FUNDAÇÃO EDSON QUEIROZ
UNIVERSIDADE DE FORTALEZA
CENTRO DE CIÊNCIAS TECNOLÓGICAS
MESTRADO EM INFORMÁTICA APLICADA

MARCUS FÁBIO FONTENELLE DO CARMO

CLASTRIN – UM CLASSIFICADOR DE TRÁFEGO DE APLICAÇÕES INTERNET UTILIZANDO A ABORDAGEM "UM-CONTRA-TODOS"

Fortaleza
Agosto de 2009

FUNDAÇÃO EDSON QUEIROZ
UNIVERSIDADE DE FORTALEZA
CENTRO DE CIÊNCIAS TECNOLÓGICAS
MESTRADO EM INFORMÁTICA APLICADA

MARCUS FÁBIO FONTENELLE DO CARMO

CLASTRIN – UM CLASSIFICADOR DE TRÁFEGO DE APLICAÇÕES INTERNET UTILIZANDO A ABORDAGEM "UM-CONTRA-TODOS"

Dissertação apresentada ao Curso de Mestrado em Informática Aplicada da Universidade de Fortaleza como requisito parcial para obtenção do Título de Mestre em Informática Aplicada.

Orientador: Prof. Dr. Raimir Holanda Filho

Fortaleza
Agosto de 2009

C287c Carmo, Marcus Fábio Fontenelle do.
CLASTRIN – Um Classificador de Tráfego de Aplicações Internet Utilizando a Abordagem "Um-Contra-Todos" / Marcus Fábio Fontenelle do Carmo. - 2009.
89 f.

Dissertação (mestrado) – Universidade de Fortaleza, 2009.
"Orientação: Prof. Dr. Raimir Holanda Filho."

1.Redes de computadores. 2. Internet. 3. Estatística multivariada.
I. Título.

CDU 681.3:621.391

MARCUS FÁBIO FONTENELLE DO CARMO

CLASTRIN – UM CLASSIFICADOR DE TRÁFEGO DE APLICAÇÕES INTERNET UTILIZANDO A ABORDAGEM "UM-CONTRA-TODOS"

Data de Aprovação: 28/08/2009

Banca Examinadora:

Prof. Raimir Holanda Filho, Doctor

(Orientador – UNIFOR)

Prof. André Luís Vasconcelos Coelho, D. Sc.

(Membro – UNIFOR)

Prof. Aldri Luiz dos Santos, D.Sc.

(Membro - UFPR)

Dedico este trabalho a minha avó Terezinha.

"Iniciar é graça divina. Perseverar é graça ainda maior, mas graça das graças é nunca desistir."
(Dom Helder Câmara)

"Não foste tu, Sancho, mas eu mesmo quem tentou tirar o máximo de mim.
É o melhor que um homem pode fazer na vida!"
(Miguel de Cervantes, em Dom Quixote)

AGRADECIMENTOS

A Deus por ter me permitido realizar mais este sonho.

A meus pais, Eudes e Fátima, pelo incentivo e apoio.

A minhas tias, Lilian e Miriam, que nunca mediram esforços para me dar acesso a livros, cultura e educação. Sem vocês nenhum de meus sonhos teriam se tornado realidade. Graças a vocês eu posso fazer o que mais gosto que é estudar e compartilhar meu conhecimento.

Aos meus irmãos Juliano, Fabíola e Vinícius que sempre me apoiaram e têm orgulho do que faço, cada um a sua maneira, mesmo sem entender muito bem o que exatamente eu faço.

A família "DO CARMO" por sempre festejar com muita alegria cada uma das minhas conquistas.

Aos meus amigos, Jorge, Telmo, Denise e Hudson, que muito mais que amigos são uma família sempre disposta a me dar apoio, além de me mostrar sempre o melhor caminho a ser seguido.

A Leila, que tem se mostrado dia após dia a companheira ideal.

Aos nobres amigos baianos, em especial, Fábio, Áurea, Jair, Marcelo, Eduardo, Carlinhos, Ramai, Myllene e Mallu que souberam aliviar meu stress e me deram força e incentivo nos momentos mais difíceis desta jornada.

A Jane Prado, que no começo do mestrado teve uma participação importante em minha vida.

Aos meus companheiros de mestrado, em especial Gabriel, Geneflides e Leandro, pela amizade e ajuda.

Aos colegas do grupo de pesquisa por compartilharem seus comentários e sugestões para a melhoria constante deste trabalho.

Ao grande amigo Idelfonso por ter me mostrado com sua empolgação o quanto pode ser divertido estudar Redes de Computadores e fazer pesquisa.

Ao amigo Tales Benigno por seu apoio e torcida.

A todos os professores do MIA por terem em algum momento compartilhado seus conhecimentos, em especial, ao professor Porfírio. Nossas conversas informais tomando cafezinho sempre foram muito esclarecedoras.

Ao professor Plácido Pinheiro, por ter acreditado e investido em meu potencial desde a graduação. Sem seu apoio não teria sido possível o término deste trabalho. Tenha certeza que a jornada apenas começou.

Ao amigo e orientador, professor Raimir Holanda pela sua dedicação na condução deste trabalho. Sua objetividade e clareza nos esclarecimentos das questões científicas foram fundamentais para minha formação. Sempre saio de nossas conversas tendo a certeza de ter crescido como pessoa e profissional. Espero que nossa parceria apenas tenha se iniciado.

Ao amigo e co-orientador, professor Everardo Bessa. Nossas discussões, muitas vezes acaloradas, geraram idéias, críticas e sugestões que sempre trouxeram melhorias a este trabalho.

Aos professores Ricardo Colares e Wellington Brito por me permitirem usar o laboratório de redes convergentes até altas horas sem nunca me impor restrições.

Aos professores André Luís Vasconcelos e Aldri Santos por terem se disponibilizado a participar da minha banca de defesa. Suas sugestões enriqueceram sobremaneira este trabalho.

Aos funcionários do MIA e CCT, Tânia, Taciana, Carlos Eduardo, Celi, Jair, Neuman e Edmilson, por sempre estarem dispostos a ajudar. Sem vocês o mestrado em informática aplicada não funcionaria.

Aos meus alunos da FIC (Faculdade Integrada do Ceará) por me proporcionarem a oportunidade de compartilhar o que aprendi. Vocês serão sempre uma grande motivação para que eu continue a aprender.

A CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) por financiar integralmente meu mestrado e minhas pesquisas.

Resumo da dissertação apresentada ao Curso de Mestrado em Informática Aplicada (MIA) da Universidade de Fortaleza, como parte dos requisitos necessários para a obtenção do grau de Mestre em Informática Aplicada.

CLASTRIN – UM CLASSIFICADOR DE TRÁFEGO DE APLICAÇÕES INTERNET UTILIZANDO A ABORDAGEM "UM-CONTRA-TODOS"

Autor: Marcus Fábio Fontenelle do Carmo
Orientador: Raimir Holanda Filho, Doctor

Neste trabalho, apresenta-se uma proposta de um classificador de aplicações presentes no tráfego Internet. A proposta deste classificador é utilizar informações estatísticas coletadas dos fluxos de dados e identificar a menor quantidade de discriminantes estatísticos capazes de distinguir os fluxos de determinada classe de aplicação dos demais, separando-os em grupos.

Para a realização desta classificação a metodologia apresentada se baseia na divisão de um problema de classificação de 1 para N em N problemas de classificação 1 para 1 (abordagem um-contra-todos – *one-against all*). A geração dos conglomerados de dados é realizada através da análise de agrupamentos (método de estatística multivariada) utilizando-se de um método não-hierárquico (K-Médias – *K-Means*) em conjunto com técnicas de aprendizagem de máquina supervisionada.

A metodologia apresentada parte do princípio de que o melhor conjunto de variáveis para classificar uma determinada aplicação não é necessariamente o mesmo para classificar N aplicações. Comparativamente a outros métodos estudados, este trabalho inovou ao apresentar uma redução do número de variáveis (*features*) a serem analisadas através de um método estatístico computacionalmente simples, que pode ser utilizado em outros conjuntos de dados (*traces*). Foi observado que para a maioria das classes apenas uma única variável foi suficiente para discriminar o tráfego da aplicação, gerando uma taxa média de acerto na classificação dos fluxos das classes sob análise de 74,40%.

Palavras-chave: Redes de Computadores; Discriminantes Estatísticos; Classificação de Tráfego; Estatística Multivariada; Análise de Agrupamentos; Aprendizagem de Máquina.

Abstract of the dissertation presented to the board of faculties of the Master Program in Applied Informatics at the University of Fortaleza, as partial fulfillment of the requirements for the Master's degree in Applied Informatics.

.

CLASTRIN – A CLASSIFIER OF INTERNET TRAFFIC APPLICATIONS USING THE "ONE-AGAINST-ALL" APPROACH

Author: Marcus Fábio Fontenelle do Carmo
Advisor: Raimir Holanda Filho, Doctor

In this work, is proposed of a classifier for applications present the Internet traffic. The purpose of this classifier is to use statistical information collected from the data flow and identify the least amount of statistical discriminators able to distinguish the flow of a given application class from the others, separating them into groups.

To achieve this classification, the proposed method is based on the decomposition of 1-to-N classification problem into N problems of 1-to-1 classification (one against all approach). The generation of data clusters is performed by cluster analysis (a method of multivariate statistics) using a non-hierarchical method (K-Means) together with techniques of supervised machine learning.

The methodology presented claims that the best set of variables used to classify a given application is not the same of that to classify N applications. Compared to other methods, this work innovates in providing a reduction in the number of variables (features) to be analyzed by a computationally simple statistical method, which can be used in other data sets (traces). It was observed that for most classes only one variable was sufficient to discriminate the traffic of the application, yielding an average rate of 74,40% of correct classification of flows in the class under analysis.

Keywords: Computer Networks; Statistical Discriminators; Traffic Classification; Multivariate Statistics; Cluster Analysis; Machine Learning.

Sumário

Lista de Figuras

Lista de Tabelas

Acrônimos

3WS	Three Way Handshake
ACK	Acknowledgement (Mensagem de Reconhecimento)
DNS	Domain Name System
FIN	Finalize (Bit FIN do cabeçalho TCP)
FTP	File Transfer Protocol
HTTP	Hypertext Transfer Protocol
IAT	Inter Arrival Time
IMAP	Internet Message Access Protocol
IP	Internet Protocol / Protocolo Internet
IRC	Internet Relay Chat
LAN	Local Area Network
LDAP	Lightweight Directory Access Protocol
MIB	Management Information Base
ML	Machine Learning
MSS	Maximum Segment Size
NTP	Network Time Protocol
P2P	Peer-to-Peer
PCA	Principal Component Analysis
POP3	Post Office Protocol 3
PSH	Push (Bit PSH do cabeçalho TCP)
QoS	Quality of Service / Qualidade de Serviço
RST	Reset (Bit RST do cabeçalho TCP)
RTT	Round Trip Time
SACK	Seletive Acknowledgement
SMTP	Simple Mail Transfer Protocol
SNMP	Simple Network Management Protocol
SYN	Synchronize (Bit SYN do cabeçalho TCP)
TCP	Transmission Control Protocol
TSL	Theoretical Stream Lenght
UDP	User Datagram Protocol
URG	Urgent (bit URG do cabeçalho TCP)

WAN	Wide Area Network
WWW	World Wide Web

Capítulo 1 – Introdução

Este capítulo aborda as questões que motivaram a realização do presente trabalho, como também seus objetivos e sua organização.

A evolução da Internet durante os últimos quarenta anos tem sido acompanhada pelo surgimento, crescimento e uso de uma variedade de aplicações. Estas aplicações evoluíram de *softwares* bem rústicos para enviar mensagens de correio eletrônico, realizar transferência de arquivos ou *login* remoto para *softwares* mais elaborados e complexos com o intuito de realizar transações de comércio eletrônico, videoconferência e acesso a conteúdo multimídia via WWW (*World Wide Web*).

Muitos esforços têm sido dedicados para gerenciar redes de computadores na tentativa de identificar e caracterizar o tráfego destas aplicações (CLAFFY, 1994), (PAXSON, 1994), (THOMPSON, et al., 1997), (FRALEIGH, et al., 2003), (MOORE, et al., 2003), (FOMENKOV, et al., 2004) e (KIM, et al., 2008). Nos últimos anos, uma nova tendência para a modelagem de tráfego surgiu baseada em métodos capazes de examinar propriedades estatísticas individuais dos fluxos (MARCHETTE, 1999), (PORTNOY, et al., 2001), (TAYLOR e ALVES-FOSS, 2002), (LAKHINA, et al., 2004c), (HOLANDA, 2005), (MOORE, et al., 2005a), (MOORE e ZUEV, 2005b), (ZUEV e MOORE, 2005), (ERMAN, et al., 2006a), (HOLANDA, et al., 2006), (AULD, et al., 2007) e (SIQUEIRA, 2008). Tais métodos têm mostrado resultados bem expressivos ao ser aplicados a uma grande variedade de aplicações Internet, além de terem demonstrado não serem afetados pelos problemas apresentados por outros métodos de classificação, tais como a utilização de portas de serviço ou a análise do conteúdo dos pacotes.

O classificador de aplicações Internet proposto neste trabalho utiliza-se de técnicas de aprendizagem de máquina supervisionada (Seção 3.1) para, de forma *offline*, analisar as propriedades estatísticas dos cabeçalhos dos pacotes tendo como base a abordagem "um-contra-todos" (*one-against-all*) (BEYGELZIMER, et al., 2005) que até então não havia sido utilizada para resolução de problemas de classificação de fluxos de tráfego da Internet.

Esta abordagem, em relação às demais abordagens estudadas, mostra que a divisão de um problema de classificação de 1 para N em N problemas de classificação 1 para 1, ou seja, um problema de classificação de N classes de objetos em vários problemas binários de classificação, resulta em uma complexidade computacional menor com igual ou maior poder de discriminação.

1.1. Estado da Arte

A expressão "classificação de tráfego" é usada para descrever métodos de classificação baseados nas características observadas em um determinado tráfego de acordo com objetivos específicos (CAIDA, 2009). Grandes esforços têm sido dedicados em pesquisas relacionadas à classificação de aplicações Internet. Este fato decorre do aumento considerável com que as atividades de ordem pessoal, empresarial e governamental dependem das redes de computadores. Williamson (WILLIAMSON, 2001) apresenta um breve tutorial sobre a medição de tráfego na Internet explicando as diversas formas de classificar esta medição. Em (KIM, et al., 2007) e (SALGARELLI, et al., 2007) são analisados os principais métodos e ferramentas de classificação de tráfego existentes atualmente.

O método mais tradicional para identificar e classificar o tráfego de aplicações Internet é através dos números das portas de serviços (KEYS, et al., 2001), (MOORE, et al., 2001), (SEN e WANG, 2002), (JUNG, et al., 2004) e (SCHECHTER, et al., 2004). Logo se percebeu que a identificação baseada em portas era ineficaz (SAROIU, et al., 2002), (KARAGIANNIS, et al., 2004), (SEN, et al., 2004) e (MOORE e PAPAGIANNAKI, 2005c). Essa ineficácia se deve ao fato de que alguns tipos de aplicações, tais como, jogos, multimídia, P2P (*Peer-to-Peer*), fazem uso de números dinâmicos para associar as portas de seus serviços. Além disso, outros tipos de tráfego fazem uso compartilhado de números de portas que originalmente foram definidas para outros serviços, e.g., alguns tipos de ataque se utilizam, por exemplo, da porta 80 (oitenta) definida para o serviço Web (protocolo HTTP – *Hypertext Transfer Protocol*)[1].

Em outra abordagem a classificação das aplicações é realizada baseada no uso de assinaturas (KIM, et al., 2004), (SEN, et al., 2004) e (MOORE e PAPAGIANNAKI,

[1] Maiores informações a respeito dos números de portas associados aos principais serviços da Internet poderão ser encontradas em (IANA, 2009).

2005c). Neste caso, o conteúdo dos pacotes é aberto e os dados são lidos. Pode-se apontar pelo menos dois sérios problemas nesta abordagem, a saber: o primeiro está relacionado com a privacidade, uma vez que a análise não se restringe às informações contidas nos cabeçalhos dos pacotes. O segundo, por sua vez, é consequência da utilização cada vez mais frequente de dados criptografados, inviabilizando desta maneira a análise dos conteúdos. Na pesquisa encontrada em (KARAGIANNIS, et al., 2005) foi utilizada uma coleta de dados (*trace*) real e foi feita uma análise dos fluxos sem acesso ao conteúdo dos pacotes, sem o prévio conhecimento das portas dos serviços e sem nenhuma informação adicional além das fornecidas nos cabeçalhos dos pacotes.

Métodos de classificação baseados em anomalia têm mostrado um alto grau de eficiência devido ao seu baixo custo de operação na rede (TAYLOR e ALVES-FOSS, 2000). Usando um *trace* real, (TAYLOR e ALVES-FOSS, 2001) e (TAYLOR e ALVES-FOSS, 2002) apresentam uma análise de eventos anormais de tráfego. Uma abordagem comumente utilizada tem sido tratar anomalias como desvios de volume de tráfego (BRUTLAG, 2000), (BARFORD, et al., 2002), (HUSSAIN, et al., 2003), (ROUGHAN, et al., 2004) e (LAKHINA, et al., 2005).

Em (LAKHINA, et al., 2004a), são tratadas anomalias em redes de *backbone* analisando a quantidade de bytes através de um enlace, enquanto em (LAKHINA, et al., 2004b) e (LAKHINA, et al., 2004c) é analisado o volume de tráfego em fluxos[2] de Origem-Destino (OD). A abordagem de detecção de anomalias baseada em volume tem tido sucesso em identificar grandes mudanças no perfil do tráfego tal como em ataques conhecidos como *bandwidth flooding attacks*. Entretanto, existem várias classes de anomalias que não causam alterações significativas no volume de tráfego. Outras abordagens têm sido utilizadas baseadas na exploração de correlação de padrões entre diferentes variáveis da MIB[3] SNMP[4] (CABRERA, et al., 2002), (RHODEN, et al.,

[2] Um fluxo é identificado como uma sequência de pacotes que apresentam o mesmo conjunto de valores contidos nos seguintes campos dos cabeçalhos TCP e IP: endereço IP de origem, endereço IP de destino, porta TCP de origem, porta TCP de destino e tipo de protocolo.

[3] MIB (*Management Information Base*) é uma base de dados que procura abranger todas as informações necessárias para a gerência da rede.

2002) e (THOTTAN e JI, 2003), baseadas em heurísticas para identificar tipos específicos de anomalias em fluxos de pacotes IP (KIM, et al., 2004) ou baseadas na análise dos protocolos de uma determinada classe de aplicação a fim de se identificar padrões de comportamento específicos (SPOGNARDI, et al., 2005).

Nos últimos anos, estudos sobre algumas características do tráfego Internet, tais como tamanho do fluxo em bytes, tempo de vida do fluxo, direção do fluxo e estrutura do endereço IP, têm sido publicados. Em (GUO e MATTA, 2001), a distribuição do tamanho de um fluxo é estudada, introduzindo uma classificação dos fluxos baseada no número de bytes, e.g., *mice* (fluxo com pequena quantidade de bytes) e *elephant* (fluxo com grande quantidade de bytes). Já em (BROWNLEE e CLAFFY, 2002) os fluxos são classificados por tempo de vida, demonstrando que a maioria dos fluxos possuem um curto ciclo de vida (menos de dois segundos)[5]. Eddie Kohler, Jinyang Li, Vern Paxson e Scott Shenker (KOHLER, et al., 2006) investigam em seu artigo a estrutura dos endereços contidos em um tráfego IP. No artigo de (BERNAILLE, et al., 2006), é proposta uma técnica para a identificação de uma aplicação considerando-se a observação dos cinco primeiros pacotes de uma conexão TCP (*Transmission Control Protocol*). No trabalho de (ERMAN, et al., 2007) é proposto um arcabouço (*framework*) que classifica os fluxos fazendo uso apenas de informações estatísticas de uma das direções do tráfego ("servidor→cliente" ou "cliente→servidor"). Os autores constataram que as informações provenientes dos fluxos de uma conexão TCP no sentido "servidor→cliente" geram uma precisão maior na classificação do que as informações obtidas dos fluxos no sentido "cliente→servidor".

Uma metodologia que utiliza o conceito de aprendizagem de máquina (ML – *Machine Learning*) é utilizada em (MCGREGOR, et al., 2004) para agrupar fluxos em um conjunto de classes de aplicações: WWW, DNS (*Domain Name System*), IMAP (*Internet Message Access Protocol*), SMTP (*Simple Mail Transfer Protocol*) e BULK. Para realizar este agrupamento os autores fizeram uso de estatísticas da camada de transporte, tais como: tamanho do pacote, contadores de bytes, duração da conexão, número de transições e tempo de inatividade – *idle time*. Na pesquisa de (ERMAN, et

[4] SNMP (*Simple Network Management Protocol*) é um protocolo de gerência típico das redes TCP/IP que facilita o intercâmbio de informação entre um dispositivo chamado gerente e os demais dispositivos de uma rede.

[5] Brownlee classifica os fluxos em *dragonfly* (fluxo com curta duração de tempo) e *tortoise* (fluxo com longa duração de tempo).

al., 2006b) os autores, através de *traces* coletados da Universidade de Auckland, agrupam os fluxos utilizando os atributos: número total de pacotes, média de tamanho de pacotes, duração do fluxo e tempo médio entre pacotes. Tais fluxos foram classificados nas seguintes classes de aplicações: DNS, FTP (*File Transfer Protocol*), HTTP, IRC (*Internet Relay Chat*), LIMEWARE, POP3 (*Post Office Protocol 3*), SMTP e SOCKS. Recentemente, técnicas de aprendizagem de máquina baseadas em redes neurais Bayesianas foram utilizadas para discriminar dados em categorias derivadas de informações contidas nos pacotes provendo uma classificação dos mesmos sem, contudo, acessar seus conteúdos (AULD, et al., 2007).

Em (SEN e WANG, 2002) foi feita uma classificação do tráfego em um número pequeno de categorias baseada em qualidade de serviços (QoS – *Quality of Service*). Segundo (AULD, et al., 2007), esta abordagem restringe o número de aplicações possíveis de serem identificadas e utiliza um conjunto pequeno de variáveis. Também é avaliado por (AULD, et al., 2007) que a abordagem de (SEN e WANG, 2002) baseia sua identificação principalmente nas informações de portas TCP ou UDP (*User Datagram Protocol*) para agrupar o tráfego em suas classes, introduzindo desta forma uma margem de erro considerável.

1.2. Objetivos

Este trabalho tem como principal objetivo propor um classificador de tráfego de aplicações Internet e assim contribuir com as pesquisas relacionadas a classificação de aplicações contidas no tráfego Internet.

A metodologia adotada, baseada no trabalho de (BEYGELZIMER, et al., 2005) e (SIQUEIRA, 2008), identifica os fluxos utilizando um pequeno número de discriminantes estatísticos e classifica as aplicações através da análise de agrupamento (Seção 2.4).

As principais referências do presente trabalho foram as pesquisas de (ZUEV e MOORE, 2005), (MOORE e ZUEV, 2005b) e (AULD, et al., 2007). A metodologia de (AULD, et al., 2007) não fez uso de informações de carga útil (dados dos pacotes, apenas informações dos cabeçalhos TCP) e nem de informações de portas de serviço.

A recente pesquisa de (AULD, et al., 2007) mostra ser possível a identificação de aplicações utilizando somente informações e estatísticas obtidas dos cabeçalhos TCP, tornando a metodologia independente do local de coleta.

Têm-se como objetivos específicos:

- A busca da menor quantidade possível de variáveis estatísticas que melhor discriminem padrões de fluxos do tráfego de uma determinada aplicação;
- A obtenção de taxas de acerto equivalentes ou mesmo mais elevadas, com relação às mais recentes pesquisas em classificação de tráfego.

1.3. Contribuições

Foram obtidos, como resultados deste trabalho, valores médios de exatidão (Equações 3.2 e 3.3) equivalentes às mais recentes pesquisas na área de classificação de aplicações Internet e baixas taxas de falsos negativos (Equação 3.5).

Utilizando um conjunto de fluxos ao qual já foram aplicados diversos algoritmos de classificação, se conseguiu mostrar que é possível através de uma reduzida quantidade de variáveis estatísticas obter altos níveis de classificação das aplicações contidas em um tráfego Internet.

Com relação à contribuição à comunidade científica, foi alcançada a publicação dos resultados obtidos em eventos internacionais (CARMO, et al., 2007a), (CARMO, et al., 2007b) e (HOLANDA, et al., 2008), nacionais (HOLANDA, et al., 2007b) e (HOLANDA, et al., 2007c), e em periódico nacional (HOLANDA, et al., 2007a).

1.4. Organização do Trabalho

Este trabalho está organizado em cinco capítulos, incluindo esta introdução. No Capítulo 2, os métodos estatísticos que foram aplicados a classificação das aplicações são descritos destacando-se as técnicas estatísticas utilizadas para realizar a seleção das variáveis e os conceitos de análise de agrupamento.

A metodologia utilizada pelo classificador e inicialmente descrita em (SIQUEIRA, 2008) é apresentada no Capítulo 3. Os resultados obtidos pelo

classificador através da aplicação da metodologia em um conjunto de *traces* real e a comparação dos resultados obtidos com as principais referências utilizadas neste trabalho (ZUEV e MOORE, 2005), (MOORE e ZUEV, 2005b) e (AULD, et al., 2007) são expostos no Capítulo 4.

As principais contribuições, bem como a indicação de possíveis linhas de trabalhos futuros, são apresentadas no Capítulo 5.

Capítulo 2 – Métodos Estatísticos Aplicados na Classificação

A seguir são descritos de forma breve os métodos estatísticos que foram aplicados à classificação das aplicações.

2.1. Introdução

Quando se considera a utilização de variáveis discriminantes, é essencial que se tenha medido nos elementos amostrais variáveis que possam realmente distinguir as populações[6], caso contrário, a qualidade da classificação estará comprometida. Os métodos mais largamente usados para identificação de variáveis estatísticas com intuito de classificar aplicações Internet são baseados em análise de variância (ANDERSON, 2003). Neste trabalho é aplicada uma análise de variância chamada *Razão F*. Optou-se por este método, pelo fato do mesmo ser largamente utilizado em outras áreas do conhecimento para a seleção de variáveis estatísticas.

Em conjunto com a Razão F é utilizada a análise da correlação linear. A análise da correlação compreende a análise de dados amostrais para saber se e como duas variáveis estão relacionadas uma com a outra em uma população. Esta análise é útil em um trabalho exploratório, quando se procura determinar quais variáveis são potencialmente importantes e qual o grau de relacionamento existente entre as mesmas. Escolheu-se fazer uso desta técnica neste trabalho devido à necessidade de se eliminar redundância de variáveis que possuam características (medidas estatísticas) semelhantes.

Muitas são as situações nas quais se tem um conjunto de dados e se busca uma divisão desses dados em grupos, de modo que os grupos tenham coesão interna e sejam heterogêneos entre si. A técnica de análise de agrupamento constitui uma ferramenta indispensável neste processo de partição dos dados em grupos e tem sido uma das técnicas mais utilizadas na prática (HOLANDA, 2005) e (ERMAN, et al., 2006a). É através desta técnica que os dados apresentados neste trabalho serão particionados.

[6] Neste trabalho, as populações são as classes de tráfego a serem analisadas.

2.2. Razão F

A análise de variância, como o nome indica, envolve o cálculo de variâncias. A variância de uma amostra é a média dos quadrados dos desvios em relação à média do grupo (STEVENSON, 2001). Simbolicamente,

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} \tag{2.1}$$

Note-se que se deve usar $n - 1$, pois está se lidando com dados amostrais (ANDERSON, et al., 2007).

A Razão F utiliza a razão de duas estimativas, dividindo a estimativa da variância entre amostras (σ_E^2) pela estimativa da variância dentro das amostras (σ_D^2).

$$Razão\ F = \frac{\sigma_E^2}{\sigma_D^2} \tag{2.2}$$

A estimativa do numerador (variância entre as amostras) envolve a variação populacional baseada na variabilidade entre as médias amostrais (TRIOLA, 2005).

$$\sigma_E^2 = n\frac{\sum_{i=1}^{k}(\bar{x}_i - \bar{\bar{x}})^2}{(k-1)} \tag{2.3}$$

A média das médias dos valores das variáveis é calculada a partir da seguinte equação:

$$\bar{\bar{x}} = \frac{\sum_{i=1}^{k}\bar{x}_i}{k} \tag{2.4}$$

Nas equações (2.3), (2.4), k representa o número de grupos amostrais e n o tamanho do grupo amostral.

A estimativa do denominador (variância dentro das amostras) envolve a variância populacional baseada nas variâncias amostrais (TRIOLA, 2005).

$$\sigma_D^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x}_1)^2 + \sum_{i=1}^{n}(x_i - \bar{x}_2)^2 + \sum_{i=1}^{n}(x_i - \bar{x}_3)^2 + \cdots + \sum_{i=1}^{n}(x_i - \bar{x}_k)^2}{k(n-1)} \quad (2.5)$$

Assim como nas Equações (2.3) e (2.4), na Equação (2.5) *k* representa o número de grupos amostrais e *n* o tamanho (quantidade de elementos) do grupo amostral[7].

Como o numerador mede a variação entre elementos de amostras distintas, pode-se afirmar que quanto **maior** o seu valor melhor será a separação entre os grupos amostrais. Quanto ao denominador, o mesmo mede a variação entre elementos de uma mesma amostra. Neste caso, pode-se afirmar que quanto **menor** o seu valor, mais semelhantes entre si são os elementos de um mesmo grupo amostral.

A partir das afirmações acima se pode deduzir que valores elevados no numerador em conjunto com valores reduzidos no denominador proporcionam um valor elevado da Razão F.

Como o objetivo da Razão F é indicar as variáveis que melhor discriminarão os elementos amostrais, pode-se afirmar que, quanto maior o seu valor para uma determinada variável, mais apropriada será esta variável para realizar a separação dos elementos amostrais, caso em que se obtém um elevado valor no numerador e um reduzido valor no denominador ao se aplicar a Equação (2.2).

2.3. Correlação Linear

A correlação é uma técnica que envolve uma forma de estimação. O termo "correlação" significa literalmente "co-relacionamento", pois indica até que ponto os valores de uma variável estão relacionados com os de outra. A forma mais comum de análise de correlação envolve dados contínuos. Através da análise de correlação obtém-se um número que resume o grau de relacionamento entre duas variáveis, ou seja, a correlação mede a força, ou o nível de intensidade, de relacionamento entre duas variáveis (STEVENSON, 2001).

[7] Neste trabalho, *k* representa a quantidade de grupos de classes de tráfego a serem analisados (grupos amostrais) e *n* o número de fluxos amostrados em cada grupo (quantidade de elementos no grupo amostral).

O grau de relacionamento entre duas variáveis contínuas é sintetizado por um coeficiente conhecido como **coeficiente de correlação**. O coeficiente de correlação é muitas vezes chamado de **coeficiente momento-produto de Pearson** ou **r de Pearson**, em homenagem ao matemático Karl Pearson (1857 – 1936), que desenvolveu a técnica (TRIOLA, 2005).

O coeficiente de correlação é dado por:

$$r = \frac{n\sum_{i=1}^{n} x_i\, y_i - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{\sqrt{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}\sqrt{n\sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i)^2}} \qquad (2.6)$$

onde *n* é o número de observações (quantidade de elementos noo grupo amostral). Os valores limite de *r* são -1 e +1, isto é, o valor de *r* pertence ao intervalo [-1,+1].

Representando, em um sistema coordenado cartesiano ortogonal, os pares ordenados das variáveis x_i e y_i (x_i,y_i), obtêm-se uma nuvem de pontos que é denominada **diagrama de dispersão**. Este diagrama fornece uma idéia visual da correlação existente entre duas variáveis.

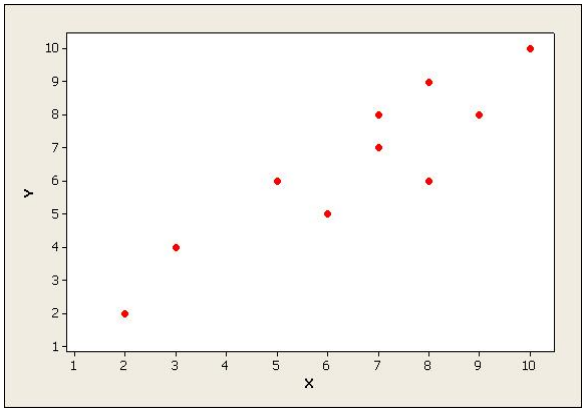

Figura 2.1. Diagrama de Dispersão.

Os pontos obtidos na Figura 2.1, vistos em conjunto, formam uma elipse em diagonal. Quanto mais fina for esta elipse, mais ela se aproximará de uma reta. Diz-se, então, que a correlação de forma elíptica tem como "imagem" uma reta imaginária, sendo, por isso, denominada **correlação linear** (CRESPO, 1996).

Uma correlação pode ser:

a) **Linear positiva** se os pontos do diagrama têm como "imagem" uma reta ascendente (Figura 2.2);
b) **Linear negativa** se os pontos do diagrama têm como "imagem" uma reta descendente (Figura 2.3);
c) **Não-linear** se os pontos do diagrama têm como "imagem" uma curva (Figura 2.4).

Se os pontos apresentam-se dispersos, não oferecendo uma "imagem" definida, pode-se concluir que não há relação alguma ou existe uma relação muito fraca entre as variáveis em estudo (Figura 2.5).

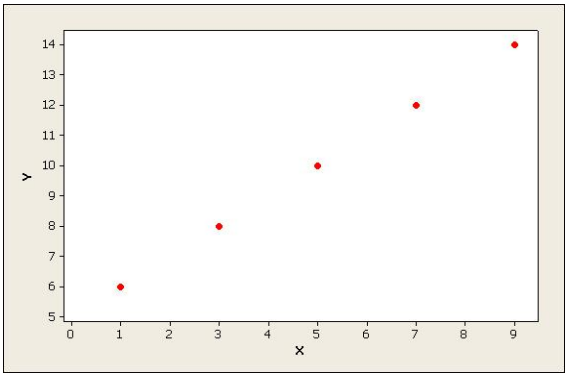

Figura 2.2. Correlação Linear Positiva.

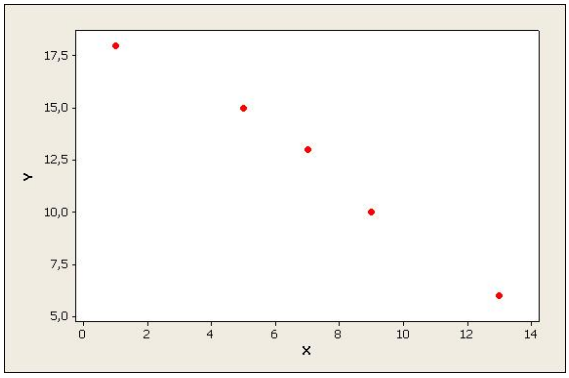

Figura 2.3. Correlação Linear Negativa.

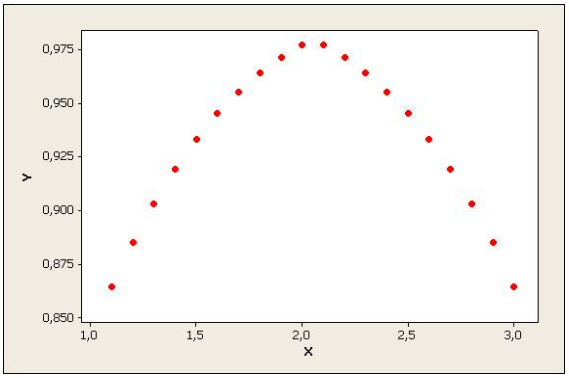

Figura 2.4. Correlação Não Linear.

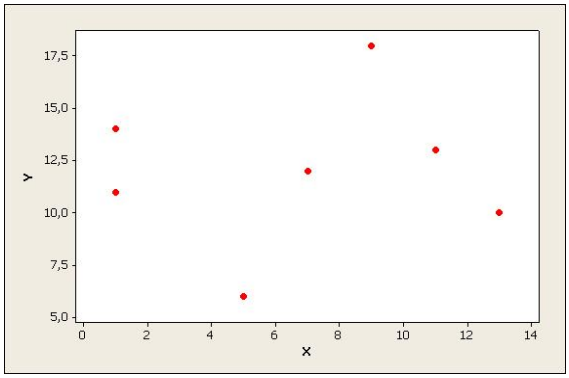

Figura 2.5. Ausência de Correlação.

O coeficiente de correlação possui duas propriedades que caracterizam a natureza de uma relação entre duas variáveis. Uma é o seu sinal (positivo ou negativo) e a outra é sua magnitude. O sinal de *r* indica o tipo de relação linear, se positiva ou

negativa. Um número com sinal positivo (ou nenhum sinal) indica uma correlação positiva (Figura 2.2) e um número com sinal negativo indica uma correlação negativa (Figura 2.3). Segundo (WITTE e WITTE, 2005):

a) À medida que valores relativamente baixos formam par com valores relativamente baixos e valores relativamente altos formam par com valores relativamente altos, tem-se uma relação **positiva** (Figura 2.2);
b) À medida que valores relativamente baixos formam par com valores relativamente altos e valores relativamente altos formam par com valores relativamente baixos, tem-se uma relação **negativa** (Figura 2.3);
c) Se alguns valores relativamente altos formam par com valores relativamente baixos e outros formam par com valores relativamente altos, tem-se uma relação **nula** ($r = 0$)[8].

A magnitude (valor) de *r* indica a intensidade da correlação entre as variáveis e o quão próximos da reta imaginária estão os pontos individuais. Valores de *r* próximos de -1 ou +1 indicam que os valores estão muito próximos da reta, ou mesmo sobre a reta, e mais forte (mais regular) torna-se a correlação. Reciprocamente, quanto mais próximo o valor de *r* fica em relação a zero maior a dispersão dos valores em relação à reta e mais fraca (menos regular) torna-se a correlação. Segundo (CRESPO, 1996):

a) Se $0{,}6 \leq |r| < 1$, há uma correlação relativamente forte entre as variáveis;
b) Se $0{,}3 \leq |r| < 0{,}6$, há uma correlação relativamente fraca entre as variáveis;
c) Se $0 < |r| < 0{,}3$, a correlação é muito fraca e, praticamente, nada pode ser concluído sobre a relação entre as variáveis em estudo.

2.4. Análise de Agrupamento

A análise de agrupamento pertence a um conjunto de técnicas utilizadas na análise estatística multivariada. A estatística multivariada consiste em um conjunto de métodos estatísticos utilizados em situações nas quais diversas variáveis são medidas simultaneamente, em cada elemento amostral (MINGOTI, 2005). Técnicas de estatística

[8] Quando *r* for igual a zero, não necessariamente se terá uma ausência de correlação. Poderá existir uma correlação entre as variáveis, mas esta correlação é não-linear.

multivariada, em geral, analisam a estrutura de correlação entre diversas variáveis, podendo revelar resultados mais completos do que se as variáveis fossem analisadas separadamente (JOHNSON, 1998).

A estatística multivariada se divide em dois grupos: um primeiro, consistindo em técnicas exploratórias de sintetização (ou simplificação) da estrutura de variabilidade dos dados, e um segundo, consistindo em técnicas de inferência estatística. Fazem parte do primeiro grupo a análise de agrupamentos (*Cluster Analysis*), análise de componentes principais (PCA – *Principal Component Analysis*), análise fatorial, análise de correlações canônicas, análise discriminante (*Discriminant Analysis*) e análise de correspondência. No segundo grupo, encontram-se os métodos de estimação de parâmetros, testes de hipóteses, análise de variância, de covariância e de regressão multivariada.

A análise de agrupamento, portanto, pode ser utilizada para encontrar grupos nos dados sob análise (KAUFMAN e ROUSSEEUW, 1990). Esta técnica compreende um conjunto de diferentes algoritmos e métodos para agrupar objetos de tipos similares em respectivas categorias. O problema enfrentado por muitos pesquisadores em diferentes áreas consiste exatamente em como organizar os dados sob análise em estruturas que sejam suficientemente representativas. Em outras palavras, a análise de agrupamento é uma ferramenta exploratória que busca particionar os componentes em diferentes grupos tal que membros de um mesmo grupo sejam os mais similares possíveis e membros de grupos distintos sejam os mais dissimilares possíveis (JAIN, 1991). Estatisticamente, isso implica que a variância intra-grupo deve ser a menor possível e que a variância inter-grupo deve ser a maior possível.

Uma questão importante refere-se ao critério a ser utilizado para se decidir até que ponto dois elementos do conjunto de dados podem ser considerados como semelhantes ou não. Para isto é necessário considerar medidas que descrevam a similaridade entre elementos amostrais de acordo com as características que neles foram medidas. Considerando que para cada elemento amostral têm-se informações de *x*-variáveis armazenadas em um vetor, a comparação entre diferentes elementos amostrais poderá ser feita através de métricas, que possibilitem a comparação de vetores, como as medidas de distância. Para cada atributo *f*, tem-se, portanto, o vetor de medidas *f* definido por:

$$X_f = \left[X_{1f}, X_{2f}, X_{3f}, \ldots, X_{if}\right]' \quad (2.7)$$

onde *i* é a quantidade de elementos amostrais.

Assim, pode-se calcular as distâncias entre os vetores de observações dos elementos amostrais e agrupar aqueles de menor distância.

Diversas medidas de distância têm sido utilizadas para formar agrupamentos. As mais comuns são: distância euclidiana, distância ponderada, distância de Minkowski, coeficiente de concordância simples e o coeficiente de concordância de Jaccard (MINGOTI, 2005). Neste trabalho foi utilizada a distância euclidiana por se tratar de uma métrica simples, largamente utilizada e de fácil implementação. A distância euclidiana entre dois elementos é dada por:

$$dist(x,y) = \sqrt{\sum(x_i - y_i)^2} \quad (2.8)$$

onde x_i e y_i são as coordenadas dos pontos *x* e *y* (Figura 2.6), ou seja, os dois elementos amostrais são comparados em cada variável pertencente ao vetor de observações.

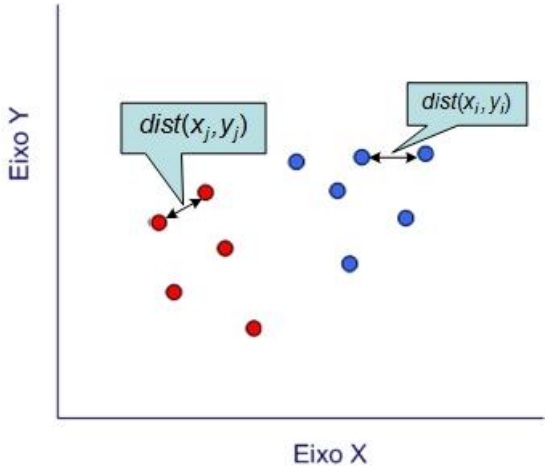

Figura 2.6. Distância Euclidiana.

Na literatura, a análise de agrupamento tem sido realizada através de várias técnicas (KAUFMAN e ROUSSEEUW, 1990) e (EVERITT, et al., 2001). Entretanto, todas essas técnicas frequentemente são classificadas em dois tipos: técnicas hierárquicas e não-hierárquicas. Segundo (MINGOTI, 2005), as técnicas hierárquicas podem ser implementadas de duas maneiras: aglomerativa ou divisiva (Figura 2.7 e Figura 2.8).

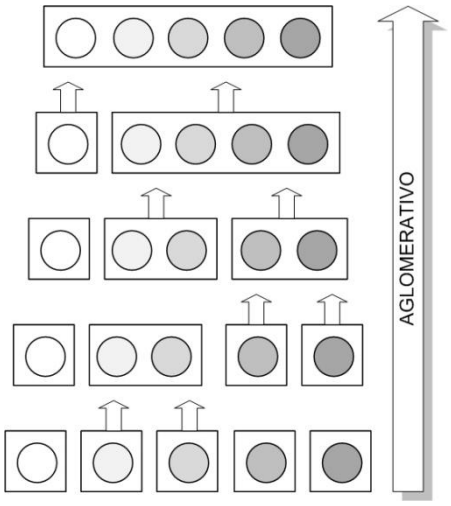

Figura 2.7. Técnica Hierárquica Aglomerativa. (MINGOTI, 2005)

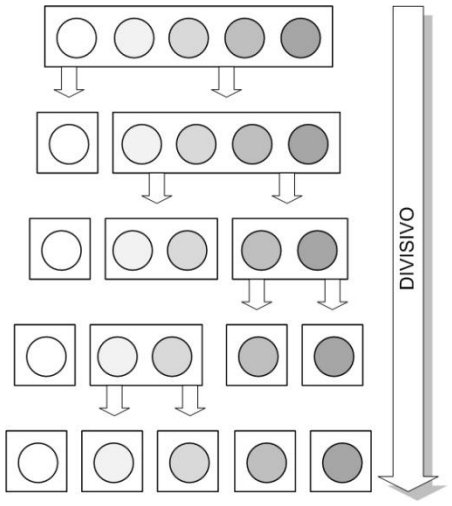

Figura 2.8. Técnica Hierárquica Divisiva. (MINGOTI, 2005)

Utilizando-se a técnica hierárquica aglomerativa, dados *n* componentes no início do processo, tem-se *n* agrupamentos, ou seja, cada elemento do conjunto de dados é considerado como sendo um agrupamento isolado. Em cada passo do algoritmo, os elementos amostrais vão sendo agrupados, formando novos agrupamentos até o momento no qual todos os elementos considerados estarão em um único grupo (*cluster*). Em cada estágio do procedimento de agrupamento, os grupos são comparados através de alguma medida de similaridade (ou dissimilaridade) previamente definida. De acordo com (MINGOTI, 2005), os principais passos para a aplicação das técnicas hierárquicas aglomerativas podem ser resumidos da seguinte forma:

1. Inicialmente cada elemento constitui um agrupamento de tamanho 1 (um). Portanto, têm-se *n* agrupamentos;
2. Em cada estágio do algoritmo, os pares de agrupamentos mais "similares" são combinados e passam a constituir um único agrupamento (Figura 2.7). Apenas um novo agrupamento pode ser formado em cada estágio;
3. Cada novo agrupamento formado é um "agrupamento de agrupamentos" formados nos estágios anteriores. Se dois elementos amostrais aparecem juntos em um mesmo *cluster* em algum estágio do processo, eles aparecerão juntos em todos os estágios subsequentes, ou seja, uma vez unidos estes elementos não poderão ser separados. Esta característica é conhecida como **propriedade de hierarquia**;

4. A partir da propriedade de hierarquia é possível construir um gráfico chamado **dendograma** ou **dendrograma** que representa as etapas de agrupamento. O dendograma é um gráfico em forma de árvore no qual a escala vertical indica o nível de similaridade (ou dissimilaridade). No eixo horizontal, são marcados os elementos amostrais numa ordem conveniente relacionada aos agrupamentos. As linhas verticais, partindo dos elementos amostrais agrupados, têm altura correspondente ao nível em que os elementos foram considerados semelhantes, isto é, a distância do agrupamento ou o nível de similaridade (Figura 2.9).

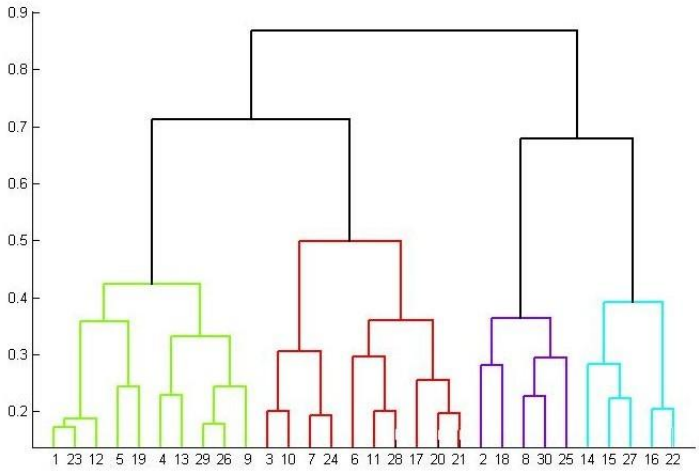

Figura 2.9. Dendograma.

Os passos para a aplicação das técnicas hierárquicas divisivas são semelhantes aos apresentados anteriormente, sendo que se inicia o procedimento com um único agrupamento (de *n* componentes) e então se divide os agrupamentos sucessivamente até se obter um número desejado de *clusters* (Figura 2.8).

As técnicas não-hierárquicas diferem das hierárquicas em vários aspectos. Primeiramente, é necessário que o usuário especifique previamente o número de agrupamentos desejados, ou seja, inicia-se com um conjunto arbitrário de agrupamentos e os membros dos agrupamentos são movidos até que a variância intra-grupo seja mínima. Em cada estágio do agrupamento, os novos grupos podem ser formados através da divisão ou junção de grupos já combinados em passos anteriores. Isto significa que, se em algum passo do algoritmo dois elementos tiverem sido colocados em um mesmo *cluster*, não necessariamente eles estarão juntos na partição final. Como consequência, não é mais possível a construção de dendogramas.

Geralmente, os algoritmos computacionais utilizados nos métodos não-hierárquicos são simples, do tipo iterativo e, em comparação com os métodos hierárquicos, têm uma maior capacidade de análise de conjuntos de dados de maior porte, ou seja, com um grande número de observações. Devido a este fato, optou-se neste trabalho por utilizar um método não-hierárquico de agrupamento.

Os métodos das K-Médias (*K-Means*) e o *Fuzzy* c-Médias (*Fuzzy c-Means*) são alguns exemplos de métodos não-hierárquicos, assim como as redes neurais artificiais aplicadas à análise de agrupamentos (EVERITT, et al., 2001).

2.4.1. Método das K-Médias

O método das K-Médias é provavelmente um dos mais conhecidos e mais utilizados métodos de análise de agrupamentos em problemas práticos (DUDA e HART, 2000). Basicamente, cada elemento amostral é alocado para aquele agrupamento cujo centróide[9] (vetor de médias amostrais) é o mais próximo do vetor de valores observados para o respectivo elemento. Originalmente, o método é composto por quatro passos (MINGOTI, 2005):

1. Primeiramente são escolhidos *k* centróides, chamados de "sementes" ou "protótipos", para iniciar o processo de partição;

2. Cada elemento do conjunto de dados é, então, comparado com cada centróide inicial, através de uma medida de distância que, em geral, é a distância euclidiana. O elemento é alocado ao grupo cuja distância é menor;

3. Depois de aplicar o passo 2 para cada um dos *n* elementos amostrais, recalcula-se os valores dos centróides para cada novo grupo formado, e repete-se o passo 2, considerando os centróides destes novos grupos;

[9] Neste trabalho, o centróide é definido como o ponto que contém como coordenadas as médias dos valores de cada uma das variáveis dos fluxos presentes no agrupamento, ou seja, tendo-se escolhidos f discriminantes para determinada classe de aplicações, o centróide de um agrupamento terá como variáveis os valores $(\bar{x}_1, \bar{x}_2, \bar{x}_3, \dots, \bar{x}_f)$.

4. Os passos 2 e 3 devem ser repetidos até que todos os elementos amostrais estejam alocados em seus grupos de acordo com uma medida de verossimilhança ou número de agrupamentos pré-definidos, isto é, até que nenhuma realocação de elementos seja necessária.

Uma grande vantagem deste algoritmo é a sua simplicidade computacional, que o torna candidato para diversas aplicações. Segundo (DUDA e HART, 2000), a ordem de complexidade deste algoritmo é de $O(nkf)$, onde n é o número de elementos no grupo amostral, k o número de agrupamentos (grupos amostrais) e f o número de variáveis (*features*) analisadas. Ainda de acordo com (DUDA e HART, 2000), o esforço computacional $O(nkf)$ é muito menor do que os métodos hierárquicos se k for muito menor que n.

O algoritmo das K-Médias é sensível a valores extremos (*outliers*) (Seção 3.3.2). Os valores extremos, sendo elementos do conjunto de dados X, são necessariamente associados a um dos k agrupamentos. Então, estes têm influência direta nos valores das médias e, por consequência, no conjunto final dos agrupamentos. Levando-se em conta de que, em geral, agrupamentos com pequeno número de elementos geralmente são formados por valores extremos, alguns algoritmos simplesmente descartam agrupamentos considerados pequenos em número de elementos.

Os trabalhos de (MCGREGOR, et al., 2004), (ERMAN, et al., 2006b) e (BERNAILLE, et al., 2006) utilizaram o método das K-Médias para agrupar diferentes aplicações com base em informações obtidas dos cabeçalhos TCP.

2.5. Resumo do Capítulo

Ao lidar com problemas de classificação de tráfego Internet é essencial a identificação de características do tráfego sob análise que possam distinguir uma aplicação que o compõe das demais.

Neste capítulo foram descritos métodos estatísticos aplicados a classificação de aplicações Internet. Tais métodos têm por objetivos determinar quais variáveis são potencialmente importantes, descobrir as relações entre diversas variáveis e indicar as que melhor discriminam um determinado tipo de aplicação das demais através de suas principais características.

No próximo capítulo, será apresentada a metodologia de classificação de classes de tráfego de aplicações Internet, com a análise das variáveis estatísticas, seleção dos discriminantes, aplicação da análise de agrupamentos e as fases de treinamento, validação e classificação, além das métricas desenvolvidas para mensurar a eficácia da classificação.

Capítulo 3 – Metodologia de Classificação

Neste capítulo é apresentada a metodologia utilizada para a identificação das classes de tráfego Internet. Esta metodologia, baseada nos trabalhos de (BEYGELZIMER, et al., 2005) *e* (SIQUEIRA, 2008)*, realiza a seleção de um subconjunto de variáveis estatísticas através do uso da Razão F e da correlação linear. Após a escolha das variáveis, as mesmas são utilizadas em uma fase de treinamento com o intuito de determinar os discriminantes estatísticos que serão utilizados para identificar uma determinada classe de aplicação. Além disso, a fase de treinamento busca definir um conjunto de agrupamentos e seus respectivos centróides. Em seguida é utilizada a análise de agrupamentos para definir à qual classe de aplicação um determinado fluxo de dados pertence. A geração dos agrupamentos se baseia na divisão de um problema de classificação de 1 para N em N problemas de classificação 1 para 1 (abordagem um-contra-todos).*

Devido à grande quantidade de informações e complexidade inerentes às tarefas de análise e classificação de aplicações Internet, buscou-se neste trabalho, através da utilização de métodos estatísticos, encontrar as melhores variáveis para identificar cada tipo de aplicação, partindo do pressuposto de que o melhor conjunto de variáveis para identificar uma classe de tráfego não é necessariamente o mesmo para identificar todas as classes.

3.1. Introdução

A tarefa de identificação de tráfego é na realidade uma tarefa de classificação. Segundo (DUDA, et al., 2000), de forma geral, para qualquer método de classificação que incorpore informação de amostras de treinamento no projeto do classificador, diz-se que utiliza aprendizagem de máquina. O termo “aprendizagem” refere-se a algum algoritmo que venha a reduzir o erro da medição (estimativa) em um conjunto de dados de

treinamento (DUDA, et al., 2000). Existem duas formas gerais de aprendizagem de máquina: **supervisionada** e **não-supervisionada**.

Na aprendizagem supervisionada, o algoritmo de aprendizagem recebe um conjunto de exemplos de treinamento para os quais os rótulos da classe associada são conhecidos. Cada exemplo (instância) é descrito por um vetor de valores (atributos) e pelo rótulo da classe associada. O objetivo é construir um classificador que possa determinar corretamente a classe de novos exemplos ainda não rotulados. Algumas pesquisas que utilizaram a abordagem supervisionada para classificação de fluxos de tráfego foram as de (ABBES, et al., 2004), (SU e ZHANG, 2006) e (WILLIAMS, et al., 2006).

Na aprendizagem não-supervisionada, o conjunto de treinamento é classificado sem a utilização da informação das classes, ou seja, os agrupamentos não são previamente conhecidos. O algoritmo analisa os exemplos fornecidos e tenta determinar se alguns deles podem ser agrupados de alguma maneira, formando agrupamentos. Após a determinação dos agrupamentos, em geral, é necessária uma análise para determinar o que cada agrupamento significa no contexto do problema sendo analisado. Exemplos da utilização de aprendizagem não-supervisionada para a classificação de fluxos são as pesquisas de (MCGREGOR, et al., 2004), (BERNAILLE, et al., 2006) e (ERMAN, et al., 2006b).

A utilização de aprendizagem de máquina através de métodos estatísticos multivariados tem alcançado alto grau de exatidão na classificação de aplicações. Notadamente a análise de agrupamentos através do método das K-médias (MCGREGOR, et al., 2004), (BERNAILLE, et al., 2006) e (ERMAN, et al., 2006b) tem sido aplicada em recentes pesquisas, para separar os fluxos em grupos de aplicações, devido a sua menor complexidade computacional quando comparada a dos algoritmos hierárquicos (Seção 2.4.1).

A metodologia proposta neste trabalho utiliza aprendizagem supervisionada e baseia-se na classificação dos fluxos. O problema basicamente consiste em explorar as propriedades estatísticas dos fluxos de dados para identificar os diferentes tipos de aplicações. Um conjunto de fluxos de treinamento é previamente classificado sabendo-se a *priori* à qual classe de aplicação um determinado fluxo pertence (caracterizando a aprendizagem supervisionada). Sua função é servir como uma coleção de exemplos que

são usados para informar ao algoritmo de classificação onde ele está errando e onde ele deve melhorar - daí a expressão “supervisionada” no nome desse tipo de aprendizagem. Os dados classificados são utilizados para rotular os agrupamentos. Os centróides dos grupos representam as classes. A descrição da metodologia é apresentada em detalhes nas seções a seguir, sendo dividida nas seguintes fases:

- Reconstrução dos fluxos e classificação manual;
- Seleção das variáveis candidatas a discriminantes estatísticos;
- Treinamento e definição dos discriminantes;
- Validação;
- Classificação.

3.2. Reconstrução dos Fluxos e Classificação Manual

A Figura 3.1 apresenta a primeira fase da metodologia, que consiste na coleta dos pacotes, reconstrução dos fluxos, cálculo das variáveis estatísticas e classificação manual dos fluxos que posteriormente serão utilizados nas fases de treinamento (Seção 3.4), validação (Seção 3.5) e classificação (Seção 3.6).

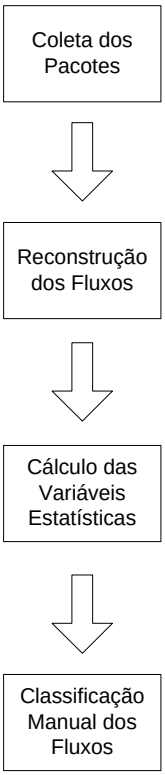

Figura 3.1. Coleta dos Fluxos e Pré-Processamento das Variáveis.

Na primeira parte da reconstrução dos conjuntos de fluxos (*traces*), os pacotes são agrupados de acordo com as informações obtidas a partir dos seus cabeçalhos

(endereço IP de origem, endereço IP de destino, porta de origem, porta de destino e tipo de protocolo).

O pré-processamento citado acima se refere ao procedimento de classificação manual executado sobre todos os fluxos do *trace* para identificar à qual aplicação os mesmos pertencem, bem como a definição de um conjunto de variáveis relacionadas a cada fluxo[10] de acordo com o propósito da classificação.

Uma análise baseada em fluxos apresenta-se mais eficaz do que a baseada em pacotes, pois o fluxo é um desdobramento do comportamento de sua aplicação. Agrupar os pacotes em fluxos permite um processamento mais efetivo das informações coletadas e a aquisição do contexto necessário para a identificação apropriada da aplicação responsável pelo fluxo (MOORE, et al., 2005c).

Durante a fase de coleta dos pacotes e reconstrução dos fluxos, diversas informações estatísticas devem ser calculadas. Alguns exemplos dessas informações são descritos na Tabela 3.1. Estas informações serão úteis para a tarefa de classificação dos fluxos (Seção 3.6), visto que a metodologia, aqui apresentada, se propõe a não utilizar informações do conteúdo dos pacotes.

Tabela 3.1. Exemplos de Variáveis Estatísticas.

Média do tempo de chegada entre pacotes
Variância do total de bytes no pacote IP
Total de pacotes ACK na direção "cliente→servidor"
Tamanho máximo da janela na direção "servidor→ cliente"
Número mínimo de RTT na direção "cliente→servidor
Tempo decorrido desde a última conexão entre estes *hosts*
Tempo ocioso em segundos

3.3. Seleção das Variáveis Candidatas a Discriminantes Estatísticos

A etapa de seleção das variáveis candidatas a discriminantes estatísticos, que serão utilizados para a identificação das aplicações Internet, é provavelmente a de maior importância. A exatidão da classificação está diretamente relacionada com a correta

[10] A definição de cada uma das 249 variáveis definidas em (MOORE, et al., 2005a) e utilizadas neste trabalho encontra-se no Anexo A.

escolha dessas variáveis e, consequentemente, dos discriminantes. Esta etapa consiste na normalização dos dados, retirada dos *outliers* e na utilização da Razão F em conjunto com a correlação linear para definir quais as variáveis candidatas a melhor identificar uma classe de aplicações.

3.3.1. Normalização dos Dados

Em situações reais, somos confrontados com variáveis que estão dentro de diferentes ordens de grandeza. Deste modo, variáveis com grandes valores teriam maior influência do que aquelas com menores valores, muito embora não necessariamente tenham maior importância durante a classificação.

Este problema é resolvido através da normalização das variáveis, colocando seus valores dentro de uma mesma escala. Uma técnica direta de normalização dos dados é utilizar a média $\bar{z}_p$ e o desvio padrão σ_p (THEODORIDIS e KOUTROUMBAS, 2006). Deste modo, todas as variáveis normalizadas apresentarão média igual a zero e desvio padrão igual a um. Esse cálculo de normalização é mais conhecido como padronização ou *z-score*. Para *m* variáveis disponíveis e *n* observações, a *p*-ésima variável será:

$$z_{ip} = \frac{z_{ip} - \bar{z}_p}{\sigma_p};\ i = 1, 2, 3, \dots, n;\ p = 1, 2, 3, \dots, m \qquad (3.1)$$

3.3.2. Eliminação dos *Outliers*

Um *outlier* (valor extremo) é definido em (THEODORIDIS e KOUTROUMBAS, 2006) como um ponto que reside demasiadamente distante da média da variável correspondente. Pontos com valores muito diferentes da média introduzem grande erro durante o treinamento de um algoritmo de aprendizagem de máquina e podem produzir efeitos desastrosos. Devido a isto, tais valores devem ser analisados e, se for o caso, eliminados da análise.

Uma forma gráfica de visualizar valores extremos é através de diagramas em caixa ou *boxplots*, também conhecidos como diagramas esqueletais (TRIOLA, 2005). Os b*oxplots* são convenientes para revelar tendências centrais, dispersão, distribuição dos dados e, como citado anteriormente, a presença de *outliers* (valores extremos). A construção de um b*oxplot* exige que se tenha o valor mínimo, o primeiro quartil Q1, a mediana (ou segundo quartil Q2), o terceiro quartil Q3 e o valor máximo.

Um *boxplot* é um gráfico de dados que consiste em uma reta que se prolonga do menor ao maior valor, e um retângulo com retas traçadas no primeiro quartil Q1, na mediana Q2 e no terceiro quartil Q3 (Figura 3.2).

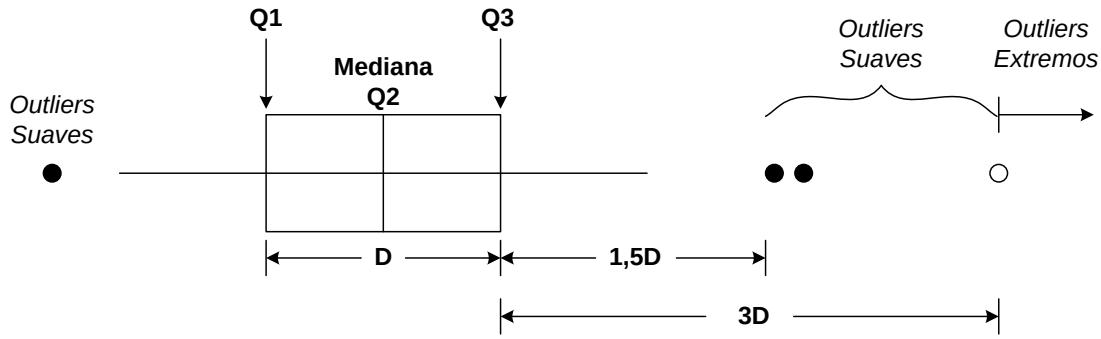

Figura 3.2. *Boxplot*. (TRIOLA, 2005)

A distância D é calculada pela diferença entre os quartis Q3 e Q1. Os *outliers* suaves são os valores que superam Q3 em 1,5D a 3D, ou estão 1,5D a 3D abaixo de Q1. Os *outliers* extremos são valores que excedem Q3 em mais de 3D ou estão a mais de 3D abaixo de Q1. Na metodologia apresentada, foram removidos os fluxos com valores acima da distância de 3D.

3.3.3. Razão F e Correlação Linear

Como citado no início da Seção 3.3, o método utilizado em nossa metodologia para seleção das variáveis candidatas baseia-se no cálculo da Razão F (Seção 2.2) e na correlação linear (Seção 2.3).

Após o cálculo da Razão F, as variáveis são ordenadas de forma decrescente. De acordo com a Seção 2.3, será realizada a análise da correlação linear destas variáveis. Buscou-se, aplicando a Equação (2.6), encontrar um subconjunto de variáveis que possuíssem uma correlação relativamente fraca ($0,3 \leq |r| < 0,6$) ou muito fraca entre si ($0 < |r| < 0,3$). Tais variáveis serão as candidatas a serem escolhidas como discriminantes estatísticos das classes de aplicações.

A análise conjunta da Razão F e da correlação linear neste momento é de extrema importância para eliminar a redundância de variáveis que possuam características (medidas estatísticas) semelhantes.

Após a análise dos resultados do processamento da Razão F e da análise de correlação linear para cada aplicação, terão sido selecionadas as variáveis com maiores

chances de melhor discriminar uma determinada aplicação das demais. O resultado final desta fase da metodologia é um conjunto individual de variáveis para cada classe de aplicação que, eventualmente, pode apresentar variáveis comuns para as diferentes classes.

3.4. Treinamento e Definição dos Discriminantes Estatísticos

Nesta fase, busca-se definir os discriminantes estatísticos que serão utilizados para identificar uma determinada classe de aplicação, além de definir um conjunto de agrupamentos e seus respectivos centróides. Estas informações serão utilizadas para separar os novos conjuntos de fluxos.

A fase de treinamento se inicia utilizando fluxos amostrados e classificados manualmente (Seção 3.2). Para cada tipo de aplicação a ser identificada, emprega-se o método de análise de agrupamentos (Seção 2.4) através do algoritmo das K-médias (Seção 2.4.1) utilizando as variáveis candidatas definidas na etapa de seleção (Seção 3.3). Através de um processo iterativo, mostrado a seguir, são definidos os discriminantes estatísticos.

1. Seleciona-se uma das variáveis candidatas (previamente ordenadas de forma decrescente);
2. Aplica-se o algoritmo das K-médias, onde *k* sempre será igual a dois (agrupamento da classe sob análise e agrupamento das demais classes);
3. Calcula-se a taxa de acerto de classificação dos fluxos;
4. Adiciona-se mais uma variável ao conjunto de variáveis inicialmente selecionadas como entrada do algoritmo das K-médias;
5. Repete os passos 2 e 3;
6. Repete os passos 4, 2 e 3 até que a taxa de acerto diminua ou permaneça inalterada, ou seja, até que não haja melhoria no poder de separação dos fluxos da classe sob análise.

Durante e após a definição dos discriminantes estatísticos, são compostos grupos de amostragem (um para cada aplicação sob análise) onde existam fluxos amostrados da classe a ser analisada e fluxos amostrados das demais classes, ou seja, um problema de classificação de 1 para N é dividido em N problemas de classificação 1 para 1

caracterizando a abordagem um-contra-todos (*one-against-all*) utilizada neste trabalho (Figura 3.3).

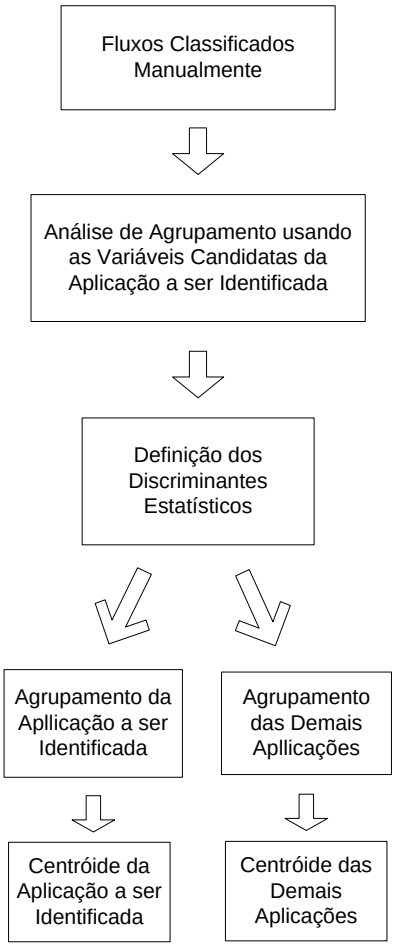

Figura 3.3. Fase de Treinamento.

O principal objetivo desta fase é determinar quais serão os discriminantes estatísticos utilizados para identificar uma determinada classe de aplicação e obter os centróides dos agrupamentos que representam cada uma das aplicações. Após ser aplicada a análise de agrupamentos o valor de cada centróide será armazenado para ser utilizado na fase de validação a seguir.

3.5. Validação

A fase de validação de nossa metodologia é dedicada a confirmar se novos fluxos serão vinculados corretamente aos agrupamentos obtidos na fase de treinamento. Para realizar esta tarefa, deve-se selecionar um conjunto de fluxos classificados manualmente diferente daquele que foi usado na etapa de treinamento (Seção 3.4).

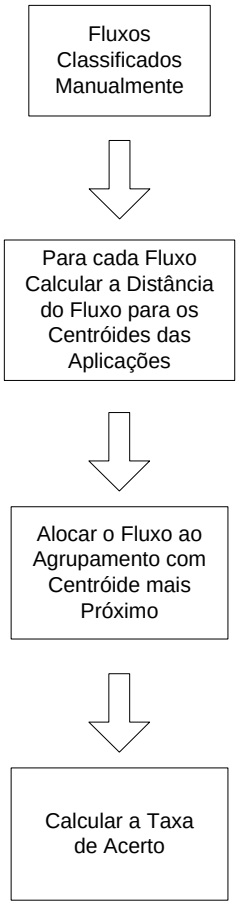

Figura 3.4. Fase de Validação.

A fase de validação é de grande importância, pois mostra se realmente foi realizada uma escolha acertada das variáveis utilizadas como discriminantes e se o algoritmo realmente será eficaz em realizar uma classificação precisa com um número de fluxos diferente do que foi utilizado na fase de treinamento.

Utilizando os discriminantes escolhidos, agrupamentos de fluxos serão gerados. Deve-se calcular a distância euclidiana de cada fluxo para os centróides dos grupos de aplicações definidos na fase de treinamento e associar o fluxo ao centróide mais próximo (Figura 3.4). A análise dos resultados obtidos leva em consideração se o fluxo foi alocado ao agrupamento correto. Para mensurar os resultados, foram definidos os seguintes parâmetros: ***exatidão média***, ***exatidão média da aplicação***, ***falso positivo*** *e* ***falso negativo***, definidos a seguir:

- **Exatidão Média**: Mede a porcentagem de fluxos corretamente classificados. Corresponde à porcentagem de fluxos da aplicação a ser identificada que estão nos agrupamentos corretos somados aos fluxos das demais aplicações que estão nos agrupamentos corretos dividido pelo total de fluxos analisados.

$$\%\ \begin{matrix}\textit{Exatidão}\\ \textit{Média}\end{matrix} = \frac{\text{Nº de Fluxos Corretamente Classificados nos Agrupamentos}}{\text{Total de Fluxos do Trace}} \qquad (3.2)$$

- **Exatidão Média da Aplicação**: Corresponde ao total de fluxos da aplicação corretamente classificados dividido pelo total de fluxos da aplicação presentes em todo o conjunto de fluxos.

$$\%\ \begin{matrix}\textit{Exatidão}\\ \textit{Média da}\\ \textit{Aplicação}\end{matrix} = \frac{\begin{matrix}\text{Nº De Fluxos da Aplicação Corretamente}\\ \text{Classificados nos Agrupamentos}\end{matrix}}{\text{Total de Fluxos da Aplicação no Trace}} \qquad (3.3)$$

- **Falso Positivo**: Indica o percentual de fluxos das aplicações que foram erroneamente identificados como sendo parte do agrupamento da aplicação sob análise.

$$\%\ \begin{matrix}\textit{Falso}\\ \textit{Positivo}\end{matrix} = \frac{\begin{matrix}\text{Nº de Fluxos das demais Aplicações Classificados nos}\\ \text{Agrupamentos da Aplicação sob Análise}\end{matrix}}{\text{Total de Fluxos nos Agrupamentos da Aplicação sob Análise}} \qquad (3.4)$$

- **Falso Negativo**: Indica o percentual de fluxos que foram classificados erroneamente como não sendo parte do agrupamento da aplicação sob análise.

$$\%\ \begin{matrix}\textit{Falso}\\ \textit{Negativo}\end{matrix} = \frac{\begin{matrix}\text{Nº de Fluxos da Aplicação sob Análise Classificados nos}\\ \text{Agrupamentos das demais Aplicações}\end{matrix}}{\text{Total de Fluxos da Aplicação sob Análise}} \qquad (3.5)$$

Outra forma de se visualizar as métricas acima é através de uma tabela conhecida como **matriz de confusão**. Nesta matriz se obtém todas as informações das métricas apresentadas para todo e qualquer conjunto de fluxos de uma classe sob análise em relação a todas as outras classes (Tabela 3.2).

Tabela 3.2. Matriz de Confusão.

	Classe sob Análise	**Demais Classes**	**Total de Fluxos da Classe**
Classe sob Análise	TAC	FN	TAC+ FN
Demais Classes	FP	TADC	FP + TADC
Total de Fluxos do Agrupamento	TAC + FP	FN + TADC	**TOTAL DE FLUXOS DO *TRACE***

A partir da Tabela 3.2 temos:

- TAC: Quantidade de fluxos da classe sob análise corretamente classificados;
- TADC: Quantidade de fluxos das demais classes corretamente classificados;
- FP: Falso Positivo;
- FN: Falso Negativo.

Analisando as Equações (3.2), (3.3), (3.4) e (3.5) e a Tabela 3.2 é possível deduzir:

$$\%\ Exatidão\ Média = \frac{TAC + TADC}{TAC + FN + TADC + FP} \tag{3.6}$$

$$\%\ Exatidão\ Média\ da\ Aplicação = \frac{TAC}{TAC + FN} \tag{3.7}$$

$$\%\ Falso\ Positivo = \frac{FP}{TAC + FP} \tag{3.8}$$

$$\%\ Falso\ Negativo = \frac{FN}{TAC + FN} \tag{3.9}$$

Estas métricas são importantes não somente para avaliar o resultado da metodologia proposta, como também para compará-la com outros métodos utilizados para realizar a mesma tarefa de classificação de tráfego por classes de aplicação.

3.6. Classificação

Considerando que os resultados obtidos na fase de validação tenham sido satisfatórios, os centróides obtidos na fase de treinamento podem ser utilizados para classificar qualquer tráfego pelo tipo de aplicação escolhido. Isto é feito calculando a distância euclidiana de cada fluxo de tráfego para os centróides dos agrupamentos desta aplicação.

Se as distâncias euclidianas das variáveis estiverem mais próximas dos valores dos centróides da aplicação, o fluxo será classificado como pertencente à classe. Se por acaso um fluxo for associado a mais de um agrupamento, a classe ao qual o fluxo pertencerá será definida baseada na taxa de acerto obtida na fase de treinamento, ou seja, se na fase de treinamento o fluxo tiver sido associado a uma das classes conflitantes, o mesmo será novamente associado a esta mesma classe na etapa de classificação. No caso do fluxo não ser associado a nenhuma classe, a classe ao qual o fluxo será associado será escolhida baseada na frequência dos fluxos das classes dentro da amostra de treinamento, caso contrário, pode-se optar por classificar o fluxo como não pertencente a nenhuma classe.

A Figura 3.5 mostra a sequência de passos para a classificação dos fluxos.

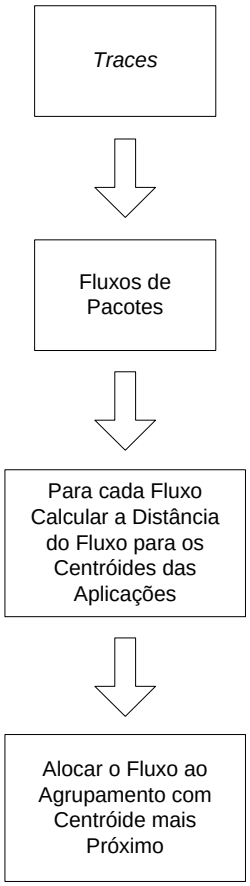

Figura 3.5. Fase de Classificação.

É importante destacar que o contexto corrente da metodologia proposta é aquele no qual o administrador da rede observa em modo *off-line* o comportamento do tráfego da rede.

3.7. Resumo do Capítulo

Neste capítulo apresentou-se a metodologia a ser utilizada pelo classificador proposto com o intuito de classificar as aplicações presentes em um conjunto de fluxos coletados da Internet. A metodologia utiliza aprendizagem de máquina supervisionada e baseia-se nos trabalhos de (BEYGELZIMER, et al., 2005) e (SIQUEIRA, 2008) para realizar a seleção de um subconjunto de variáveis estatísticas que melhor discriminem os fluxos de uma determinada classe de aplicação sob análise dos demais.

Foram definidas e descritas as fases de reconstrução dos fluxos e classificação manual, seleção das variáveis candidatas a discriminantes estatísticos, treinamento e

definição dos discriminantes, validação e classificação, bem como as métricas utilizadas para a verificação da exatidão da classificação.

A aplicação desta metodologia será tema do Capítulo 4 onde a mesma foi utilizada em um conjunto de *traces* previamente coletados e disponibilizados a comunidade científica. Logo após, os resultados obtidos são comparados com os resultados de outros métodos aplicados sobre o mesmo conjunto de fluxos.

Capítulo 4 – Resultados Obtidos

Este capítulo discorre sobre a aplicação da metodologia proposta em um conjunto real de fluxos de aplicações Internet. Os resultados obtidos foram comparados a recentes pesquisas em identificação de aplicações.

A partir de um conjunto de *traces* classificados manualmente, aplicou-se a metodologia descrita no capítulo anterior. Após a análise das variáveis estatísticas, na fase de treinamento, foram definidos os discriminantes e os centróides dos agrupamentos que representam cada uma das classes de aplicações. Logo após, foram aplicadas as fases de validação e classificação. Ao final foi feito um comparativo entre a metodologia proposta e outros métodos de classificação aplicados ao mesmo conjunto de fluxos.

4.1. Introdução

Com o objetivo de avaliar a metodologia proposta e os discriminantes selecionados, foi utilizado um conjunto de *traces* de tráfego. Os *traces* utilizados neste trabalho foram coletados e pré-processados por (MOORE e ZUEV, 2005b) e disponibilizados para a comunidade científica. O método de coleta do tráfego é descrito detalhadamente em (MOORE, et al., 2003). O tráfego refere-se a uma rede de 1.000 usuários conectados por um período de 24 horas através de uma conexão *full-duplex Gigabit Ethernet*. O *trace* completo foi dividido em 10 arquivos (Tabela 4.1), cada um deles equivalente a um período de 1.680 segundos (28 minutos).

Tabela 4.1. Número de Fluxos em cada *Trace* Coletado. (MOORE, et al., 2005a)

	NÚMERO DE FLUXOS
TRACE 01	24.863
TRACE 02	23.801
TRACE 03	22.932
TRACE 04	22.285
TRACE 05	21.648
TRACE 06	19.384
TRACE 07	55.835
TRACE 08	55.494
TRACE 09	66.248
TRACE 10	65.036
TOTAL	**377.526**

Os *traces* foram analisados em (ZUEV e MOORE, 2005), (MOORE e ZUEV, 2005b) e (AULD, et al., 2007), e também nesta pesquisa com a finalidade de se comparar os resultados obtidos. Durante o processo de classificação manual (Seção 3.2), para cada fluxo coletado foi identificada a aplicação à qual ele está associado. A Tabela 4.2 mostra as aplicações encontradas nos *traces* e suas classes correspondentes e a Tabela 4.3 indica o número de fluxos presentes em cada uma das classes de aplicações.

Tabela 4.2. Tipos de Aplicações contidas em cada Classe de Tráfego.

Classe da Aplicação	Exemplos de Aplicações Encontradas nos *Traces*
ATTACK	Virus e *Worm*
BULK	FTP
DATABASE	Postgres, Oracle e Ingres
GAMES	*Half-Life* e *Microsoft Direct Play*
INTERACTIVE	SSH, Telnet e Rlogin
MAIL	SMTP, IMAP e POP3
MULTIMEDIA	*Windows Media Player* e *Real Palyer*
P2P	KaZaA, BitTorrent e GnuTella
SERVICES	X11, DNS, LDAP e NTP
WWW	WWW

Tabela 4.3. Quantidade de Fluxos de cada Classe de Aplicações Presentes nos *Traces.*

CLASSES	NÚMERO DE FLUXOS
ATTACK	1.793
BULK	11.538
DATABASE	2.648
GAMES	8
INTERACTIVE	110
MAIL	28.567
MULTIMEDIA	576
P2P	2.094
SERVICES	2.099
WWW	328.093
TOTAL	**377.526**

Além das categorias de aplicações, durante o pré-processamento foi gerado para cada fluxo um conjunto de estatísticas relacionadas ao fluxo. Um total de 249 variáveis foi definido por (MOORE, et al., 2005a), incluindo estatísticas simples sobre o tamanho do pacote e o tempo entre pacotes, e informações derivadas do protocolo de transporte (TCP), tais como contadores de pacotes *Syn* e *Ack*. A definição de cada uma das 249 variáveis definidas em (MOORE, et al., 2005a) e utilizadas neste trabalho encontra-se no Anexo A.

A Figura 4.1 mostra a largura de banda e o tempo de coleta de cada um dos 10 arquivos de *traces*.

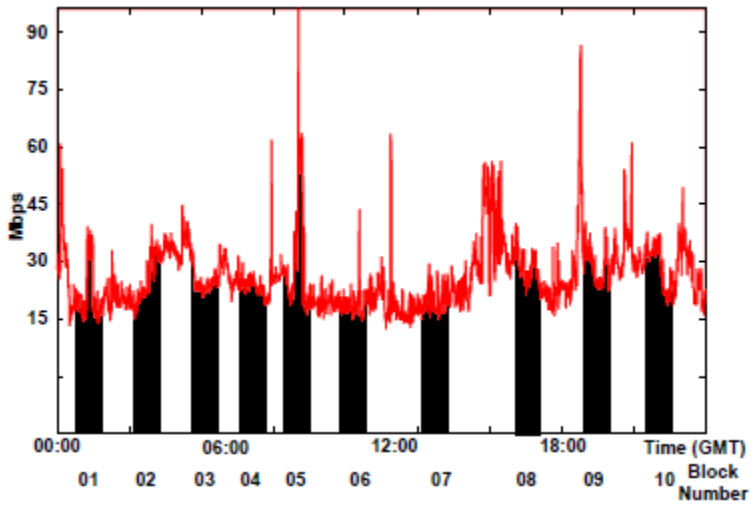

Figura 4.1. Banda de Transmissão Durante o Dia de Coleta. (MOORE, et al., 2005a)

4.2. Seleção das Variáveis Candidatas a Discriminantes Estatísticos

Após a fase de normalização (Seção 3.3.1) e remoção dos *outliers* (Seção 3.3.2), buscou-se entre o grupo de variáveis aquelas que melhor identificassem o tráfego da aplicação sob análise considerando-se a variância de seus valores com relação às outras aplicações. Conforme as Seções 4.2.1 e 4.2.2, esta tarefa foi realizada a partir da normalização dos dados, retirada dos *outliers* e do cálculo das Razões F dos grupos de amostras escolhidas aleatoriamente entre os fluxos da aplicação sob análise e das demais aplicações.

4.2.1. Normalização dos Dados e Eliminação dos *Outliers*

Para cada uma das 248 variáveis dos fluxos (a 249ª variável corresponde ao tipo de classe ao qual o fluxo pertence), foi realizado o processo de normalização dos dados e remoção dos *outliers* conforme definido nas Seções 3.3.1 e 3.3.2. Com relação aos *outliers*, foram removidos os fluxos com valores acima da distância de três desvios padrão, o que resultou na remoção de menos de 1% do total de fluxos.

4.2.2. Cálculo da Razão F nas Amostras de Dados

Utilizando então cada uma das 248 variáveis, nossa meta foi descobrir quais variáveis eram as melhores para se identificar o tráfego das aplicações listadas na Tabela 4.2. Conforme pode se observar na Tabela 4.3, as aplicações dos tipos GAMES, INTERACTIVE e MULTIMEDIA foram retiradas da análise devido à quantidade de fluxos das mesmas terem se mostrado irrelevantes em comparação ao total de fluxos das outras aplicações.

Foram agregados todos os fluxos em um único *trace*. Logo após foram selecionados grupos de amostras de fluxos para cada classe de aplicação sob análise (1500 fluxos) e demais aplicações (1500 fluxos), ou seja, foram compostos sete grupos de amostragem (um para cada aplicação sob análise) onde existiam 1500 fluxos amostrados da classe a ser analisada e 1500 fluxos amostrados das demais classes.

A seleção das amostras e separação das variáveis foi feita utilizando o software estatístico MINITAB (MINITAB, 2004). Foram listadas as 20 (vinte) variáveis que apresentaram maiores valores de Razão F (Anexo B), calculados através de um programa codificado na linguagem MATLAB (CHAPMAN, 2003). Na referência mais recente utilizada neste trabalho (AULD, et al., 2007) são utilizadas 28 (vinte e oito)

variáveis estatísticas para classificar em conjunto todas as classes de tráfego. Como um dos objetivos deste trabalho é a busca da menor quantidade possível de variáveis estatísticas que melhor discriminem padrões de fluxos do tráfego de uma determinada aplicação, optou-se pela seleção inicial das 20 (vinte) variáveis que apresentaram maiores valores de Razão F por se tratar de uma quantidade considerável de variáveis.

4.2.3. Cálculo da Correlação Linear entre as Variáveis de Maior Razão F

Após a escolha das 20 variáveis que apresentaram maiores valores de Razão F aplicou-se a Equação (2.6) aos 1500 fluxos de cada classe, através de um programa codificado na linguagem MATLAB (CHAPMAN, 2003), para calcular a correlação linear entre cada uma dessas variáveis e as demais. Esta análise se torna necessária para eliminar a redundância de variáveis que possuam características (medidas estatísticas) semelhantes.

A relação completa dos valores da correlação linear entre as 20 (vinte) variáveis candidatas para cada classe de aplicação se encontra listada no Anexo C. O resultado da análise da correlação linear destas 20 variáveis para cada uma das classes é mostrado a seguir (valores absolutos).

Tabela 4.4. Correlação Linear – ATTACK.

	VAR002	**VAR081**	**VAR171**	**VAR192**	**VAR231**
VAR002	1	0,36549	0,14477	0,42144	0,31179
VAR081	0,36549	1	0,12339	0,36841	0,40555
VAR171	0,14477	0,12339	1	0,36354	0,03515
VAR192	0,42144	0,36841	0,36354	1	0,04808
VAR231	0,31179	0,40555	0,03515	0,04808	1

Pode-se observar na Tabela 4.4 que as variáveis menos correlacionadas em relação à classe ATTACK são as variáveis 171 e 231, descritas na Tabela 4.5.

Tabela 4.5. Variáveis Candidatas – ATTACK.

VAR002	Número da porta de conexão no cliente
VAR081	Tamanho máximo de segmento observado durante o tempo da conexão (direção "cliente→servidor")
VAR171	Terceiro quartil dos bytes de controle no pacote
VAR192	Terceiro quartil dos bytes de controle no pacote na direção "servidor→cliente"
VAR231	Arco-tangente das dez maiores frequências escolhidas pela magnitude de suas contribuições (direção "cliente→servidor") – Frequência #3

Tabela 4.6. Correlação Linear – DATABASE.

	VAR082	**VAR083**	**VAR085**	**VAR089**	**VAR090**	**VAR190**
VAR082	1	0,15917	0,26695	0,37443	0,46459	0,50723
VAR083	0,15917	1	0,53569	0,15403	0,31846	0,29694
VAR085	0,26695	0,53569	1	0,23385	0,30168	0,34139
VAR089	0,37443	0,15403	0,23385	1	0,34735	0,35746
VAR090	0,46459	0,31846	0,30168	0,34735	1	0,43543
VAR190	0,50723	0,29694	0,34139	0,35746	0,43543	1

Na Tabela 4.6 é mostrado que as variáveis menos correlacionadas em relação à classe DATABASE são as variáveis 82 e 83. As mesmas estão descritas na Tabela 4.7.

Tabela 4.7. Variáveis Candidatas – DATABASE.

VAR082	Tamanho máximo de segmento observado durante o tempo da conexão (direção "servidor→cliente")
VAR083	Tamanho mínimo de segmento observado durante o tempo da conexão (direção "cliente→servidor")
VAR085	Média do tamanho de segmento observado durante o tempo de vida da conexão calculado como o valor total dos bytes de dados dividido pelo número de pacotes (direção "cliente→servidor")
VAR089	Tamanho mínimo de janela na direção "cliente→servidor"
VAR090	Tamanho mínimo de janela na direção "servidor→cliente"
VAR190	Mediana dos bytes de controle no pacote na direção "servidor→cliente"

Tabela 4.8. Correlação Linear – FTP.

	VAR001	VAR095	VAR189	VAR213
VAR001	1	0,21323	0,16925	0,136
VAR095	0,21323	1	0,55716	0,50471
VAR189	0,16925	0,55716	1	0,56539
VAR213	0,136	0,50471	0,56539	1

A Tabela 4.8 apresenta as variáveis menos correlacionadas para a classe FTP (variáveis 01 e 213) e a Tabela 4.9 descreve seus significados.

Tabela 4.9. Variáveis Candidatas – FTP.

VAR001	Número da porta de conexão no servidor
VAR095	Número total de bytes enviados na janela inicial, i.e, o número de bytes de dados enviados antes de receber o primeiro pacote ACK. Pacotes retransmitidos não serão levados em conta (direção "cliente→servidor")
VAR189	Primeiro quartil dos bytes de controle no pacote na direção "servidor→cliente"
VAR213	Arco-tangente das dez maiores frequências escolhidas pela magnitude de suas contribuições (direção "cliente→servidor") – Frequência #3

Tabela 4.10. Correlação Linear – MAIL.

	VAR026	VAR083	VAR164	VAR170	VAR179	VAR231
VAR026	1	0,31777	0,29919	0,591	0,11822	0,5361
VAR083	0,31777	1	0,02384	0,17003	0,19267	0,22639
VAR164	0,29919	0,02384	1	0,53906	0,23107	0,31049
VAR170	0,591	0,17003	0,53906	1	0,21242	0,33113
VAR179	0,11822	0,19267	0,23107	0,21242	1	0,09879
VAR231	0,5361	0,22639	0,31049	0,33113	0,09879	1

Na Tabela 4.10, pode-se notar que as variáveis menos correlacionadas em relação à classe MAIL são as variáveis 83 e 164. Tais variáveis são descritas na Tabela 4.11.

Tabela 4.11. Variáveis Candidatas – MAIL.

VAR026	Mediana dos bytes de controle no pacote (tamanho do cabeçalho IP/TCP)
VAR083	Tamanho mínimo de segmento observado durante o tempo da conexão (direção "cliente→servidor")
VAR164	Terceiro quartil do número total de bytes no pacote IP
VAR170	Média dos bytes de controle no pacote
VAR179	Máximo de bytes no pacote Ethernet na direção "servidor→cliente"
VAR231	Arco-tangente das dez maiores frequências escolhidas pela magnitude de suas contribuições (direção "cliente→servidor") – Frequência #3

Tabela 4.12. Correlação Linear – P2P.

	VAR110	**VAR117**	**VAR166**	**VAR172**	**VAR221**
VAR110	1	0,26143	0,13102	0,06086	0,35341
VAR117	0,26143	1	0,0018	0,05696	0,28233
VAR166	0,13102	0,0018	1	0,12005	0,12586
VAR172	0,06086	0,05696	0,12005	1	0,07012
VAR221	0,35341	0,28233	0,12586	0,07012	1

Ao observar a Tabela 4.12, nota-se que as variáveis menos correlacionadas em relação à classe P2P são as variáveis 117 e 166. A Tabela 4.13 apresenta o significado destas variáveis.

Tabela 4.13. Variáveis Candidatas – P2P.

VAR110	Máximo tempo ocioso, calculado como o tempo máximo entre dois pacotes consecutivos na mesma direção (direção "servidor→cliente")
VAR117	Número máximo de RTT na direção "cliente→servidor"
VAR166	Variância do total de bytes no pacote IP
VAR172	Máximo dos bytes de controle no pacote
VAR221	Arco-tangente das dez maiores frequências escolhidas pela magnitude de suas contribuições (todo o tráfego) – Frequência #3

Tabela 4.14. Correlação Linear – SERVICES.

	VAR158	VAR190	VAR194	VAR221
VAR158	1	0,52525	0,32513	0,48973
VAR190	0,52525	1	0,5328	0,5462
VAR194	0,32513	0,5328	1	0,38836
VAR221	0,48973	0,5462	0,38836	1

As variáveis menos correlacionadas em relação à classe SERVICES são mostradas na Tabela 4.14 (variáveis 158 e 194). Na Tabela 4.15 é descrito o significado de cada uma delas.

Tabela 4.15. Variáveis Candidatas – SERVICES.

VAR158	Número máximo de bytes no pacote Ethernet
VAR190	Mediana dos bytes de controle no pacote na direção "servidor→cliente"
VAR194	Variância dos bytes de controle no pacote na direção "servidor→cliente"
VAR221	Arco-tangente das dez maiores frequências escolhidas pela magnitude de suas contribuições (todo o tráfego) – Frequência #3

Tabela 4.16. Correlação Linear – WWW.

	VAR017	VAR083	VAR084	VAR096
VAR017	1	0,29477	0,2958	0,32786
VAR083	0,29477	1	0,0482	0,26695
VAR084	0,2958	0,0482	1	0,53645
VAR096	0,32786	0,26695	0,53645	1

Na Tabela 4.16 as variáveis menos correlacionadas em relação à classe WWW (variáveis 83 e 84) são mostradas, enquanto a Tabela 4.17 apresenta o significado de cada uma delas.

Tabela 4.17. Variáveis Candidatas – WWW.

VAR017	Mínimo do total de bytes no pacote IP
VAR083	Tamanho mínimo de segmento observado durante o tempo da conexão (direção "cliente→servidor")
VAR084	Tamanho mínimo de segmento observado durante o tempo da conexão (direção "servidor→cliente")
VAR096	Número total de bytes enviados na janela inicial, i.e, o número de bytes de dados enviados antes de receber o primeiro pacote ACK. Pacotes retransmitidos não serão levados em conta (direção "servidor→cliente")

4.3. Treinamento e Definição dos Discriminantes Estatísticos

Na fase de treinamento, foram utilizadas as amostras de fluxos geradas na etapa de seleção das variáveis candidatas (Seção 4.2) com o intuito de descobrir as coordenadas dos centróides relativos às classes de aplicações a serem analisadas e os centróides das demais classes de aplicações.

Depois de identificadas as variáveis candidatas, utilizando o algoritmo das K-médias e baseado nos valores da correlação linear, as mesmas foram testadas de forma iterativa, de acordo com a Seção 3.4, para medir o grau de separação causado aos fluxos de cada classe de aplicação sob análise. A partir daí foi identificado e determinado o conjunto de variáveis discriminantes para cada classe de aplicação.

Tabela 4.18. Variáveis Discriminantes – ATTACK.

Variável Candidata	**Descrição**	**% Exatidão Média da Aplicação**
171	Terceiro quartil dos bytes de controle no pacote	98,07
231	Arco-tangente das dez maiores frequências escolhidas pela magnitude de suas contribuições (direção "cliente→servidor") – Frequência #3	98,07
192	Terceiro quartil dos bytes de controle no pacote na direção "servidor→cliente"	87,40

Pode-se perceber através da Tabela 4.18 que a variável 171 individualmente conseguiu discriminar 98,07% do tráfego da classe ATTACK em relação às demais

classes. Ao acrescentar a variável 231 o poder de discriminação não foi alterado. E ao se acrescentar uma terceira variável, obteve-se uma queda no poder de discriminação dos fluxos da classe.

Tabela 4.19. Variáveis Discriminantes – DATABASE.

Variável Candidata	Descrição	% Exatidão Média da Aplicação
83	Tamanho mínimo de segmento observado durante o tempo da conexão (direção "cliente→servidor")	100%
82	Tamanho máximo de segmento observado durante o tempo da conexão (direção "servidor→cliente")	70%

A Tabela 4.19 mostra que a variável 83 individualmente foi suficiente para discriminar 100% do tráfego da classe DATABASE em relação às demais classes de aplicações. Ao acrescentar a variável 82, o que se obteve foi uma queda considerável no poder de discriminação.

Tabela 4.20. Variáveis Discriminantes – FTP.

Variável Candidata	Descrição	% Exatidão Média da Aplicação
213	Arco-tangente das dez maiores frequências escolhidas pela magnitude de suas contribuições (direção "cliente→servidor") – Frequência #3	100%
01	Número da porta de conexão no servidor	99,47%

Na Tabela 4.20 pode-se observar que uma única variável foi suficiente para discriminar 100% do tráfego da classe FTP. Ao acrescentar outra variável, o poder de discriminação diminuiu para 99,47%.

Tabela 4.21. Variáveis Discriminantes – MAIL.

Variável Candidata	Descrição	% Exatidão Média da Aplicação
83	Tamanho mínimo de segmento observado durante o tempo da conexão (direção "cliente→servidor")	100%
164	Terceiro quartil do número total de bytes no pacote IP	92,47%

Ao observar a Tabela 4.21, nota-se que uma única variável foi suficiente para discriminar 100% do tráfego da classe MAIL. O acréscimo de outra variável diminuiu o poder de discriminação para 92,47%.

Tabela 4.22. Variáveis Discriminantes – P2P.

Variável Candidata	Descrição	% Exatidão Média da Aplicação
166	Variância do total de bytes no pacote IP	92,93%
117	Número máximo de RTT na direção "cliente→servidor"	92,93%
221	Arco-tangente das dez maiores frequências escolhidas pela magnitude de suas contribuições (todo o tráfego) – Frequência #3	92,93%

Observando a Tabela 4.22, nota-se que uma única variável foi suficiente para discriminar 92,93% do tráfego da classe P2P e que o acréscimo de variáveis não implicou em alteração no poder de discriminação dos fluxos da classe.

Tabela 4.23. Variáveis Discriminantes – SERVICES.

Variável Candidata	Descrição	% Exatidão Média da Aplicação
158	Número máximo de bytes no pacote Ethernet	99,47%
194	Variância dos bytes de controle no pacote na direção "servidor→cliente"	99,47%
221	Arco-tangente das dez maiores frequências escolhidas pela magnitude de suas contribuições (todo o tráfego) – Frequência #3	99,47%

Na Tabela 4.23 é possível observar que uma única variável foi suficiente para discriminar 99,47% do tráfego da classe SERVICES e que o acréscimo de variáveis não implicou em alteração no poder de discriminação dos fluxos da classe.

Tabela 4.24. Variáveis Discriminantes – WWW.

Variável Candidata	Descrição	% Exatidão Média da Aplicação
83	Tamanho mínimo de segmento observado durante o tempo da conexão (direção "cliente→servidor")	79,47%
84	Tamanho mínimo de segmento observado durante o tempo da conexão (direção "servidor→cliente")	85,33%
17	Mínimo do total de bytes no pacote IP	85,33%
96	Número total de bytes enviados na janela inicial, i.e, o número de bytes de dados enviados antes de receber o primeiro pacote ACK. Pacotes retransmitidos não serão levados em conta (direção "servidor→cliente")	43,40%

Pode ser observado na Tabela 4.24 a evolução do poder de discriminação das variáveis candidatas em relação à classe WWW. Ao agregar as variáveis 83 e 84 ocorreu um aumento de 79,47% para 85,33% no percentual de acerto na separação dos fluxos da classe. Ao adicionar mais uma variável o poder de discriminação não sofreu impacto, mas ao acrescentar uma quarta variável ao conjunto de discriminantes percebeu-se uma sensível redução no poder de separação dos fluxos.

Após chegar ao conjunto de variáveis que melhor discriminassem os fluxos da classe sob análise, as coordenadas dos centróides para cada conjunto de problemas de classificação 1 para 1 foram geradas. Estas coordenadas são mostradas nas tabelas a seguir (Tabelas 4.25 a 4.31).

Tabela 4.25. Centróides dos Agrupamentos – ATTACK.

	VAR171
AGRUPAMENTO DA CLASSE	-0,0335
AGRUPAMENTO DAS DEMAIS CLASSES	0,1165

Tabela 4.26. Centróides dos Agrupamentos – DATABASE.

	VAR083
AGRUPAMENTO DA CLASSE	-0,3546
AGRUPAMENTO DAS DEMAIS CLASSES	0,8853

Tabela 4.27. Centróides dos Agrupamentos – FTP.

	VAR213
AGRUPAMENTO DA CLASSE	-0,3667
AGRUPAMENTO DAS DEMAIS CLASSES	1,5268

Tabela 4.28. Centróides dos Agrupamentos – MAIL.

	VAR083
AGRUPAMENTO DA CLASSE	-0,3297
AGRUPAMENTO DAS DEMAIS CLASSES	1,0792

Tabela 4.29. Centróides dos Agrupamentos – P2P.

	VAR166
AGRUPAMENTO DA CLASSE	-0,2777
AGRUPAMENTO DAS DEMAIS CLASSES	0,6369

Tabela 4.30. Centróides dos Agrupamentos – SERVICES.

	VAR158
AGRUPAMENTO DA CLASSE	-0,4207
AGRUPAMENTO DAS DEMAIS CLASSES	0,4657

Tabela 4.31. Centróides dos Agrupamentos – WWW.

	VAR083	**VAR084**
AGRUPAMENTO DA CLASSE	0,7499	0,6856
AGRUPAMENTO DAS DEMAIS CLASSES	-0,6597	-0,6031

4.4. Validação

Na fase de validação foi selecionado um conjunto de fluxos classificados manualmente diferente daquele que foi usado na etapa de treinamento (1500 fluxos amostrados da

classe a ser analisada e 1500 fluxos amostrados das demais classes). Logo após, aplicou-se um algoritmo para associar cada fluxo ao centróide (definido na fase de treinamento) mais próximo através da distância euclidiana. O cálculo da distância de cada fluxo para os centróides e a associação a cada um dos agrupamentos foi realizado utilizando um programa codificado na linguagem MATLAB (CHAPMAN, 2003).

Foram gerados dois agrupamentos, um para a classe a ser analisada e outro para as demais classes. Os resultados desta fase são mostrados na Tabela 4.32.

Tabela 4.32. Resultados da Fase de Validação.

	ATTACK	DATABASE	FTP	MAIL	P2P	SERVICES	WWW	MÉDIA
% EXATIDÃO MÉDIA	63,93%	78,60%	33,83%	71,40%	71,30%	86,13%	85,40%	**70,08%**
% EXATIDÃO MÉDIA DA APLICAÇÃO	91,60%	100,00%	64,47%	98,00%	90,93%	88,67%	82,20%	**87,98%**
% FALSOS POSITIVOS	41,03%	29,97%	60,02%	36,03%	34,71%	15,61%	12,18%	**32,79%**
% FALSOS NEGATIVOS	8,40%	0,00%	35,53%	2,00%	9,07%	11,33%	17,80%	**12,02%**

A Tabela 4.32 mostra que a fase de validação confirmou a escolha acertada das variáveis utilizadas como discriminantes. Isto pode ser percebido ao analisar os valores do percentual de exatidão média da aplicação para cada classe de tráfego que se assemelham aos valores apresentados nas Tabelas 4.18 a 4.24 (com exceção a classe FTP).

4.5. Classificação

Na fase de classificação, aplica-se o programa classificador a um novo conjunto de fluxos (1500 fluxos amostrados da classe a ser analisada e 1500 fluxos amostrados das demais classes) associando cada fluxo do *trace* ao agrupamento com menor valor de distância euclidiana para seu centróide. Logo após verifica-se a precisão da classificação através das métricas definidas na Seção 3.5. Os resultados da classificação nos agrupamentos são apresentados através das matrizes de confusão mostradas no Anexo D e resumidos na Tabela 4.33.

Tabela 4.33. Resultados da Fase de Classificação.

	ATTACK	DATABASE	FTP	MAIL	P2P	SERVICES	WWW	MÉDIA
% EXATIDÃO MÉDIA	71,67%	74,73%	75,37%	79,47%	59,50%	78,07%	82,00%	**74,40%**
% EXATIDÃO MÉDIA DA APLICAÇÃO	97,93%	100,00%	100,00%	100,00%	84,93%	99,40%	75,87%	**94,02%**
% FALSOS POSITIVOS	35,80%	33,57%	33,01%	29,11%	43,70%	30,33%	13,53%	**31,29%**
% FALSOS NEGATIVOS	2,07%	0,00%	0,00%	0,00%	15,07%	0,60%	24,13%	**5,98%**

Pode-se observar pelos resultados da Tabela 4.33 que os valores das taxas de exatidão média, exatidão média da aplicação, falsos positivos e falsos negativos se mantiveram próximos e em alguns casos superiores aos valores obtidos pelas análises do conjunto de fluxos analisados na fase de validação utilizando-se os mesmos discriminantes estatísticos.

4.6. Comparação dos Resultados

Com relação aos resultados do método NAIVE BAYES aplicado ao mesmo conjunto de fluxos por (ZUEV e MOORE, 2005), do método FCBF+NAIVE BAYES utilizado por (MOORE e ZUEV, 2005b) e do método BAYESIAN NEURAL aplicado por (AULD, et al., 2007), é apresentado na Tabela 4.34 um comparativo com os resultados da abordagem apresentada neste trabalho (UM-CONTRA-TODOS).

Tabela 4.34. Percentual de Exatidão Média e Exatidão Média das Aplicações Comparadas entre Si.

MÉTODO	Exatidão Média	Exatidão Média ATTACK	Exatidão Média DATABASE	Exatidão Média FTP	Exatidão Média MAIL	Exatidão Média P2P	Exatidão Média SERVICES	Exatidão Média WWW	Média da Exatidão Média das Aplicações
NAIVE BAYES	66,71%	58,08%	20,20%	89,26%	56,85%	45,59%	91,19%	65,97%	61,02%
FCBF + NAIVE BAYES	96,29%	NA[11]	68,63%	90,03%	87,54%	55,18%	44,54%	98,06%	74,00%
BAYESIAN NEURAL	68,80%	68,60%	97,60%	97,40%	99,60%	62,00%	96,00%	99,80%	88,72%
UM-CONTRA-TODOS	74,40%	97,93%	100,00%	100,00%	100,00%	84,93%	99,40%	75,87%	94,02%

Nota-se analisando a Tabela 4.34 que para a maioria das classes de tráfego foram obtidos valores de exatidão média da aplicação superiores aos dos demais métodos. É importante frisar que as classes ATTACK e P2P constituem os tipos de tráfego mais difíceis de serem identificados devido às características e comportamentos dos mesmos serem semelhantes muitas vezes ao tráfego WWW. Devido a esta semelhança de comportamento os fluxos destas três aplicações muitas vezes se misturam entre si ao serem agrupados.

Vale ressaltar que apesar do método FCBF+NAIVE BAYES utilizado por (MOORE e ZUEV, 2005b) obter uma exatidão média significativamente maior (96,29%), em seu trabalho foram utilizadas 10 variáveis para discriminar os fluxos de todas as aplicações analisadas em conjunto. Neste trabalho as classes de aplicações foram analisadas separadamente e, para a maioria delas, foi utilizada apenas uma única variável discriminante (com exceção da classe WWW que fez uso de duas variáveis) e obtido um alto percentual médio de acerto na classificação. A Tabela 4.35 mostra o número de discriminantes utilizados em cada método comparado.

[11] Não aplicável.

Tabela 4.35. Número de Discriminantes Utilizados em Cada Método Analisado.[12]

MÉTODO	NÚMERO DE DISCRIMINANTES UTILIZADOS
BAYESIAN NEURAL	28
FCBF + NAIVE BAYES	10
UM-CONTRA-TODOS	6

Nas Tabelas 4.36 e 4.37 são comparadas as informações sobre falsos positivos e falsos negativos das pesquisas de (AULD, et al., 2007). As pesquisas de (ZUEV e MOORE, 2005) e (MOORE e ZUEV, 2005b) não apresentaram tais informações.

Tabela 4.36. Número de Falsos Positivos.

Falsos Positivos								
MÉTODO	**ATTACK**	**DATABASE**	**FTP**	**MAIL**	**P2P**	**SERVICES**	**WWW**	**Média**
BAYESIAN NEURAL	10,60%	0,40%	2,00%	2,90%	53,00%	1,20%	13,40%	11,93%
UM-CONTRA-TODOS	35,80%	33,57%	33,01%	29,11%	43,70%	30,33%	13,53%	31,29%

Tabela 4.37. Número de Falsos Negativos.

Falsos Negativos								
MÉTODO	**ATTACK**	**DATABASE**	**FTP**	**MAIL**	**P2P**	**SERVICES**	**WWW**	**Média**
BAYESIAN NEURAL	58,90%	1,00%	1,80%	2,30%	5,20%	1,20%	10,40%	11,54%
UM-CONTRA-TODOS	2,07%	0,00%	0,00%	0,00%	15,07%	0,60%	24,13%	5,98%

Embora os valores de falsos positivos não tenham se mostrados satisfatórios, foram obtidos valores de falsos negativos abaixo da metodologia de (AULD, et al., 2007), com exceção as classes P2P e WWW.

[12] No método NAIVE BAYES apresentado por (ZUEV e MOORE, 2005) não foi mostrada a quantidade de variáveis utilizadas na classificação.

4.7. Resumo do Capítulo

Neste capítulo utilizou-se a metodologia descrita no capítulo 3, onde a mesma foi aplicada a um conjunto de fluxos reais para testar a sua viabilidade prática.

Os resultados mostraram-se compatíveis com as recentes pesquisas na área de classificação de aplicações Internet com as vantagens de um menor número de variáveis necessárias para a identificação de uma determinada classe de aplicações e a não utilização do número de porta de serviços e informações do conteúdo dos pacotes. Tais vantagens são decorrência de uma abordagem até então nunca utilizada na classificação de aplicações Internet (abordagem um-contra-todos – *one-against all*).

O próximo capítulo é dedicado às conclusões da presente pesquisa e sugestões de trabalhos futuros.

Capítulo 5 – Conclusão e Trabalhos Futuros

Este capítulo apresenta as conclusões deste trabalho, incluindo suas principais contribuições e propostas para trabalhos futuros.

Este trabalho propôs uma nova abordagem para a criação de um classificador de classes de aplicações presentes no tráfego Internet. A metodologia foi dividida em cinco fases:

- Reconstrução dos fluxos e classificação manual;
- Seleção das variáveis candidatas a discriminantes estatísticos;
- Treinamento e definição dos discriminantes;
- Validação;
- Classificação.

No capítulo de resultados foi apresentada a classificação das diferentes classes de aplicações e realizou-se uma comparação com outros métodos encontrados na literatura. Mostrou-se que, com a redução do número de variáveis necessárias para a identificação dos fluxos, a exatidão média da classificação de cada aplicação não sofreu grandes perdas, tendo, na realidade, melhorado em alguns casos em relação aos outros métodos.

Pesquisas anteriores utilizavam um mesmo conjunto de variáveis para discriminar todas as classes de tráfego, o que levava a uma menor exatidão quando da classificação específica de uma classe. A metodologia apresentada parte do princípio de que o melhor conjunto de variáveis para classificar uma determinada aplicação não é o mesmo para classificar N aplicações. A principal contribuição deste trabalho, comparativamente a outros métodos estudados, foi apresentar a utilização de uma abordagem (um-contra-todos – *one-against-all*) que, até então, não havia sido utilizada para realizar classificação de tráfego de aplicações Internet.

A consequência da utilização desta abordagem foi a redução do número de variáveis (*features*) a serem analisadas obtendo uma média de acerto na classificação dos fluxos das classes sob análise de 74,40% e média de falsos negativos de 5,98%. Isso

pode ser constatado ao adicionar mais variáveis ao conjunto de variáveis para se gerar os agrupamentos. Notou-se que a adição de novas variáveis em geral não aumentou o poder de separação dos fluxos.

Este trabalho também se diferencia de diversas pesquisas anteriores que utilizavam análise de portas de serviço como o único discriminante para identificação de tráfego de aplicações Internet e de métodos de aprendizagem de máquina que utilizavam conjuntos de dados sem classificação prévia.

Como citado anteriormente, foram obtidos resultados equivalentes às mais recentes pesquisas na área de identificação de aplicações com um número menor de informações. A metodologia descrita neste trabalho apresentou uma redução do número de variáveis (*features*) a serem analisadas através de um método estatístico computacionalmente simples, que pode ser utilizado em outros conjuntos de dados (*traces*).

Como trabalhos futuros iniciados por esta pesquisa incluem-se a aplicação da metodologia em *traces* coletados de outros *backbones*, a aplicação de outras medidas de distância (distância ponderada, distância de Minkowski, coeficiente de concordância simples, coeficiente de concordância de Jaccard, etc.) e a aplicação de outros métodos de estatística multivariada para a separação dos fluxos, como a análise discriminante.

Referências Bibliográficas

ABBES, T., BOUHOULA, A. e RUSINOWITCH, M. 2004. Protocol Analysis in Intrusion Detection Using Decision Tree. *International Conference on Information Technology: Coding and Computing - ITCC.* Washington, EUA : IEEE Computer Society Press, 2004. pp. 404-409.

ANDERSON, D. R., SWEENEY, D. J. e WILLIAMS, T. A. 2007. *Estatística Aplicada à Administração e Economia.* 2ª Ed. São Paulo : Thomson, 2007. p. 616.

ANDERSON, T. W. 2003. *An Introduction to Multivariate Statistical Analysis.* 3ª Ed. Nova York : John Wiley & Sons, 2003. p. 752.

AULD, T., MOORE, A. W. e GULL, S. F. 2007. Bayesian Neural Networks for Internet Traffic Classification. *IEEE Transactions on Neural Networks.* Janeiro de 2007, Vol. 18, pp. 223-239.

BARFORD, P., et al. 2002. A Signal Analysis of Network Traffic Anomalies. *Workshop on Internet Measurement.* Marselha, França : ACM Special Interest Group on Data Communications - SIGCOMM, 2002. pp. 71-82.

BERNAILLE, L., et al. 2006. Traffic Classification on the Fly. *Computer Communication Review.* Abril de 2006, Vol. 36, pp. 23-26.

BEYGELZIMER, A., LANGFORS, J. e ZADROZNY, B. 2005. Weighted One-Against-All. *Conference on Innovative Applications of Artificial Intelligence.* Pittsburgh, EUA : AAAI-05 / IAAI-05, Julho de 2005. pp. 720-725.

BROWNLEE, N. e CLAFFY, K. 2002. Understanding Internet Traffic Streams: Dragonflies and Tortoises. *IEEE Communications Magazine.* 2002, Vol. 40, pp. 110-117.

BRUTLAG, J. 2000. Aberrant Behavior Detection in Timeseries for Network Monitoring. *USENIX Conference on System Administration.* Nova Orleans, EUA : USENIX Association, 2000. pp. 139-146.

CABRERA, J. B. D., et al. 2002. Proactive Intrusion Detection and Distributed Denial of Service Attacks - A Case Study in Security Management. *Journal of Network and Systems Management.* 2002, Vol. 10, pp. 225-254.

CAIDA. 2009. *CAIDA: The Cooperative Association for Internet Data Analysis.* [Online] 2009. [Citado em: 07 de Maio de 2009.] http://www.caida.org.

CARMO, M. F. F., et al. 2007a. Using Statistical Discriminators and Cluster Analysis to P2P and Attack Traffic Monitoring. *Latin American Network Operations and Management Symposium - LANOMS.* Petrópolis, Brasil : IEEE/IFIP, 2007a. pp. 67-75.

CARMO, M.F. F., et al. 2007b. Attack Detection based on Statistical Discriminators. *Global Information Infrastructure Symposium - GIIS.* Marrakech, Marrocos : IEEE Computer Society Press, Julho de 2007b. pp. 181-186.

CHAPMAN, S. J. 2003. *Programação em MATLAB para Engenheiros.* São Paulo : Thomson, 2003. p. 481.

CLAFFY, K. 1994. Internet Traffic Characterization. *Tese de Doutorado.* San Diego, EUA : Universidade da Califórnia, 1994. p. 135.

CRESPO, A. A. 1996. *Estatística Fácil.* 14ª Ed. São Paulo : Saraiva, 1996. p. 224.

DUDA, R.O. e HART, P.E. 2000. *Pattern Classification.* 2ª Ed. Nova York : John Wiley & Sons, 2000. p. 654.

ERMAN, J., ARLITT, M. e MAHANTI, A. 2006a. Traffic Classification using Cluster Algorithms. *International Conference on Measurement and Modeling of Computer Systems - Workshop on Mining Network Data.* Pisa, Itália : ACM Special Interest Group on Data Communications - SIGCOMM, 2006a. pp. 281-286.

ERMAN, J., et al. 2007. Identifying and Discrimination between Web and Peer-to-Peer Traffic in the Network Core. *International World Wide Web Conference - WWW.* Banff, Canadá : ACM, Maio de 2007. pp. 883-892.

ERMAN, J., MAHANTI, A. e ARLITT, M. 2006b. Internet Traffic Identification using Machine Learning. *Global Telecommunications - GLOBECOM.* San Francisco, EUA : IEEE Computer Society Press, Novembro de 2006b. pp. 1-6.

EVERITT, B, LANDAU, S. e LEESE, M. 2001. *Cluster Analysis.* 4ª Ed. Nova York : Arnold, 2001. p. 237.

FOMENKOV, M., et al. 2004. Longitudinal Study of Internet Traffic in 1998-2003. *Winter International Synposium on Information and Communication Technologies - WISICT.* Cancun, México : Trinity College Dublin, 2004. pp. 1-6.

FRALEIGH, C., et al. 2003. Packet-level Traffic Measurements from the Sprint IP Backbone. *IEEE Network.* Junho de 2003, Vol. 17, pp. 6-16.

GUO, L. e MATTA, I. 2001. The War Between Mice and Elephants. *Technical Report BU-CS-2001-005.* Boston, EUA : Boston University, 2001.

HOLANDA, R. 2005. A New Methodology for Packet Trace Classification and Compression based on Semantic Traffic Characterization. *Tese (Doutorado em Ciência da Computação).* Catalunha, Espanha : Universidade Politécnica da Catalunha - UPC, Setembro de 2005.

HOLANDA, R., et al. 2008. An Internet Traffic Classification Methodology based on Statistical Discriminators. *Network Operations and Management Symposium - NOMS.* Salvador, Brasil : IEEE, 2008. pp. 907-910.

HOLANDA, R., MAIA, J. E. B. e CARMO, M. F. F. 2007a. Detecting Computer Network Attacks using Statistical Discriminators and Cluster Analysis. *Revista Tecnologia (Universidade de Fortaleza - UNIFOR).* Junho de 2007a, Vol. 28, pp. 33-41.

—. 2007b. Identificação da Componente de Tráfego de Ataque baseada em Discriminantes Estatísticos. *Workshop em Desempenho de Sistemas Computacionais e de Comunicação - WPerformance.* Rio de Janeiro, Brasil : Anais do XXVII Congresso da Sociedade Brasileira de Computação - CSBC, 2007b. pp. 690-702.

—. 2007c. Seleção de Discriminantes Estatísticos para Identificação de Tráfego de Ataques. *Workshop de Gerência e Operação de Redes e Serviços - WGRS.* Belém, Brasil : Anais do XXV Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos - SBRC, 2007c. pp. 63-74.

HOLANDA, R., MAIA, J. E. B. e SIQUEIRA, G. P. 2006. Broadband Network Traffic Characterization and Classification using a Multivariate Statistical Method. *Revista Tecnologia (Universidade de Fortaleza - UNIFOR).* Dezembro de 2006, Vol. 27, pp. 113-122.

HUSSAIN, A., HEIDEMANN, J. e PAPADOPOULOS, C. 2003. A Framework for Classifying Denial of Service Attacks. *Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication.* Karlsruhe, Alemanha : ACM Special Interest Group on Data Communication - SIGCOMM, 2003. pp. 99-110.

IANA. 2009. *Internet Assigned Numbers Authority - IANA.* [Online] 2009. [Citado em: 07 de Maio de 2009.] http://www.iana.org/.

JAIN, R. 1991. *The Art of Computer Systems Performance Analysis.* Nova York : John Wiley & Sons, 1991. p. 720.

JOHNSON, D. 1998. *Applied Multivariate Methods for Data Analysis.* Belmont : Brooks/Cole, 1998. p. 567.

JUNG, J., et al. 2004. Fast Portscan Detection Using Sequential Hypothesis Testing. *Symposium on Security and Privacy.* Oakland, EUA : IEEE Computer Society Press, Maio de 2004. pp. 211-225.

KARAGIANNIS, T., et al. 2004. Transport Layer Identification of P2P Traffic. *Internet Measurement Conference - IMC.* Taormina, Itália : ACM Special Interest Group on Data Communications - SIGCOMM, 2004. pp. 121-134.

KARAGIANNIS, T., PAPAGIANNAKI, K. e FALOUTSOS, M. 2005. BLINC: Multilevel Traffic Classification in the Dark. *Conference on Applications, Technologies, Architectures and Protocols for Computer Communication.* Filadélfia, EUA : ACM Special Interest Group on Data Communication - SIGCOMM, Agosto de 2005. pp. 229-240.

KAUFMAN, L. e ROUSSEEUW, P. 1990. *Finding Groups in Data: An Introduction to Cluster Analysis.* Nova York : John Wiley & Sons, 1990. p. 335.

KEYS, K., et al. 2001. The Architecture of CoralReef: An Internet Traffic Monitoring Software Suite. *Passive and Active Measurement Workshop - PAM.* Amsterdam, Holanda : Springer, Abril de 2001.

KIM, H. A. e KARP, B. 2004. Autograph: Toward Automated Distributed Worm Signature Detection. *USENIX Security Symposium.* San Diego, EUA : USENIX, 2004. pp. 271-286.

KIM, H., et al. 2007. Comparison of Internet Traffic Classification Tools. *Workshop on Application Classification and Identification - WACI.* Califórnia, EUA : Internet Measurement Research Group - IMRG, Outubro de 2007.

KIM, H., et al. 2008. Internet Traffic Classification Demystified: Myths, Caveats and the Best Practices. *CoNEXT 2008.* Madrid, Espanha : ACM, Dezembro de 2008.

KIM, M. S., et al. 2004. A Flow-Based Method for Abnormal Network Traffic Detection. *Network Operations and Management Symposium - NOMS.* Seoul, Coréia do Sul : IEEE Computer Society Press, 2004. pp. 599-612.

KOHLER, E., et al. 2006. Observed Structure of Address in IP Traffic. *IEEE/ACM Transactions on Networking - TON.* Dezembro de 2006, Vol. 14, pp. 1207-1218.

LAKHINA, A., CROVELLA, M. e DIOT, C. 2004a. Characterization of Network-Wide Anomalies in Traffic Flows. *Internet Measurement Conference.* Taormina, Itália : ACM Special Interest Group on Data Communication - SIGCOMM, 2004a. pp. 201-206.

—. 2004b. Diagnosing Network-Wide Traffic Anomalies. *Conference on Applications, Technologies, Architectures and Protocols for Computer Communication.* Portland, EUA : ACM Special Interest Group on Data Communication - SIGCOMM, 2004b. pp. 219-230.

—. 2005. Mining Anomalies using Traffic Feature Distributions. *Conference on Applications, Technologies, Architectures and Protocols for Computer Communication.* Filadélfia, EUA : ACM Special Interest Group on Data Communication - SIGCOMM, 2005. pp. 217-228.

LAKHINA, A., et al. 2004c. Structural Analysis of Network Traffic Flows. *International Conference on Measurement and Modeling of Computer Systems.* Nova York, EUA : ACM Special Interest Group on Measurement and Evaluation - SIGMETRICS, 2004c. pp. 61-72.

MARCHETTE, D. 1999. A Statistical Method for Profiling Network Traffic. *Workshop on Intrusion Detection and Network Monitoring.* Santa Clara, EUA : USENIX Association - Berkeley, 1999. p. 13.

MCGREGOR, A., et al. 2004. Flow Clustering using Machine Learning Techniques. *Passive and Active Measurement Workshop - PAM.* Antibes Juan-lesPins, França : Springer, Maio de 2004. pp. 205-214.

MINGOTI, S. 2005. *Análise de Dados através de Métodos de Estatística Multivariada.* Belo Horizonte : UFMG - Universidade Federal de Minas Gerais, 2005. p. 297.

MINITAB. 2004. *Meet MINITAB - Release 14 for Windows.* EUA : Minitab Inc, 2004.

MOORE, A. e PAPAGIANNAKI, K. 2005c. Toward the Accurate Identification of Network Applications. *Passive and Active Measurement Workshop - PAM.* Boston, EUA : Springer, 2005c. pp. 41-54.

MOORE, A. e ZUEV, D. 2005b. Internet Traffic Classification using Bayesian Analysis Techniques. *International Conference on Measurements and Modeling of Computer Systems.* Banff, Canadá : ACM Special Interest Group on Performance Evaluation - SIGMETRICS, 2005b. pp. 50-60.

MOORE, A., et al. 2003. Architecture of a Network Monitor. *Passive and Active Measurement Workshop - PAM.* La Jolla, EUA : Springer, 2003. pp. 77-86.

MOORE, A., ZUEV, D. e CROGAN, M. 2005a. Discriminators for Use in Flow-Based Classification. *RR-05-13.* Londres, Inglaterra : University of London - Queen Mary, Agosto de 2005a. p. 16.

MOORE, D., et al. 2001. CoralReef Software Suite as a Tool for System and Network Administrators. *USENIX Conference on System Administration.* San Diego, EUA : USENIX Association, Dezembro de 2001. pp. 133-144.

PAXSON, V. 1994. Empirically Derived Analytic Models of Wide-Area TCP Connections. *IEEE/ACM Transactions on Networking - TON.* Agosto de 1994, Vol. 2, pp. 316-336.

PORTNOY, L., ESKIN, E. e STOLFO, S. 2001. Intrusion Detection with Unlabeled Data using Clustering. *ACM Workshop on Data Mining Applied to Security - DMAS.* 2001.

RHODEN, G. E., MELO, E. T. L. e WESTPHALL, C. B. 2002. Detecção de Intrusões em Backbones de Redes de Computadores através da Análise de Comportamento com SNMP. *Workshop em Segurança de Sistemas Computacionais.* Búzios, Brasil : Anais do XX Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos - SBRC, 2002. pp. 9-16.

ROUGHAN, M., et al. 2004. Combining Routing and Traffic Data for Detection of IP Forwarding Anomalies. *International Conference on Measurement and Modeling of Computer Systems.* Nova York, EUA : ACM Special Interest Group on Measurement and Evaluation - SIGMETRICS, 2004. pp. 416-417.

SALGARELLI, L., GRINGOLI, F. e KARAGIANNIS, T. 2007. Comparing Traffic Classifiers. *ACM SIGCOMM Computer Communication Review.* Julho de 2007, Vol. 37, pp. 65-68.

SAROIU, S., et al. 2002. An Analysis of Internet Content Delivery Systems. *Symposium on Operating Systems Design and Implementation - OSDI.* Boston, EUA : ACM Special Interest Group on Operating Systems - SIGOPS, Dezembro de 2002. pp. 315-327.

SCHECHTER, S., JUNG, J. e BERGER, A. 2004. Fast Detection of Scanning Worm Infections. *International Symposium on Recent Advances in Intrusion Detection - RAID.* Sophia Antipolis, França : Springer, 2004. pp. 59-81.

SEN, S. e WANG, J. 2002. Analyzing Peer-to-Peer Traffic Across Large Networks. *Worshop on Internet Measurement.* Marselha, França : ACM Special Interest Group on Data Communication - SIGCOMM, 2002. pp. 137-150.

SEN, S., SPATSCHECK, O. e WANG, D. 2004. Accurate, Scalable In-Network Identification on P2P Traffic using Application Signatures. *International World Wide Web Conference - WWW.* Nova York, EUA : ACM, 2004. pp. 512-521.

SIQUEIRA, G. P. 2008. Uma Metodologia para Identificação de Classes de Tráfego baseada em Discriminantes Estatísticos e Análise de Agrupamentos. *Dissertação (Mestrado em Informática Aplicada).* Fortaleza, Brasil : Universidade de Fortaleza - UNIFOR, Agosto de 2008. p. 111.

SPOGNARDI, A., LUCARELLI, A. e PIETRO, R. 2005. A Methodology for P2P File-Sharing Traffic Detection. *International Workshop on Hot Topics in Peer-to-Peer Systems - HOT-P2P.* La Jolla, EUA : IEEE Computer Society Press, 2005. pp. 52-61.

STEVENSON, W. J. 2001. *Estatística Aplicada à Administração.* São Paulo : Harbra, 2001. p. 498.

SU, J. e ZHANG, H. 2006. Full Bayesian Network Classifiers. *International Conference on Machine Learning.* Pittsburgh, EUA : ACM, 2006. pp. 897-904.

TAYLOR, C. e ALVES-FOSS, J. 2002. An Empirical Analysis of NATE: Network Analysis of Anomalous Traffic Events. *New Security Paradigms Workshop.* Virgínia Beach, EUA : ACM Special Interest Group on Security, Audit, and Control - SIGSAC, 2002. pp. 18-26.

—. 2000. Low Cost Network Intrusion Detection. Moscou : University of Idaho, 2000. p. 15.

—. 2001. NATE: Network Analysis of Anomalous Traffic Events. *New Security Paradigms Workshop.* Novo México : ACM Special Interest Group on Security, Audit and Control - SIGSAC, 2001. pp. 89-96.

THEODORIDIS, S. e KOUTROUMBAS, K. 2006. *Pattern Recogniton.* 3ª Ed. San Diego : Academic Press, 2006. p. 856.

THOMPSON, K., MILLER, G. J. e WILDER, R. 1997. Wide-Area Internet Traffic Patterns and Characteristics. *IEEE Network.* Novembro/Dezembro de 1997, Vol. 11, pp. 10-23.

THOTTAN, M. e JI, C. 2003. Anomaly Detection in IP Networks. *IEEE Transactions on Signal Processing.* Agosto de 2003, Vol. 51, pp. 2191-2204.

TRIOLA, M. F. 2005. *Introdução a Estatística.* 7ª Ed. Rio de Janeiro : LTC, 2005. p. 682.

WILLIAMS, N., ZANDER, S. e ARMITAGE, G. 2006. A Preliminary Performance Comparision of Five Machine Learning Algorithms for Practical IP Traffic Flow Classification. *ACM SIGCOMM Computer Communication Review.* Outubro de 2006, Vol. 36, 5 Ed., pp. 5-16.

WILLIAMSON, C. 2001. Internet Traffic Measurement. *IEEE Internet Computing.* Piscataway, EUA : IEEE Computer Society Press, Novembro de 2001. Vol. 5, pp. 70-74.

WITTE, R. e WITTE, J. 2005. *Estatística.* 7ª Ed. Rio de Janeiro : LTC, 2005. p. 506.

ZUEV, D. e MOORE, A. 2005. Traffic Classification using a Statistical Approach. *Passive and Active Measurement Workshop - PAM.* Boston, EUA : Springer, Março de 2005. pp. 321-324.

Anexos

Anexo A – Variáveis Analisadas

1. Número da porta de conexão no servidor
2. Número da porta de conexão no cliente
3. Tempo mínimo de chegada de todos os pacotes do fluxo (considerando ambas as direções)
4. Primeiro quartil do tempo de chegada entre pacotes
5. Mediana do tempo de chegada entre pacotes
6. Média do tempo de chegada entre pacotes
7. Terceiro quartil do tempo de chegada entre pacotes
8. Máximo de tempo de chegada entre os pacotes
9. Variância do tempo de chegada entre os pacotes
10. Mínimo dos bytes no pacote ethernet
11. Primeiro quartil dos bytes no pacote ethernet
12. Mediana dos bytes no pacote ethernet
13. Média dos bytes no pacote ethernet
14. Terceiro quartil dos bytes no pacote ethernet
15. Máximo de bytes no pacote ethernet
16. Variância dos bytes no pacote ethernet
17. Mínimo do total de bytes no pacote IP
18. Primeiro quartil do total de bytes no pacote IP
19. Mediana do total de bytes no pacote IP
20. Média do total de bytes no pacote IP
21. Terceiro quartil do total de bytes no pacote IP
22. Máximo do total de bytes no pacote IP
23. Variância do total de bytes no pacote IP
24. Mínimo dos bytes de controle no pacote (tamanho do cabeçalho IP/TCP)
25. Primeiro quartil dos bytes de controle no pacote (tamanho do cabeçalho IP/TCP)
26. Mediana dos bytes de controle no pacote (tamanho do cabeçalho IP/TCP)
27. Média dos bytes de controle no pacote (tamanho do cabeçalho IP/TCP)
28. Terceiro quartil dos bytes de controle no pacote (tamanho do cabeçalho IP/TCP)

29. Máximo dos bytes de controle no pacote (tamanho do cabeçalho IP/TCP)
30. Variância dos bytes de controle no pacote (tamanho do cabeçalho IP/TCP)
31. Total de bytes transferidos na direção “cliente→servidor”
32. Total de bytes transferidos na direção “servidor→cliente”
33. Total de pacotes ACK na direção “cliente→servidor”
34. Total de pacotes ACK na direção “servidor→cliente”
35. O numero total de pacotes de confirmação observados que não carregavam dados (apenas o cabeçalho TCP, sem nenhum dado) e não tinham nenhum dos *flags* SYN/FIN/RST (direção “cliente→servidor”)
36. O numero total de pacotes de confirmação observados que não carregavam dados (apenas o cabeçalho TCP, sem nenhum dado) e não tinham nenhum dos *flags* SYN/FIN/RST (direção “servidor→cliente”)
37. Número total de pacotes ACK carregando blocos TCP SACK (direção “cliente→servidor”)
38. Número total de pacotes ACK carregando blocos TCP SACK (direção “servidor→cliente”)
39. Número total de pacotes ACK carregando blocos TCP D-SACK (direção “cliente→servidor”)
40. Número total de pacotes ACK carregando blocos TCP D-SACK (direção “servidor→cliente”)
41. Máximo número de blocos SACK em um pacote SACK (direção “cliente→servidor”)
42. Máximo número de blocos SACK em um pacote SACK (direção “servidor→cliente”)
43. O número de bytes distintos enviados (i.e., o total de bytes enviados excluindo bytes retransmitidos e bytes enviados devido à verificação da janela (direção “cliente→servidor”)
44. O número de bytes distintos enviados (i.e., o total de bytes enviados excluindo bytes retransmitidos e bytes enviados devido à verificação da janela (direção “servidor→cliente”)
45. Contagem de todos os pacotes com no mínimo um byte de dados TCP (direção “cliente→servidor”)
46. Contagem de todos os pacotes com no mínimo um byte de dados TCP (direção “servidor→cliente”)

47. O total de bytes observados. Note que isto inclui bytes de retransmissões e os de verificação de janela, se existir (direção “cliente→servidor”)
48. O total de bytes observados. Note que isto inclui bytes de retransmissões e os de verificação de janela, se existir (direção “servidor→cliente”)
49. Contador de todos os pacotes que precisaram ser retransmitidos (direção “cliente→servidor”)
50. Contador de todos os pacotes que precisaram ser retransmitidos (direção “servidor→cliente”)
51. Total de bytes de dados contidos nos pacotes retransmitidos (direção “cliente→servidor”)
52. Total de bytes de dados contidos nos pacotes retransmitidos (direção “servidor→cliente”)
53. Contador de todos os pacotes de verificação de janela observados (pacotes de verificação de janela são normalmente enviados pelo transmissor quando o receptor sinaliza uma janela de recepção de tamanho zero, para ver se a janela já está aberta (direção “cliente→servidor”)
54. Contador de todos os pacotes de verificação de janela observados (pacotes de verificação de janela são normalmente enviados pelo transmissor quando o receptor sinaliza uma janela de recepção de tamanho zero, para ver se a janela já está aberta (direção “servidor→cliente”)
55. Total de bytes de dados enviados nos pacotes de verificação da janela (direção “cliente→servidor”)
56. Total de bytes de dados enviados nos pacotes de verificação da janela (direção “servidor→cliente”)
57. Contador dos pacotes que foram recebidos fora de ordem (direção “cliente→servidor”)
58. Contador dos pacotes que foram recebidos fora de ordem (direção “servidor→cliente”)
59. Contador de todos os pacotes com o bit PSH habilitado no cabeçalho TCP (direção “cliente→servidor”)
60. Contador de todos os pacotes com o bit PSH habilitado no cabeçalho TCP (direção “servidor→cliente”)

61. Contador de todos os pacotes com o bit SYN habilitado no cabeçalho TCP (direção “cliente→servidor”)
62. Contador de todos os pacotes com o bit FIN habilitado no cabeçalho TCP (direção “cliente→servidor”)
63. Contador de todos os pacotes com o bit SYN habilitado no cabeçalho TCP (direção “servidor→cliente”)
64. Contador de todos os pacotes com o bit FIN habilitado no cabeçalho TCP (direção “servidor→cliente”)
65. Se o ponto de parada requisitado *Window Scaling Options* está especificado na RFC 1323 um “Y” é colocado no respectivo campo. Se a opção não foi requisitada, um “N” é impresso no segmento SYN (direção “cliente→servidor”)
66. Se o ponto de parada requisitado *Timestamp Options* está especificado na RFC 1323 um “Y” é colocado no respectivo campo. Se a opção não foi requisitada, um “N” é impresso. Por exemplo, um “Y/N” no campo significaria que o *Timestamp* não foi especificado no segmento SYN (direção “cliente→servidor”)
67. Se o ponto de parada requisitado *Window Scaling Options* está especificado na RFC 1323 um “Y” é colocado no respectivo campo. Se a opção não foi requisitada, um “N” é impresso no segmento SYN (direção “servidor→cliente”)
68. Se o ponto de parada requisitado *Timestamp Options* está especificado na RFC 1323 um “Y” é colocado no respectivo campo. Se a opção não foi requisitada, um “N” é impresso. Por exemplo, um “Y/N” no campo significaria que o *Timestamp* não foi especificado no segmento SYN (direção “cliente→servidor”)
69. Fator de escala de janela utilizado. Este campo somente é válido se a conexão for totalmente capturada para incluir os pacotes SYN na direção “cliente→servidor”
70. Fator de escala de janela utilizado. Este campo somente é válido se a conexão for totalmente capturada para incluir os pacotes SYN na direção “servidor→cliente”
71. Se o ponto de parada enviando um SACK permitido no pacote SYN abriu a conexão, um “Y” é impresso, de outro modo um “N” é impresso (direção “cliente→servidor”)
72. Se o ponto de parada enviando um SACK permitido no pacote SYN abriu a conexão, um “Y” é impresso, de outro modo um “N” é impresso (direção “servidor→cliente”)

73. Número total de pacotes ACK carregando informações SACK (direção “cliente→servidor”)
74. Número total de pacotes ACK carregando informações SACK (direção “servidor→cliente”)
75. Número total de pacotes com o bit URG ligado no cabeçalho TCP (direção “cliente→servidor”)
76. Número total de pacotes com o bit URG ligado no cabeçalho TCP (direção “servidor→cliente”)
77. Número total de dados urgentes enviados. O campo é calculado pela soma dos valores de *offset* (deslocamento) encontrados nos pacotes que tem o bit URG ligado no cabeçalho TCP (direção “cliente→servidor”)
78. Número total de dados urgentes enviados. O campo é calculado pela soma dos valores de *offset* (deslocamento) encontrados nos pacotes que tem o bit URG ligado no cabeçalho TCP (direção “servidor→cliente”)
79. Tamanho máximo de segmento (MSS – *Maximum Segment Size*) requisitado como uma opção TCP no pacote SYN abrindo a conexão (direção “cliente→servidor”)
80. Tamanho máximo de segmento (MSS – *Maximum Segment Size*) requisitado como uma opção TCP no pacote SYN abrindo a conexão (direção “servidor→cliente”)
81. Tamanho máximo de segmento observado durante o tempo da conexão (direção “cliente→servidor”)
82. Tamanho máximo de segmento observado durante o tempo da conexão (direção “servidor→cliente”)
83. Tamanho mínimo de segmento observado durante o tempo da conexão (direção “cliente→servidor”)
84. Tamanho mínimo de segmento observado durante o tempo da conexão (direção “servidor→cliente”)
85. Média do tamanho de segmento observado durante o tempo de vida da conexão calculado como o valor total dos bytes de dados dividido pelo número de pacotes (direção “cliente→servidor”)
86. Média do tamanho de segmento observado durante o tempo de vida da conexão calculado como o valor total dos bytes de dados dividido pelo número de pacotes (direção “servidor→cliente”)

87. Tamanho máximo da janela na direção “cliente→servidor”
88. Tamanho máximo da janela na direção “servidor→cliente”
89. Tamanho mínimo de janela na direção “cliente→servidor”
90. Tamanho mínimo de janela na direção “servidor→cliente”
91. Número de vezes que uma janela de tamanho zero foi anunciada (direção “cliente→servidor”)
92. Número de vezes que uma janela de tamanho zero foi anunciada (direção “servidor→cliente”)
93. Tamanho médio de janela calculada como a soma de todas as janelas dividida pelo número de pacotes (direção “cliente→servidor”)
94. Tamanho médio de janela calculada como a soma de todas as janelas dividida pelo número de pacotes (direção “servidor→cliente”)
95. Número total de bytes enviados na janela inicial, i.e, o número de bytes de dados enviados antes de receber o primeiro pacote ACK. Pacotes retransmitidos não serão levados em conta (direção “cliente→servidor”)
96. Número total de bytes enviados na janela inicial, i.e, o número de bytes de dados enviados antes de receber o primeiro pacote ACK. Pacotes retransmitidos não serão levados em conta (direção “servidor→cliente”)
97. Número total de pacotes enviados na janela inicial (direção “cliente→servidor”)
98. Número total de pacotes enviados na janela inicial (direção “servidor→cliente”)
99. TSL (*Theoretical Stream Lenght*), calculado como a diferença entre as seqüências de números entre os pacotes SYN e FIN, dando o tamanho da *stream* de dados (direção “cliente→servidor”)
100. TSL (*Theoretical Stream Lenght*), calculado como a diferença entre as seqüências de números entre os pacotes SYN e FIN, dando o tamanho da *stream* de dados (direção “servidor→cliente”)
101. Número de dados perdidos dado pela diferença entre o TSL e os bytes enviados. Se a conexão não foi completada, o cálculo é inválido e um “NA” (*Not Available*) é impresso (direção “cliente→servidor”)
102. Número de dados perdidos dado pela diferença entre o TSL e os bytes enviados. Se a conexão não foi completada, o cálculo é inválido e um “NA” (*Not Available*) é impresso (direção “servidor→cliente”)

103. Número de dados truncados para a captura. Leva em conta o total de bytes capturados para análise diminuído do total de bytes transmitido (direção “cliente→servidor”)
104. Número de dados truncados para a captura. Leva em conta o total de bytes capturados para análise diminuído do total de bytes transmitido (direção “servidor→cliente”)
105. Número de pacotes truncados para a captura. Leva em conta o total de pacotes capturados para análise diminuído do total de pacotes transmitidos (direção “cliente→servidor”)
106. Número de pacotes truncados para a captura. Leva em conta o total de pacotes capturados para análise diminuído do total de pacotes transmitido (direção “servidor→cliente”)
107. Total de dados transmitidos, calculado pela diferença entre os tempos de captura do primeiro e último pacotes carregando dados TCP não zero (direção “cliente→servidor”)
108. Total de dados transmitidos, calculado pela diferença entre os tempos de captura do primeiro e último pacotes carregando dados TCP não zero (direção “servidor→cliente”)
109. Máximo tempo ocioso, calculado como o tempo máximo entre dois pacotes consecutivos na mesma direção (direção “cliente→servidor”)
110. Máximo tempo ocioso, calculado como o tempo máximo entre dois pacotes consecutivos na mesma direção (direção “servidor→cliente”)
111. Vazão média calculada entre os bytes enviados dividido pelo tempo decorrido sendo o tempo decorrido como a diferença entre a captura do primeiro e último pacote na direção “cliente→servidor”
112. Vazão média calculada entre os bytes enviados dividido pelo tempo decorrido sendo o tempo decorrido como a diferença entre a captura do primeiro e último pacote na direção “servidor→cliente”
113. Número total de RTT (*Round Trip Time*), não sendo levados em conta as retransmissões na direção “cliente→servidor”
114. Número total de RTT (*Round Trip Time*), não sendo levados em conta as retransmissões na direção “servidor→cliente”
115. Número mínimo de RTT na direção “cliente→servidor”
116. Número mínimo de RTT na direção “servidor→cliente”

117. Número máximo de RTT na direção “cliente→servidor”
118. Número máximo de RTT na direção “servidor→cliente”
119. Média dos valores de RTT encontrados calculados como a soma de todos os valores de RTT divido pelo número de RTT na direção “cliente→servidor”
120. Média dos valores de RTT encontrados calculados como a soma de todos os valores de RTT divido pelo número de RTT na direção “servidor→cliente”
121. Desvio padrão dos valores de RTT na direção “cliente→servidor”
122. Desvio padrão dos valores de RTT na direção “servidor→cliente”
123. Valor RTT do 3WS (*Three Way Handshake* – abertura de conexão), assumindo que os pacotes SYN da conexão foram capturados na direção “cliente→servidor”
124. Valor RTT do 3WS (*Three Way Handshake* – abertura de conexão), assumindo que os pacotes SYN da conexão foram capturados na direção “servidor→cliente”
125. Número total de amostras RTT *full-size segments*. *Full-size segments* são definidos como os de maior tamanho na conexão na direção “cliente→servidor”
126. Número total de amostras RTT *full-size segments*. *Full-size segments* são definidos como os de maior tamanho na conexão na direção “servidor→cliente”
127. Número mínimo de amostras RTT *full-size* na direção “cliente→servidor”
128. Número mínimo de amostras RTT *full-size* na direção “servidor→cliente”
129. Número máximo de amostras RTT *full-size* na direção “cliente→servidor”
130. Número máximo de amostras RTT *full-size* na direção “servidor→cliente”
131. Valor médio das amostras RTT *full-size* na direção “cliente→servidor”
132. Valor médio das amostras RTT *full-size* na direção “servidor→cliente”
133. Desvio padrão das amostras RTT *full-size* na direção “cliente→servidor”
134. Desvio padrão das amostras RTT *full-size* na direção “servidor→cliente”
135. Número total de pacotes ACK recebidos após as perdas serem detectadas e as retransmissões ocorrerem. O pacote ACK é recebido após ser percebido um evento de perda e o mesmo ser recuperado na direção “cliente→servidor”
136. Número total de pacotes ACK recebidos após as perdas serem detectadas e as retransmissões ocorrerem. O pacote ACK é recebido após ser percebido um evento de perda e o mesmo ser recuperado na direção “servidor→cliente”
137. Contador do número de segmentos que foram cumulativamente reconhecidos e não diretamente reconhecidos na direção “cliente→servidor”

138. Contador do número de segmentos que foram cumulativamente reconhecidos e não diretamente reconhecidos na direção “servidor→cliente”
139. Número total de reconhecimentos duplicados recebidos na direção “cliente→servidor”
140. Número total de reconhecimentos duplicados recebidos na direção “servidor→cliente”
141. Número total de reconhecimentos triplamente duplicados recebidos (três reconhecimentos duplicados para o mesmo segmento), uma condição comumente usada para disparar a fase de *fast-retransmit*/*fast-recovery* do TCP na direção “cliente→servidor”
142. Número total de reconhecimentos triplamente duplicados recebidos (três reconhecimentos duplicados para o mesmo segmento), uma condição comumente usada para disparar a fase de *fast-retransmit*/*fast-recovery* do TCP na direção “servidor→cliente”
143. Número máximo de retransmissões detectado durante o tempo de vida da conexão na direção “cliente→servidor”
144. Número máximo de retransmissões detectado durante o tempo de vida da conexão na direção “servidor→cliente”
145. Tempo mínimo entre duas retransmissões de um segmento entre todas as retransmissões na direção “cliente→servidor”
146. Tempo mínimo entre duas retransmissões de um segmento entre todas as retransmissões na direção “servidor→cliente”
147. Tempo máximo entre duas retransmissões de um segmento na direção “cliente→servidor”
148. Tempo máximo entre duas retransmissões de um segmento na direção “servidor→cliente”
149. Tempo médio entre duas retransmissões de um segmento calculado entre todas as retransmissões na direção “cliente→servidor”
150. Tempo médio entre duas retransmissões de um segmento calculado entre todas as retransmissões na direção “servidor→cliente”
151. Desvio padrão dos tempos de retransmissão das amostras obtido entre todas as retransmissões na direção “cliente→servidor”
152. Desvio padrão dos tempos de retransmissão das amostras obtido entre todas as retransmissões na direção “servidor→cliente”

153. Número mínimo de bytes no pacote Ethernet na direção “cliente→servidor”
154. Primeiro quartil de bytes no pacote Ethernet
155. Mediana dos bytes no pacote Ethernet
156. Média dos bytes no pacote Ethernet
157. Terceiro quartil dos bytes no pacote Ethernet
158. Número máximo de bytes no pacote Ethernet
159. Variância dos bytes no pacote Ethernet
160. Número mínimo do total de bytes no pacote IP
161. Primeiro quartil do número total de bytes no pacote IP
162. Mediana do número total de bytes no pacote IP
163. Média do número total de bytes no pacote IP
164. Terceiro quartil do número total de bytes no pacote IP
165. Máximo do número total de bytes no pacote IP
166. Variância do total de bytes no pacote IP
167. Número mínimo dos bytes de controle no pacote
168. Primeiro quartil dos bytes de controle no pacote
169. Mediana dos bytes de controle no pacote
170. Média dos bytes de controle no pacote
171. Terceiro quartil dos bytes de controle no pacote
172. Máximo dos bytes de controle no pacote
173. Variância dos bytes de controle no pacote
174. Número mínimo de bytes no pacote Ethernet na direção “servidor→cliente”
175. Primeiro quartil dos bytes no pacote Ethernet na direção “servidor→cliente”
176. Mediana dos bytes no pacote Ethernet na direção “servidor→cliente”
177. Média dos bytes no pacote Ethernet na direção “servidor→cliente”
178. Terceiro quartil dos bytes no pacote Ethernet na direção “servidor→cliente”
179. Máximo de bytes no pacote Ethernet na direção “servidor→cliente”
180. Variância dos bytes no pacote Ethernet na direção “servidor→cliente”
181. Número mínimo do total de bytes no pacote IP na direção “servidor→cliente”
182. Primeiro quartil do total de bytes no pacote IP na direção “servidor→cliente”
183. Mediana do total de bytes no pacote IP na direção “servidor→cliente”
184. Média do total de bytes no pacote IP na direção “servidor→cliente”
185. Terceiro quartil do total de bytes no pacote IP na direção “servidor→cliente”
186. Máximo do total de bytes no pacote IP na direção “servidor→cliente”

187. Variância do total de bytes no pacote IP na direção “servidor→cliente”
188. Número mínimo de bytes de controle na direção “servidor→cliente”
189. Primeiro quartil dos bytes de controle no pacote na direção “servidor→cliente”
190. Mediana dos bytes de controle no pacote na direção “servidor→cliente”
191. Média dos bytes de controle no pacote na direção “servidor→cliente”
192. Terceiro quartil dos bytes de controle no pacote na direção “servidor→cliente”
193. Máximo de bytes de controle no pacote na direção “servidor→cliente”
194. Variância dos bytes de controle no pacote na direção “servidor→cliente”
195. Mínimo do tempo entre pacotes (IAT – *Inter Arrival Time*) na direção “cliente→servidor”
196. Primeiro quartil do tempo entre pacotes (IAT – *Inter Arrival Time*) na direção “cliente→servidor”
197. Mediana do tempo entre pacotes (IAT – *Inter Arrival Time*) na direção “cliente→servidor”
198. Média do tempo entre pacotes (IAT – *Inter Arrival Time*) na direção “cliente→servidor”
199. Terceiro quartil do tempo entre pacotes (IAT – *Inter Arrival Time*) na direção “cliente→servidor”
200. Máximo do tempo entre pacotes (IAT – *Inter Arrival Time*) na direção “cliente→servidor”
201. Variância do tempo entre pacotes (IAT – *Inter Arrival Time*) na direção “cliente→servidor”
202. Mínimo de tempo entre pacotes (IAT – *Inter Arrival Time*) na direção “servidor→cliente”
203. Primeiro quartil do tempo entre pacotes (IAT – *Inter Arrival Time*) na direção “servidor→cliente”
204. Mediana do tempo entre pacotes (IAT – *Inter Arrival Time*) na direção “servidor→cliente”
205. Média do tempo entre pacotes (IAT – *Inter Arrival Time*) na direção “servidor→cliente”
206. Terceiro quartil do tempo entre pacotes (IAT – *Inter Arrival Time*) na direção “servidor→cliente”
207. Máximo do tempo entre pacotes (IAT – *Inter Arrival Time*) na direção “servidor→cliente”

208. Variância do tempo entre pacotes (IAT – *Inter Arrival Time*) na direção "servidor→cliente"
209. Tempo decorrido desde a última conexão entre estes *hosts*
210. Número de transições entre *transaction mode* e *bulk transfer mode*
211. Tempo gasto no *bulk transfer mode*
212. Duração da conexão
213. Percentual de tempo gasto em *bulk transfer mode*
214. Tempo ocioso em segundos
215. Percentual de tempo ocioso
216. Largura de banda efetiva baseada na entropia (ambas as direções)
217. Largura de banda efetiva baseada na entropia (direção "cliente→servidor")
218. Largura de banda efetiva baseada na entropia direção "servidor→cliente")
219. Arco-tangente das dez maiores frequências escolhidas pela magnitude de suas contribuições (todo o tráfego) – Frequência #1
220. Arco-tangente das dez maiores frequências escolhidas pela magnitude de suas contribuições (todo o tráfego) – Frequência #2
221. Arco-tangente das dez maiores frequências escolhidas pela magnitude de suas contribuições (todo o tráfego) – Frequência #3
222. Arco-tangente das dez maiores frequências escolhidas pela magnitude de suas contribuições (todo o tráfego) – Frequência #4
223. Arco-tangente das dez maiores frequências escolhidas pela magnitude de suas contribuições (todo o tráfego) – Frequência #5
224. Arco-tangente das dez maiores frequências escolhidas pela magnitude de suas contribuições (todo o tráfego) – Frequência #6
225. Arco-tangente das dez maiores frequências escolhidas pela magnitude de suas contribuições (todo o tráfego) – Frequência #7
226. Arco-tangente das dez maiores frequências escolhidas pela magnitude de suas contribuições (todo o tráfego) – Frequência #8
227. Arco-tangente das dez maiores frequências escolhidas pela magnitude de suas contribuições (todo o tráfego) – Frequência #9
228. Arco-tangente das dez maiores frequências escolhidas pela magnitude de suas contribuições (todo o tráfego) – Frequência #10
229. Arco-tangente das dez maiores frequências escolhidas pela magnitude de suas contribuições (direção "cliente→servidor") – Frequência #1

247. Arco-tangente das dez maiores frequências escolhidas pela magnitude de suas contribuições (direção “servidor→cliente”) – Frequência #9
248. Arco-tangente das dez maiores frequências escolhidas pela magnitude de suas contribuições (direção “servidor→cliente”) – Frequência #10
249. Classe da aplicação (ATTACK, BULK, DATABASE, GAMES, INTERACTIVE, MAIL, MULTIMEDIA, P2P, SERVICES, WWW)

Anexo B – Valores da Razão F por Aplicação

Tabela B.1. ATTACK – Razão F.

VARIÁVEL	RAZÃO F
VAR081	1990,3545418030
VAR165	1978,2146974728
VAR158	1978,1704080238
VAR166	1853,1714243473
VAR159	1844,1471362076
VAR085	1044,2215678227
VAR022	910,1861934654
VAR015	910,1703527234
VAR171	678,7527038273
VAR173	527,3368540768
VAR194	512,0931282309
VAR002	424,6700259342
VAR231	400,7182119039
VAR030	397,8885233505
VAR233	376,1880188328
VAR192	376,1599709810
VAR221	373,8782734994
VAR028	370,6256471237
VAR223	356,3220690719
VAR225	339,2921404887

Tabela B.2. DATABASE – Razão F.

VARIÁVEL	RAZÃO F
VAR082	1519,304732
VAR190	1519,269878
VAR186	1508,955293
VAR179	1508,682816
VAR090	1254,327298
VAR180	1231,785833
VAR187	1230,290585
VAR094	1054,141557
VAR088	915,7424881
VAR083	908,7441266
VAR089	880,4908543
VAR085	755,4946636
VAR093	686,7009427
VAR192	677,9725473
VAR191	640,0983539
VAR022	585,154857
VAR015	585,1295784
VAR087	484,1274709
VAR095	480,7932385
VAR084	425,3654733

Tabela B.3. FTP – Razão F.

VARIÁVEL	RAZÃO F
VAR001	1247,448613
VAR189	581,610905
VAR213	515,450726
VAR079	468,747546
VAR095	403,8669608
VAR098	307,5468199
VAR220	290,4879685
VAR222	280,3096288
VAR224	271,2627484
VAR154	269,1153204
VAR161	268,9559564
VAR155	263,0309613
VAR162	262,9563991
VAR230	252,9555435
VAR111	251,6708675
VAR226	249,5135577
VAR232	242,7452176
VAR234	235,1742468
VAR012	230,4186588
VAR019	229,6028614

Tabela B.4. MAIL – Razão F.

VARIÁVEL	RAZÃO F
VAR083	979,1579603
VAR026	773,9702744
VAR231	644,6678312
VAR233	629,9754602
VAR169	626,4008571
VAR221	615,023137
VAR223	597,9876874
VAR170	590,1668913
VAR225	573,622384
VAR227	528,4181893
VAR095	515,9074645
VAR179	501,4456594
VAR186	501,0053006
VAR082	498,5924108
VAR086	472,4914973
VAR164	470,8328785
VAR157	469,2956577
VAR187	428,5380598
VAR180	428,3167235
VAR171	409,0242777

Tabela B.5. P2P – Razão F.

VARIÁVEL	RAZÃO F
VAR221	433,1450395
VAR223	423,4951495
VAR225	409,5159417
VAR227	405,2534525
VAR110	375,259139
VAR109	370,5797614
VAR166	332,9299988
VAR159	332,7424415
VAR081	317,428031
VAR165	316,8855322
VAR158	316,6358234
VAR006	297,0833905
VAR198	297,0833905
VAR205	297,0833905
VAR117	294,1577521
VAR172	272,5968405
VAR119	272,3123911
VAR022	261,5124228
VAR015	261,3557944
VAR233	259,7500981

Tabela B.6. SERVICES – Razão F.

VARIÁVEL	RAZÃO F
VAR221	4378,917192
VAR231	4047,241791
VAR026	2981,948364
VAR223	2916,649983
VAR233	2706,553728
VAR190	2155,089741
VAR015	2145,975187
VAR022	2145,123969
VAR169	1991,45469
VAR225	1978,805174
VAR227	1766,316056
VAR194	924,9918483
VAR027	920,9047797
VAR020	918,641362
VAR023	907,3993168
VAR013	906,7237156
VAR016	906,2927173
VAR158	788,4763561
VAR165	788,1054272
VAR081	780,3117211

Tabela B.7. WWW – Razão F.

VARIÁVEL	RAZÃO F
VAR083	1761,549842
VAR084	885,8076078
VAR095	739,5476415
VAR086	725,0143223
VAR017	571,619687
VAR181	560,2244646
VAR096	483,9619696
VAR160	481,9358497
VAR098	447,369012
VAR010	445,4580721
VAR174	440,53737
VAR029	412,2728053
VAR030	411,7030908
VAR193	400,2929845
VAR194	392,5811284
VAR172	381,4119048
VAR085	336,6021841
VAR153	333,5566961
VAR191	302,2913487
VAR028	302,2200133

Anexo C – Matrizes de Correlação por Aplicação

Tabela C.1. Matriz de Correlação (20 Variáveis) – ATTACK.

	VAR002	VAR015	VAR022	VAR028	VAR030	VAR081	VAR085	VAR158	VAR159	VAR165	VAR166	VAR171	VAR173	VAR192	VAR194	VAR221	VAR223	VAR225	VAR231	VAR233
VAR002	1	-0,10053	-0,10052	-0,27066	-0,32346	-0,36549	-0,26528	-0,36394	-0,35861	-0,36392	-0,35901	-0,14477	-0,24013	-0,42144	-0,29337	0,32934	0,31	0,29454	0,31179	0,29719
VAR015	-0,10053	1	1	0,14351	0,10895	0,71929	0,5909	0,71979	0,68153	0,7198	0,68154	0,17622	0,11857	-0,0789	0,09155	-0,38047	-0,3398	-0,29565	-0,4015	-0,3559
VAR022	-0,10052	1	1	0,1435	0,10899	0,7193	0,5909	0,7198	0,68153	0,71981	0,68154	0,17621	0,1186	-0,07891	0,09158	-0,38047	-0,33979	-0,29564	-0,40149	-0,35589
VAR028	-0,27066	0,14351	0,1435	1	0,86501	0,17575	0,09165	0,17254	0,20684	0,17253	0,20792	0,75904	0,70492	0,59264	0,68714	0,03615	0,03414	0,03054	0,02442	0,0238
VAR030	-0,32346	0,10895	0,10899	0,86501	1	0,18841	0,07327	0,18527	0,23345	0,18531	0,23452	0,67718	0,68633	0,61927	0,68712	0,03	0,02331	0,0173	0,08213	0,05957
VAR081	-0,36549	0,71929	0,7193	0,17575	0,18841	1	0,83926	0,99956	0,94926	0,99956	0,94935	0,12339	0,22302	0,36841	0,30743	-0,39356	-0,39069	-0,37732	-0,40555	-0,39803
VAR085	-0,26528	0,5909	0,5909	0,09165	0,07327	0,83926	1	0,83862	0,74356	0,83862	0,74384	-0,03848	0,02227	0,27364	0,09107	-0,29354	-0,29057	-0,28389	-0,31152	-0,3025
VAR158	-0,36394	0,71979	0,7198	0,17254	0,18527	0,99956	0,83862	1	0,94948	1	0,94957	0,12007	0,21973	0,36612	0,30479	-0,39487	-0,39204	-0,37874	-0,40664	-0,3992
VAR159	-0,35861	0,68153	0,68153	0,20684	0,23345	0,94926	0,74356	0,94948	1	0,94947	1	0,19378	0,31451	0,3921	0,38823	-0,35113	-0,34378	-0,32489	-0,35726	-0,34673
VAR165	-0,36392	0,7198	0,71981	0,17253	0,18531	0,99956	0,83862	1	0,94947	1	0,94956	0,12005	0,21976	0,36611	0,30481	-0,39487	-0,39203	-0,37873	-0,40663	-0,39919
VAR166	-0,35901	0,68154	0,68154	0,20792	0,23452	0,94935	0,74384	0,94957	1	0,94956	1	0,19512	0,31583	0,39286	0,38925	-0,351	-0,34366	-0,32479	-0,35717	-0,34663
VAR171	-0,14477	0,17622	0,17621	0,75904	0,67718	0,12339	-0,03848	0,12007	0,19378	0,12005	0,19512	1	0,85993	0,36354	0,73629	0,07748	0,0736	0,06984	0,03515	0,04152
VAR173	-0,24013	0,11857	0,1186	0,70492	0,68633	0,22302	0,02227	0,21973	0,31451	0,21976	0,31583	0,85993	1	0,56733	0,87305	0,14615	0,14419	0,14101	0,10362	0,11224
VAR192	-0,42144	-0,0789	-0,07891	0,59264	0,61927	0,36841	0,27364	0,36612	0,3921	0,36611	0,39286	0,36354	0,56733	1	0,7306	0,03933	0,01273	-0,01243	0,04808	0,02021
VAR194	-0,29337	0,09155	0,09158	0,68714	0,68712	0,30743	0,09107	0,30479	0,38823	0,30481	0,38925	0,73629	0,87305	0,7306	1	0,10426	0,08233	0,06635	0,06442	0,05202
VAR221	0,32934	-0,38047	-0,38047	0,03615	0,03	-0,39356	-0,29354	-0,39487	-0,35113	-0,39487	-0,351	0,07748	0,14615	0,03933	0,10426	1	0,97204	0,921	0,97498	0,95087
VAR223	0,31	-0,3398	-0,33979	0,03414	0,02331	-0,39069	-0,29057	-0,39204	-0,34378	-0,39203	-0,34366	0,0736	0,14419	0,01273	0,08233	0,97204	1	0,98112	0,95228	0,98166
VAR225	0,29454	-0,29565	-0,29564	0,03054	0,0173	-0,37732	-0,28389	-0,37874	-0,32489	-0,37873	-0,32479	0,06984	0,14101	-0,01243	0,06635	0,921	0,98112	1	0,90602	0,96697
VAR231	0,31179	-0,4015	-0,40149	0,02442	0,08213	-0,40555	-0,31152	-0,40664	-0,35726	-0,40663	-0,35717	0,03515	0,10362	0,04808	0,06442	0,97498	0,95228	0,90602	1	0,97186
VAR233	0,29719	-0,3559	-0,35589	0,0238	0,05957	-0,39803	-0,3025	-0,3992	-0,34673	-0,39919	-0,34663	0,04152	0,11224	0,02021	0,05202	0,95087	0,98166	0,96697	0,97186	1

Tabela C.2. Matriz de Correlação (20 Variáveis) – DATABASE.

	VAR015	VAR022	VAR082	VAR083	VAR084	VAR085	VAR087	VAR088	VAR089	VAR090	VAR093	VAR094	VAR095	VAR179	VAR180	VAR186	VAR187	VAR190	VAR191	VAR192
VAR015	1	1	0,80704	-0,11956	0,00834	0,15447	-0,05651	0,30446	-0,24135	0,33312	-0,14183	0,31962	-0,08651	0,80943	0,75672	0,80948	0,7567	-0,38862	-0,3807	-0,54222
VAR022	1	1	0,80703	-0,11955	0,00834	0,15448	-0,0565	0,30446	-0,24134	0,33312	-0,14182	0,31962	-0,0865	0,80943	0,75671	0,80948	0,75669	-0,38864	-0,38074	-0,54222
VAR082	0,80704	0,80703	1	-0,15917	0,05658	-0,26695	-0,15315	0,43465	-0,37443	0,46459	-0,25833	0,45108	-0,19296	0,99749	0,92393	0,99749	0,92398	-0,50723	-0,50881	-0,67135
VAR083	-0,11956	-0,11955	-0,15917	1	0,4495	0,53569	0,14884	-0,20297	0,15403	-0,31846	0,15374	-0,24682	0,70437	-0,15883	-0,12054	-0,15879	-0,12026	0,29694	0,11494	0,10863
VAR084	0,00834	0,00834	0,05658	0,4495	1	0,19201	0,09431	-0,11889	0,06635	-0,16825	0,07969	-0,13752	0,25845	0,05689	0,11038	0,05692	0,11088	0,03802	-0,09556	-0,03049
VAR085	0,15447	0,15448	-0,26695	0,53569	0,19201	1	0,29054	-0,21802	0,23385	-0,30168	0,2814	-0,24528	0,56544	-0,26753	-0,26623	-0,26745	-0,2663	0,34139	0,23246	0,18483
VAR087	-0,05651	-0,0565	-0,15315	0,14884	0,09431	0,29054	1	-0,22628	0,88508	-0,26448	0,98233	-0,2415	0,0893	-0,14872	-0,17622	-0,14872	-0,1762	0,25745	0,0759	0,07936
VAR088	0,30446	0,30446	0,43465	-0,20297	-0,11889	-0,21802	-0,22628	1	-0,30862	0,91881	-0,27061	0,98984	-0,15557	0,43226	0,37778	0,43236	0,37802	-0,33914	-0,11559	-0,17355
VAR089	-0,24135	-0,24134	-0,37443	0,15403	0,06635	0,23385	0,88508	-0,30862	1	-0,34735	0,94002	-0,32673	0,1031	-0,36969	-0,34553	-0,36969	-0,34527	0,35746	0,23332	0,24823
VAR090	0,33312	0,33312	0,46459	-0,31846	-0,16825	-0,30168	-0,26448	0,91881	-0,34735	1	-0,31038	0,96157	-0,23625	0,46262	0,41789	0,4627	0,41796	-0,43543	-0,19662	-0,22741
VAR093	-0,14183	-0,14182	-0,25833	0,15374	0,07969	0,2814	0,98233	-0,27061	0,94002	-0,31038	1	-0,28745	0,1014	-0,25375	-0,2714	-0,25375	-0,27129	0,30892	0,14291	0,15545
VAR094	0,31962	0,31962	0,45108	-0,24682	-0,13752	-0,24528	-0,2415	0,98984	-0,32673	0,96157	-0,28745	1	-0,18263	0,4488	0,39623	0,44889	0,39641	-0,37769	-0,14289	-0,19568
VAR095	-0,08651	-0,0865	-0,19296	0,70437	0,25845	0,56544	0,0893	-0,15557	0,1031	-0,23625	0,1014	-0,18263	1	-0,19201	-0,1686	-0,19197	-0,16855	0,28776	0,20349	0,09869
VAR179	0,80943	0,80943	0,99749	-0,15883	0,05689	-0,26753	-0,14872	0,43226	-0,36969	0,46262	-0,25375	0,4488	-0,19201	1	0,92782	1	0,92788	-0,50884	-0,51227	-0,67517
VAR180	0,75672	0,75671	0,92393	-0,12054	0,11038	-0,26623	-0,17622	0,37778	-0,34553	0,41789	-0,2714	0,39623	-0,1686	0,92782	1	0,92781	1	-0,4491	-0,42641	-0,59666
VAR186	0,80948	0,80948	0,99749	-0,15879	0,05692	-0,26745	-0,14872	0,43236	-0,36969	0,4627	-0,25375	0,44889	-0,19197	1	0,92781	1	0,92786	-0,50897	-0,51246	-0,6752
VAR187	0,7567	0,75669	0,92398	-0,12026	0,11088	-0,2663	-0,1762	0,37802	-0,34527	0,41796	-0,27129	0,39641	-0,16855	0,92788	1	0,92786	1	-0,44888	-0,42584	-0,59602
VAR190	-0,38862	-0,38864	-0,50723	0,29694	0,03802	0,34139	0,25745	-0,33914	0,35746	-0,43543	0,30892	-0,37769	0,28776	-0,50884	-0,4491	-0,50897	-0,44888	1	0,82994	0,6177
VAR191	-0,3807	-0,38074	-0,50881	0,11494	-0,09556	0,23246	0,0759	-0,11559	0,23332	-0,19662	0,14291	-0,14289	0,20349	-0,51227	-0,42641	-0,51246	-0,42584	0,82994	1	0,8188
VAR192	-0,54222	-0,54222	-0,67135	0,10863	-0,03049	0,18483	0,07936	-0,17355	0,24823	-0,22741	0,15545	-0,19568	0,09869	-0,67517	-0,59666	-0,6752	-0,59602	0,6177	0,8188	1

Tabela C.3. Matriz de Correlação (20 Variáveis) – FTP.

	VAR001	VAR012	VAR019	VAR079	VAR095	VAR098	VAR111	VAR154	VAR155	VAR161	VAR162	VAR189	VAR213	VAR220	VAR222	VAR224	VAR226	VAR230	VAR232	VAR234
VAR001	1	0,16941	0,16739	-0,22863	0,21323	-0,08573	-0,05598	0,11825	0,12632	0,11832	0,12472	0,16925	0,136	0,12399	0,11628	0,11508	0,11277	0,10217	0,09381	0,09388
VAR012	0,16941	1	0,99997	-0,20102	0,47008	-0,10322	0,4016	0,75911	0,68072	0,75942	0,68047	0,4034	0,75408	0,63029	0,61912	0,6063	0,58323	0,54657	0,53624	0,52856
VAR019	0,16739	0,99997	1	-0,20067	0,469	-0,1017	0,40214	0,75924	0,6808	0,75957	0,6806	0,40197	0,75506	0,63081	0,61968	0,60683	0,58372	0,54706	0,53676	0,52906
VAR079	-0,22863	-0,20102	-0,20067	1	-0,23478	0,5213	-0,44491	-0,30529	-0,2537	-0,30671	-0,25384	-0,61821	-0,40743	-0,34256	-0,3328	-0,33028	-0,32239	-0,34354	-0,33196	-0,32687
VAR095	0,21323	0,47008	0,469	-0,23478	1	-0,46505	0,38094	0,66646	0,74083	0,66684	0,74034	0,55716	0,50471	0,35967	0,34824	0,33965	0,32098	0,31522	0,30513	0,30109
VAR098	-0,08573	-0,10322	-0,1017	0,5213	-0,46505	1	-0,34061	-0,39192	-0,40291	-0,39164	-0,40279	-0,61427	-0,15878	-0,2004	-0,19142	-0,18752	-0,17787	-0,19954	-0,1906	-0,18749
VAR111	-0,05598	0,4016	0,40214	-0,44491	0,38094	-0,34061	1	0,50862	0,45569	0,50985	0,45639	0,5158	0,59863	0,6059	0,61087	0,60964	0,60365	0,61037	0,61607	0,61435
VAR154	0,11825	0,75911	0,75924	-0,30529	0,66646	-0,39192	0,50862	1	0,91177	0,99997	0,91182	0,59556	0,70344	0,56979	0,55521	0,54254	0,51724	0,50066	0,48784	0,4813
VAR155	0,12632	0,68072	0,6808	-0,2537	0,74083	-0,40291	0,45569	0,91177	1	0,91229	0,99997	0,61974	0,65806	0,50632	0,49161	0,48015	0,45737	0,44522	0,43172	0,42586
VAR161	0,11832	0,75942	0,75957	-0,30671	0,66684	-0,39164	0,50985	0,99997	0,91229	1	0,91237	0,59664	0,70505	0,57078	0,55623	0,54355	0,51825	0,50169	0,48889	0,48233
VAR162	0,12472	0,68047	0,6806	-0,25384	0,74034	-0,40279	0,45639	0,91182	0,99997	0,91237	1	0,61927	0,65854	0,50678	0,49211	0,48063	0,45781	0,44569	0,43223	0,42634
VAR189	0,16925	0,4034	0,40197	-0,61821	0,55716	-0,61427	0,5158	0,59556	0,61974	0,59664	0,61927	1	0,56539	0,44714	0,44143	0,43686	0,42502	0,4411	0,43259	0,42538
VAR213	0,136	0,75408	0,75506	-0,40743	0,50471	-0,15878	0,59863	0,70344	0,65806	0,70505	0,65854	0,56539	1	0,74497	0,73055	0,71862	0,69341	0,68666	0,67194	0,6612
VAR220	0,12399	0,63029	0,63081	-0,34256	0,35967	-0,2004	0,6059	0,56979	0,50632	0,57078	0,50678	0,44714	0,74497	1	0,9885	0,98636	0,97531	0,98234	0,96979	0,9639
VAR222	0,11628	0,61912	0,61968	-0,3328	0,34824	-0,19142	0,61087	0,55521	0,49161	0,55623	0,49211	0,44143	0,73055	0,9885	1	0,99862	0,98893	0,97104	0,98332	0,97778
VAR224	0,11508	0,6063	0,60683	-0,33028	0,33965	-0,18752	0,60964	0,54254	0,48015	0,54355	0,48063	0,43686	0,71862	0,98636	0,99862	1	0,99477	0,97505	0,98801	0,98439
VAR226	0,11277	0,58323	0,58372	-0,32239	0,32098	-0,17787	0,60365	0,51724	0,45737	0,51825	0,45781	0,42502	0,69341	0,97531	0,98893	0,99477	1	0,97376	0,98812	0,9889
VAR230	0,10217	0,54657	0,54706	-0,34354	0,31522	-0,19954	0,61037	0,50066	0,44522	0,50169	0,44569	0,4411	0,68666	0,98234	0,97104	0,97505	0,97376	1	0,98627	0,98418
VAR232	0,09381	0,53624	0,53676	-0,33196	0,30513	-0,1906	0,61607	0,48784	0,43172	0,48889	0,43223	0,43259	0,67194	0,96979	0,98332	0,98801	0,98812	0,98627	1	0,99848
VAR234	0,09388	0,52856	0,52906	-0,32687	0,30109	-0,18749	0,61435	0,4813	0,42586	0,48233	0,42634	0,42538	0,6612	0,9639	0,97778	0,98439	0,9889	0,98418	0,99848	1

Tabela C.4. Matriz de Correlação (20 Variáveis) – MAIL.

	VAR026	VAR082	VAR083	VAR086	VAR095	VAR157	VAR164	VAR169	VAR170	VAR171	VAR179	VAR180	VAR186	VAR187	VAR221	VAR223	VAR225	VAR227	VAR231	VAR233
VAR026	1	-0,11746	0,31777	-0,01057	0,11971	-0,29875	-0,29919	0,80632	0,591	0,44759	-0,11822	-0,05569	-0,11834	-0,0549	0,53213	0,47033	0,41381	0,39363	0,5361	0,47651
VAR082	-0,11746	1	0,19041	0,82435	0,03657	-0,22995	-0,23073	0,07712	0,21581	0,17118	0,99974	0,90737	0,99974	0,90741	0,02942	0,09318	0,14169	0,13872	0,02448	0,0891
VAR083	0,31777	0,19041	1	0,39075	0,72866	-0,09811	-0,09879	0,31734	0,17003	0,0526	0,19267	0,22595	0,19268	0,22618	0,2284	0,22526	0,21518	0,2083	0,22639	0,22687
VAR086	-0,01057	0,82435	0,39075	1	0,15497	-0,22655	-0,22788	0,19645	0,22334	0,09458	0,82528	0,82913	0,82529	0,82935	-0,00496	0,03532	0,0589	0,06119	-0,00493	0,03641
VAR095	0,11971	0,03657	0,72866	0,15497	1	0,11924	0,1189	0,06866	-0,04749	-0,1132	0,0396	0,07433	0,03961	0,0743	0,08895	0,10044	0,10778	0,10374	0,08574	0,10034
VAR157	-0,29875	-0,22995	-0,09811	-0,22655	0,11924	1	1	-0,41389	-0,53783	-0,59224	-0,23029	-0,21672	-0,23022	-0,21697	-0,3037	-0,33663	-0,34434	-0,3508	-0,31072	-0,34199
VAR164	-0,29919	-0,23073	-0,09879	-0,22788	0,1189	1	1	-0,41499	-0,53906	-0,59265	-0,23107	-0,21742	-0,231	-0,21767	-0,30345	-0,33638	-0,34408	-0,35053	-0,31049	-0,34176
VAR169	0,80632	0,07712	0,31734	0,19645	0,06866	-0,41389	-0,41499	1	0,81401	0,61289	0,07529	0,09867	0,07527	0,09971	0,45631	0,42144	0,38039	0,36775	0,44525	0,41544
VAR170	0,591	0,21581	0,17003	0,22334	-0,04749	-0,53783	-0,53906	0,81401	1	0,90696	0,21242	0,19105	0,21243	0,19233	0,34888	0,35853	0,35248	0,35144	0,33113	0,34585
VAR171	0,44759	0,17118	0,0526	0,09458	-0,1132	-0,59224	-0,59265	0,61289	0,90696	1	0,16798	0,15045	0,16803	0,15151	0,3174	0,34322	0,34951	0,35448	0,29806	0,3283
VAR179	-0,11822	0,99974	0,19267	0,82528	0,0396	-0,23029	-0,23107	0,07529	0,21242	0,16798	1	0,90774	1	0,90777	0,02862	0,09236	0,14082	0,13776	0,02384	0,08841
VAR180	-0,05569	0,90737	0,22595	0,82913	0,07433	-0,21672	-0,21742	0,09867	0,19105	0,15045	0,90774	1	0,90772	1	0,06607	0,12257	0,16507	0,16249	0,06471	0,12135
VAR186	-0,11834	0,99974	0,19268	0,82529	0,03961	-0,23022	-0,231	0,07527	0,21243	0,16803	1	0,90772	1	0,90775	0,02857	0,0923	0,14075	0,13769	0,02365	0,08825
VAR187	-0,0549	0,90741	0,22618	0,82935	0,0743	-0,21697	-0,21767	0,09971	0,19233	0,15151	0,90777	1	0,90775	1	0,06607	0,12261	0,1651	0,16253	0,06471	0,12138
VAR221	0,53213	0,02942	0,2284	-0,00496	0,08895	-0,3037	-0,30345	0,45631	0,34888	0,3174	0,02862	0,06607	0,02857	0,06607	1	0,96898	0,91517	0,88331	0,98209	0,95293
VAR223	0,47033	0,09318	0,22526	0,03532	0,10044	-0,33663	-0,33638	0,42144	0,35853	0,34322	0,09236	0,12257	0,0923	0,12261	0,96898	1	0,98178	0,96641	0,95617	0,98639
VAR225	0,41381	0,14169	0,21518	0,0589	0,10778	-0,34434	-0,34408	0,38039	0,35248	0,34951	0,14082	0,16507	0,14075	0,1651	0,91517	0,98178	1	0,9947	0,90609	0,97229
VAR227	0,39363	0,13872	0,2083	0,06119	0,10374	-0,3508	-0,35053	0,36775	0,35144	0,35448	0,13776	0,16249	0,13769	0,16253	0,88331	0,96641	0,9947	1	0,8759	0,95827
VAR231	0,5361	0,02448	0,22639	-0,00493	0,08574	-0,31072	-0,31049	0,44525	0,33113	0,29806	0,02384	0,06471	0,02365	0,06471	0,98209	0,95617	0,90609	0,8759	1	0,969
VAR233	0,47651	0,0891	0,22687	0,03641	0,10034	-0,34199	-0,34176	0,41544	0,34585	0,3283	0,08841	0,12135	0,08825	0,12138	0,95293	0,98639	0,97229	0,95827	0,969	1

Tabela C.5. Matriz de Correlação (20 Variáveis) - P2P.

	VAR006	VAR015	VAR022	VAR081	VAR109	VAR110	VAR117	VAR119	VAR158	VAR159	VAR165	VAR166	VAR172	VAR198	VAR205	VAR221	VAR223	VAR225	VAR227	VAR233
VAR006	1	-0,20925	-0,20921	-0,1605	0,71922	0,71436	0,26084	0,25304	-0,15262	-0,10772	-0,15257	-0,10794	-0,09994	1	1	-0,29836	-0,36241	-0,39937	-0,43032	-0,34785
VAR015	-0,20925	1	1	0,6027	-0,19973	-0,19908	-0,14892	-0,1727	0,6058	0,50442	0,60588	0,50421	-0,003	-0,20925	-0,20925	-0,13819	-0,08213	-0,03263	-0,01967	-0,13264
VAR022	-0,20921	1	1	0,60272	-0,19966	-0,19901	-0,14893	-0,17273	0,60582	0,50441	0,60591	0,5042	-0,00306	-0,20921	-0,20921	-0,13821	-0,08216	-0,03266	-0,01971	-0,13266
VAR081	-0,1605	0,6027	0,60272	1	-0,1501	-0,14958	0,00613	-0,02716	0,99807	0,8704	0,99807	0,87048	0,11521	-0,1605	-0,1605	-0,18267	-0,17949	-0,17268	-0,16196	-0,2099
VAR109	0,71922	-0,19973	-0,19966	-0,1501	1	0,99692	0,2594	0,22944	-0,1523	-0,13062	-0,15222	-0,13095	-0,06173	0,71922	0,71922	-0,35427	-0,43573	-0,48375	-0,52459	-0,44608
VAR110	0,71436	-0,19908	-0,19901	-0,14958	0,99692	1	0,26143	0,23007	-0,1518	-0,1307	-0,15172	-0,13102	-0,06086	0,71436	0,71436	-0,35341	-0,43482	-0,48286	-0,52374	-0,44523
VAR117	0,26084	-0,14892	-0,14893	0,00613	0,2594	0,26143	1	0,94287	0,00421	0,00175	0,00417	0,0018	0,05696	0,26084	0,26084	-0,28233	-0,32426	-0,34451	-0,35832	-0,27698
VAR119	0,25304	-0,1727	-0,17273	-0,02716	0,22944	0,23007	0,94287	1	-0,02869	-0,00297	-0,02875	-0,00292	0,03867	0,25304	0,25304	-0,26202	-0,29858	-0,31572	-0,3267	-0,2379
VAR158	-0,15262	0,6058	0,60582	0,99807	-0,1523	-0,1518	0,00421	-0,02869	1	0,87468	1	0,87474	0,10833	-0,15262	-0,15262	-0,1838	-0,18019	-0,17317	-0,16232	-0,21008
VAR159	-0,10772	0,50442	0,50441	0,8704	-0,13062	-0,1307	0,00175	-0,00297	0,87468	1	0,87464	0,99999	0,11817	-0,10772	-0,10772	-0,12629	-0,11676	-0,10544	-0,09412	-0,13346
VAR165	-0,15257	0,60588	0,60591	0,99807	-0,15222	-0,15172	0,00417	-0,02875	1	0,87464	1	0,87469	0,10822	-0,15257	-0,15257	-0,18383	-0,18022	-0,1732	-0,16235	-0,21011
VAR166	-0,10794	0,50421	0,5042	0,87048	-0,13095	-0,13102	0,0018	-0,00292	0,87474	0,99999	0,87469	1	0,12005	-0,10794	-0,10794	-0,12586	-0,11634	-0,10507	-0,09374	-0,13312
VAR172	-0,09994	-0,003	-0,00306	0,11521	-0,06173	-0,06086	0,05696	0,03867	0,10833	0,11817	0,10822	0,12005	1	-0,09994	-0,09994	0,07012	0,03819	0,0222	0,01236	-0,03279
VAR198	1	-0,20925	-0,20921	-0,1605	0,71922	0,71436	0,26084	0,25304	-0,15262	-0,10772	-0,15257	-0,10794	-0,09994	1	1	-0,29836	-0,36241	-0,39937	-0,43032	-0,34785
VAR205	1	-0,20925	-0,20921	-0,1605	0,71922	0,71436	0,26084	0,25304	-0,15262	-0,10772	-0,15257	-0,10794	-0,09994	1	1	-0,29836	-0,36241	-0,39937	-0,43032	-0,34785
VAR221	-0,29836	-0,13819	-0,13821	-0,18267	-0,35427	-0,35341	-0,28233	-0,26202	-0,1838	-0,12629	-0,18383	-0,12586	0,07012	-0,29836	-0,29836	1	0,96878	0,91574	0,88432	0,91427
VAR223	-0,36241	-0,08213	-0,08216	-0,17949	-0,43573	-0,43482	-0,32426	-0,29858	-0,18019	-0,11676	-0,18022	-0,11634	0,03819	-0,36241	-0,36241	0,96878	1	0,98098	0,96552	0,95491
VAR225	-0,39937	-0,03263	-0,03266	-0,17268	-0,48375	-0,48286	-0,34451	-0,31572	-0,17317	-0,10544	-0,1732	-0,10507	0,0222	-0,39937	-0,39937	0,91574	0,98098	1	0,99469	0,94552
VAR227	-0,43032	-0,01967	-0,01971	-0,16196	-0,52459	-0,52374	-0,35832	-0,3267	-0,16232	-0,09412	-0,16235	-0,09374	0,01236	-0,43032	-0,43032	0,88432	0,96552	0,99469	1	0,93631
VAR233	-0,34785	-0,13264	-0,13266	-0,2099	-0,44608	-0,44523	-0,27698	-0,2379	-0,21008	-0,13346	-0,21011	-0,13312	-0,03279	-0,34785	-0,34785	0,91427	0,95491	0,94552	0,93631	1

Tabela C.6. Matriz de Correlação (20 Variáveis) – SERVICES.

	VAR013	VAR015	VAR016	VAR020	VAR022	VAR023	VAR026	VAR027	VAR081	VAR158	VAR165	VAR169	VAR190	VAR194	VAR221	VAR223	VAR225	VAR227	VAR231	VAR233
VAR013	1	0,65941	0,89843	0,99997	0,65941	0,89904	-0,49034	-0,36606	0,46592	0,46696	0,46697	-0,36196	-0,27338	-0,49014	-0,42899	-0,41453	-0,39535	-0,39479	-0,43494	-0,41941
VAR015	0,65941	1	0,80134	0,66059	1	0,80116	-0,63833	-0,39293	0,67063	0,67128	0,6713	-0,56131	-0,52525	-0,32513	-0,48973	-0,43282	-0,36981	-0,35905	-0,49785	-0,43848
VAR016	0,89843	0,80134	1	0,89828	0,80133	0,99999	-0,46872	-0,31247	0,51396	0,51508	0,51509	-0,34364	-0,36883	-0,37283	-0,4074	-0,37946	-0,34718	-0,34378	-0,41174	-0,38266
VAR020	0,99997	0,66059	0,89828	1	0,66059	0,89887	-0,49534	-0,37288	0,46655	0,46762	0,46763	-0,36725	-0,27831	-0,49517	-0,43077	-0,41613	-0,39679	-0,39617	-0,43669	-0,42098
VAR022	0,65941	1	0,80133	0,66059	1	0,80115	-0,63834	-0,39303	0,67064	0,67129	0,67131	-0,56132	-0,52526	-0,32502	-0,48969	-0,43279	-0,36978	-0,35903	-0,49782	-0,43845
VAR023	0,89904	0,80116	0,99999	0,89887	0,80115	1	-0,46833	-0,31117	0,51425	0,51536	0,51537	-0,34288	-0,36785	-0,37244	-0,40755	-0,37963	-0,34736	-0,34398	-0,41189	-0,38283
VAR026	-0,49034	-0,63833	-0,46872	-0,49534	-0,63834	-0,46833	1	0,81568	-0,43364	-0,43549	-0,43552	0,89491	0,77963	0,67333	0,6155	0,57439	0,52608	0,51289	0,61256	0,57124
VAR027	-0,36606	-0,39293	-0,31247	-0,37288	-0,39303	-0,31117	0,81568	1	-0,25275	-0,25611	-0,25625	0,80571	0,72619	0,81903	0,38353	0,35579	0,32856	0,31999	0,38313	0,35388
VAR081	0,46592	0,67063	0,51396	0,46655	0,67064	0,51425	-0,43364	-0,25275	1	0,99996	0,99996	-0,48709	-0,18494	-0,16887	-0,40651	-0,40417	-0,39355	-0,38699	-0,4137	-0,4094
VAR158	0,46696	0,67128	0,51508	0,46762	0,67129	0,51536	-0,43549	-0,25611	0,99996	1	1	-0,48911	-0,18719	-0,17253	-0,40896	-0,40639	-0,39562	-0,38896	-0,41595	-0,41143
VAR165	0,46697	0,6713	0,51509	0,46763	0,67131	0,51537	-0,43552	-0,25625	0,99996	1	1	-0,48913	-0,18723	-0,17239	-0,40891	-0,40635	-0,39559	-0,38892	-0,4159	-0,41139
VAR169	-0,36196	-0,56131	-0,34364	-0,36725	-0,56132	-0,34288	0,89491	0,80571	-0,48709	-0,48911	-0,48913	1	0,6101	0,64377	0,55256	0,51505	0,47005	0,45602	0,5375	0,50102
VAR190	-0,27338	-0,52525	-0,36883	-0,27831	-0,52526	-0,36785	0,77963	0,72619	-0,18494	-0,18719	-0,18723	0,6101	1	0,5328	0,5462	0,49322	0,4392	0,42264	0,55223	0,49643
VAR194	-0,49014	-0,32513	-0,37283	-0,49517	-0,32502	-0,37244	0,67333	0,81903	-0,16887	-0,17253	-0,17239	0,64377	0,5328	1	0,38836	0,35583	0,3262	0,31753	0,36729	0,33746
VAR221	-0,42899	-0,48973	-0,4074	-0,43077	-0,48969	-0,40755	0,6155	0,38353	-0,40651	-0,40896	-0,40891	0,55256	0,5462	0,38836	1	0,98204	0,94365	0,92437	0,98976	0,97026
VAR223	-0,41453	-0,43282	-0,37946	-0,41613	-0,43279	-0,37963	0,57439	0,35579	-0,40417	-0,40639	-0,40635	0,51505	0,49322	0,35583	0,98204	1	0,9857	0,97534	0,97515	0,99097
VAR225	-0,39535	-0,36981	-0,34718	-0,39679	-0,36978	-0,34736	0,52608	0,32856	-0,39355	-0,39562	-0,39559	0,47005	0,4392	0,3262	0,94365	0,9857	1	0,99665	0,93932	0,97986
VAR227	-0,39479	-0,35905	-0,34378	-0,39617	-0,35903	-0,34398	0,51289	0,31999	-0,38699	-0,38896	-0,38892	0,45602	0,42264	0,31753	0,92437	0,97534	0,99665	1	0,92173	0,97117
VAR231	-0,43494	-0,49785	-0,41174	-0,43669	-0,49782	-0,41189	0,61256	0,38313	-0,4137	-0,41595	-0,4159	0,5375	0,55223	0,36729	0,98976	0,97515	0,93932	0,92173	1	0,98157
VAR233	-0,41941	-0,43848	-0,38266	-0,42098	-0,43845	-0,38283	0,57124	0,35388	-0,4094	-0,41143	-0,41139	0,50102	0,49643	0,33746	0,97026	0,99097	0,97986	0,97117	0,98157	1

Tabela C.7. Matriz de Correlação (20 Variáveis) – WWW.

	VAR010	VAR017	VAR028	VAR029	VAR030	VAR083	VAR084	VAR085	VAR086	VAR095	VAR096	VAR098	VAR153	VAR160	VAR172	VAR174	VAR181	VAR191	VAR193	VAR194
VAR010	1	0,89728	-0,53843	-0,56392	-0,54171	0,28334	0,26494	0,21355	0,15025	0,32193	0,01784	-0,06197	0,93904	0,87758	-0,55682	0,99901	0,89513	-0,40333	-0,56111	-0,43494
VAR017	0,89728	1	-0,83122	-0,87057	-0,83628	0,29477	0,2958	0,19197	0,17709	0,34632	0,0482	-0,02809	0,83755	0,96977	-0,85961	0,89604	0,9971	-0,62265	-0,86623	-0,67145
VAR028	-0,53843	-0,83122	1	0,9548	0,9465	-0,19289	-0,22283	-0,11904	-0,13971	-0,25775	-0,05086	-0,00076	-0,49885	-0,80286	0,94911	-0,5398	-0,8331	0,70069	0,95618	0,78476
VAR029	-0,56392	-0,87057	0,9548	1	0,96062	-0,23562	-0,25805	-0,12109	-0,16381	-0,28903	-0,07028	-0,01652	-0,52013	-0,83598	0,98741	-0,56271	-0,86753	0,71522	0,99502	0,77128
VAR030	-0,54171	-0,83628	0,9465	0,96062	1	-0,22459	-0,2315	-0,13342	-0,15289	-0,28757	-0,05441	-0,00775	-0,49138	-0,78668	0,92683	-0,54155	-0,83459	0,68812	0,95701	0,80433
VAR083	0,28334	0,29477	-0,19289	-0,23562	-0,22459	1	0,32786	0,69151	0,31251	0,66925	0,26695	0,20792	0,24884	0,27166	-0,22346	0,28201	0,2926	-0,13195	-0,23299	-0,19837
VAR084	0,26494	0,2958	-0,22283	-0,25805	-0,2315	0,32786	1	0,12408	0,65654	0,22875	0,53645	0,33422	0,23315	0,27477	-0,24862	0,26369	0,29312	-0,27134	-0,25434	-0,21255
VAR085	0,21355	0,19197	-0,11904	-0,12109	-0,13342	0,69151	0,12408	1	0,11807	0,67787	0,10883	0,05528	0,18268	0,1698	-0,10654	0,21184	0,18889	0,06854	-0,11727	-0,17498
VAR086	0,15025	0,17709	-0,13971	-0,16381	-0,15289	0,31251	0,65654	0,11807	1	0,18199	0,85022	0,62464	0,12728	0,15892	-0,15303	0,14887	0,17408	-0,42798	-0,15964	-0,28647
VAR095	0,32193	0,34632	-0,25775	-0,28903	-0,28757	0,66925	0,22875	0,67787	0,18199	1	0,14023	0,00405	0,28428	0,32187	-0,27762	0,32066	0,34396	-0,06985	-0,28595	-0,35457
VAR096	0,01784	0,0482	-0,05086	-0,07028	-0,05441	0,26695	0,53645	0,10883	0,85022	0,14023	1	0,79325	0,00993	0,03887	-0,06331	0,01696	0,04623	-0,32175	-0,06751	-0,22068
VAR098	-0,06197	-0,02809	-0,00076	-0,01652	-0,00775	0,20792	0,33422	0,05528	0,62464	0,00405	0,79325	1	-0,07395	-0,0474	0,00171	-0,06338	-0,0314	-0,35671	-0,01185	-0,09845
VAR153	0,93904	0,83755	-0,49885	-0,52013	-0,49138	0,24884	0,23315	0,18268	0,12728	0,28428	0,00993	-0,07395	1	0,89833	-0,52269	0,93812	0,83554	-0,36506	-0,51751	-0,40778
VAR160	0,87758	0,96977	-0,80286	-0,83598	-0,78668	0,27166	0,27477	0,1698	0,15892	0,32187	0,03887	-0,0474	0,89833	1	-0,84408	0,87637	0,96693	-0,58383	-0,83176	-0,65808
VAR172	-0,55682	-0,85961	0,94911	0,98741	0,92683	-0,22346	-0,24862	-0,10654	-0,15303	-0,27762	-0,06331	0,00171	-0,52269	-0,84408	1	-0,5556	-0,85656	0,68738	0,98242	0,7793
VAR174	0,99901	0,89604	-0,5398	-0,56271	-0,54155	0,28201	0,26369	0,21184	0,14887	0,32066	0,01696	-0,06338	0,93812	0,87637	-0,5556	1	0,89697	-0,40501	-0,56345	-0,43675
VAR181	0,89513	0,9971	-0,8331	-0,86753	-0,83459	0,2926	0,29312	0,18889	0,17408	0,34396	0,04623	-0,0314	0,83554	0,96693	-0,85656	0,89697	1	-0,62595	-0,87063	-0,67505
VAR191	-0,40333	-0,62265	0,70069	0,71522	0,68812	-0,13195	-0,27134	0,06854	-0,42798	-0,06985	-0,32175	-0,35671	-0,36506	-0,58383	0,68738	-0,40501	-0,62595	1	0,71906	0,61624
VAR193	-0,56111	-0,86623	0,95618	0,99502	0,95701	-0,23299	-0,25434	-0,11727	-0,15964	-0,28595	-0,06751	-0,01185	-0,51751	-0,83176	0,98242	-0,56345	-0,87063	0,71906	1	0,7755
VAR194	-0,43494	-0,67145	0,78476	0,77128	0,80433	-0,19837	-0,21255	-0,17498	-0,28647	-0,35457	-0,22068	-0,09845	-0,40778	-0,65808	0,7793	-0,43675	-0,67505	0,61624	0,7755	1

Anexo D – Matrizes de Confusão por Aplicação

Tabela D.1. Matriz de Confusão – ATTACK.

	ATTACK	Demais Classes	Total de Fluxos da Classe
ATTACK	1469	31	1500
Demais Classes	819	681	1500
Total de Fluxos do Agrupamento	2288	712	3000

Tabela D.2. Matriz de Confusão – DATABASE.

	DATABASE	Demais Classes	Total de Fluxos da Classe
DATABASE	1500	0	1500
Demais Classes	758	742	1500
Total de Fluxos do Agrupamento	2258	742	3000

Tabela D.3. Matriz de Confusão – FTP.

	FTP	Demais Classes	Total de Fluxos da Classe
FTP	1500	0	1500
Demais Classes	739	761	1500
Total de Fluxos do Agrupamento	2239	761	3000

Tabela D.4. Matriz de Confusão – MAIL.

	MAIL	Demais Classes	Total de Fluxos da Classe
MAIL	1500	0	1500
Demais Classes	616	884	1500
Total de Fluxos do Agrupamento	2116	884	3000

Tabela D.5. Matriz de Confusão – P2P.

	P2P	Demais Classes	Total de Fluxos da Classe
P2P	1274	226	1500
Demais Classes	989	511	1500
Total de Fluxos do Agrupamento	2263	737	3000

Tabela D.6. Matriz de Confusão – SERVICES.

	SERVICES	Demais Classes	Total de Fluxos da Classe
SERVICES	1491	9	1500
Demais Classes	649	851	1500
Total de Fluxos do Agrupamento	2140	860	3000

Tabela D.7. Matriz de Confusão – WWW.

	WWW	Demais Classes	Total de Fluxos da Classe
WWW	1138	362	1500
Demais Classes	178	1322	1500
Total de Fluxos do Agrupamento	1316	1684	3000

Livros Grátis

(http://www.livrosgratis.com.br)

Milhares de Livros para Download:

Baixar livros de Literatura
Baixar livros de Literatura de Cordel
Baixar livros de Literatura Infantil
Baixar livros de Matemática
Baixar livros de Medicina
Baixar livros de Medicina Veterinária
Baixar livros de Meio Ambiente
Baixar livros de Meteorologia
Baixar Monografias e TCC
Baixar livros Multidisciplinar
Baixar livros de Música
Baixar livros de Psicologia
Baixar livros de Química
Baixar livros de Saúde Coletiva
Baixar livros de Serviço Social
Baixar livros de Sociologia
Baixar livros de Teologia
Baixar livros de Trabalho
Baixar livros de Turismo