Avaliação de Classificação de Tráfego IP baseado em Aprendizagem de Máquina Restrita à Arquitetura de Serviços Diferenciados

Michael Taynnan Barros[†], Reinaldo Cézar de Morais Gomes [†], Marcelo S. Alencar[†],
Paulo Ribeiro L. Júnior[†] e Anderson Costa [§]
[†]Instituto de Estudos Avançados em Comunicações, Campina Grande, PB
[†]Universidade Federal de Campina Grande, Campina Grande, PB
[§]Instituto Federal de Educação, Ciência e Tecnologia da Paraiba, Campina Grande, PB
email: {michael.taob,paulo,malencar}@iecom.org.br, reinaldo@dsc.ufcg.edu.br, anderson@ifpb.edu.br

***Resumo*— A classificação de tráfego na arquitetura DiffServ é feita utilizando técnicas que não apresentam acurácia elevada. Esse fato pode ser a causa do baixo desempenho da arquitetura, pois com ela é preciso negociar recursos de uma determinada aplicação, que deve ser conhecida. Técnicas baseadas em aprendizagem de máquina para classificação de tráfego IP são as abordagens usadas, atualmente, para aumentar a acurácia na etapa de classificação. Este artigo apresenta uma avaliação de desempenho dessas técnicas (Árvores de Decisão, *Naive* Bayes, Redes Bayseanas e Redes Neurais) com base nas restrições de complexidade do DiffServ. Os resultados apresentam uma indicação para as técnicas: Àrvores de Decisão e Redes Bayseanas.**

***Palavras-chave*— Classificação, tráfego, redes bayseanas**

I. Introdução

BILHÕES de pessoas no mundo inteiro podem se comunicar hoje por uma estrutura chamada Internet, que tem causado efeitos econômicos e sociais [1]. Porém, a Internet pode estar à beira de uma crise [2]. O número de usuários está crescendo enormemente, podendo atingir cerca de cinco bilhões de usuários no ano de 2020 [3], e as tecnologias envolvidas para disponibilização de conexões de qualidade não acompanham esse ritmo. Estima-se, também, que mais de 50 bilhões de entidades (sensores, programas de segurança e controle, etc.) podem usar a Internet nos próximos nove anos [3].

Os dados precedentes preocupam a comunidade científica de redes de computadores e das grandes empresas de telecomunicações, que direcionam suas atividades para o desenvolvimento das redes de próxima geração (NGNs). O objetivo dessa nova linha de pesquisa é o desenvolvimento de tecnologias, arquiteturas e protocolos que incrementem os recursos da infra-estrutura de Internet do futuro. Com essas preocupações em mente, o número de tecnologias propostas para aumentar a taxa de transmissão em redes de computadores cresceu bastante, mas com custo elevado, e além de não representar tudo, por exemplo, aplicações multimídias classificadas como não elásticas apresentam baixa tolerância à atraso. Uma solução para esse problema é a utilização das arquiteturas de qualidade de serviço (QoS), e a abordagem mais citada é o modelo de Diferenciação de Serviços (DiffServ). Contudo, muitos obstáculos ainda têm impedido a implementação em larga escala dessas arquiteturas na Internet [4]. Segundo os ISPs (*Internet Service Providers*), essas tecnologias não estão maduras o suficiente, em termos de desempenho e segurança, para encorajar implementações comerciais.

O primeiro bloco da arquitetura DiffServ é o classificador de tráfego. Esse bloco é responsável pela identificação do tráfego atual dos roteadores que implementam a arquitetura. Os resultados da classificação devem ser os mais acurados possíveis, pois, tráfegos identificados erroneamente podem levar à atribuição de níveis de QoS errados, afetando o desempenho da rede e da aplicação, e os outros blocos necessitam desses resultados. As abordagens usadas para classificação são baseadas em porta e carga [5]. Como mostrado em [6][7][8][9], esses tipos de abordagens falham em classificar fluxos de P2P, jogos, Streaming e FTP. Esses fluxos correspondem a mais de 50% de todo o tráfego de Internet, indicando que, a classificação não consegue identificar grande parte dos fluxos que passam pelos roteadores. O problema se resume em identificar um classificador com maior desempenho e que não afete o desempenho da rede e das aplicações.

O objetivo deste artigo é avaliar o desempenho de quatro classificadores de tráfego IP baseados em aprendizagem de máquina (Árvores de Decisão, *Naive* Bayes, Redes Bayseanas e Redes Neurais), com respeito à acuracia, à qualidade na classificação e à complexidade, do ponto de vista da arquitetura de Serviços Diferenciados da Internet e no contexto de Qualidade de Serviço. Há intenção de definir um espaço de classificadores habilitados à implementação nessa arquitetura. O espaço de classificadores é o espaço de soluções para o problema de classificação no modelo DiffServ. A implementação é direta, ou seja, não é necessária alteração no modelo, mas a permutação da técnica atual por uma contida no espaço de soluções. É esperado um aumento no desempenho da arquitetura, possibilitando um aumento de ISPs com o DiffServ.

O escopo deste artigo restringe-se à uma avaliação de desempenho dos classificadores com cinco amostras de tráfego

Artigo recebido em 26 de outubro de 2011. Artigo aceito em 26 de março de 2012.

IP com diversidade geográfica e de aplicações. Uma análise de variância dos resultados e uma análise gráfica via intervalos de confiança é apresentado.

Nos resultados é possível observar o efeito de cada classificador nas variáveis dependentes do experimento, fato comprovado com a análise de variância com o teste ANOVA. Sobre o contexto de acurácia, qualidade de classificação e complexidade a técnica Árvores de Decisão apresenta resultados conclusivos de que ela pode aumentar o desempenho da arquitetura DiffServ. O número de cinco amostras é suficiente para a análise total dos resultados (e também para WEB, DNS, Mail/News, Games, NetOp e Encryption), pois a variabilidade dos valores é relativamente baixa, entretanto, para certo tipo de aplicações (i.e., P2P, FTP, Streaming e Chat) esse número não é suficiente.

II. Trabalhos Relacionados

A. Revisão de Trabalhos sobre Classificação de Tráfego IP de Internet

Muitos classificadores foram propostos para aumentar o desempenho de classificação de tráfego IP para aplicações P2P na Internet. Por isso, Kim at al. [6] avaliaram técnicas agrupadas em: abordagem baseada em porta, abordagem baseada no comportamento de *hosts* e abordagem baseada em aprendizagem de maquina. Os resultados mostraram que a abordagem de aprendizagem de maquina apresenta a maior acurácia e qualidade de classificação, e a técnica SVM(*Support Vector Machine*) foi a melhor desse grupo apresentando 94.2-97.8% de acurácia.

Sob o mesmo contexto do trabalho precedente, Lim et al [10] respondem as seguintes perguntas: Por que os classificadores baseados em aprendizagem de máquina possuem maior precisão? E qual fator é o responsável pela diferenciação de desempenho entre essas técnicas? Os resultados do processo experimental efetuado nesse artigo afirma que a técnica de árvore de decisão (C4.5) apresenta maior desempenho. Outras contribuições indicam que os primeiros pacotes de uma conexão na camada de transporte são essenciais para a classificação do tráfego e que a discretização das estatísticas dos fluxos influi diretamente nos resultados.

B. Revisão de Trabalhos sobre Desempenho do DiffServ

As redes de *backbone* apresentam baixo desempenho por métricas de QoS (perdas, atraso, variação do atraso, vazão). É necessária a criação de um técnica que implemente classes estáticas sobre diferentes tipos de tráfego para controle de SLA (*Service Layer Agreement*), porém com a restrição de uso de vazão apertada (*tight*). Filsfils et al. [11] apresenta um estudo sobre *testbeds* Cisco com a arquitetura DiffServ padrão, obtendo: baixo atraso, baixa variação do atraso, alocação de largura de banda apertada e projeto de SLAs apertados.

Uma preocupação corrente é que a arquitetura de diferenciação de serviços não disponibiliza máxima vazão disponível. Esse fato afeta o desempenho da arquitetura degradando a qualidade de serviços de aplicações da Internet. Onali e Artozi [12] argumentam que um mapeamento ótimo de tipos de serviços em classes de serviços e uma alocação ótima de largura de banda em cada classe são as soluções. Eles propõem um modelo com essas características, apresentando um ganho no desempenho do sistema como também na utilização dos recursos da rede.

Muitos trabalhos foram desenvolvidos para QoS em peças distintas e separadas. A interação entre elas é um dos motivos para muitos ISPs não implementarem a arquitetura DiffServ. O desafio atual da comunidade de QoS, abordado por Meddeb [4], é juntar todas essas peças em uma solução robusta o suficiente para a implantação nos ISPs da internet.

C. Revisão de Trabalhos sobre Classificação de Tráfego no Desempenho de QoS da Internet

Na fase de negociação de recursos, os classificadores de tráfego inferem o comportamento do tráfego que passa nos roteadores que negociam recursos sobre prioridade. Porém, os classificadores encontrados na literatura, aplicados ao problema, não apresentam desempenho satisfatório. Sobre esse problema técnico, Park et al. apresentaram dois trabalhos abordados a seguir.

Em [13], os autores apresentam um estudo de novos classificadores de tráfego aplicado ao problema de estabelecimento de recursos e QoS na Internet. Eles demonstraram que a técnica de árvores de decisão apresenta maior precisão e menor complexidade dentre as avaliadas, conclusões muito parecidas com as deste trabalho. Em [14] eles propõem uma nova metodologia de classificação, efetuada em duas fases: seleção de recursos e classificação. Foram utilizados um algoritmo genético para fase de seleção de recursos mais uma árvore de decisão e uma árvore de decisão com uma otimização de poda mais uma rede bayseana. Essas combinações foram simuladas apresentando maior desempenho para a primeira combinação.

III. Visão global dos classificadores de tráfego baseado em aprendizagem de máquina

A. Árvores de Decisão - AD

Uma árvore de decisão é um modelo de aprendizagem de máquina que decide a categoria de uma aplicação de um nova amostra baseado em valores de várias características de fluxos do conjunto de dados. Os nós internos denotam diferentes características, os ramos entre os nós correspondem aos valores possíveis que essas características podem ter, enquanto os nós folhas correspondem aos valores da classificação das aplicações. Para realizar a classificação, uma dada tupla, que corresponde às características dos fluxos, caminha pela árvore do nó raiz até todos os folhas. O rótulo de um nó folha é o resultado da classificação. A Árvore de Decisão J.48, usada neste artigo, é uma implementação livre da árvore C.45.

B. Naive Bayes - NB

Naive Bayes é a mais simples técnica probabilística de classificação baseada no teorema de Bayes. Ela analisa as relações entre cada característica e as categorias de aplicações para cada instância, derivando a probabilidade condicional para cada relação entre os valores das características e das categorias.

C. Redes Bayseanas - RB

Pode-se definir uma rede bayseana como um grafo direcionado e acíclico que representa um conjunto de características (ou categorias) como os nós do grafo, tendo as relações probabilísticas como vértices. Se a assumpção de independência condicional não é válida, o aprendizado de Redes Bayseanas pode superar o do *Naive* Bayes.

D. Redes Neurais - RN

Uma rede neural é altamente interconectada por unidades, neurônios, do qual a saída é uma combinação de múltiplas entradas de outros neurônios. É usado neste trabalho o modelo mais simples de classificação usando redes neurais, chamado de *Multilayer Perceptron*, que consiste de uma única camada de entrada (características), uma única camada de saída dos neurônios (categorias) e um ou mais camadas escondidas entre eles.

IV. Metodologia da Avaliação

A. Objetivos da Avaliação

O objetivo específico da investigação empírica é responder à seguinte pergunta de pesquisa: *Sobre as restrições de desempenho impostas pela arquitetura DiffServ, qual(ais) é(são) o(s) classificador(res) de tráfego que apresenta(m) melhor desempenho em termos de acurácia, qualidade na classificação e análise de complexidade?*

B. Formulação de Hipóteses

Quatro hipóteses são formuladas, as primeiras hipóteses estão relacionadas com o efeito da classificação de tráfego entre diferentes técnicas. A seguir as primeiras hipóteses são apresentadas:

H1-0: Classificadores de tráfego baseados em aprendizagem de máquina não apresentam diferença de desempenho.
H1-A Classificadores de tráfego baseados em aprendizagem de máquina apresentam diferença de desempenho.

A hipótese nula deve ser rejeitada para que haja uma comparação entre as diferentes técnicas. Para isso, será feito uma análise de variância dos dados.

A escolha de um ou mais classificadores de tráfego para a arquitetura DiffServ será efetuada com a análise das hipóteses H2, formuladas abaixo:

H2-0 Classificadores de tráfego que possuem a melhor acurácia não são os escolhidos.
H2-A Classificadores de tráfego que possuem a melhor acurácia são os escolhidos.

É preferível que classificadores com menor complexidade sejam escolhidos, pois o desempenho dos classificadores não deve afetar o desempenho da rede e das aplicações. As hipóteses H2 foram formuladas com base nesse argumento e serão analisadas por meio de intervalos de confiança.

Como as hipóteses H2 devem ser analisadas se a hipótese H1-0 for rejeitada, uma árvore de hipóteses foi desenvolvida, como mostrado na Figura 1.

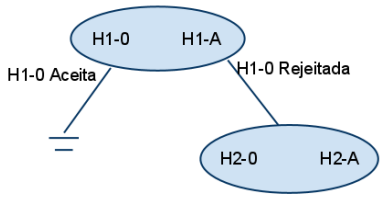

Fig. 1. Árvore de Hipóteses.

C. Definição de Variáveis

1) Variáveis Independentes: Com o aumento do número de dados para treinamento é esperado um aumento na acurácia dos classificadores. Para efetuar essa análise uma variável independente será o tamanho de dados de treinamento dos classificadores. Outra variável corresponderá os classificadores, sendo esta considerada como categórica.

2) Variáveis Dependentes: As variáveis dependentes escolhidas são: acurácia, precisão, cobertura e *f-measure*. Acurácia é a razão da soma de todos os verdadeiros positivos pela soma de todos os verdadeiros positivos e falsos positivos para todas as classes.

$$Acuracia = \frac{\sum_{i=0}^{C} \frac{VerdPositivos[i]}{VerdPositivos[i]+FalsPositivos[i]}}{C} \quad (1)$$

A precisão de um classificador é a razão de verdadeiros positivos pela soma de verdadeiros positivos e falsos positivos ou a percentagem de fluxos que são atribuídos propriamente para uma aplicação.

$$Precisao = \frac{VerdPositivos}{VerdPositivos + FalsPositivos} \quad (2)$$

Cobertura é razão entre os verdadeiros positivos e a soma de verdadeiros positivos e falsos negativos ou a percentagem de fluxos em uma classe de aplicação que é corretamente identificada.

$$Cobertura = \frac{VerdPositivos}{VerdPositivos + FalsNegativos} \quad (3)$$

Finalmente, *F-measure* considera tanto precisão quanto cobertura em única variável, sendo a média harmônica delas.

$$F - measure = \frac{2xPrecisaoxCobertura}{Precisao + Cobertura} \quad (4)$$

A acurácia é responsável por medir o nível de acerto da classificação em toda uma população. Essa métrica está relacionada com o desempenho de cada técnica. Já as outras métricas estão relacionadas com a qualidade da classificação e são as responsáveis pelas comparações per-aplicações. Essas comparações são necessárias para observar as falhas na classificação das unidades experimentais.

D. Seleção de Unidades Experimentais

As unidades experimentais selecionadas para a investigação empírica são:

- Árvores de Decisão
- *Naive* Bayes
- Redes Bayseanas
- Redes Neurais

Esses classificadores foram escolhidos por serem citados na comunidade científica. Os classificadores apresentados são promissores para a substituição dos classificadores implementados na arquitetura atual de diferenciação de serviços.

E. Definição da População

A população usada neste trabalho é o tráfego IP de Internet em redes de *backbone*. As amostras dessa população são *traces* desse tráfego, que são coletadas, geralmente, por administradores de ISPs.

Para classificar uma determinada aplicação um pacote não basta, é necessário o agrupamento de pacotes em fluxos, definidos por uma tupla de cinco elementos: endereço IP da fonte, endereço IP do destino, número da porta da fonte, número da porta do destino e protocolo da camada de transporte.

Classificadores que usam aprendizagem de máquina identificam aplicações baseando-se em características estatísticas dos fluxos. Para este trabalho serão computadas 37 características de fluxos unidirecionais: protocolo, portas fonte e destino, número de pacotes, bytes transferidos, número de pacotes sem carga da camada 4, tempo de ínicio, tempo de fim, duração, vazão de pacotes e vazão de bytes média, max/min/média/desvio padrão do tamanho de pacotes e tempo inter-chegada, número de pacotes TCP com FIN, SYN, RSTS,PUSH,ACK,URG (*Urgent*), CWE (*Congestion Window Reduced*), e ECE (*Explicit Congestion Notification Echo*) ajustados todos em zero para pacotes UDP, e o tamanho dos primeiros dez pacotes.

A técnica comumente usada para validação é a classificação por *payload*, que consiste em encontrar uma *tag* hexadecimal que identifica as aplicações. Entretanto, existe um problema para a validação, pois as amostras que são fornecidas por grandes centros de pesquisa, geralmente, são sem a *payload* ou com a *payload* criptografada, fato que tem chamado a atenção para pesquisas sobre classificação de tráfego *in the dark*. Como o intuito deste trabalho é a avaliação inicial de desempenho dos classificadores, a técnica de classificação por porta será selecionada como ferramenta para validação da classificação, ela não é a ideal mas classifica corretamente 70% das aplicações [6].

Cinco amostras de tráfego IP de Internet, coletadas em nós de borda de redes de backbone, foram obtidas em [15] e [16]. Elas apresentam variedade temporal e geográfica, sendo pertencente à um período de 2002 a 2011 e localizadas no Estados Unidos e Japão. As amostras são: Link OC48 CAIDA EUA nos anos 2002 e 2003, Rede Backbone CAIDA Chicago-EUA em 2010, Rede Backbone São Francisco-EUA CAIDA em 2011 e Rede BackBone WIDE-JP (M) em 2011.

A classificação das aplicações é feita identificando o tráfego por fluxo IP em 11 categorias. A Tabela I apresenta todas as aplicações e suas categorias pela técnica de classificação baseada em porta.

TABLE I
CATEGORIAS DAS APLICAÇÕES

Categoria	Aplicação/Protocolo
WEB	HTTP, HTTP
P2P	FastTrack, eDonkey, BitTorrent, Ares,Direct Connect, Gnutella, WinMX, OpenNap, MP2P, SoulSeek, FileBEE, GoBoogy, Soribada, PeerEnabler, Napster Blubster, FileGuri, FilePia IMESH, ROMNET, HotLine, Waste
FTP	FTP
DNS	DNS
Mail/News	BIFF, SMTP, POP, IMAP, IDENTD, NNTP
Streaming	MMS(WMP), Real, Quicktime, Shoutcast, Vbrick Strmg, Logitech Video IM,Backbone Radio, PointCast, ABACast
NetOp	Netbios, SMB, SNMP, NTP, SpamAssasin, GoToMyPc, RIP,ICMP, BGP, Bootp, Traceroute
Encryption	SSH, SSL, Kerberos, IPSec, ISAKMP
Games	Quake, HalfLife, Age of Empires, DOOM, WOW, Star Sieze, Everquest, Startcraft, Asherons,
	Battle Feld Vietnam, HALO
Chat	AIM, IRC, MSN Messenger, Yahoo messenger, IChat, QNext, MS Netmeet, PGPFone, TALK.
Unknown	-

A Figura 2 apresenta todos os fluxos em quantidade mapeando suas respectivas categorias de aplicações por cores, sendo o *ground-truth* das amostras usadas no artigo. O intuito dessa figura é apresentar a porcentagem das aplicações em cada amostra em prol de conclusões mais justas, baseando no número disponível de fluxos em cada aplicação para classificação.

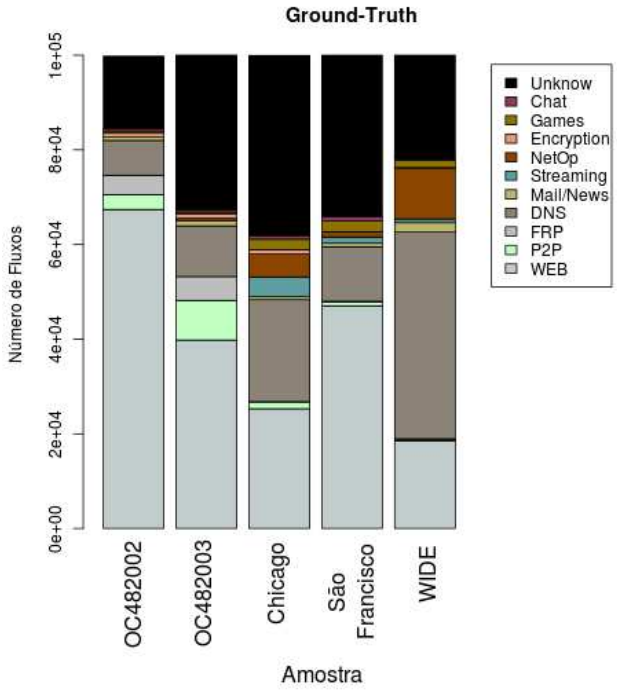

Fig. 2. Ground Truth mapeando número de fluxos, amostra e aplicação.

F. Design do Experimento

De acordo com os fatores, as unidades experimentais e o número de amostras supracitadas, é escolhido o *design* fatorial completo. Em relação aos níveis dos fatores, o tamanho dos dados de treinamento será de 1000 e de 10000 fluxos e os classificadores serão: *Naive* Bayes, Redes Bayseanas, Árvores de Decisão e Redes Neurais. As replicações dos tratamentos e a randomização serão caracterizadas pelo número de amostras usadas da população. O número de amostras escolhidas é cinco, com 10000 fluxos de dados por amostra. Com isso o número total de experimentos será 40.

Os experimentos são constituídos por três blocos. O primeiro bloco faz a captura, análise e pré-processamento das populações. O segundo bloco realiza as inferências sobre todas as amostras nas unidades experimentais seguindo os tratamentos. Finalmente, no terceiro bloco, os resultados obtidos com a medição das variáveis dependentes serão estatisticamente analisados para que conclusões sejam obtidas.

G. Instrumentação

Para a realização dos experimentos das unidades experimentais será usado o software WEKA. Essa ferramenta foi escolhida por possuir todos os classificadores implementados. O *benchmarking* NetraMark [17] é a ferramenta utilizada nos experimentos para criação de arquivos de entrada na ferramenta WEKA (.arff) e a validação da classificação usando o método baseado em porta por CoralReef [18]. As análises estatísticas serão realizadas com o auxílio da ferramenta R.

As configurações das unidades experimentais, mostradas nas Tabelas II, III, IV e V foram utilizadas no WEKA:

TABLE II
CONFIGURAÇÃO DE ÁRVORES DE DECISÃO

Parâmetro	Valor
BinarySplits	False
minObjNUmber	2
numFolds	3
reducedErrNum	False
saveInstaceData	False
subtreeRaising	True
unpruned	False
useLaplace	False

TABLE III
CONFIGURAÇÃO DE REDES NEURAIS

Parâmetro	Valor
decay	False
hiddenlayers	a
learningRate	0.3
momentum	0.2
nominaltoBFilter	TRUE
normalizesAtt	TRUE
normalizesNClass	TRUE
trainingTime	100
validationThers	20

TABLE IV
CONFIGURAÇÃO DE NAIVE BAYES

Parâmetro	Valor
displayModel InOldFormat	FALSE
useKEstimation	FALSE
useSupevised Discretation	FALSE

TABLE V
CONFIGURAÇÃO DE REDES BAYSEANAS

Parâmetro	Valor
BiffFIle	None
estimator	SE -A 0.5
SearchAlg	K2 -P 1 -S BAYES
useADTree	FALSE

H. Execução dos Experimentos

A execução dos experimentos pode ser detalhada com o auxílio da Figura 3. Primeiramente, as populações serão pré-processadas, para formatação dos dados de entrada para a ferramenta WEKA. Então, cada amostra será inferenciada por cada unidade experimental em cada nível do fator tamanho de treinamento, seguindo a *design* fatorial completo.

Em seguida, ocorre a medição das variáveis dependentes, que será feita pela ferramenta. Depois, uma análise estatística será efetuada para as avaliações das primeiras hipóteses por meio de um teste estatístico. Caso a primeira hipótese nula seja refutada, a avaliação das hipóteses H2 será feita por meio de intervalos de confiança. Depois que a avaliação das hipóteses estiver terminada, as conclusões poderão ser definidas.

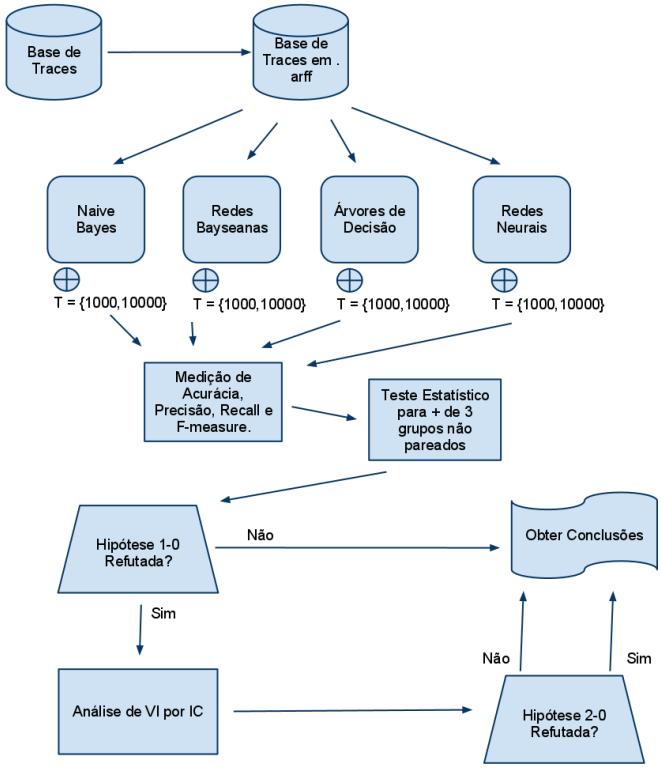

Fig. 3. Diagrama do procedimento de execução dos experimentos.

I. Ameaças à Validade

As validades de conclusão e interna aparentemente são as mais fortes, pois, existe uma relação entre fator e efeito, sendo que ela é considerada causal. Essa relação já foi confirmada com a revisão bibliográfica (dados em [6] e [10]). Porém outros fatores particulares a cada classificador, que os configuram, podem invalidar os resultados, fazendo com que as conclusões sobre o desempenho sejam invalidadas.

Em relação à validade de construção, pode-se considerar os resultados do mundo teórico para o mundo prático, pois, as restrições impostas pelos roteadores são baseadas nos equipamentos do mundo real. Entretanto, não se pode generalizar os resultados para um escopo fora do trabalho, pois, é necessário um modelo de simulação da arquitetura DiffServ e a avaliação dela com as unidades experimentais estudadas. Só desse modo, o efeito real dos classificadores na arquitetura poderá ser estudado.

V. RESULTADOS

A. Resultados envolvendo a hipótese H1-0

Como todas as variáveis de resposta do experimento são médias faz-se o uso do teorema central do limite para justificar a normalidade dos dados analisados, pois como o teorema afirma, as médias tendem a pertencer cada vez mais a uma distribuição normal quando o número de médias das amostras cresce, sendo o número cinco de médias das amostras para este tipo de problema, e como os resultados mostram, suficiente para ter significância estatística. Para avaliar se essa diferença de desempenho é significantivo entre os classificadores, utiliza-se a análise de variância para dados normais com mais de três grupos, chamado ANOVA. O α escolhido é 0,05 para o teste.

TABLE VI
CATEGORIAS DAS APLICAÇÕES

Métrica	P-valor
Acurácia	$< 2.2e^{-16}$
Precisão	$= 3.21e^{-12}$
Cobertura	$< 2.2e^{-16}$
F-measure	$< 2.2e^{-16}$

Na Tabela VI são apresentados todos os P-valores referentes às métricas relacionadas ao desempenho dos classificadores. Nota-se que todos os P-valores têm valores abaixo do α do teste estatístico escolhido, tendo a hipótese H1-0 como rejeitada. Pode-se concluir que existe diferença estatística significativa entre os classificadores e eles apresentam diferença de desempenho. A partir da rejeição da hipótese H1-0 os resultados envolvendo as hipóteses H2 podem ser calculados.

B. Resultados envolvendo a hipótese

A análise em questão foi dividida em três partes: acurácia, qualidade na classificação e análise de complexidade. O objetivo dessas análises é verificar a quantidade de acertos dos classificadores (acurácia), a qualidade desse acerto (qualidade na classificação) e se a complexidade irá afetar o desempenho da arquitetura DiffServ (análise de complexidade). O valor de α para todos os gráficos precedentes possui o valor de 0,05. A seguir, nas subseções, são apresentados os resultados para cada análise separada.

1) Acurácia: As Figuras 4 e 5 apresentam os resultados para acurácia dos classificadores. O tamanho de 1000 fluxos de treinamento é suficiente para que a técnica AD apresente aproximadamente 90% de acurácia média para todas as amostradas usadas na Figura 4. O mesmo desempenho não foi obtido com os outros classificadores. Vale destacar o baixo desempenho do classificador NB que alcançou em média 30% de acurácia.

Para 10000 fluxos de treinamento, na Figura 5, o mesmo comportamento é observado em relação à 1000 fluxos. O classificador AD apresenta, novamente, o maior valor de acurácia em relação aos classificadores estudados. O NB continua com desempenho baixo, apresentando valores de acurácia inferiores a 20%.

Observa-se que as técnicas apresentaram um aumento relativo de acurácia, isso porque o número de dados de treinamento

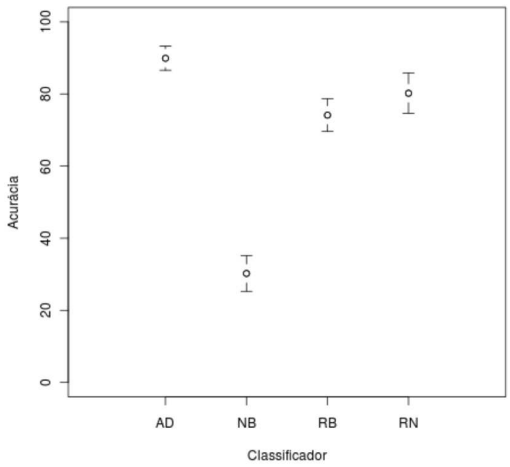

Fig. 4. Acurácia dos classificadores para 1000 fluxos de treinamento. Árvore de Decisão (AD), Naive Bayes (NB), Redes Bayseanas (RB) e Redes Neurais (RN).

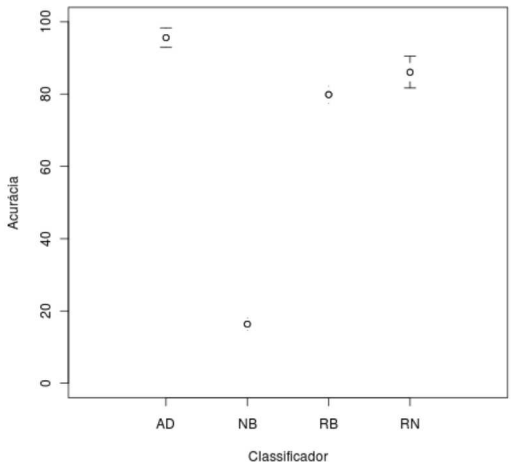

Fig. 5. Acurácia dos classificadores para 10000 fluxos de treinamento. Árvore de Decisão (AD), Naive Bayes (NB), Redes Bayseanas (RB) e Redes Neurais (RN).

aumentou, contemplando novas e diferentes aplicações para treinamento, permitindo que elas sejam identificadas na fase de teste. Lembrando que os dados de treinamento foram obtidos aleatoriamente a partir das amostras do tráfego de Internet que possuem 100000 fluxos. A técnica NB apresentou uma queda de desempenho na acurácia mesmo com o aumento do número de dados de treinamento. Pode-se concluir, então, que a técnica não apresenta resultados que satisfaçam o problema, pois a negociação de recursos e atribuição de valores de QoS pelo DiffServ necessitam de alta precisão na identificação dos fluxos e suas aplicações.

2) Qualidade na classificação: A qualidade na classificação está ligada às métricas: Precisão, Cobertura e *F-measure*. Primeiramente verifica-se a qualidade total dos classificadores. Como se pode observar na Figura 6, o desempenho

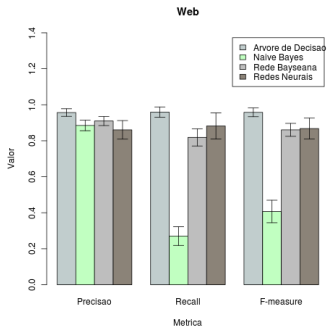

Fig. 7. Precisão, Cobertura e *F-measure* total para todos os classificadores sobre todas as amostras para WEB.

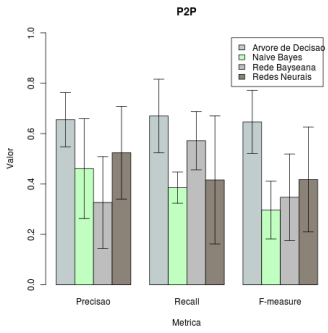

Fig. 8. Precisão, Cobertura e *F-measure* total para todos os classificadores sobre todas as amostras para P2P.

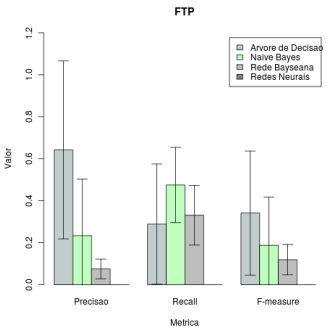

Fig. 9. Precisão, Cobertura e *F-measure* total para todos os classificadores sobre todas as amostras para FTP.

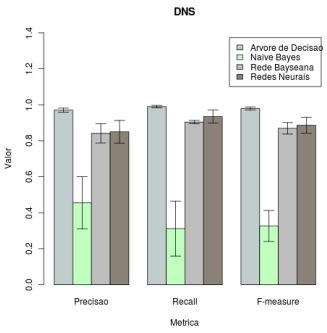

Fig. 10. Precisão, Cobertura e *F-measure* total para todos os classificadores sobre todas as amostras para DNS.

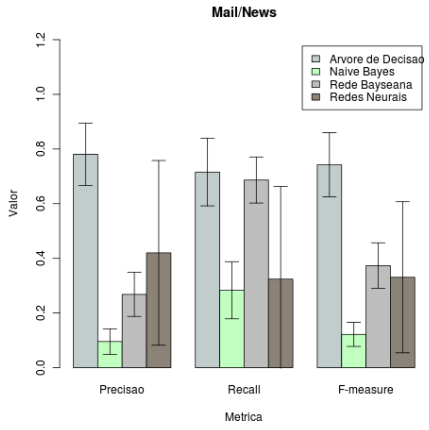

Fig. 11. Precisão, Cobertura e *F-measure* total para todos os classificadores sobre todas as amostras para Mail/News.

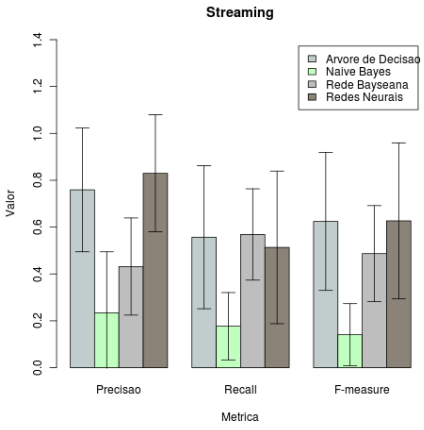

Fig. 12. Precisão, Cobertura e *F-measure* total para todos os classificadores sobre todas as amostras para Streaming.

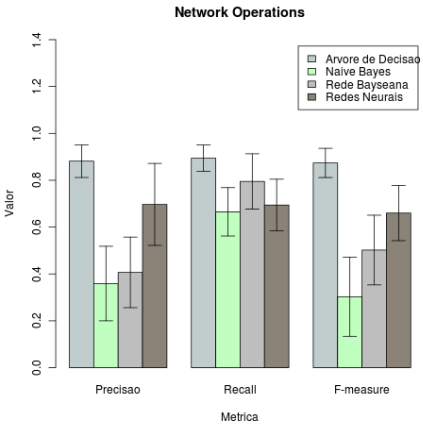

Fig. 13. Precisão, Cobertura e *F-measure* total para todos os classificadores sobre todas as amostras para Network Operations.

sobre todas as métricas, dos classificadores AD e RB são os mesmos, sendo que um tem maior desempenho e o outro menor, respectivamente.

Com o objetivo de investigar quais são as aplicações que prejudicam o desempenho na qualidade da classificação é apresentada uma análise per-aplicação, em que, se observam os resultados para todas as categorias de aplicações separadamente. As Figuras 7-16 apresentam os gráficos dos resultados.

As Figuras 7 e 10, respectivamente para WEB e DNS, apresentam resultados parecidos com os obtidos na análise total. Esses gráficos apresentam baixa variabilidade dos dados, pois o número de fluxos de WEB e DNS são os maiores para todas as amostras. Esse resultado é diferente para todos os outros gráficos, em que a variação dos valores das variáveis de resposta aumenta para todas as outras categorias. Isso acontece pelo fato de o número de fluxos dessas categorias não ser suficiente como em aplicações WEB e DNS. Os intervalos de confiança de todos os classificadores conseguem ultrapassar (em casos como P2P, Streaming, Chat e FTP) as médias das outras métricas. Para conseguir uma avaliação per-aplicação mais justa, nesses casos é necessário que os números de amostras aumentem, ou que amostras com um número maior das outras categorias de aplicações sejam encontradas, ou que uma base de dados sintética e justa seja criada.

A queda no desempenho de todos os classificadores ocorre em categorias das aplicações que apresentam características de fluxo muito variáveis, como por exemplo, aplicações P2P, Streaming, Chat e FTP. Essas aplicações apresentam números de fluxos mais próximos, mesmo assim a diferença no desempenho é notória. Isso pode ser causada pelo baixo número de fluxos dessas categorias nas bases de dados de treinamento, por que o classificador pode não estar pronto para a classificação. É importante notar também que essas categorias de aplicações possuem atribuição aleatória de portas para comunicação, podendo ser a causa da variabilidade dos resultados, isso porque é mais complicado para o classificador verificar esses campos com um rigor maior.

Mesmo com o problema mencionado sobre o tamanho das amostras ou número de fluxos, é possível observar que essa variação em questão é maior para a técnica de Redes Neurais e menor para Árvore de Decisão, em um âmbito geral. Isso pode ser explicado pelo mecanismo de desenvolvimento de modelos para classificação, como posteriormente será melhor retratado, Redes Neurais precisam de um tempo maior para treinamento, e o experimento pode não ter sido o suficiente. Então, para aplicações com características mais previsíveis de fluxo como:

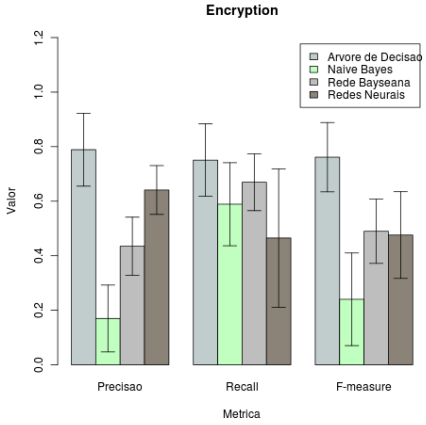

Fig. 14. Precisão, Cobertura e *F-measure* total para todos os classificadores sobre todas as amostras para Encryption.

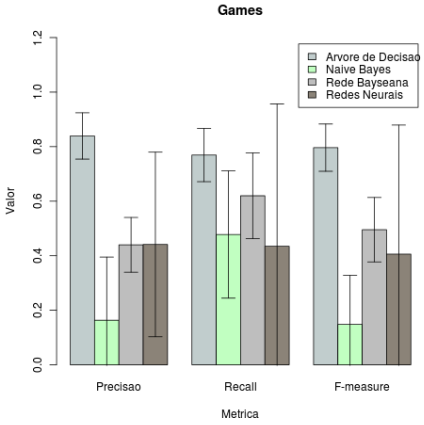

Fig. 15. Precisão, Cobertura e *F-measure* total para todos os classificadores sobre todas as amostras para Games.

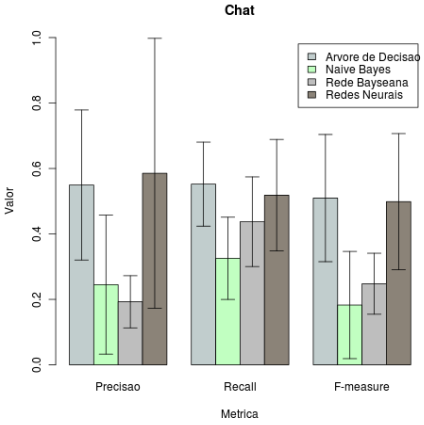

Fig. 16. Precisão, Cobertura e *F-measure* total para todos os classificadores sobre todas as amostras para Chat.

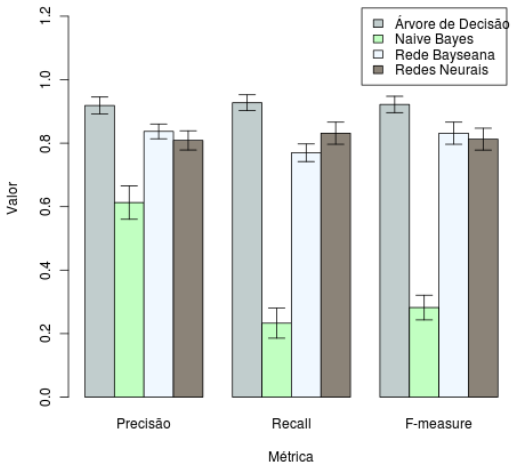

Fig. 6. Precisão, Cobertura e F-measure total para todos os classificadores sobre todas as amostras.

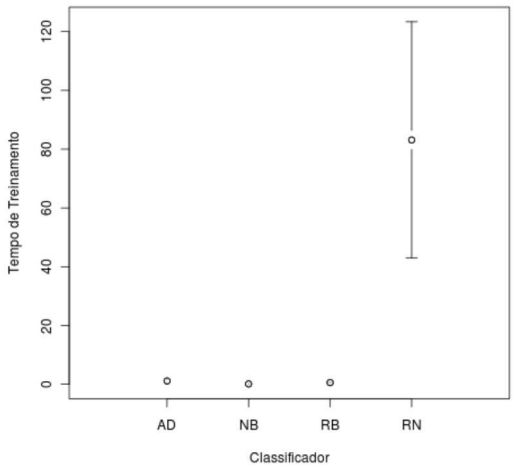

Fig. 17. Tempo de Treinamento médio em segundos dos classificadores: Árvore de Decisão (AD), *Naive* Bayes (NB), Redes Bayseanas (RB) e Redes Neurais (RN).

Games, *Network Operations*, *Encryption* e *Mail/News*, pode-se observar um desempenho superior em todas as métricas do classificador Árvores de Decisão.

3) Análise de Complexidade: A análise de complexidade é o ponto final para que se possa aceitar ou rejeitar a hipótese H2-0. Cada classificador possui um tempo para desenvolver um modelo na fase de treinamento, que é responsável pela classificação real dos fluxos de tráfego IP, representado pela análise de tempo de treinamento. O tempo necessário para que um classificador efetue seu trabalho é inferido neste trabalho ao tempo de classificação por fluxo.

A Figura 17 apresenta os resultados do tempo de treinamento médio para os classificadores estudados. O classificador RN apresenta um largo intervalo de tempo com variação para o seu treinamento. Esse desempenho do RN é inviável para o DiffServ, visto que em roteadores de borda de redes de backbone que utilizarem um modelo de classificação *online*, no qual esses classificadores podem atuar, passam milhares de fluxos por segundo, e o DiffServ não deve ter um tempo de *back-off* como esse, até que o classificador esteja pronto.

O tempo de classificação é muito importante, porque se pode aumentar o tempo de processamento de pacotes e degradar o desempenho da aplicação em relação ao atraso fim-a-fim. Para realizar essa análise baseia-se nos fluxos que possuem menor atraso possível, que são aplicações de vídeo-conferência, categorizadas neste trabalho como *Streaming*. Em [19], os autores apresentam um valor máximo de $100ms$ de atraso fim a fim para esse tipo de aplicação, então, pode-se inferir que os classificadores devem identificar o fluxo em um tempo bem menor que o valor mencionado, pois deve-se ainda contabilizar valores de atraso de propagação e *over-head* de roteamento multiplicado pelo número de saltos entre cliente e servidor. A Figura 18 apresenta os valores de tempo de classificação por fluxo para cada classificador avaliado.

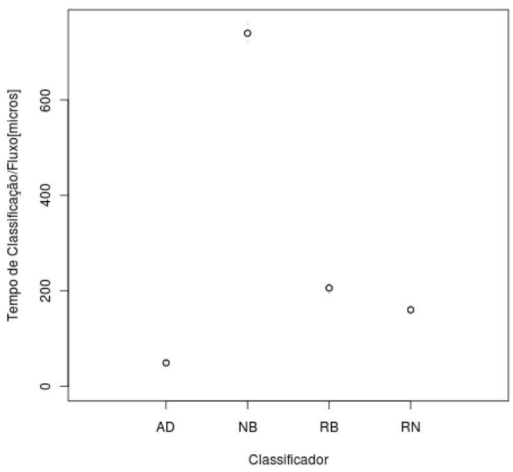

Fig. 18. Tempo de Classificação/Fluxo médio em microsegundos dos classificadores: Árvore de Decisão (AD), *Naive* Bayes (NB), Redes Bayseanas (RB) e Redes Neurais (RN).

O gráfico dessa figura mostra que todos os classificadores apresentam valores bem inferiores a 100ms, estando em ordem de microsegundos e aproximadamente 1000 vezes menor que o valor estabelecido pelas aplicações não-eslásticas, como *Streaming*.

Toda a análise de resultados indica que os classificadores adequados ao DiffServ são Árvore de Decisão e Rede Bayseana. A Árvore de Decisão apresenta um maior desempenho em relação a todas as variáveis dependentes envolvidas nos experimentos. Com isso a hipótese H2-0 é rejeitada, pois o desempenho da Árvore de Decisão é estritamente viável ao DiffServ. O desempenho da Rede Bayseana também é satisfatório, entretanto é inferior à Árvore de Decisão com relação à Acurácia, Precisão, Cobertura e *F-measure*.

VI. Conclusões e Trabalhos Futuros

Neste artigo é apresentada uma avaliação experimental com o *design* fatorial completo que objetiva indicar o classificador de tráfego mais adequado para o modelo de QoS DiffServ. Foram tratados o tamanho dos dados de treinamento (1000, 10000) e os classificadores (*Naive* Bayes, Rede Bayseana, Árvore de Decisão e Rede Neural) como variáveis independentes. As seguintes variáveis dependentes foram escolhidas: Acurácia, Precisão, Cobertura e *F-measure*. Foram escolhidas cinco amostras de tráfego IP com diversidade geográfica e de aplicações para uma análise de variância dos resultados com o teste ANOVA e uma análise gráfica via intervalos de confiança.

Foi concluído que existe um efeito entre as variáveis dependentes e independentes. O número de cinco amostras é suficiente para a análise total dos resultados (e também para WEB, DNS, Mail/News, Games, NetOp e Encryption), pois a variabilidade dos valores é relativamente baixa, entretanto, para certo tipo de aplicações (i.e., P2P, FTP, *Streaming* e *Chat*) esse número não é suficiente. A Rede Neural não é adequada ao problema, pois precisa de tempo de treinamento longo. Porém todos os classificadores, depois que o modelo de classificação está pronto, estão aptos para atuarem no modelo DiffServ em se tratando do tempo de processamento de fluxos para a classificação do tráfego. Pode-se concluir que a Árvore de Decisão pode aumentar o desempenho da arquitetura DiffServ. Esses resultados podem ser invalidados por causa do pequeno número de amostras, da técnica de validação escolhida ser a baseada em porta e a avaliação ser *off-line*.

O número de amostras, como já citado, pode ser um ponto a favor da invalidação dos resultados, para que seja realizados novos experimentos com maior confiabilidade estatística para todas as categorias de aplicações deve-se fazer um estudo de validação com classificação de tráfego *in the dark*. A avaliação de fatores internos dos classificadores pode aumentar o desempenho da Rede Bayseana, e pode ser um ponto a ser melhor investigado. E por último, uma implementação dos classificadores em um modelo DiffServ é desejado para medir o efeito do novo sistema.

Referências

[1] S. Richards, *FutureNet: The Past, Present, and Future of the Internet as Told by Its Creators and Visionaries*, Wiley, Ed. Wiley, 2002.
[2] J. Zittrain, *The Future of the Internet–And How to Stop It*, Caravan, Ed. Caravan, 2008.
[3] T. Tronco, *New Network Architectures: The path to the future internet*, Springer, Ed. Springer, 2010.
[4] A. Meddeb, "Internet qos: pieces of the puzzle," *Comm. Mag.*, vol. 48, pp. 86–94, January 2010. [Online]. Available: http://portal.acm.org/citation.cfm?id=1820984.1821000
[5] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, and W. Weiss, "An Architecture for Differentiated Service," RFC 2475 (Informational), Internet Engineering Task Force, December 1998, updated by RFC 3260. [Online]. Available: http://www.ietf.org/rfc/rfc2475.txt
[6] H. Kim, D. Barman, M. Faloutsos, M. Fomenkov, and K. Lee, "Internet traffic classification demystified: The myths, caveats and best practices," in *In Proc. ACM CoNEXT*, 2008.
[7] A. Callado, C. Kamienski, G. Szabo, B. Gero, J. Kelner, S. Fernandes, and D. Sadok, "A survey on internet traffic identification," *Communications Surveys Tutorials, IEEE*, vol. 11, no. 3, pp. 37 –52, quarter 2009.
[8] L. Jun, Z. Shunyi, L. Yanqing, and Z. Zailong, "Internet traffic classification using machine learning," in *Communications and Networking in China, 2007. CHINACOM '07. Second International Conference on*, aug. 2007, pp. 239 –243.
[9] T. Nguyen and G. Armitage, "A survey of techniques for internet traffic classification using machine learning," *Communications Surveys Tutorials, IEEE*, vol. 10, no. 4, pp. 56 –76, quarter 2008.
[10] Y.-s. Lim, H.-c. Kim, J. Jeong, C.-k. Kim, T. T. Kwon, and Y. Choi, "Internet traffic classification demystified: on the sources of the discriminative power," in *Proceedings of the 6th International COnference*, ser. Co-NEXT '10. New York, NY, USA: ACM, 2010, pp. 9:1–9:12. [Online]. Available: http://doi.acm.org/10.1145/1921168.1921180
[11] C. Filsfils and J. Evans, "Deploying diffserv in backbone networks for tight sla control," *IEEE Internet Computing*, vol. 9, pp. 58–65, January 2005. [Online]. Available: http://portal.acm.org/citation.cfm?id=1053547.1053594
[12] T. Onali and L. Atzori, "Traffic classification and bandwidth management in diffserv-aware traffic engineering architectures," in *ICC*, 2008, pp. 70–74.
[13] J. Park, H.-R. Tyan, and C.-C. Kuo, "Internet traffic classification for scalable qos provision," in *Multimedia and Expo, 2006 IEEE International Conference on*, july 2006, pp. 1221 –1224.
[14] J. Park, H.-R. Tyan, and C.-C. J. Kuo, "Ga-based internet traffic classification technique for qos provisioning," in *Proceedings of the 2006 International Conference on Intelligent Information Hiding and Multimedia*, ser. IIH-MSP '06. Washington, DC, USA: IEEE Computer Society, 2006, pp. 251–254. [Online]. Available: http://dx.doi.org/10.1109/IIH-MSP.2006.103
[15] CAIDA, http://www.caida.org/home/. Acessado em Agosto 2011, 2011.
[16] WIDE, http://mawi.wide.ad.jp/mawi/. Acessado em Agosto 2011, 2011.

[17] S. Lee, H. Kim, D. Barman, S. Lee, C.-k. Kim, T. Kwon, and Y. Choi, "Netramark: a network traffic classification benchmark," *SIGCOMM Comput. Commun. Rev.*, vol. 41, pp. 22–30. [Online]. Available: http://doi.acm.org/10.1145/1925861.1925865

[18] C. R. (CAIDA), http://www.caida.org/tools/measurement/coralreef/, 2011.

[19] M. Baldi and Y. Ofek, "End-to-end delay analysis of videoconferencing over packet switched networks," *IEEE/ACM Transactions on Networking*, vol. 8, pp. 479–492, 1998.