória de fluxos UDP formam a segunda estratégia.

As variáveis estatísticas são extraídas de sub-fluxos e são utilizadas para classificar todo o fluxo. Nos casos em que o número de pacotes no fluxo é menor do que N, todo o fluxo foi utilizado para extrair as características e usado como sub-fluxo inicial e aleatório. Como resultado da etapa final, cada sub-fluxo será representado por um vetor de variáveis. O processo de avaliação utilizado foi o de validação cruzada com 10 partições.

5. Resultados Obtidos

A abordagem 1ct foi aplicada aos dados extraídos da Tabela 3. Os resultados foram comparados com os obtidos utilizando um classificador *Naïve Bayes* multi-classe. Usando as informações de protocolo obtidas no cabeçalho do pacote IP, os classificadores podem ser constituídos de duas partes, as quais foram treinadas separadamente: uma para os fluxos TCP e outra para os fluxos UDP. As métricas abaixo foram utilizadas para avaliar o desempenho dos classificadores:

$$\text{precisão} = \frac{\text{verdadeiro positivo}}{\text{verdadeiro positivo} + \text{falso positivo}} \tag{3}$$

$$\text{revocação} = \frac{\text{verdadeiro positivo}}{\text{verdadeiro positivo} + \text{falso negativo}} \tag{4}$$

As Tabelas 4 e 5 apresentam as matrizes de confusão para as classes TCP, utilizando as variáveis extraídas de sub-fluxos iniciais, para os classificadores 1ct e *Naïve Bayes* multi-classe, respectivamente. A precisão global, em termos de fluxos, foi de 98,43% para a abordagem 1ct e de 97,05% para o classificador *Naïve Bayes* multi-classe.

Tabela 4. Matriz de confusão para 1ct (TCP)

	CHAT	***MAIL***	***P2P***	***SSL***	***WWW***
CHAT	4929	2	1	25	43
MAIL	1	4997	0	0	2
P2P	0	2	4886	22	90
SSL	0	1	6	4958	35
WWW	5	1	82	73	4839

Tabela 5. Matriz de confusão para Naïve Bayes multi-classe (TCP)

	CHAT	*MAIL*	*P2P*	*SSL*	*WWW*
CHAT	4865	9	10	50	66
MAIL	0	4995	5	0	0
P2P	0	1	4687	9	303
SSL	0	3	31	4923	43
WWW	89	0	44	74	4793

As Tabelas 6 e 7 apresentam as matrizes de confusão para as duas classes de tráfego UDP investigadas. No tráfego UDP não é possível garantir a aquisição de pacotes iniciais dos fluxos, portanto esta análise baseia-se apenas nos sub-fluxos aleatórios. Observe que, novamente, a estratégia 1ct tem um desempenho médio superior, neste caso, 98,75% de precisão contra 96,35% para o *Naïve Bayes* multi-classe, superior portanto a 2%.

Tabela 6. Matriz de confusão de 1ct (UDP)

	DNS	**VOIP**
DNS	992	8
VOIP	17	983

Tabela 7. Matriz de confusão de Naïve Bayes multi-classe (UDP)

	DNS	**VOIP**
DNS	998	2
VOIP	71	929

A avaliação de desempenho global pode ser vista nas Tabelas 8 e 9, em termos de fluxos, bytes e pacotes. De acordo com as informações apresentadas nestas tabelas, observa-se que os resultados de revocação e precisão são sempre melhores ao se utilizar a abordagem 1ct.

Tabela 8. Resultados (Revocação)

	Fluxos	**Bytes**	**Pacotes**
Multi-classe TCP	97,05%	97,04%	97,09%
1ct TCP	98,43%	98,59%	98,53%
Multi-classe UDP	96,35%	99,69%	99,33%
1ct UDP	98,75%	99,53%	99,51%

Tabela 9. Resultados (Precisão)

	Fluxos	Bytes	Pacotes
Multi-classe TCP	97,1%	94,51%	97,27%
1ct TCP	98,44%	96,93%	98,53%
Multi-classe UDP	96,57%	99,04%	98,31%
1ct UDP	98,75%	99,65%	99,46%

Observe que a precisão média em fluxos, bytes e pacotes podem ter variações significativas dependendo da aplicação. Para as classes analisadas essa variação não foi significativa mas em outros contextos isso poderia ser relevante. Observe também que os classificadores binários foram especializados para cada classe em termos das variáveis utilizadas e do tamanho da janela. O número de variaveis afeta principalmente o tempo de processamento e o tamanho da janela afeta o atraso. Observou-se que mesmo otimizando estes dois parâmetros, a precisão e a revocação foram consistentemente favoráveis ao método 1ct.

6. Conclusões

Neste trabalho avaliou-se a abordagem 1ct baseada no classificador *Naïve Bayes* para classificação on-line de tráfegos TCP e UDP, a qual permite a utilização de um classificador especializado para cada classe, com as seguintes optimizações: tamanho de sub-fluxo e variáveis específicas para cada classe.

O desempenho foi testado para sub-fluxos formados por pacotes iniciais de cada fluxo TCP, assim como para sub-fluxos extraídos de posições aleatórias de fluxos UDP. Os resultados da classificação, utilizando-se medidas de desempenho (precisão e revocação), também são comparados com os produzidos pelo classificador *Naïve Bayes* multi-classe. Os resultados mostram a superioridade da abordagem proposta.

Finalmente, quando se trata de classificação de tráfego online, a escalabilidade do método proposto é um aspecto relevante em redes de alta velocidade. Neste trabalho não foram realizados testes de escalabilidade. Nesta situação, amostragem do tráfego é a técnica comumente proposta para enfrentar este problema.

No momento, a continuidade deste trabalho segue em três direções: investigar o desempenho da abordagem proposta para novas classes de aplicações; investigar o desempenho de outros classificadores diferentes do *Naïve Bayes* dentro da abordagem 1ct e investigar novas variáveis para a classificação online de tráfego.

7. Agradecimentos

Este trabalho foi financiado pelo programa PROSUP da CAPES.

Referências

Bernaille L., Teixeira, R., Akodkenou, I., Soule, A., and Salamatian, K. (2006) "Traffic classification on the fly". ACM SIGCOMM Computer Communication Review, v.36 n.2.

Duda, R. O., Hart, P. E., and Stork, D. G. (2000), Pattern Classification, Wiley-Interscience, 2nd edition.

Erman, J., Arlitt, M., and Mahanti, A. (2006) "Traffic Classification Using Clustering Algorithms". SIGCOMM Workshops September 11-15, 2006, pp.281-286.

Este, A., Gringoli, F., and Salgarelli, L. (2009) "Support Vector Machines for TCP traffic classification". Elsevier Computer Network, 53(14), pp. 2476 - 2490.

Hall, M. A. (1998) "Correlation-based feature selection for machine learning". Ph.D. thesis, Department of Computer Science, University of Waikato, Hamilton, New Zealand.

Jpcap: Java library for capturing and sending network packets. http://netresearch.ics.uci.edu/kfujii/jpcap/doc/index.html (as of September 2010).

Li, W., Abdin, K., Dann, R., Moore, A. (2006) "Approaching real- time network traffic classification". Technical Report RR-06-12, Department of Computer Science, Queen Mary, University of London.

Moore, A.W., Zuev, D., and Crogan, M. (2005). "Discriminators for use in flow-based classification". In: Passive & Measurement Workshop.

Nguyen, T.T.T., and Armitage, G. (2006) "Training on multiple sub-flows to optimise the use of machine learning classifiers in real-world ip networks". In: Proc. IEEE 31st Conference on Local Computer Networks, Tampa, Florida, USA (2006).

Nguyen, T.T.T., and Armitage, G. (2008) "Clustering to Assist Supervised Machine Learning for Real-Time IP Traffic Classification". IEEE International Conference on Communications, pp. 5857-5862, Beijing, China.

OpenDPI: Ipoque's DPI software's Open Source Version. http://www.opendpi.org/ (as of September 2010).

Tavallaee, M., Lu, W., and Ghorbani, A. A. (2009) "Online Classification of Network Flows". In Proceedings of the 2009 Seventh Annual Communication Networks and Services Research Conference (CNSR '09). IEEE Computer Society, Washington, DC, USA, 78-85.

Wang, Y., and Yu, S.-Z. (2009) "Supervised Learning Real-time Traffic Classifiers". Journal of Networks, Vol 4, No 7.

WEKA 3.6, http://www.cs.waikato.ac.nz/ml/weka/ (as of September 2010).

Zhang, J., Qian, Z., Shou, G., and Hu, Y. (2009) "An automated on-line traffic flow classification scheme". Fifth International Conference on Intelligent Information Hiding and Multimedia Signal Processing, Kyoto - Japan.

Tabela 10. Variáveis para 1ct (TCP)

CHAT	MAIL	P2P	SSL	WWW
Máximo de bytes IP (CS)	Variação maxima de bytes entre pacotes (CS)	Desvio Padrão do intervalo entre pacotes (CS)	Mediana da variação de bytes entre pacotes (CS)	Desvio padrão do intervalo entre pacotes (CS)
Flags SYN (CS)	Maximo de bytes IP (CS)	Variação média de bytes entre pacotes (CS)	Variação mínima de bytes entre pacotes (CS)	Intervalo mínimo entre pacotes (CS)
Tamanho mínimo de janela (SC)	Mediana de bytes IP (CS)	Variação mediana de bytes entre pacotes (CS)	Variação máxima de bytes entre pacotes (CS)	Variação minima de bytes entre pacotes (CS)
Tamanho máximo de janela (SC)	Mediana de bytes IP (SC)	Variação mínima de bytes entre pacotes (CS)	Flag FIN (CS)	Variação maxima de bytes entre pacotes (CS)
	Tamanho máximo de janela (SC)	Mediana de bytes IP (CS)	Flag PSH (CS)	Flag PSH (CS)
	Flag ACK (SC)	Máximo de bytes IP (CS)	Variação máxima de bytes entre pacotes (SC)	Tamanho mediano de janela (SC)
		Tamanho mínimo de janela (CS)	Mínimo de bytes IP (SC)	Flag PSH (SC)
		Tamanho máximo de janela (CS)	Tamanho máximo de janela (SC)	Flag ACK (SC)
		Média do intervalo entre pacotes (SC)	Flag FIN (SC)	
		Desvio padrão do intervalo entre pacotes (SC)	Flag PSH (SC)	
		Mediana do intervalo entre pacotes (SC)	Flag RST (SC)	
		Intervalo mínimo entre pacotes (SC)		
		Variação máxima de bytes entre pacotes (SC)		
		Mediana de bytes IP (SC)		
		Tamanho médio de janela (SC)		
		Desvio padrão do tamanho de janela (SC)		
		Tamanho máximo de janela (SC)		
		Flag RST (SC)		

Tabela 11. Variáveis para 1ct (UDP)

DNS	VOIP
Desvio padrão do intervalo entre pacotes	Média do intervalo entre pacotes
Mediana do intervalo entre pacotes	Média de bytes IP
Média de bytes IP	Mínimo de bytes IP
Mínimo de bytes IP	