Um Novo Método de Classificação de Tráfego Através de Processos Multifractais

Yulios Zavala
Universidade Estadual de Campinas, Faculdade de Engenharia Elétrica e de Computação – FEEC
Avenida Albert Einstein nº 400, CEP 13.083-852, Campinas, SP – Brasil.
Email: yulios@decom.fee.unicamp.br

Jeferson Wilian de Godoy Stênico
Universidade Estadual de Campinas, Faculdade de Engenharia Elétrica e de Computação – FEEC
Avenida Albert Einstein nº 400, CEP 13.083-852, Campinas, SP – Brasil.
Email: jeferson@decom.fee.unicamp.br

Lee Luan Ling
Universidade Estadual de Campinas, Faculdade de Engenharia Elétrica e de Computação – FEEC
Avenida Albert Einstein nº 400, CEP 13.083-852, Campinas, SP – Brasil.
Email: lee@decom.fee.unicamp.br

Resumo

Neste artigo, através de processos multifractais, nós propomos um novo método de classificação de tráfego. Para isto, utilizamos da cascata conservativa para o desenvolvimento da proposta. Este novo método não considera a analise da carga útil dos pacotes ou dos números de portas das aplicações. No entanto os resultados demostraram que o método proposto é viável e eficiente na classificação de tráfego, conseguindo taxas com precisão bastante elevadas. O novo método faz uso de algoritmos de aprendizagem de máquinas junto com estatísticas de fluxo de pacotes na determinação da classe do tráfego. Foram realizados vários testes para validar a eficiência e a precisão da abordagem proposta e seu desempenho mostrou-se superior comparado com outra abordagem na literatura.

PALAVRAS-CHAVE. **Cascata Conservativa, Processos Multifractais, Classificação de Tráfego.**

ABSTRACT

In this paper, through multifractal processes, we propose a new method of traffic classification. For this, we use conservative cascade for the development of the proposal. This new method does not consider the analysis of the payload of the packet or port numbers of applications. However, the results show that the proposed method is feasible and efficient in traffic classification, achieving very high rates accurately. The new method uses machine learning algorithms together with the statistical of packet flow in determining the class of traffic. Conducted several tests to validate the efficiency and accuracy of the proposed approach and its performance was superior compared with one well-known approach in the literature.

KEYWORDS. Conservativa Cascata, Multifractal Process, Traffic Classification

1. Introdução

Segundo levantamento feito por Cisco (2011), o tráfego mundial de dados na internet teve um aumento rápido e considerável e há previsões de quadruplicar nos próximos anos. Isto ocorre principalmente por vários fatores: a grande quantidade de dispositivos que se conectam a rede todos os dias, o crescimento de usuários, a maior velocidade de banda larga e o incremento de aplicações na rede. Baseado neste panorama a correta classificação dos tipos de tráfego que fluem na rede exerce um papel de extrema importância.

A caracterização de carga de trabalho, a provisão de rotas, modelagem, policiamento de tráfego e planejamento de capacidade estão inseridos nas tarefas de gestão de redes e seu bom funcionamento depende da identificação e classificação do tráfego de redes, Alshammari, R and Zincir-Heywood, A. N. (2007).

Ter o conhecimento do que está fluindo nas redes em tempo real, faz com que os operadores tenham a possibilidade de reagir de forma rápida, evitando problemas e conseguindo seus diversos objetivos de negócio. Assim se estes operadores quiserem bloquear a entrada de tráfego de algum protocolo em sua rede ou se algum ISP (*Internet Server Provider*) tenta processar diferentes tipos de conexões com prioridade diferentes (por exemplo limitação dos atrasos dos dados em tempo real), a identificação do protocolo em uso é chave, Hurley,J.; Garcia-Palacios, E. and Sezer, S. (2011).

A classificação do tráfego pode ser uma parte essencial dos Sistemas de Detecção de Intrusão utilizada para detectar padrões de ataques de negação de serviço, ou identificar o mau uso dos recursos de rede, por parte de um cliente que, de alguma forma contraria os termos de serviço do operador Nguyen,T.T.T. and Armitage,G. (2008).

A intercepção legal do tráfego de dados IP tornou-se mais uma obrigação dos ISP porque assim como as empresas de telefonia devem oferecer suporte para a intercepção da utilização do telefone, os provedores de serviços de internet estão cada vez mais sujeitos aos pedidos do governo para obter informações sobre o uso da rede pelos usuários em determinados pontos no tempo Nguyen,T.T.T. and Armitage,G. (2008). Para satisfazer esta obrigação as soluções dos ISP têm como parte essencial a classificação de tráfego.

Dessa forma, a precisa classificação de tráfego de rede é fundamental para várias atividades relacionas às redes, desde monitoramento de segurança até auditorias, e desde provisão de qualidade de serviço até previsões de fornecimento de longo prazo Moore, A.W. and Zuev, D. (2005).

Cada vez mais, novos aplicativos estão sendo implantados na internet (ex. p2p, voip, vídeo streaming, redes sociais), aplicativos que tornaram-se populares rapidamente e que incrementam o uso de portas imprevisíveis. Com esta evolução do tráfego as técnicas de classificação tradicionais, tais como aquelas baseadas nos números de portas bem conhecidas ou análise do payload dos pacotes Dehghani, F.; Movahhedinia, N.; Khayyambashi, M.R.; Kianian, S.; (2010) não são eficazes para todos os tipos de tráfego de rede, ou são incapazes de ser implantados por causa de preocupações de segurança ou privacidade para os dados.

Pensando nisto, neste artigo nós propomos um método simples e eficiente de classificação de tráfego, que leva em consideração as características multifractais dos dados, obtendo resultados bastante significativos e demostrando que o novo esquema é uma alternativa eficaz na identificação dos tráfegos.

O artigo está organizado da seguinte forma: na Seção 2 faremos uma explanação de alguns dos principais trabalhos relacionados com classificação e identificação de tráfegos realizados nos últimos anos. Na seção 3, discutiremos aspectos da teoria multifractal relacionado ao analise do tráfego, introduzindo o conceito de cascatas multiplicativas conservativa. Na Seção 4 é a apresentada o método de classificação proposto. Na Seção 5 apresentamos as configurações para a realização dos experimentos. Na seção 6 são mostrados os resultados dos experimentos. Na Seção 6 concluímos o nosso trabalho.

2. Trabalhos Relacionados

Existem um número considerável de trabalhos que estudam mecânismos de classificação de tráfegos de redes e internet. Esta seção fornece uma visão geral dos mecânismos e sistemas que estão relacionados ao nosso trabalho.

Nos últimos anos as pesquisas realizadas nesta área têm aplicado nas estatísticas dos fluxos de tráfego algoritmos de aprendizado de maquina para fazer as classificações. Assim o uso destas técnicas têm se mostrado promissoras no problema de classificação de tráfego. Nguyen, T.T.T. and Armitage, G. (2008) apresenta um completo levantamento e comparação da literatura ate o ano 2008 da classificação de tráfego com maquinas de aprendizado.

Moore, A.W. and Zuev, D. (2005) propuseram 249 discriminadores de fluxo e de aprendizagem de máquina usada para selecionar os melhores para classificar novos fluxos. Estratégias semelhantes foram aplicados em Zander S., Hguyen T., Armitage G. (2005), Junior, G.P.S.; Maia, J. E.B.; Holanda, R. B. and Sousa, J.N. (2007), para determinar o protocolo ou a classe de protocolo do fluxo que está envolvido. No entanto, estes métodos requerem informações completas sobre o fluxo (por exemplo, total de *bytes* passados) e não são aplicáveis em tempo real.

Bernaille, L.; Teixeira, R.; Akodjenou, I.; Soule, A. and Salamatian, K.(2006),mostraram que as classificações podem ser atingidas se a aprendizagem de máquinas é aplicada apenas para os comprimentos iniciais dos pacotes de um fluxo. O comprimento do primeiro pacote se torna a primeira entrada para o algoritmo de aprendizado de máquina, o comprimento do segundo pacote de entrada se torna o segundo e assim por diante, até o ponto onde o comprimento do pacote N entra e podem ser agrupados em um espaço N-dimensional.

Huang, N.; Jai, G. and Chao H. (2008) ,analisam a troca de pacotes iniciais em um formato semelhante ao Bernaille, L.; Teixeira, R.; Akodjenou, I.; Soule, A. and Salamatian, K. (2006) mas mostram que, considerando os pacotes em grupos pode melhorar a precisão. No entanto, ambos os métodos utilizam números de portas bem conhecidas em suas estratégias de classificação deixando possíveis falhas na segurança do sistema.

Carela-Español,V.; Barlet-Ros, P.;Cabellos-Aparicio, A. and Solé-Pareta, J. (2011) estudaram a classificação de tráfego com informações de fluxos ip de roteadores e switches exportados com o protocolo Netflow desenvolvido pelo Cisco . Aplicou nos dados coletados a popular técnica de arvore de decisão C4.5. Fez o analise da classificação usando a teoria de amostragem (sampling) a qual usa uma porcentagem do total de amostras coletadas, assim variando a taxa de amostras mostrou que ela tem um grave impacto sobre o desempenho do método de classificação.

Em Wang, Y. and Yu, S.Z. (2009) se realiza 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.

Embora haja um amplo trabalho relacionado com o campo da classificação de tráfego, a maioria das técnicas baseadas em ML tem relativamente um limitado sucesso na pratica entre os operadores de rede.

3. O Tráfego Multifractal de Rede

O tráfego nas redes de comunicações é analisado mediante processos probabilísticos que representam o uso que os usuários impõem sobre os recursos da rede. Assim consideram-se variáveis como o tempo entre chegadas dos pacotes, tempo entre conexões, duração das conexões, comprimentos dos pacotes, duração entre sessões.

No início das pesquisas sobre teoria de filas, se considerava que os tempos entre chegadas eram independentes entre si, assim como a quantidade da demanda. Posteriormente se

precisou incluir um efeito de correlação existente entre elas. Dessa forma começou-se a usar de modelos de tráfego Poissonianos onde a correlação cai exponencialmente com o tempo.

Conceitos teóricos importantes para a análise das redes aparecem com Kolmogorov, A.N. (1962). quem introduziu o conceito de auto-similaridade para nomear processos escalonados sem alterações de suas propriedades estatísticas e em 1977 Mandelbrot, B. B. (1977) propõe o termo fractal para descrever objetos muito irregulares. Com estes conceitos é que Leland W., Taqqu M., Willinger W. and Wilson D. (1994). pesquisadores de Bellcore Morristown *Research and Engineering Center*, mostraram que o tráfego sobre redes LAN e WAN é estatisticamente auto-similar possuindo comportamento fractal. Assim a analise multifractal introduzida pela primeira vez, através da utilização de cascatas multiplicativas, no contexto de turbulências por (Kolmogorov, 1941) utilizasse para analisar o tráfego de redes.

A partir do trabalho de Leland et. al teve uma intensificação nos estudos dos fractais, autores como Riedi ,R.H. and Véhel, J.L. (1997) mostraram características multifractais do tráfego TCP, em Feldmann, A., Gilbert, A.C. and Willinger, W. (1997) se mostrou o comportamento multifractal do tráfego de redes por meio do fenômeno escala e em 2003, Krishna, P. M.; Gadre, V. M. and Desai, U. B. (2003) propuseram o modelo VVGM (*Variable Variance Gaussian Multiplier*) baseado em cascata multiplicativa.

A. Cascatas Multiplicativas

Definição 1: Uma cascata multiplicativa é um processo iterativo que fragmenta um determinado conjunto em partes menores de acordo com alguma regra geométrica, e ao mesmo tempo, distribui a massa total do conjunto dado de acordo com outra regra.

A cascata denominada binomial, ou seja, onde a divisão de um determinado conjunto ocorre de dois em dois, é a forma mais simples de se obter um processo multifractal, constituindo de um procedimento iterativo no intervalo fechado $[0,1]$. Sejam m_0 e m_1 denominados multiplicadores da cascata, dois números positivos cuja soma é 1. No estágio $k = 0$ da cascata, obtemos a medida inicial μ_0 do processo com valor aleatório entre o intervalo fechado $[0,1]$. No estágio $k = 1$ a medida μ_1 distribui massa, sendo, $m_0$no subintervalo $[0,1/2]$ e massa igual a m_1 em $[1/2,1]$. Já para o estágio $k = 2$, o intervalo $[0,1/2]$ é novamente dividido em dois subintervalos $[0,1/4]$ e $[1/4,1/2]$, da mesma forma é repetido o procedimento para o intervalo $[1/2,1]$, obtendo então as seguintes medidas: Mandelbrot, B. B. Calvet, L. and Fisher, A. (1997).

$$\mu_2[0,1/4] = m_0 m_0 \;\; \mu_2[1/4,1/2] = m_0 m_1$$
$$\mu_2[1/2,/3/4] = m_1 m_0 \;\; \mu_2[3/4,1] = m_1 m_1 \qquad (1)$$

Repetindo esse procedimento, podemos gerar uma sequência de medidas μ_k convergindo para um processo multiplicativo, ou multifractal μ.

A Figura 1 apresenta o processo de construção de uma cascata conservativa multiplicativa conservativa como descrito acima.

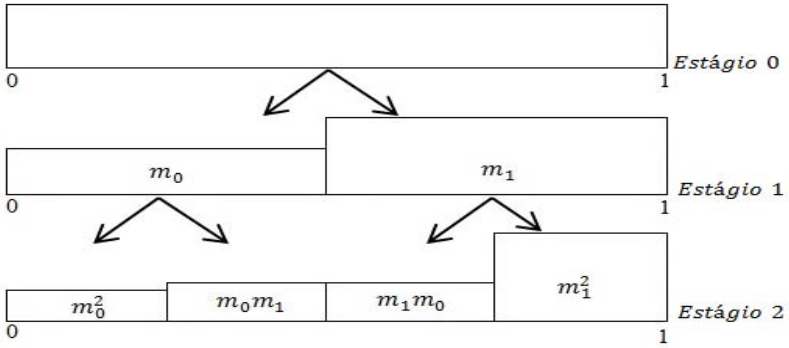


Figura 1 - Processo de Construção da Cascata Multiplicativa Conservativa.

B. Estimação da Densidade de Probabilidade dos Multiplicadores

Nesta seção, apresentamos o algoritmo para a estimação dos multiplicadores (geradores das cascatas), propostos por Feldmann, A., Gilbert, A.C. and Willinger, W. (1997) e uma extensão sugerida com a qual se pode obter funções analíticas para as densidades de probabilidade dos multiplicadores.

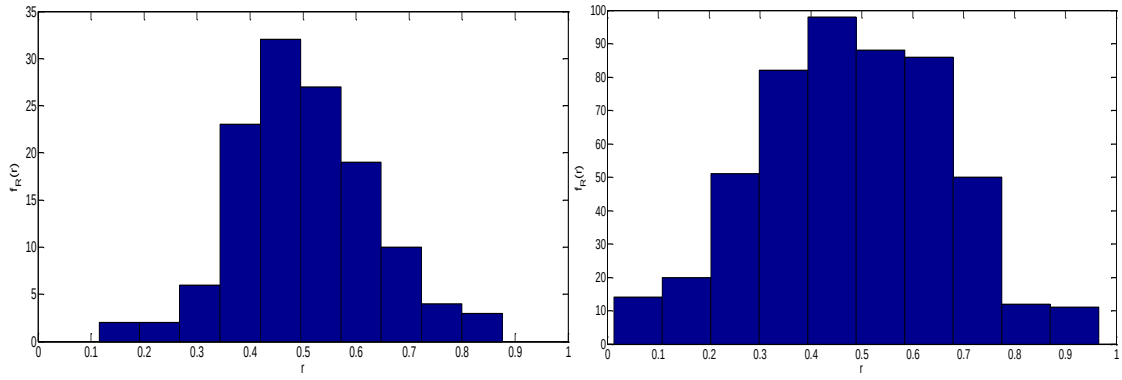
Dessa forma, sejam os dados de tráfego no estágio N da cascata X_i^N, A série de tráfego no estágio $N-1$ da cascata pode ser obtida agregando valores consecutivos do estágio posterior em blocos não sobrepostos de tamanho 2. De forma análoga, dada à série na escala, $(N-j)$, $X_i^{N-j}, (i = 1, \dots, 2^{N-j})$ obtemos os dados na escala $(N-j-1)$ pela soma consecutiva dos valores do estágio $(N-j)$ da seguinte forma:

$$X_i^{N-j-1} = X_{2i-1}^{N-j} + X_{2i}^{N-j} \tag{2}$$

para $i = 1, \dots, 2^{N-j-1}$. Este procedimento termina quando a agregação dos valores forma apenas um ponto na última escala da cascata. Uma estimativa $r_j^{(i)}$ dos multiplicadores pode ser obtida pela seguinte equação adaptada de Feldmann, A., Gilbert, A.C. and Willinger, W. (1997):

$$r_j^{(i)} = \frac{X_i^{N-j}}{X_{2i-1}^{N-j-1}} \tag{3}$$

para $i = 1, \dots, 2^{N-j-1}$. Podemos considerar $r_j^{(i)}$como amostras da distribuição de multiplicadores $f_{R_j}(r)$ no estágio j. A distribuição dos multiplicadores na escala j pode ser obtida pelos histogramas de $r_j^{(i)}$. O modelo multiplicador Gaussiano de variância variável (*Variable Variance Gaussian Multiplier,VVGM*) proposto por Krishna, P. M.; Gadre, V. M. and Desai, U. B. (2003), é uma cascata multiplicativa que aproxima os histogramas obtidos por gaussianas. A distribuição dos multiplicadores $f_{R_j}(r)$ neste modelo é gaussiana centrada em 0.5 com variâncias que mudam a cada escala. Essas variâncias são estimadas a partir dos histogramas para processos de tempo de chegada de pacote. Na Figura 2 mostra os histogramas para uma série de tráfego HTTP, para os estágios N = 5 e N = 8. Podemos observar que a distribuição dos multiplicadores é aproximadamente gaussiana.



(a) (b)

Figura 2 - (a) Estágio N = 5, (b) Estágio N = 8

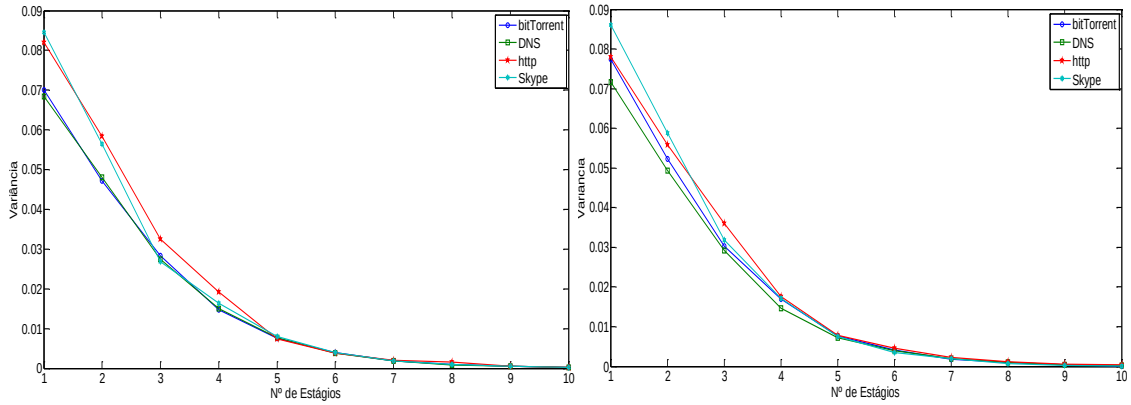
4. Proposta Para a Classificação do Tráfego

Nessa seção será apresentada a metodologia proposta para classificação dos fluxos de tráfego de rede baseada em cascatas multiplicativas. Também descreveremos o conjunto de dados usados nos experimentos.

A. Método de Classificação Proposto

O aprendizado de maquina é efetuado a partir de raciocínio sobre exemplos fornecidos por um processo externo ao algoritmo de aprendizado, assim permite tomar decisões baseado em experiências acumuladas por meio da solução bem-sucedida de problemas anteriores. Oliveira R. (2005).Neste trabalho, nós utilizamos o algoritmo de aprendizado de máquina C4.5 Quinlan, J.R. (1993) em combinação com um processo de cascata para extração de características de grupo de fluxos. Este processo extrai as características por meio da aplicação do método de estimação de densidade de probabilidade dos multiplicadores como citado na Seção 3. Em geral, cada exemplo é descrito por um vetor de valores de características e pelo rótulo da classe associada. O objetivo do algoritmo é construir um classificador que possa determinar corretamente a classe de novos exemplos ainda não rotulados, ou seja, exemplos que não tenham o rotulo da classe.

O objetivo do processo de cascata está na obtenção dos multiplicadores associados a um grupo de fluxos analisados. Os valores dos multiplicadores de acordo com Krishna, P. M.; Gadre, V. M. and Desai, U. B. (2003) assume uma distribuição normal com média igual a ½ como visto na Figura 2 e as variações dos multiplicadores em cada estágios possuem característica exponencial, o que de fato pode ser visto na Figura 3. Dessa forma, a quantidade do número de variâncias é igual à quantidade do número de estágios da cascata. Finalmente os valores obtidos das variâncias são colocados em um vetor em que nós denotamos por "vetor de características".



(a) (b)

Figura 3 - Medida das Variâncias dos Multiplicadores para as Características (a) Tempo de Duração, (b) Quantidade de Bytes.

A Figura 4 apresenta a fase de treinamento de nossa metodologia, a qual se inicia com a coleta dos traços de tráfego. Depois são extraídos os fluxos de acordo com as informações obtidas a partir dos seus cabeçalhos <protocolo, ip origem, ip destino, porta origem, porta destino >. Construímos o conjunto de dados de treinamento rotulando cada fluxo com o nome da aplicação que gerou, para isso usamos a ferramenta de software livre *L7-filter*. Depois continuamos com os passos do processo de cascatas descrito anteriormente, obtendo as variâncias dos multiplicadores para cada estágio.

Os valores das variâncias extraídas são usados pelo algoritmo C4.5 para criar uma árvore de decisão ou modelo de classificação, nós usamos a implementação J48 feita com java do

algoritmo C4.5, esta implementação forma parte do *software open source Weka* (http://www.cs.waikato.ac.nz/ml/weka/). Finalmente esse modelo de decisão construído será usado no processo de classificação.

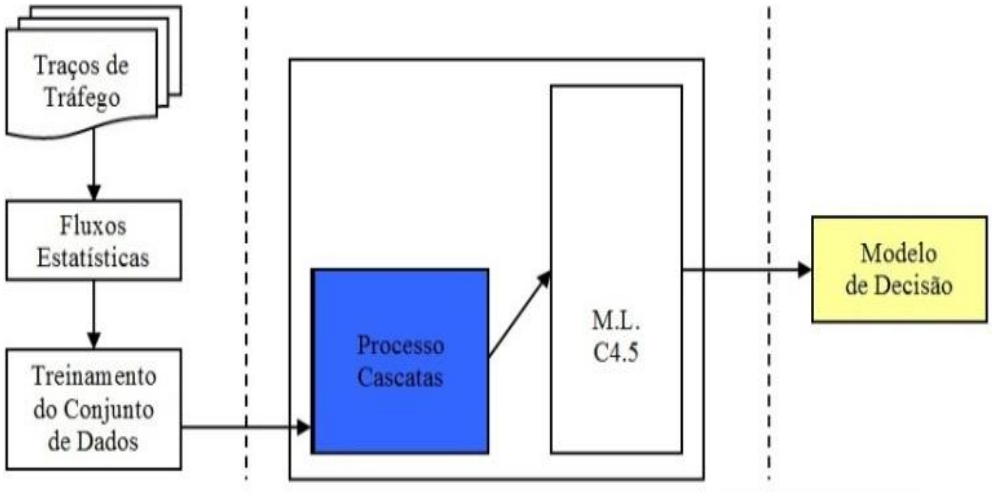


Figura 4 - Fase de Treinamento.

O esquema de classificação é mostrado na Figura 5, inicialmente o traço de tráfego a ser classificado é agrupado em fluxos, depois obtemos estatísticas das variáveis de tráfego usadas para a análise, com estas estatísticas cria-se o conjunto de dados respectivo ao traço inicial. O próximo passo é extrair as variâncias dos multiplicadores do conjunto de dados aplicando o processo de cascatas enunciado no início desta seção. Finalmente obtemos a predição da classificação.

O intuito de nossa metodologia é mostrar uma classificação sem análise da carga útil (*payload*) ou dos números de portas *Well-Known*. Nossa metodologia está baseada especificamente na análise com cascatas multiplicativas de três variáveis dos fluxos: número de pacotes, total de bytes e tempo de duração.

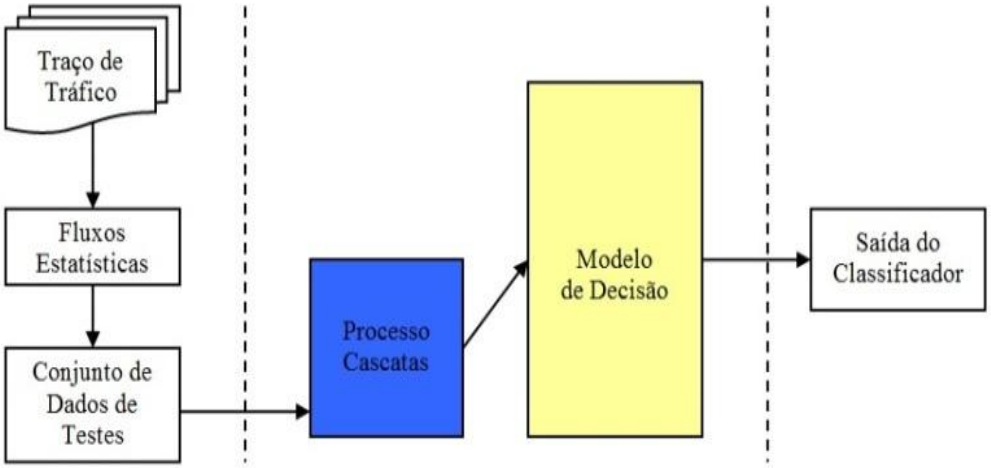


Figura 5 - Fase de Classificação

Na Figura 6 é mostrado em síntese os processos descritos nas Figuras 4 e 5.

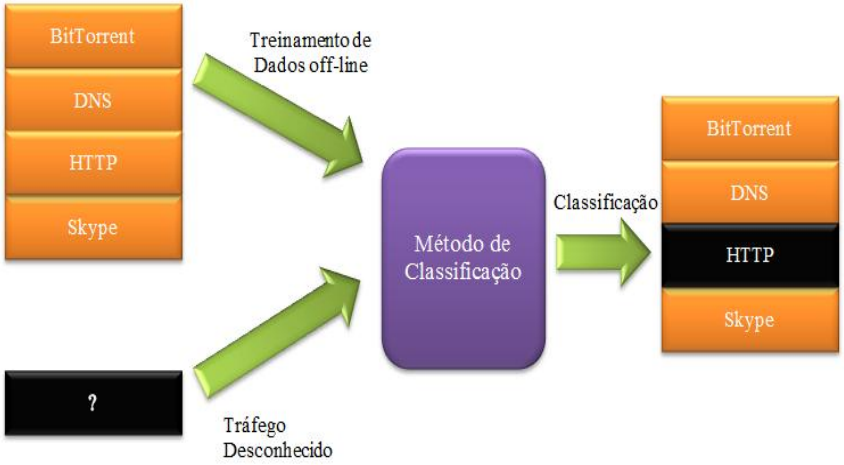


Figura 6 - Esquema de Classificação

B. Base de Referência (Base-truth)

Estabelecimento da *Base-truth* ou base de dados de referência é uma fase critica de qualquer método de classificação de tráfego que use *machine learning* – (ML), porque todo o desempenho da classificação depende da precisão na identificação dos fluxos que serão usados como referências Carela-Español,V.; Barlet-Ros, P.;Cabellos-Aparicio, A. and Solé-Pareta, J. (2011).

A Tabela I mostra os traços de utilizamos neste trabalho, traços II e IV do conjunto de dados de avaliação utilizados no trabalho de Carela-Español,V.; Barlet-Ros, P.; Cabellos-Aparicio, A. and Solé-Pareta, J. (2011). Esses traços foram recolhidos no link de acesso Gigabit da Universidade Politécnica da Catalunha (UPC) (http://loadshedding.ccaba.upc.edu/trafic_classification), que conecta cerca de 25 faculdades e 40 departamentos (geograficamente distribuídas em 10 campus) à Internet através da rede espanhola de pesquisa e educação (RedIRIS).

TABELA I. **Conjunto de Dados de Avaliação**

Nome Traço	Data	Fluxos Utilizados	Classe	Fluxos x Classe
UPC-II (Treinamento)	11-12-08	1314284	P2P	414829
			DNS	260595
			HTTP	194086
			VoIP	444774
UPC-IV (Teste)	12-12-08	1495302	P2P	380320
			DNS	286337
			HTTP	129994
			VoIP	698651

Em nossos experimentos, nós analisamos 1.314.284 fluxos do traço UPC-II como conjunto de dados de treinamento, esses fluxos analisados são os fluxos das quatro maiores classes existentes no traço, de igual maneira utilizamos 1.495.302 fluxos das maiores classes do traço UPC-IV como conjunto de dados de testes. O formato dos traços consiste em um arquivo de texto simples onde cada linha contém diversas informações relacionadas a um fluxo, nós utilizamos apenas três informações as quais são mostradas na Tabela II, assim os conjuntos de dados ficaram com o seguinte formato: <duração, pacotes, *bytes*, rotulo>.

O processo de rotulo dos fluxos foi feito com uma ferramenta baseada no software livre (http://l7-filter.sourceforge.net/). O *software L7-filter* é do tipo DPI (*Deep Packet Inspection*),

assim ele procura por padrões característicos no *payload* dos pacotes e os rotula com o aplicativo correspondente.

TABELA II. **Variáveis Características dos Fluxos**

Características	Descrição
Duração	Duração do Fluxo em Segundos
Pacotes	Número Total de Pacotes no Fluxo
Bytes	Total de *Bytes* por Fluxo

C. Métricas de Desempenho

Caracterizamos o desempenho de nosso método de classificação com as seguintes métricas de uso frequente na literatura: Precisão, Acurácia e Recall, Carela-Español,V.; Barlet-Ros, P.;Cabellos-Aparicio, A. and Solé-Pareta, J. (2011), Dehghani, F.; Movahhedinia, N.; Khayyambashi, M.R.; Kianian, S. (2010)

Precisão é o número de verdadeiros positivos TP *(True Positives* ou número de itens corretamente classificados como pertencentes à classe positiva), dividido pelo número total de elementos classificados como pertencentes à classe positiva, ou seja, a soma de verdadeiros positivos TP e falsos positivos FP (*False Positives* ou número de itens incorretamente rotulados como pertencentes à classe positiva). Esta medida é calculada através da seguinte expressão:

$$Precisão = \frac{TP}{TP+FP} \tag{4}$$

Acurácia é o acerto do sistema considerando a proporção de instâncias corretamente classificadas no total dos registos classificados, apresenta a mesma expressão de Precisão, mas com a diferença de que a precisão refere-se a apenas uma classe de fluxo e acurácia refere-se a todas as classes de fluxos.

Recall é definido como o número de verdadeiros positivos TP dividido pelo número total de elementos que realmente pertencem à classe positiva, ou seja, a soma dos verdadeiros positivos TP e falsos negativos FN (*False Negatives* ou número de itens que não foram identificados como pertencentes à classe positiva, mas deveria ter sido). Sua expressão é a seguinte:

$$Recall = \frac{TP}{TP+FN} \tag{5}$$

5. Resultados

Nesta seção serão apresentados os resultados utilizando a metodologia proposta para quatro classes de tráfegos:

1. P2P (*bittorrent*), protocolo para troca de arquivos;
2. DNS sistema usado para resolver os nomes de domínio em endereços de rede;
3. HTTP, protocolo utilizado para a passagem de documentos de hipertexto como paginas web;
4. VoIP (*Skype*) telefonia sobre internet.

Além disso, comparamos os nossos resultados com o método proposto por Dehghani, F.; Movahhedinia, N.; Khayyambashi, M.R.; Kianian, S. (2010), que utiliza uma abordagem baseada nas estatísticas dos conteúdos de carga útil (*payload*) e em estatísticas das características dos pacotes, para os seguintes traços de tráfego HTTP e DNS. É importante mencionar que a comparação de taxas HTTP e DNS está grandemente limitada devido ao uso de diferentes conjuntos de dados, a falta de um traço de referência comum é uma dificuldade para as pesquisas nesta área.

Na Tabela III mostramos os resultados referentes à métrica Precisão para os tráfegos analisados. Podemos observar que a classe de tráfego HTTP apresenta a melhor taxa de 98% quando utilizamos cascatas com 9 estágios. Para o método proposto por Dehghani, F.; Movahhedinia, N.; Khayyambashi, M.R.; Kianian, S. (2010) está precisão é de 91%. Como

podemos observar na Tabela IV. Já o tráfego do tipo DNS o método proposto apresenta uma taxa de 99%, sendo que o método proposto por Dehghani, F.; Movahhedinia, N.; Khayyambashi, M.R.; Kianian, S. (2010) obtém uma taxa de 80% (vide Tabela IV).

TABELA III. **Precisão dos Resultados no Conjunto de Dados Teste**

Applicação	N=6	N=7	N=8	N=9	N=10
Bittorrent	0.873	0.916	0.929	0.965	0.944
DNS	0.879	0.943	0.977	0.991	0.996
HTTP	0.885	0.948	0.957	0.981	0.913
Skype	0.975	0.989	0.999	0.999	0.999

Na Tabela IV mostramos os resultados referentes ao nível $N = 9$ de uma cascata, onde obtemos nossos melhores resultados, como podemos observar a nossa proposta apresenta resultados superiores ao método comparado Dehghani, F.; Movahhedinia, N.; Khayyambashi, M.R.; Kianian, S.; (2010), para os traços de tráfegos HTTP e DNS, sendo uma alternativa eficiente na classificação de tráfego.

TABELA IV. **Comparação de Métodos (Precisão)**

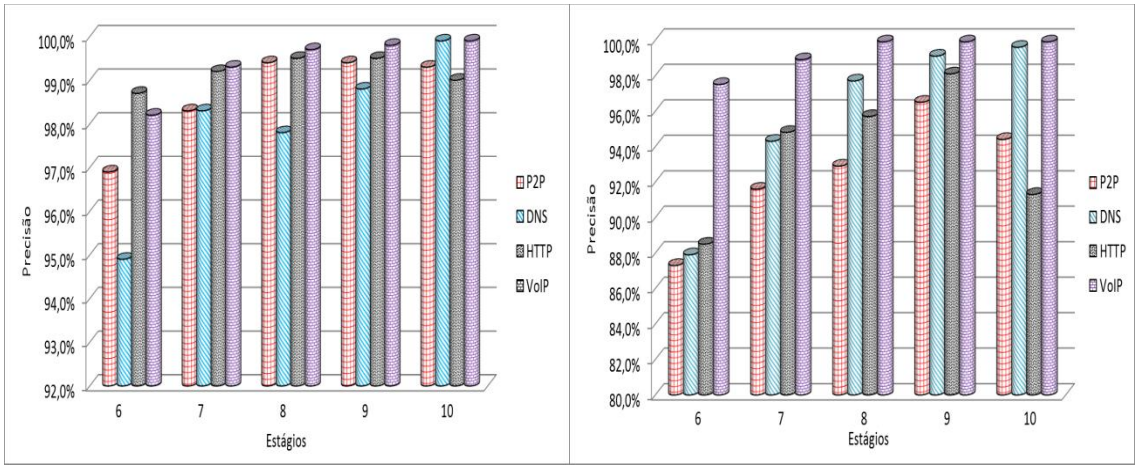
Tráfego/Método	Proposto (N = 9)	Proposto por Dehghani et.al.(2010)
HTTP	0.981	0.910
DNS	0.991	0.800

Nas Figuras 7 e 8 observamos a métrica Precisão e Recall na classificação dos quatro tipos de classes, utilizando 6, 7, 8, 9 e 10 estágios. Observa-se que as taxas de treinamento e teste são todas próximas a 100%, no entanto, a melhor precisão para todas as classes é alcançada com o uso de cascatas com 9 estágios.

Na Tabela V mostramos os resultados da comparação dos resultados da classificação para a métrica Recall, na fase de testes, usando o método proposto por e Dehghani, F.; Movahhedinia, N.; Khayyambashi, M.R.; Kianian, S.; (2010) novamente observamos que o nosso método apresenta melhores resultados.

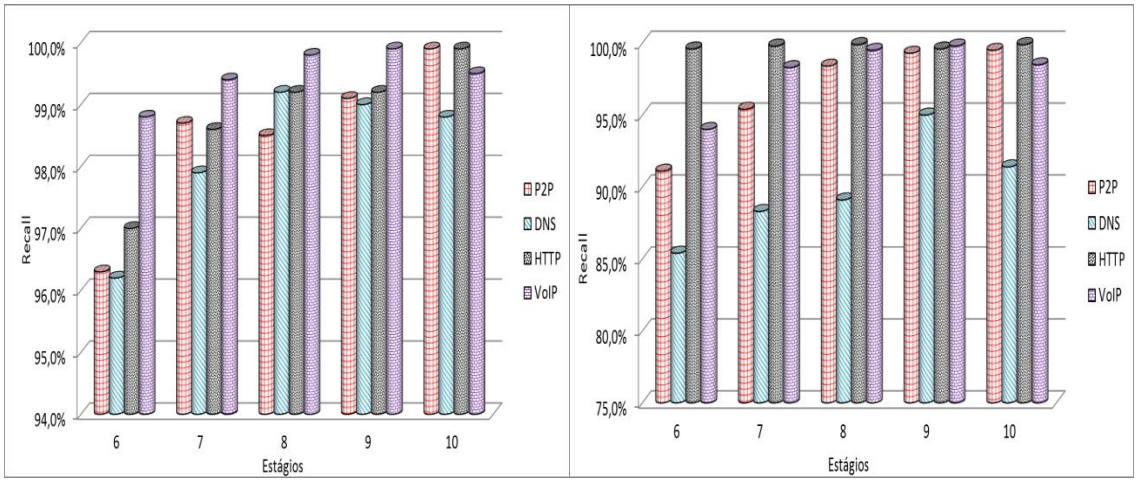
TABELA V. **Comparação de Métodos (Recall)**

Tráfego/Método	Proposto (N = 9)	Proposto por Dehghani et.al.(2010)
HTTP	0.996	0.931
DNS	0.950	0.832



(a) **(b)**

Figura 7- Precisão: (a) Conjunto de Dados, (b) Conjunto de Dados de Testes.



(a) (b)

Figura 8 - Recall: (a) Conjunto de Dados, (b) Conjunto de Dados de Testes.

Podemos notar também na Tabela VI, que a metodologia proposta baseada nas variâncias dos multiplicadores dos tráfegos, consegue taxas de Acurácia tanto no conjunto de dados de treinamento como no conjunto de dados de testes acima dos valores usando apenas o algoritmo C.45, Quinlan,J.R. (1993) o que mostra que o método proposto é robusto e eficiente.

TABELA VI. **Acurácia Total em %**

Conjuto de Dados	C.45	N=6	N=7	N=8	N=9	N=10
Treinamento	93.7	97.2	98.8	99.2	99.4	99.6
Teste	91.1	92.1	95.8	97.2	98.7	97.5

6. Conclusões

Neste trabalho foi proposto utilizando cascata multiplicativa, um novo método de classificação de tráfego. Utilizamos das variâncias obtidas em cada estágio da cascata para construir um vetor denominado "vetor característica", com isto, aplicamos em um algoritmo de aprendizado de máquina como o C4.5. Vários testes foram realizados com o intuito de validar a nossa proposta. Comparamos os resultados do método proposto com os resultados obtidos por outro método existente na literatura, para mostrar a eficiência da metodologia.

Os resultados obtidos demostraram que o método proposto é eficiente e robusto, pois conseguiu classificações de tráfego em um nível de precisão bastante elevado, considerando diferentes estágios da cascata. Porém nossa técnica pode ser mais bem apurada por meio da seleção de um número ideal de estágios das cascatas utilizadas para a análise. Acreditamos que isso deve ocorrer se conseguirmos analisar de forma mais minuciosa os padrões de comportamento de cada tipo de tráfego existes na rede.

Referências

Alshammari,R. and Zincir-Heywood,A. N. (2007), A Flow Based Approach for SSH Traffic Detection, in Systems, Man and Cybernetics, 2007. ISIC. IEEE International Conference on, 2007, pp. 296-301.

Bernaille, L.; Teixeira, R.; Akodjenou, I.; Soule, A. and Salamatian, K. (2006), Traffic Classification on The Fly, ACM Sigcomm comput.commun. rev., 2006, 36, (2), pp. 23–26.

Carela-Español,V.; Barlet-Ros, P.;Cabellos-Aparicio, A. and Solé-Pareta, J. (2011), Analysis of the Impact of Sampling on NetFlow Traffic Classification, Computer Networks, vol. 55, pp. 1083-1099, 2011.

Cisco, (2011), Global Internet Traffic Projected to Quadruple by 2015. Available: http://newsroom.cisco.com/dlls/2011/prod_060111.html.

Dehghani, F.; Movahhedinia, N.; Khayyambashi, M.R.; Kianian, S. (2010). Real-Time Traffic Classification Based on Statistical and Payload Content Features. Intelligent Systems and Applications (ISA), 2010 2nd International Workshop, pp.1-4, 22-23.

Erman, J.; Mahanti, A. and Arlitt, M. (2006), Internet Traffic Identification using Machine Learning," in Global Telecommunications Conference, 2006. GLOBECOM '06. IEEE, 2006, pp. 1-6.

Feldmann, A., Gilbert, A.C. and Willinger, W. (1997), Data Networks as Cascades: Investigating the Multifractal Nature of Internet WAN Traffic. Proc. Of 35th Annual Allerton Conf. on Communications, Control, and Computing, pp 269-280.

Hurley,J.; Garcia-Palacios, E. and Sezer, S. (2011), Classifying Network Protocols: A 'Two-Way' Flow Approach, Communications, IET, vol. 5, pp. 79-89.

Junior, G.P.S.; Maia, J. E.B.; Holanda, R. B. and Sousa, J.N. (2007), P2P Traffic Identification Using Cluster Analysis. Global Information Infrastructure symp., giis 2007, july 2007, pp. 128–133.

Krishna, P. M.; Gadre, V. M. and Desai, U. B. (2003), Multifractal Based Network Traffic Modeling, Kluwer Academic Publishers, Boston, MA.

Kolmogorov, A.N. (1962), A Refinement of Previous Hypothses Concerning the Local Structure of Turbulence in a Viscous Incompressible Fluid a High Reynolds number, 13, 82–85.

Leland W., Taqqu M., Willinger W. and Wilson D. (1994), On the self-similar nature of Ethernet Traffic (extended version), IEEE/ACM Transactions on Networking, v.2, n.1, pp 1-15.

Mandelbrot, B. B. (1977), The Fractal Geometry of Nature. Nova York: W.H.Freeman.

Mandelbrot, B. B. Calvet, L. and Fisher, A. (1997), Large Deviations and The Distribution of Price Changes. Discussion paper No 1165 of the Cowles Foundation for Economics at Yale University.

Moore, A.W. and Zuev, D. (2005), Internet Traffic Classification Using Bayesian Analysis Techniques. SIGMETRICS Perform. Eval. Rev., vol. 33, pp. 50-60, 2005.

Moore, A. Crogan, M; Moore, A. W.; Mary Q. and Zuev, D. (2005), Discriminators for Use in Flow-Based Classification. Technical Report, Intel Research, 2005.

Nguyen, T.T.T. and Armitage, G. (2008), A Survey of Techniques for Internet Traffic Classification Using Machine Learning. Communications Surveys & Tutorials, IEEE, vol. 10, pp. 56-76.

Oliveira R.,S.(2005), Sistemas Inteligentes Fundamentos e Aplicações. Baureri, SP, 2005.

Quinlan, J.R. (1993), C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers.

Riedi ,R.H. and Véhel,J.L. (1997), Multifractal Properties of TCP Traffic: A Numerical Study , Technical Report 3129, INRIA Rocquencourt.

Traffic Classification at the Universitat Politècnica de Catalunya (UPC). <http://loadshedding.ccaba.upc.edu/trafic_classification>.

Zander S., Hguyen T., Armitage G. (2005), Automated Traffic Classification and Application Identification Using Machine Learning. proc. IEEE conf. on local computer networks 30th anniversary, lcn.

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

Weka, Data mining software in java.< http://www.cs.waikato.ac.nz/ml/weka/>.