FUNDAÇÃO DE ENSINO “EURÍPIDES SOARES DA ROCHA”
CENTRO UNIVERSITÁRIO EURÍPIDES DE MARÍLIA – UNIVEM
TRABALHO DE CONCLUSÃO DE CURSO

HELEN PAZINI FERREIRA

PROPOSTA DE UMA REDE NEURAL SOM PARA CLASSIFICAÇÃO DE PACOTES DE DADOS QUE TRAFEGAM POR REDE DE COMPUTADORES

Marília
2011

HELEN PAZINI FERREIRA

PROPOSTA DE UMA REDE NEURAL SOM PARA CLASSIFICAÇÃO DE PACOTES DE DADOS QUE TRAFEGAM POR REDE DE COMPUTADORES

TCC – Trabalho de Conclusão de Curso apresentado ao Curso de Bacharelado em Ciência da Computação da Fundação de Ensino "Eurípides Soares da Rocha", mantenedora do Centro Universitário Eurípides de Marília – UNIVEM. Sob orientação do Prof. Ms. Rodolfo Barros Chiaramonte, como requisito parcial para obtenção do título de Bacharel em Ciência da Computação.

Marília
2011

FERREIRA, Helen Pazini

Proposta de uma rede neural SOM para classificação de pacotes de dados que trafegam por rede de computadores / Helen Pazini Ferreira; orientador: Ms. Rodolfo Barros Chiaramonte. Marília, SP: [s.n.], 2011.

49 f.

Trabalho de Curso (Bacharelado em Ciência da Computação) – Fundação de Ensino "Eurípides Soares da Rocha", mantenedora do Centro Universitário Eurípides de Marília – UNIVEM, Marília, 2011.

CDD: 004.6

1. RNA 2. Redes de Computadores 3. SOM

CENTRO UNIVERSITÁRIO EURÍPIDES DE MARÍLIA
BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

TRABALHO DE CONCLUSÃO DE CURSO – AVALIAÇÃO FINAL

Helen Pazini Ferreira

PROPOSTA DE UMA REDE NEURAL SOM PARA CLASSIFICAÇÃO DE PACOTES DE DADOS QUE TRAFEGAM POR REDE DE COMPUTADORES

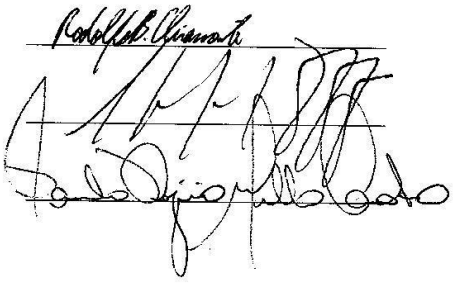
Banca examinadora da monografia apresentada ao Curso de Bacharelado em Ciência da Computação do UNIVEM/F.E.E.S.R., para obtenção do Título de Bacharel em Ciência da Computação.

Nota: 8,5 (oito e meio)

Orientador: Rodolfo Barros Chiaramonte

1°. Examinador: Emerson Alberto Marconato

2°. Examinador: Paulo Rogerio de Mello Cardoso



Marília, 28 de novembro de 2011.

Dedicatória

Dedico este trabalho:

Aos meus pais, Constantino e Alexandra Ferreira, os quais acreditaram e investiram em mim para que essa vitória fosse alcançada.

A minha amiga, muito mais que irmã, Andressa Ferreira, pelo apoio, paciência e incentivo.

Aos amigos e familiares, que de alguma forma estiveram ao meu lado, seja com palavras ou gestos.

E ao meu futuro marido, Diogo Pereira, que esteve sempre ao meu lado e nunca me deixou desistir.

Grata a todos.

Agradecimentos

Agradeço em primeiro lugar a Deus pela sabedoria que me deste para chegar até aqui.

Agradeço grandemente aos meus pais pelo custeio e esforço para minha formação.

Agradeço aos meus amigos de sala, de estudos, de sorrisos e atritos, que de me acompanharam por estes anos.

Grata aos docentes e em especial ao meu orientador e amigo Rodolfo pela paciência, orientação, esclarecimentos e por acreditar nos frutos deste trabalho.

"Para realizar grandes conquistas,
devemos não apenas agir,
mas também sonhar;
não apenas planejar,
mas também acreditar."

Anatole France

FERREIRA, Helen Pazini. **Proposta de uma rede neural SOM para classificação de pacotes de dados que trafegam por rede de computadores**. 2011. 49 f. Trabalho de Conclusão de Curso (Bacharelado em Ciência da Computação) – Centro Universitário Eurípides de Marília, Fundação de Ensino "Eurípides Soares da Rocha", Marília, 2011.

Resumo

O número de usuários que acessam informações pela maior rede de computadores mundialmente conhecida, a Internet, dobrará até 2012 chegando a dois bilhões. Contudo, como saber se as informações acessadas através de uma rede de computador são seguras?

Este trabalho aborda conceitos sobre Redes de Computadores no que diz respeito ao modelo ISO/OSI (*International Organization for Standardization/Open Systems Interconnection*), os tipos de protocolos por camada e captura de pacotes utilizando software verificador de tráfego de rede. Paralelamente é realizado o estudo referente a Redes Neurais Artificiais (RNA) e abordado os conceitos gerais sobre o modelo MCP (sigla para McCulloch e Pitts), arquitetura de redes neurais, tipos de aprendizado e por fim enfatizar as características de uma rede neural auto-organizável do tipo SOM (do inglês *Self-Organizing Maps*).

O principal objetivo deste trabalho é validar uma metodologia capaz de analisar alguns exemplos de protocolos pertencentes à camada de Aplicação e ordená-los segundo os conceitos de treinamento de uma rede neural artificial do tipo SOM.

O resultado obtido da ordenação, através do reconhecimento de relações entre padrões apresentados à entrada de uma RNA, possibilita a um administrador de rede de computador verificar pacotes de dados que apresentem algum tipo de anormalidade com relação aos demais pacotes da amostra de rede.

FERREIRA, Helen Pazini. **Proposta de uma rede neural SOM para classificação de pacotes de dados que trafegam por rede de computadores**. 2011. 49 f. Trabalho de Conclusão de Curso (Bacharelado em Ciência da Computação) – Centro Universitário Eurípides de Marília, Fundação de Ensino "Eurípides Soares da Rocha", Marília, 2011.

Abstract

The number of users accessing information through the largest network of computers known worldwide, the Internet, will double by 2012 reaching two billion. However, how know if the information accessed through a computer network are secure?

This document discusses concepts about Computer Networks with respect to the model ISO/OSI (International Organization for Standardization/Open Systems Interconnection), the types of protocols by layer using packet capture and checker traffic software of network. Alongside the study about Artificial Neural Networks (ANN) are discussed the general concepts of the model MCP (stands for McCulloch and Pitts), neural network architecture, types of learning and finally to emphasize the characteristics of a neural network self-organizing the type SOM (English Self-Organizing Maps).

The main objective of this study is to validate a methodology able to analyze some examples of protocols belonging to the Application layer and sort them according to the concepts of training a artificial neural network SOM type.

The result of sorting through the recognition of relationships between patterns presented to the entry of an ANN, allows that network administrator verifies computer data packets that show some kind of abnormality in relation to other packages in the sample network.

Lista de Figuras

Lista de Tabelas

LISTA DE ABREVIATURAS

CSBC: Congresso da Sociedade Brasileira de Computação

DNS: Domain Name System

ENIA: Encontro Nacional de Inteligência Artificial

HTTP: HyperText Transfer Protocol

IA: Inteligência Artificial

IP: Internet Protocol

ISO: International Organization for Standardization

MCP: McCulloch e Pitts

OSI: Open Systems Interconnection

PPP: ponto-a-ponto

RNA: Rede Neural Artificial

SOM: Self-Organizing Maps

SSDP: Simple Service Discovery Protocol

WWW: World Wide Web

TCP: Transmission Control Protocol

SUMÁRIO

Capítulo 1

Introdução

É crescente o volume de usuários que trocam informações na maior rede heterogenia de computadores mundialmente conhecida atualmente, a Internet, como mostra os dados estatísticos do Ibope Nielsen Online do dia 18 de Março de 2011. O acesso à rede no Brasil cresce 9,6% e chega a 73,9 milhões de internautas e todo esse crescimento se traduz em maior tráfego de informações e conseqüentemente compartilhamento de recursos. O tráfego de rede, por onde circulam milhares de informações pelos canais da rede desperta tanto interesse indireto aos usuários que buscam obter informações por meio da rede, quanto a pesquisadores e administradores de sistema ligados a área de redes de computadores.

Pesquisadores buscam analisar comportamentos e prováveis impactos na infra-estrutura da rede, com a finalidade de propor novas soluções que garantam a integridade dos serviços e compartilhamento dos recursos. Ao realizar análise do tráfego de rede, pesquisadores podem compreender melhor o comportamento de novas aplicações como, por exemplo, compartilhamento de arquivos em redes ponto-a-ponto (Arthur & Panigrahy, 2006), ou mesmo identificar o comportamento de disseminadores de mensagens de *spam* por correio eletrônico (Steding-Jessen et al., 2008). Toda informação obtida por monitoramento de funções da rede e disponibilizada para a comunidade científica, são importantes para a evolução da pesquisa na área de rede, pois tendo grande volume de dados para testar novas tecnologias, gera-se maior credibilidade à pesquisa (Melo, 2009).

No que diz respeito a administradores de sistema de rede, uma vez coletadas e armazenadas as informações contidas no tráfego da rede, é possível registrar, para fins históricos, as atividades da rede. Desta forma permite-se futura análise para identificação de comportamentos maliciosos na rede que possam indicar abusos ou ataques à infra-estrutura e serviços sob sua responsabilidade e para fins de auditoria (Bishop et al., 2006).

Exigências de coleta por parte dos administradores de redes vêm sendo, inclusive, objeto de algumas propostas de legislação em vários países e inclusive no Congresso Nacional Brasileiro (Senado Federal, 2008). Cada vez mais, elementos de auditoria interna de empresas e até mesmo perícia criminal dependem de dados coletados em máquinas ou no tráfego de redes.

Pelos motivos apresentados, torna-se importante a análise das informações contidas nos pacotes de dados que trafegam por redes de computadores. A classificação dos pacotes de dados não é uma tarefa trivial, visto que é necessário conhecer as informações descritas no cabeçalho de cada pacote, ou seja, cada pacote carrega consigo um conjunto de informações como tipo, origem e destino, tamanho, conteúdo, entre outras variações como criptografia, que os identifica.

Surge aqui a proposta de coletar uma amostra de tráfego de rede e organizar os pacotes de dados, contidos nesta amostra, em tipos comuns de protocolo. Foi desenvolvida uma aplicação, em linguagem JAVA, utilizando a lógica de treinamento de rede neural artificial tipo SOM, que recebe como entrada os dados brutos bit a bit, calcula-se a distância entre os dados de entrada e os pesos atribuídos aos neurônios e gera a ordenação dos dados por tipos de protocolos de rede. A SOM possui, em sua maioria, "a possibilidade de um desempenho superior ao dos modelos convencionais" (Braga et al., 2007) de RNA.

É nesse contexto que se insere o trabalho aqui apresentado.

1.1 Motivação

Garantir que o aprendizado da rede neural artificial identifique nos pacotes de dados "características estatisticamente relevantes" (Braga et al., 2007) é uma tarefa de alto desempenho, visto que cada pacote, que trafega na rede, contém controles que os caracterizam e devem ser identificados. Entre os controles estão os protocolos, os quais possibilitam conexão, comunicação ou transferência de dados entre no mínimo dois sistemas computacionais, realizando interação entre componentes de software e hardware dos computadores.

Considerando-se, por exemplo, um administrador de rede que deseja avaliar uma amostra do tráfego da rede de determinada instituição ou mesmo disponibilizar dados de tráfego para pesquisa, e desta forma, gerar um gráfico visual dos pacotes acessados. Nesse caso, para garantir a confiabilidade dos dados da amostra e não menos importante a privacidade dos usuários, no projeto proposto o administrador visualizará somente os dados resultantes enfatizando os protocolos pertencentes à camada de aplicação. O resultado demonstra os dados por zona de vizinhança de pacotes, ou seja, pacotes de dados organizados

por protocolos que possuem o mesmo rótulo (mesma nomenclatura padrão) ordenadamente. Espera-se que os pacotes de dados que havendo algum tipo de alteração no que diz respeito ao padrão descrito no cabeçalho ou até mesmo em seu conteúdo e não pertençam à mesma zona de vizinhança, sejam avaliados de forma mais detalhada pelo administrador da rede.

Diante de situações similares a essa, onde se torna necessária uma metodologia que valide o critério de ordenação dos pacotes de dados por rotulação de nomes de protocolos, garantindo a categorização mais próxima possível dos dados de entrada.

1.2 Objetivo

O principal objetivo desta dissertação é classificar os pacotes de dados, organizando-os por tipos comuns de protocolos.

Cada protocolo é pertencente a uma determinada camada do modelo ISO/OSI, além de serem responsáveis de modo específico por métodos de acesso e controle ao meio (conexão, comunicação ou transferência de dados).

Alguns exemplos de protocolos como HTTP (*HyperText Transfer Protocol*) e DNS (*Domain Name System*), pertencentes à camada de Aplicação do modelo ISO/OSI e têm por responsabilidade, respectivamente, efetuar a transferência utilizada em toda a *World Wide Web* (WWW) e criar um esquema hierárquico de atribuição de nomes baseado no domínio e um esquema de banco de dados distribuído para implementar esse esquema de nomenclatura (Tanenbaum, 2003).

Para validar o objetivo principal deste trabalho, é descrito, de forma detalhada, os seguintes objetivos específicos:

- Coletar amostra de tráfego de rede por meio de um software verificador de tráfego de dados. Utiliza-se a ferramenta *Wireshark*.
- Desenvolver uma RNA, em linguagem JAVA, utilizando a lógica do algoritmo proposto para redes auto-organizáveis, ou seja, “neurônios topologicamente próximos tendem a responder a padrões ou estímulos semelhantes” (Braga et al., 2007).
- As entradas da RNA recebem os pacotes de dados brutos (bit a bit), que se auto-organizarão em grupos comuns de protocolos, através do cálculo da distância

Euclidiana e escolha do neurônio vencedor, ou seja, o neurônio que possui a menor distância, a cada época de treinamento.

- Observar ao término do treinamento os resultados da organização e validar uma metodologia capaz de classificar as informações de entrada e distribuí-las de forma visual em formato texto (relatório).
- Finalmente, facilitar a análise visual do administrador de rede, ou mesmo disponibilização dos resultados para pesquisa.

1.3 Contribuição

Complementar ao objetivo descrito anteriormente, a metodologia proposta busca permitir aos administradores de rede, analisar o resultado em formato texto (relatório), após aplicação da lógica de uma RNA tipo SOM, que analisa o arquivo de dados brutos do tráfego de rede coletado.

Com base nessa análise seria possível avaliar possíveis ataques, quando detectados pacotes logicamente organizados de forma não padrão aos demais. Com o resultado observado pelo administrador, para maiores detalhes, este geraria um processo de detalhamento de acesso e controle, através de históricos da rede.

Em nível de pesquisa, os resultados disponibilizados para análise de comportamento da rede visam desenvolver novas aplicações que garantam a integridade dos serviços, compartilhamento dos recursos e integridade de infraestrutura da rede.

1.4 Organização do texto

Nos capítulos seguintes serão descritos, primeiramente no capítulo 2, os principais conceitos sobre redes de computadores, sendo apresentado o modelo ISO/OSI e os principais protocolos por camada de rede, os tipos e processos de coleta de dados, além das formas de análise que o *sniffer Wireshark* apresenta aos usuários. Em seqüência, no capítulo 3, são apresentadas as principais características e conceitos referentes a redes neurais artificiais, com breve ligação conceitual a redes biológicas quanto à disposição dos neurônios, que são as

inspirações para uma RNA. Nos subitens do capítulo 3 são apresentados: o primeiro modelo neural MCP, as formas de arquitetura e os tipos de aprendizagem em redes neurais. Seguindo, no item 3.4 é exposto o tipo de rede SOM, o qual é baseado esse trabalho, a forma de treinamento e rotulação das saídas apresentadas pela rede neural. Por fim no item 3.5 são abordados exemplos de aplicações desenvolvidas para área de redes neurais artificiais aplicadas.

Como considerações finais são apresentadas a metodologia proposta, os resultados obtidos, conclusão e trabalhos futuros, respectivamente nos capítulos 4 e 5.

Capítulo 2

Conceitos sobre Redes de Computadores

Para uma melhor compreensão dos diversos aspectos relacionados à área de rede neural artificial aplicada a resolução de problemas no reconhecimento de padrões, aqui relacionados à classificação de pacotes de dados de uma amostra de arquivos de tráfego de rede, é importante que seja discutido anteriormente conceitos gerais sobre redes de computadores, características do modelo ISO/OSI, qual o tipo de informação contida nos protocolos específicos por camada do modelo e quais os tipos de informações demonstradas por ferramentas verificadoras de tráfego de rede. Portanto, as seções a seguir descrevem com mais detalhes cada tópico citado a cima.

2.1 Redes de Computadores

Segundo Tanenbaum, o conceito de redes de computadores é definido como “um conjunto de computadores autônomos interconectados por uma única tecnologia”, visto que “dois computadores estão interconectados quando podem trocar informações”, seja por fio de cobre, fibra óptica, microondas, ondas de infravermelho ou satélites de comunicação.

As redes podem ser usadas em aplicações comerciais, aplicações domésticas ou por usuários móveis, por exemplo, onde as trocas de dados são constantes. Contudo, este meio pelo qual os cidadãos manifestam suas opiniões, buscam informações, enviam dados, mesmo de forma possivelmente incorreta, enganosa ou completamente equivocada (Tanenbaum, 2003), tem crescido de forma desenfreada. Será que estamos realmente seguros? Mesmo acessando informações por dispositivos privados ou equipamentos domésticos, pode-se confiar no conceito sobre privacidade?

A resposta para estas questões é freqüentemente “não”. Um dos maiores problemas enfrentados pelo armazenamento de informações confidencias nos computadores é o roubo de identidade, onde ladrões coletam dados pessoais suficientes para clonarem cartões de créditos e documentos pessoais. Porém toda essa questão pode ser solucionada com investimento em

segurança, tanto por parte das indústrias de tecnologia em informática, quanto aos usuários comuns no que diz respeito à boa ética.

Sabe-se que a Internet é, atualmente, a maior rede heterogênea de computadores interligada por milhões de dispositivos computacionais espalhados ao redor do mundo (Figura 2.1).

Figura 2.1. Rede corporativa – Visão do conceito de Internet (Souza, 2008).

As aplicações de rede, como paginação na *Web* e transferência de arquivos, são exemplos que executam em sistemas terminais. Sistemas terminais realizam a comunicação por protocolos de rede, que são responsáveis pela conexão, comunicação ou transferência de dados em rede de computadores. Os sistemas terminais são interconectados por meio de enlace de comunicação, que podem ser divididas em dois tipos básicos: enlace ponto-a-ponto (tipo PPP) ou multiponto (rede local Ethernet), já a técnica de troca de dados pode ser definida como comutação de pacotes ou comutação de circuito.

A comutação de circuito é comumente utilizada nos sistemas telefônicos. A comutação de pacotes é a forma como as mensagens chegam aos enlaces de entrada da rede, são armazenadas e enviadas ao enlace de saída, seguindo uma rota até seu destino. Na comutação de pacotes, as mensagens transmitidas são fragmentadas em pacotes menores, os quais são transmitidos pela rede de modo independente uns dos outros.

Os pacotes de dados são formados por seqüência de bytes, que carregam informações que os ajudarão a chegar ao seu destino. Podem ser endereço IP (*Internet Protocol*)

emissor/destino, tipos de protocolo de rede, tamanho, conteúdo, entre outras características. Cada pacote, em sua maioria, é divido em três partes:

- Cabeçalho: contém instruções sobre os dados contidos no pacote (comprimento, sincronização, número de pacotes, protocolo, endereço origem /destino).
- Corpo: são os dados propriamente ditos, ou seja, o conteúdo da mensagem.
- Rodapé: possuem bits responsáveis por delimitar o tamanho do pacote e avisar ao dispositivo que foi alcançado seu fim ou realizar verificação de erros.

Para que todo esse processo de comutação e comunicação ocorra sem zona de conflito entre linguagem e forma de envio e recepção por equipamentos, houve-se a necessidade de se padronizar os hardwares e softwares responsáveis pelo tráfego de informações (vias). As vantagens à padronização são: reduzir a complexidade, padronizar as interfaces, facilitar a engenharia modular, garantir a tecnologia interoperável, acelerar a evolução e simplificar o ensino e o aprendizado (Torres, 2009).

No tópico 2.2 serão apresentados os conceitos sobre protocolos, sobre o modelo ISO/OSI e características relevantes de cada camada pertencente ao modelo.

2.2 Modelo ISO/OSI e Protocolos por Camada

O Modelo ISO/OSI conceitua um esquema conceitual que permite o trabalho de forma produtiva e independente no desenvolvimento de padrões. Apesar de não ter sido adotado para fins comerciais, ele apresenta de forma clara um sistema de comunicação, e facilita a compreensão dos tipos de protocolos e suas funções particulares na rede de computador (Comer, 2007).

O Modelo propõe a divisão abstrata da topologia da rede em sete camadas, de forma hierárquica. As características de cada camada serão apresentadas a seguir de modo resumido.

1ª Física: Define os aspectos mecânicos eletrônicos de transferência de dados e a interface de hardware. Também responsável pela transmissão de bits de um ponto a outro.

2ª Enlace ou Ligação de Dados: Responsável pela correta transmissão de dados através da camada física, onde é lida e atribuída uma estrutura lógica. As informações vindas de camadas superiores são convertidas em quadros.

3ª Rede: Fornece endereços para os dados. Responsável por traduzir os quadros e transformá-los em pacotes de dados.

4ª Transporte: Responsável por estabelecer, manter e terminar conexões lógicas pelo endereçamento e pela transição entre endereços lógicos e físicos. A informação é dividida em pequenos segmentos que ao chegarem ao destino, com sucesso, são novamente agrupados. Caso haja envio de dados incorretos, a camada de transporte é responsável por pedir nova retransmissão da informação.

5ª Sessão: Define a ligação entre dois computadores e coordena a interação entre eles. Havendo programas que rodam em máquinas isoladas, a camada de sessão estabelece a conversação entre elas. Sua função também é de estruturar circuitos oferecidos pela camada de transporte, realizando desta forma o controle do diálogo.

6ª Apresentação: Verifica a compatibilidade entre dois equipamentos, convertendo, caso necessário, os dados através da criptografia da informação. Soluciona problemas de sintaxe.

7ª Aplicação: Define os protocolos para serem utilizados para determinadas tarefas. Responsável por definir as ações de interface da rede, oferecendo serviços de correio eletrônico, bancos de dados distribuídos, transferência de arquivos e estabelecimento de sessões entre usuários.

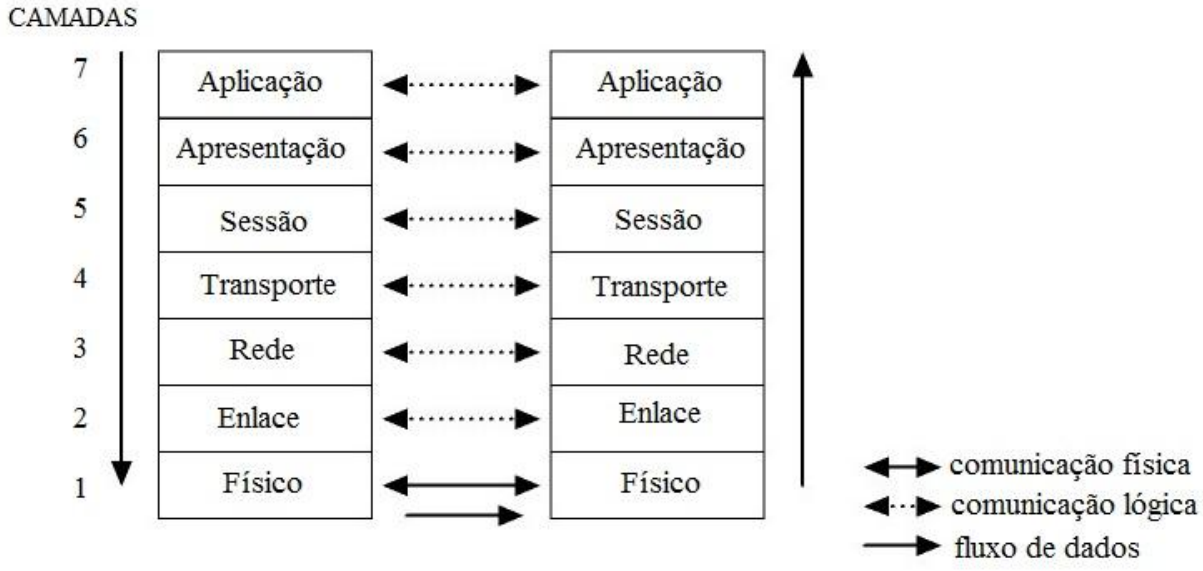


Figura 2.2. Modelo ISO/OSI conceitual em sete camadas (Schubert, 2010).

Todas as camadas, ilustradas na Figura 2.2, envolvem protocolos que possuem tarefas específicas em meio à atividade de uma rede de computadores. Os protocolos serão de extrema importância quanto à classificação de pacotes de dados e campos de verificação para auxiliar a ordenação classificatória da RNA.

Portanto faz-se necessário definir o que é um protocolo e qual sua finalidade em uma rede.

Protocolo é o padrão para que dois computadores se comuniquem, ou seja, uma linguagem comum que viabiliza a conversação entre os agentes (Cantu, 2003). É importante citar que o protocolo de rede é independente de sistema operacional, desta forma é possível acessar, por exemplo, a *Web* através de um sistema operacional *Linux* ou *Microsoft Windows*, pois estes obedecem a uma ordem de protocolos TCP/IP (*Transmission Control Protocol/Internet Protocol*) e HTTP.

Toda interação entre componentes de software e hardware dos computadores é realizada por protocolos que buscam garantir a integridade dos dados transmitidos ou determinar o caminho de um pacote de dados da fonte até seu destino.

A Tabela 1 ilustra a hierarquia dos protocolos, que visam reduzir a complexidade do projeto, organizando a rede em uma série de níveis (camadas).

Tabela 1. Hierarquia organizacional dos protocolos por camada (autoria própria).

Camada	Exemplos
7 - Aplicação	HTTP, DNS, SSDP, SMTP, SNMP, FT, NFS, NTP, BOOTP, DHCP, RMON, TFTP, POP3, IMAP, TELNET
6 - Apresentação	XDR, SSL, TLS
5 - Sessão	ASP, ADSP
4 - Transporte	TCP, UDP, RTP, SCTP, ATP, NBP, AEP, RTMP
3 - Rede	IP, ICMP, IPsec, RIP, OSPF, BGP,ARP
2 - Enlace de dados	MTP-2, MAC (Media Access Control), SDLC
1 - Físico	X.25

2.3 Processos de Coleta e Análise de Dados

O processo de coleta e análise de dados pode ser efetuado por meio de um software verificador de tráfego de rede. Softwares verificadores de tráfego também podem ser utilizados para se obter informações sobre o comportamento dos usuários e infra-estrutura de rede.

2.3.1 Tipos de coleta

Para cada tipo de dado e necessidade de determinada análise, um tipo de coleta é especificado. Ao se focar um tipo de aplicação ou serviço, administradores e pesquisadores na área de redes se utilizam de registros de *logs* (atividades da rede) resultantes do relatório gerado pelos programas servidores que realizam determinados serviços, como por exemplo, estudar a carga de requisições de um servidor *Web*. Neste caso, os registros estão mais detalhados e é possível resumir as informações coletadas com base no conhecimento de semântica de tráfego de rede.

Outra face da coleta de dados é a obtenção de dados brutos, onde todo conteúdo de qualquer comunicação que atravessa um canal pode ser monitorado. Geralmente realizadas em terminais de usuário ou servidores de aplicação, pois são os pontos que possuem dados suficientes para interpretar as requisições do usuário e respostas do servidor. Esse tipo de análise permite obter uma visão global de toda comunicação da rede, ou seja, implica na participação do administrador da rede em questão. Toda informação sobre cada aplicação é relacionada com os bytes trafegados, além dos cabeçalhos dos diversos protocolos trazerem informações que identifiquem a máquina origem/destino e até mesmo o usuário envolvido no processo.

Outro tipo de coleta é a de dados sobre fluxos, que analisa os padrões de tráfego, como volumes, origens e destinos, entre cada par origem/destino observado por unidade de tempo, sem preocupação com a semântica dos serviços. É resultante o endereço IP de origem e destino contendo informações que podem afetar a privacidade do usuário e/ ou a segurança da rede.

Este trabalho foca a coleta de uma amostra de tráfego de rede com dados brutos, por ser uma aplicação global e abrangente.

O resultado de um tráfego bruto contém todas as informações de pacotes de dados. Esse tipo de dados é obtido através de roteadores ou *switches*. Ao analisar o conteúdo dos dados, podem ser obtidas informações sobre origem e destino do mesmo, tipo de serviço usado, horário da conexão, conteúdo da comunicação, identificação do usuário, entre outras informações que trafegam pela rede (Tanenbaum, 2003).

Contudo os administradores buscam avaliar os dados de conexão e observar o uso da rede, identificando possíveis ataques, origem de determinado tipo de tráfego, buscando manter a integridade e um bom funcionamento da rede. Já os pesquisadores podem utilizar dos dados brutos para uma melhor análise quanto as característicos de fluxo de tráfego, e estudar mais detalhadamente determinados protocolos.

2.3.2 Processo de coleta

O processo de coleta pode ser dividido em três partes principais: acesso aos dados do tráfego, coleta de tráfego e definição de parâmetros (coleta propriamente dita) e finalmente a análise de resultados.

2.3.2.1 Acesso aos dados do tráfego

Primeiramente para que se possa realizar a coleta de dados brutos, faz-se necessária encontrar um método de acesso a todo conteúdo dos pacotes que trafegam por um canal de interesse. Se o canal for um servidor, onde são obtidos dados de uma máquina específica, então somente é acessada a máquina hospedeira e ativado o software verificador de tráfego de rede para realizar a coleta. Porém se é desejado coletar e analisar todo tráfego de entrada e saída de dados de uma rede se faz necessário ter acesso ao canal que conecta a rede ao restante da Internet. Nesse caso, é comum que haja apenas roteadores ou chaves Ethernet (*switches*) nas extremidades do canal. Em situações como a descrita anteriormente, onde não é

possível se realizar diretamente uma coleta, há a alternativa de se usar as técnicas de interceptação ou espelhamento do tráfego, que serão descritas a seguir.

Para realizar a interceptação de dados, um computador com duas interfaces de rede deve ser colocado no meio do fluxo de dados. É necessário usar cada interface para se conectar a um dos extremos do canal original que se deseja monitorar. A cópia do fluxo de pacotes de dados, obtido pela entrada por uma das interfaces para a outra, é gerenciado pelo sistema operacional que dever garantir que o fluxo no canal seja mantido inalterado. Paralelamente, o sistema deve copiar cada pacote recebido para um arquivo de armazenamento local, arquivo tal que constitui o registro de tráfego.

No espelhamento do tráfego, um elemento de rede (roteador ou chave Ethernet) assume essa funcionalidade. Nesse caso, a rede determina a interface de rede do canal de interesse, onde o elemento de rede pode ser programado para realizar uma cópia de cada pacote recebido ou enviado. A cópia é transmitida por outra interface do mesmo elemento de rede, à qual se pode então conectar o computador de coleta. A máquina apenas armazena cada pacote que recebe através da interface, sem nenhum outro tratamento.

A interceptação exige normalmente um computador com mais recursos, pois precisa de duas interfaces de rede, configuração para copiar os dados recebidos do tráfego e garantir um armazenamento sem perdas. Porém apresenta facilidade quanto à prática de coletas, pois não impõe grandes exigências sobre a rede a ser monitorada.

O espelhamento também reduz a demanda sobre o computador de coleta, pois necessita ser apenas capaz copiar os dados recebidos para um arquivo. Porém é dependente do elemento de rede no ponto da coleta que controla os recursos de espelhamento de tráfego.

Figura 2.3, ilustra os tipos de acesso aos dados: intercepção (a) e espalhamento (b).

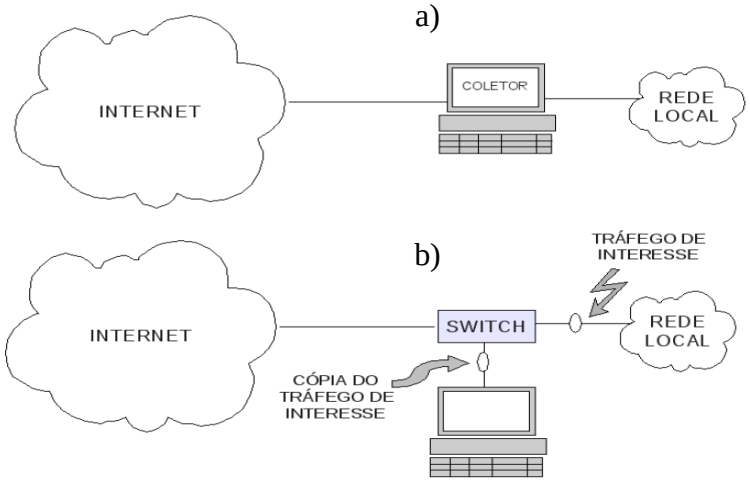


Figura 2.3. Tipos de acesso aos dados: intercepção (a) e espalhamento (b) (Melo, 2009).

2.3.2.2 Coleta de tráfego

Em união com a técnica adotada para se obter o acesso ao tráfego dos dados de rede, está o software de coleta propriamente dito. Nesse trabalho, é utilizado para se ter acesso aos dados diretamente de um servidor, dentre outros aplicativos existentes, o analisador (usualmente chamado de *sniffer* de rede) *Wireshark* (*Wireshark*, 2011).

Usualmente, o *Wireshark* monitora os dados juntamente com o driver WinPcap, que estende o sistema operacional para fornecer acesso de baixo nível e uma biblioteca utilizada para acessar as camadas de baixo nível da rede. O WinPcap é usado em sistemas operacionais *Windows*, já o LibPcap (biblioteca para processamento de *log*s) é a API disponível para sistemas operacionais *Unix*.

Por ser um *sniffer* multiplataforma, executado em ambiente *Windows* ou *Unix Like*, não havendo necessidade de uma configuração complexa para seu funcionamento e ter uma interface gráfica de boa interpretação, o *Wireshark* é muito utilizado em instituições e indústrias.

Uma de suas características mais relevantes para este trabalho é disponibilidade de se gerar um arquivo dos resultados de saída no formato texto ou no mesmo formato do próprio programa

Abaixo será descrito o processo utilizado para captura dos pacotes e armazenamento dos resultados.

- Primeiramente é necessário indicar qual interface de rede a ser rastreada, acesse o item no *menu* [Capture / Interfaces];
- Após clicar no botão [Start], se a interface estiver ativa, o rastreamento começa imediatamente e a janela principal do *Wireshark* passa a mostrar uma porção de pacotes, como mostrado na Figura 2.4;
- Para interromper a captura de pacotes, clique no botão [Stop] da barra de ferramentas ou no item de *menu* [Capture / Stop].
- Para salvar o resultado da captura de pacotes, clique no *menu* [File / Salve as], insira o nome e o tipo de formato que deseja armazenar.

O *Wireshark* disponibiliza a opção para salvar em arquivo texto e foi esta a forma utilizada neste trabalho.

A Figura 2.4 demonstra uma captura de pacotes realizada no dia 08 de Setembro de 2011, e mostra o painel contendo a lista dos pacotes capturados.

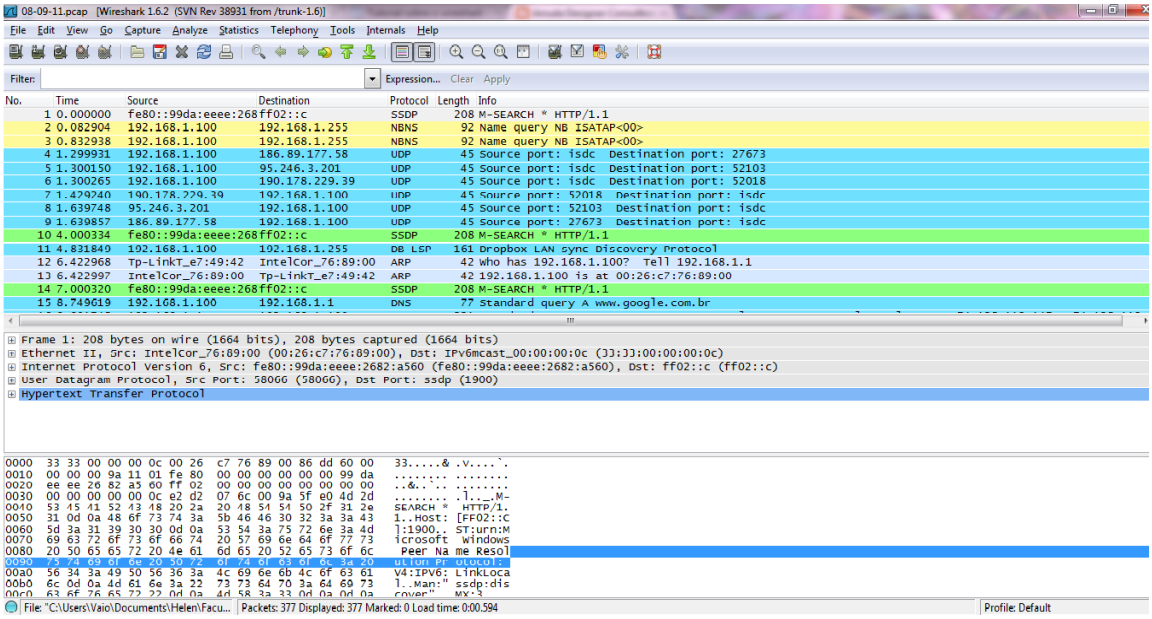


Figura 2.4. Painel contendo a lista dos pacotes capturados pelo programa *Wireshark* (autoria própria).

No tópico 4.2, serão explicados os métodos utilizados para realização da análise de resultados utilizando o programa *Wireshark*.

2.4 Conclusão

Foram descritos neste capítulo os conceitos sobre redes de computadores. Quanto a redes foram apresentadas as principais técnicas de coleta de pacotes de dados utilizando um software verificador de tráfego de rede, e as formas de análise de resultados descritiva no *Wireshark*

Capítulo 3

Conceitos sobre Redes Neurais Artificiais

Nos itens a seguir serão descritas as características e conceitos sobre redes neurais artificiais e apresentada a rede neural tipo SOM, a qual sua lógica de treinamento é baseada este projeto.

3.1 Redes Neurais Artificiais e Rede Neural Biológica

Simplificadamente, RNA pode ser descrita como um sistema composto por vários neurônios também intitulados como unidades de processamento simples (nodos). Essas unidades são interligadas por conexões chamadas sinápticas, onde são associados pesos as mesmas, as quais têm a função de ponderar e calcular determinada função matemática correspondente à entrada recebida por cada neurônio da rede. Os neurônios podem estar dispostos em uma ou mais camadas, geralmente unidirecionais (Braga et al, 2007).

3.1.1 Inspiração para Redes Neurais Artificiais

O funcionamento de uma RNA tem inspiração na estrutura física do cérebro humano, onde os neurônios de entrada da RNA correspondem aos neurônios biológicos do sentido, que recebem excitações do exterior. Os neurônios de saída correspondentes aos motoneurônios, que são os neurônios biológicos que excitam os músculos, utilizam as respostas das excitações para alterar, de alguma forma, o mundo exterior. Há ainda, os neurônios internos, conhecidos como "*hidden*", ou seja, neurônios que estão "encobertos" no interior da rede.

Como os neurônios processam as informações? Por meio de impulsos nervosos, transmissores de um sinal codificado de um estímulo dado ao longo da membrana do neurônio, a partir de seu ponto de aplicação. Os impulsos nervosos podem passar de uma célula a outra, criando assim uma cadeia de informação dentro de uma rede de neurônios (Cardoso, 2000).

Dois tipos de fenômenos estão envolvidos no processamento do impulso nervoso: os elétricos e os químicos. Os impulsos elétricos são responsáveis por propagar o sinal dentro de um neurônio, já o impulso químico transmite o sinal de um neurônio a outro ou para uma célula muscular.

O processo químico de interação entre os neurônios e entre os neurônios e células efetoras, acontecem na terminação do neurônio, em uma estrutura chamada sinapse. Aproximando-se do dendrito de outra célula (mas sem continuidade material entre ambas as células), o axônio libera substâncias químicas chamadas neurotransmissores, que se ligam aos receptores químicos do neurônio seguinte e promove mudanças excitatórias ou inibitórias em sua membrana. É desta forma que os neurotransmissores possibilitam que os impulsos nervosos de uma célula influenciem nos impulsos nervosos de outras, permitindo assim que as células nervosas se comuniquem.

O que dispara a liberação de um neurotransmissor? O potencial de ação estimula a entrada, que causa a adesão das vesículas sinápticas aos locais de liberação, sua fusão com a membrana plasmática e a descarga de seu suprimento de transmissor. O transmissor se difunde para a célula alvo, onde se liga a uma proteína receptora na superfície externa da membrana celular. Após um breve período o transmissor se dissocia do receptor e a resposta é terminada. Para impedir que o transmissor associe-se novamente a um receptor e recomece o ciclo, o transmissor, é destruído pela ação de catabolizar enzima, ou é absorvido, normalmente na terminação pré-sináptica. Cada neurônio pode produzir somente um tipo de transmissor (Cardoso, 2000). A Figura 3.1 mostra a estrutura física de uma célula nervosa.

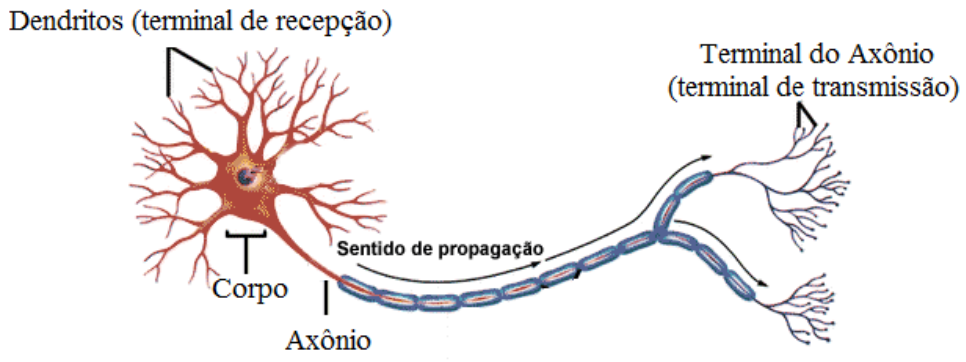


Figura 3.1. Célula nervosa e descrição das partes que a compõe (Louredo, 2011. Adaptado).

Com o objetivo de simular o comportamento de uma rede biológica, foi proposto na década de 1980 o primeiro modelo para uma RNA, chamado MCP. Esse foi o marco para início dos estudos voltados a área de RNA e Inteligência Artificial aplicadas.

No item 3.2, que se segue, será apresentada uma breve explicação sobre o modelo MCP e a importância da função de ativação para que seja gerada a saída de resultados da rede.

3.2 Rede Neural Artificial – Modelo MCP

O primeiro modelo de RNA foi proposto por Warren S. McCulloch, psiquiatra e neurologista, que por 20 anos dedicou-se a representar um evento do sistema nervoso, juntamente com Walter Pitts, matemático e recém-graduado na época. Ambos publicaram um artigo na Universidade de Chicago em 1943, intitulado "*A logical calculus of the ideas immanent in nervous activity*", *Bulltin of Mathematical Biophysics* (5: pg. 115-133).

O artigo trazia discussões sobre redes lógicas de nodos e novas idéias sobre a máquina de estados finitos (máquina de Turing), elementos de decisão de limiar lineares e representação lógica de várias formas de comportamento e memória. O trabalho enfatizava descrever um modelo artificial de neurônio e apresentar capacidade computacional, ao invés de técnicas de aprendizado (veja Figura 3.2).

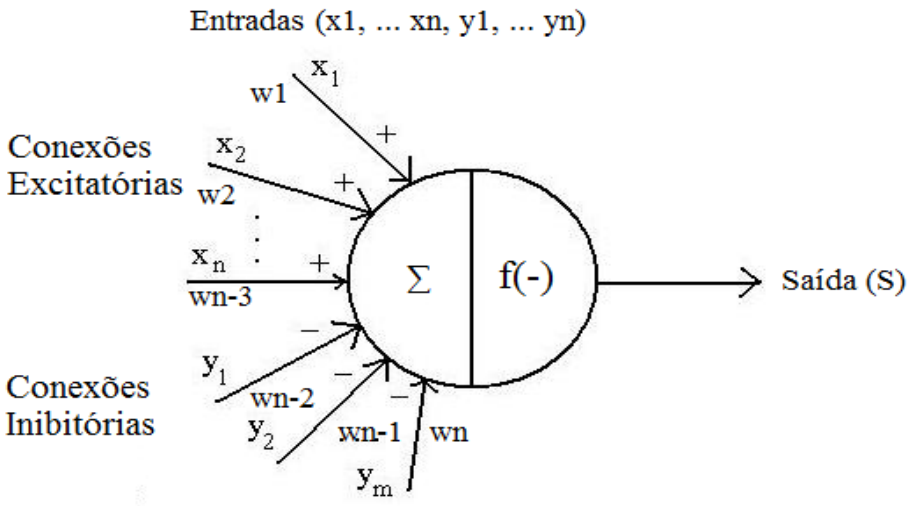


Figura 3.2. Neurônio MCP, no qual ∑ representa a soma ponderada das entradas e f (-) a função de ativação (Roque, 2008. Adaptado).

Pode-se observar no modelo uma descrição matemática com *n* terminais de entrada (dendritos) que recebem os valores de x_1, x_2, ... x_n, y_1, y_2, ..., y_n (representam as ativações dos neurônios anteriores) e um terminal de saída *S* (axônio). Para representar o comportamento

das sinapses, as entradas tem pesos acoplados $w_1, w_2, ..., w_n$ (que podem ser positivos ou negativos). O efeito de uma sinapse particular *i* no neurônio pós-sináptico é dado por $x_i w_i$.

O comportamento em um neurônio biológico chamado *threshold,* ocorre quando a soma dos impulsos ultrapassa o limite de excitações, representada no modelo pelo somatório $x_i w_i$.

A função de ativação de um neurônio MCP, responsável por gerar a saída (S) a partir dos valores do vetor de pesos e de entrada, é apresentada na Equação 3.1, sendo a mesma do tipo degrau deslocada do limiar de ativação θ em relação à origem (Braga et al, 2007).

$$f(\mathrm{u}) = \begin{cases} 1 \ \sum_{i=1}^{n} x_i w_i \geq \theta \\ 0 \ \sum_{i=1}^{n} x_i w_i \leq \theta \end{cases} \tag{3.1}$$

onde $\sum_{i=1}^{n} x_i w_i$ representa a soma ponderada das entradas e θ é o limiar (*threshold*) da função de ativação.

O tipo de função de ativação a ser aplica depende diretamente do tipo de problema a ser abordado. Para o trabalho proposto a função de ativação para rede SOM é baseada na medida de distância Euclidiana.

Além da função de ativação, há duas características relevantes referente à RNA: a arquitetura, ou seja, a forma como os neurônios ficam conectados em rede; e o tipo de aprendizado, ou seja, a capacidade de uma RNA aprender por meio de exemplos.

3.3 Arquitetura de RNAs e Tipos de Aprendizado

No requisito arquitetura (topologia), segundo Braga et al (p. 10, 2007), “um conjunto de neurônios artificiais conectados na forma de uma rede (neural) é capaz de resolver problemas de complexidade elevada”, mesmo possuindo capacidade computacional limitada, ou seja, a arquitetura é determinante na capacidade de processamento de uma RNA e pela escolha correta do número de conexões para obter um treinamento bem sucedido.

Há basicamente dois tipos de topologias: não recorrentes e recorrentes. Em redes neurais não recorrentes, não há realimentação em suas saídas para as suas entradas. Podem ser chamadas de redes neurais "sem memória". Suas estruturas geralmente aparecem em formato de camada única ou multicamadas. Em redes multicamadas existe um conjunto de neurônios de entrada, uma camada de saída e uma ou mais camadas intermediárias (encobertas). Um típico exemplo para essa arquitetura são as redes do tipo *feedfoward* (veja Figura 3.3), onde o sinal é propagado para frente, ou seja, da entrada para a saída.

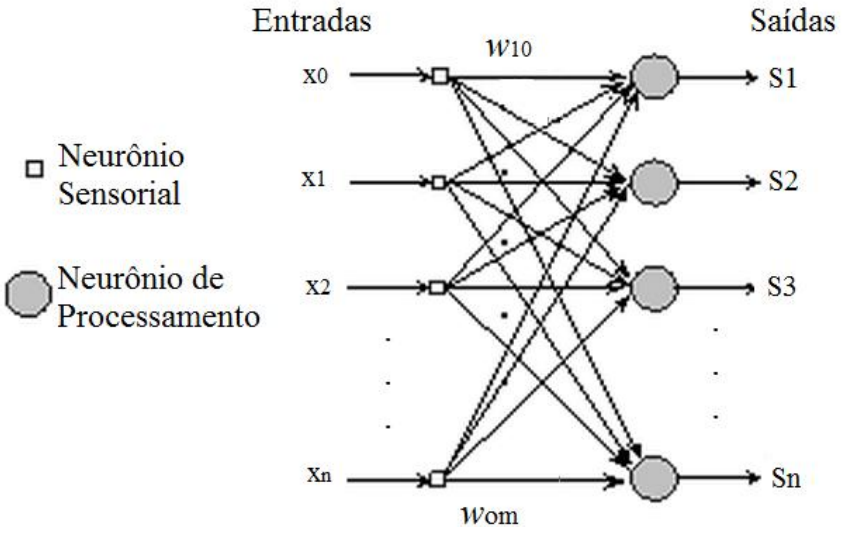


Figura 3.3. Rede *feedforward* com uma única camada (Ivcon, 2011).

Redes neurais recorrentes contêm realimentação (*feedback*) das saídas para as entradas, sendo suas saídas determinadas pelas entradas atuais e pelas saídas anteriores. Essa arquitetura pode, não necessariamente, conter camadas e apresentar interligações entre neurônios da mesma camada ou nas camadas não consecutivas, de modo dinâmico. São ditas redes neurais "com memória".

Existe ainda o caso de redes neurais parcialmente recorrentes, como por exemplo, a rede de Elman. Nessa arquitetura, a realimentação ocorre entre a saída e a entrada da primeira camada oculta, através de uma unidade de contexto (camada extra), normalmente uma estrutura de atraso, que armazena a saída da primeira camada oculta por um passo de tempo.

Um típico exemplo de redes neurais recorrentes é visto no modelo Hopfield (conhecida como auto-associativa, veja Figura 3.4).

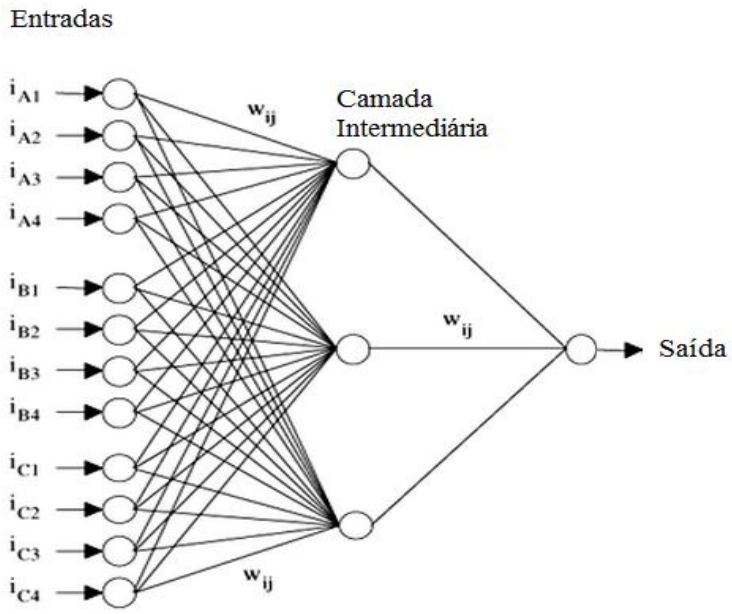


Figura 3.4. Rede Hopfield com recorrência entre entradas e camada intermediária (Segatto; Coury, 2011).

Outra característica, não menos importante, é o aprendizado da RNA. A aprendizagem ocorre através de treinamento, ou seja, a capacidade de aprender a partir de um conjunto de exemplos. Treinar uma RNA consiste em ajustar os pesos sinápticos (w_{ij}) para que se obtenha a saída (γ_i) desejada. Portanto, as formas de aprendizado podem ser divididas em dois grupos: Aprendizado supervisionado e Aprendizado não supervisionado.

No aprendizado supervisionado há necessidade de um agente supervisor que, durante o treinamento, é responsável por estimular as entradas da rede e observar as saídas desejadas. Os pesos das entradas são ajustados continuamente, por meio da função de ativação, para que as saídas se adaptem as saídas desejadas. O ajuste dos pesos é necessário caso o resultado da fórmula de correção de erro seja diferente de zero, ou seja, através da expressão genérica para o erro $e(t)$ no instante de tempo t, para $e(t) = \gamma_d(t) - \gamma(t)$, onde $\gamma_d(t)$ é a saída desejada e $\gamma(t)$ o resultado atual calculado pela saída da rede. Os exemplos mais comuns para esse tipo de aprendizado são a regra delta e o algoritmo de retropropagação (*backpropagation*) utilizado pela rede Perceptron multicamadas.

No aprendizado não supervisionado, como se pode concluir, não há um agente supervisor para acompanhar o processo de aprendizagem. Neste caso, apenas os padrões de entrada são apresentados à rede, a qual ajusta seus pesos para que entradas semelhantes gerem saídas semelhantes, todavia é essencial que haja redundância e regularidade nas entradas para

que seja possível um aprendizado não supervisionado. Um de seus principais exemplos para esse tipo de aprendizado é a lei de Hebb e o aprendizado por competição visto na rede SOM.

Há ainda o aprendizado por reforço, sendo considerado em algumas literaturas um caso particular do aprendizado supervisionado, pois utiliza tanto as entradas quanto as saídas durante o treinamento. A diferença é dada no que diz respeito ao cálculo do erro entre a saída desejada e saída resultante atual da rede, que na aprendizagem por reforço, não é calculado esse valor do erro, mas apenas é informado para a rede se a resposta está correta ou não.

Exposto até aqui o conceito geral sobre RNAs e as principais características das mesmas. No item 3.4 serão apresentadas as características de uma RNA tipo SOM, a qual esse trabalho é baseado no que diz respeito à implementação e análise de resultados de uma amostra de tráfego de rede.

3.4 RNA tipo SOM

A rede neural SOM, proposta por Teuvo Kohonen no ano de 1972, possui grande influência neurofisiológica, pois é baseado no mapa topológico presente no córtex cerebral, por esse motivo o nome sugestivo para organização em mapas (*Self-Organizing Maps)*. Assim como há áreas específicas no cérebro que desempenham determinadas funções, como fala, visão e audição, há a presença de subáreas que são responsáveis por mapear internamente as respostas das mesmas. Ao observar a organização de tais órgãos (veja Figura 3.5), é possível perceber como cada subárea está ordenada topologicamente e é provado que neurônios espacialmente próximos tendem a responder padrões ou estímulos semelhantes (Braga et al., 2007).

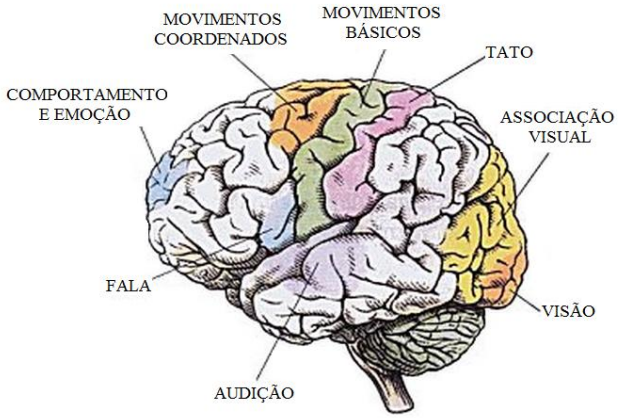


Figura 3.5. Disposição topológica das áreas do cérebro (São Francisco, 2011).

O modo como os neurônios, topologicamente ordenados, respondem aos estímulos, é chamado de *feedback*. A ocorrência de *feedback* entre os neurônios, é resultante da influência do estado de ativação dos neurônios vizinhos, e pode ser classificado em: excitatório, inibitório ou levemente excitatório.

O estímulo excitatório, ocorre quando os vizinhos estão diretamente próximos, ou seja, na primeira camada de influência (vizinhança R_1), próximos ao centro.

O estímulo inibitório ocorre ao redor dos vizinhos diretamente próximos, ou seja, fora da área de R_1, porém em uma segunda área R_2 ($R_2 > R_1$).

O estímulo levemente excitatório, ocorre quando os vizinhos estão dispostos em uma terceira área de raio R_3, onde $R_3 > R_2 > R_1$.

A Figura 3.6 mostra a disposição de dados por vizinhança, ou seja, a arquitetura, após as fases de treinamento da RNA.

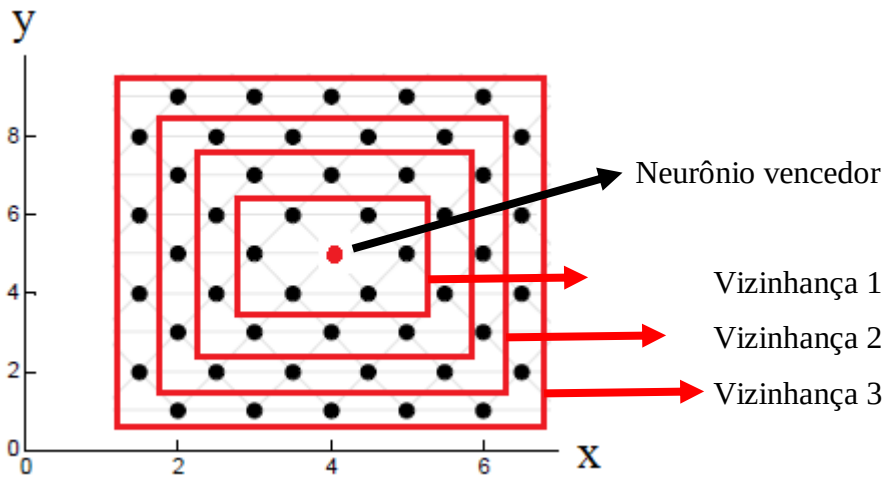


Figura 3.6. Disposição topológica de uma rede SOM por zona de vizinhança (Mousavi, 2005. Adaptado).

A arquitetura da rede SOM apresenta uma importante particularidade, pois cada neurônio da rede recebe todos os padrões de entradas apresentados a mesma, e resulta na saída de todos os neurônios. O estado de ativação de um neurônio é determinado pela distância entre o peso e o vetor de entrada, sendo a função de ativação da rede SOM determinada pela medida da distância Euclidiana mostrada na equação 3.2. Quanto mais semelhante à entrada for do vetor de pesos de um neurônio, maior o valor de sua entrada.

$$d_i(t) = \sum_{j=1}^{n} (x_j - w_{ij}(t))^2 \qquad (3.2)$$

Onde o vetor de pesos é representado por $w_{ij,}$ x_j é o valor de entrada na posição *j* e d_i o valor de saída resultante do cálculo da distância para o neurônio *i* no tempo *t*.

No item 3.4.1 será descrito os passos do algoritmo de treinamento que resulta no mapa auto-organizável, as duas fases de treinamento da rede e a técnica de rotulação dos neurônios que auxilia a classificação de padrões futuros apresentados a rede SOM.

3.4.1 Treinamento e Rotulação

O modelo de rede neural tipo SOM pertence à grade conhecida como mapas auto-organizáveis, onde o treinamento ocorre de modo competitivo (também chamado de *winner-takes-all*), o qual os neurônios de saída competem entre si para serem ativados.

A aprendizagem para esse tipo de rede ocorre de modo não supervisionado, porém não integralmente, visto que há necessidade de se rotular o neurônio de saída, entende-se por rótulo o nome do fato. O trabalho aqui proposto considera os pacotes de dados pertencentes à camada de aplicação do modelo ISO/OSI, e os rótulos dos neurônios conhecidos são os nomes de protocolos: HTTP, DNS e SSDP (*Simple Service Discovery Protocol*).

Como apresentado anteriormente, o algoritmo de treinamento busca organizar os neurônios em vizinhanças, onde padrões semelhantes pertencem à mesma região. Para que esse processo ocorra, cada vez que um novo padrão é apresentado à rede, os neurônios competem entre si para se tornar ativo, ou seja, para esse projeto é proposto escolher o neurônio que possui o menor valor de saída (menor distância Euclidiana). Escolhido o neurônio vencedor, dá inicio a atualização dos pesos dentro do raio de vizinhança do mesmo. Durante o processo de treinamento a taxa de aprendizagem e o raio de vizinhança são decrementados, e a rede cria regiões de padrões semelhantes segundo cada protocolo.

A seguir é apresentada a organização da proposta de uma rede tipo SOM sob a forma de um algoritmo computacional, em 9 passos.

1. Inicializar os pesos dos *n* neurônios de entrada com valores aleatórios entre 0 e 1.

2. Determinar o valor do raio de vizinhança *Vi* de cada neurônio *n* (inicialmente o valor de *Vi* é igual ao tamanho da rede).

3. Apresentar os dados brutos de cada pacote como entrada da rede.

4. Calcular a distância Euclidiana entre a entrada e os pesos para cada neurônio de saída, como demonstrada na equação 3.2

5. Selecionar o neurônio vencedor, ou seja, o neurônio que apresentar a menor distância $d_j(t)$.

6. Os pesos do neurônio selecionado são atualizados (mostrados na Equação 3.3) juntamente com todos os neurônios que estão dentro da vizinhança definida por *Vi(t)*.

$$w_{ij}(t+1) = w_{ij}(t) + TA(t) * (x_i(t) - w_{ij}(t)) \quad (3.3)$$

onde TA se refere a taxa de aprendizado, iniciada com valor 0,6 e multiplicado por 0,5 a cada iteração. Entende-se por iteração o final da atualização de cada valor dos neurônios vizinhos e não a cada entrada apresentada a rede.

7. Se necessário, modificar o raio de vizinhança de todos os neurônios. Essa modificação deverá implicar num decréscimo do raio de *Vi(t)*.

8. Voltar ao passo 3 caso exista algum padrão que faça parte do conjunto de treinamento e ainda não foi apresentado à rede.

9. O número de iterações também pode ser definido como limite de padrões apresentados à rede. Caso o valor da taxa não tenha alcançado o valor estipulado para término do laço de treinamento, deve ser realizada nova iteração do conjunto de treinamento retornando ao passo 3.

O treinamento da rede SOM ocorre em duas fases, sendo a fase de ordenação que busca agrupar os neurônios por meio dos padrões apresentados a entrada da rede, e a fase convergência que aprimora o agrupamento de modo que o raio de vizinhança seja reduzido.

Após o treinamento, a técnica de rotulação é aplicada aos neurônios ordenadamente agrupados para auxiliar a classificação de padrões desconhecidos, que futuramente podem ser apresentados a rede SOM.

Para finalizar os conceitos sobre redes neurais artificiais, no item 3.5 são apresentadas técnicas que utilizam redes neurais como solução de reconhecimento de padrão e semântica, aplicadas a eventos ocorridos no cotidiano baseado no comportamento humano.

3.5 Trabalhos Relacionados

"Uma das aplicações existentes em RNA é a Máquina de Escrever Fonética, que utiliza o modelo SOM (Self-Organizing Map), de Kohonen para aprender fonemas, que posteriormente serão transformados em palavras, através de regras gramaticais aprendidas automaticamente". (Kohonen, 1990, p. 1468).

Estudos nacionais voltados à área de Inteligência Artificial (IA) e RNA como apresentado no VIII ENIA (Encontro Nacional de Inteligência Artificial), ocorrido juntamente com o XXXI CSBC (Congresso da Sociedade Brasileira de Computação), entre os dias 19 a 22 de Julho de 2011, no Centro de Convenções da cidade de Natal-RN, apresenta a proposta intitulada "Aprendizado de Máquina Sem Fim e a Leitura da Web". A proposta de aprendizado, apresentada pelo Prof. Dr. Estevam Rafael Hruschka Júnior (pesquisador e professor adjunto da Universidade Federal de São Carlos - UFSCar), mostra que o ideal para se construir computadores inteligentes com capacidade de raciocinar, aprender e tomar decisões independentes, é algo que os pesquisadores da área de IA vêm desenvolvendo ao longo de anos. O projeto *ReadTheWeb* (leitura da web), desenvolvido em parceria por pesquisadores da Universidade Federal de São Carlos (UFSCar) e da Universidade Carnegie Mellon, dos Estados Unidos, busca desenvolver "um sistema computacional capaz de aprender de forma autônoma e de utilizar os conhecimentos já adquiridos para evoluir seu próprio aprendizado" (Hruschka, 2011). A proposta de um "aprendizado sem fim" é baseado no comportamento humano de aquisição de conhecimento gradativo ao longo do tempo, onde "somos cada vez mais capazes de refletir sobre o que sabemos e validar (supervisionar) o que aprendemos" (Hruschka, 2011), de forma continua e sem fim, auto-supervisionado e auto-reflexivo. O mais interessante é que o sistema busca a interação com seres humanos de forma a trocarem conhecimento.

3.6 Conclusão

Foram descritos neste capítulo os conceitos sobre redes neurais artificiais, característica sobre a arquitetura do modelo MCP. Quanto à rede neural tipo SOM, foi descrito: o tipo de aprendizado, o treinamento da rede e o conceito sobre rotulação de

neurônios. Por fim foram apresentados alguns exemplos de aplicações utilizando redes neurais.

Baseado no conhecimento, explicações e conceitos expostos até aqui, o capítulo 4 deste trabalho tem por objetivo demonstrar metodologia proposta para o desenvolvimento de uma RNA tipo SOM capaz de classificar pacotes de dados advindos da camada de aplicação baseado no conceito das sete camadas do modelo ISO.

Capítulo 4

Metodologia

A metodologia proposta tem como objetivo auxiliar administradores de rede na tarefa de coleta de amostra de tráfego em rede, forma de análise e disponibilização de dados da mesma. A disponibilização pode ser solicitada por pesquisadores, análise de auditoria interna à empresa ou até mesmo para fins de ordem judicial.

Uma amostra de dados coletada em canais de grande fluxo gera, conseqüentemente, grande quantidade de *log*s de rede. A análise dos dados gerados de forma manual para verificar, por exemplo, possíveis ataques, se tornam uma tarefa impossível. Desta forma, e após serem apresentadas as justificativas, o trabalho propõe o desenvolvimento de uma rede neural tipo SOM capaz de ordenar pacotes de dados em tipos comuns por protocolos (enfatizando a camada de aplicação do modelo ISO/OSI), mantendo a integridade dos dados coletados.

A seguir são discutidas as características da rede neural proposta.

4.1 Arquitetura

Com a necessidade de se detectar algum pacote de dados que não pertence de padrão de entrada da RNA, o trabalho propõe as seguintes fases para a metodologia:

- Captura de pacotes de dados, presente em amostra de tráfego de dados coletada por um *sniffer* de rede, aqui utilizado o *Wireshark.*
- Com auxílio da ferramenta *Wireshark,* identificar os pacotes, por descrição dos protocolos, que pertencem à camada de aplicação do modelo ISO/OSI.
- Salvar a amostra coleta com formato de extensão texto. Em linguagem JAVA ler o arquivo texto, converter os dados de entrada de hexadecimal para binário (pois a RNA manipula os dados dispostos desta forma).
- Após armazenar os dados convertidos, os pacotes são apresentados à entrada da RNA, dispostos em vetores (cada pacote, de forma integral, representa um vetor de binários).
- A fase de treinamento consiste em calcular a distância Euclidiana entre a entrada apresentada a rede e os pesos para cada neurônio de saída (como citado no item 3.4.1).
- Após o cálculo das distâncias para todas as entradas é escolhido o neurônio vencedor, ou seja, o neurônio que possui a menor distância Euclidiana.
- O neurônio vencedor é removido da disputa.
- A taxa de aprendizagem é recalculada e as demais escolhas pelo neurônio vencedor ocorrem de modo sucessivo, até que reste um único neurônio à disputa.
- No item 4.3 são apresentados os resultados obtidos referente à ordenação dos pacotes da amostra coletada.

4.2 Fases da Metodologia

Nos itens que se seguem são descritas, de forma detalhada, as fases da metodologia proposta para ordenação dos pacotes de dados por tipos comuns de protocolos e verificação de suposto ataque à rede de computadores. Tarefa que busca auxiliar administradores de rede de computadores e até mesmo disponibilizar os dados resultantes para pesquisas e melhorias com relação à infraestrutura da mesma.

4.2.1 Captura de pacotes

A captura dos dados que trafegam através de uma rede de computadores foi obtido pelo programa *Wireshark*, instalado diretamente na máquina hospedeira (servidor da rede). Ativando-se o software verificador de tráfego de rede, é descrito o método de coleta dos dados por meio dos passos indicados abaixo (citado no item 2.3.2.2):

- Selecionada a interface de rede a ser rastreada, neste caso o IP da máquina hospedeira, acessando o item no *menu* [Capture / Interfaces];
- Iniciado a captura dos pacotes de dados (clicar no botão [Start]), o rastreamento é apresentado através da janela principal do *Wireshark* (Figura 2.4);
- Interrompida a captura de pacotes (clicar no botão [Stop] da barra de ferramentas ou no item de *menu* [Capture / Stop]), foram obtidos 377 pacotes durante um total de 0.468 segundos de captura, sendo 35 do tipo HTTP, 50 do tipo DNS e 42 tipo do SSDP;
- Foi salva a cópia do resultado da captura de pacotes (clicar no *menu* [File / Salve as]) como 08-09-11 em formato texto.

Uma vez gerada e obtida uma cópia do tráfego em um canal, como demonstrado na Figura 4.1, o *Wireshark* trás informações como origem, destino, tipo de protocolo, tamanho, porta e entre outras informações contidas no pacote de dado.

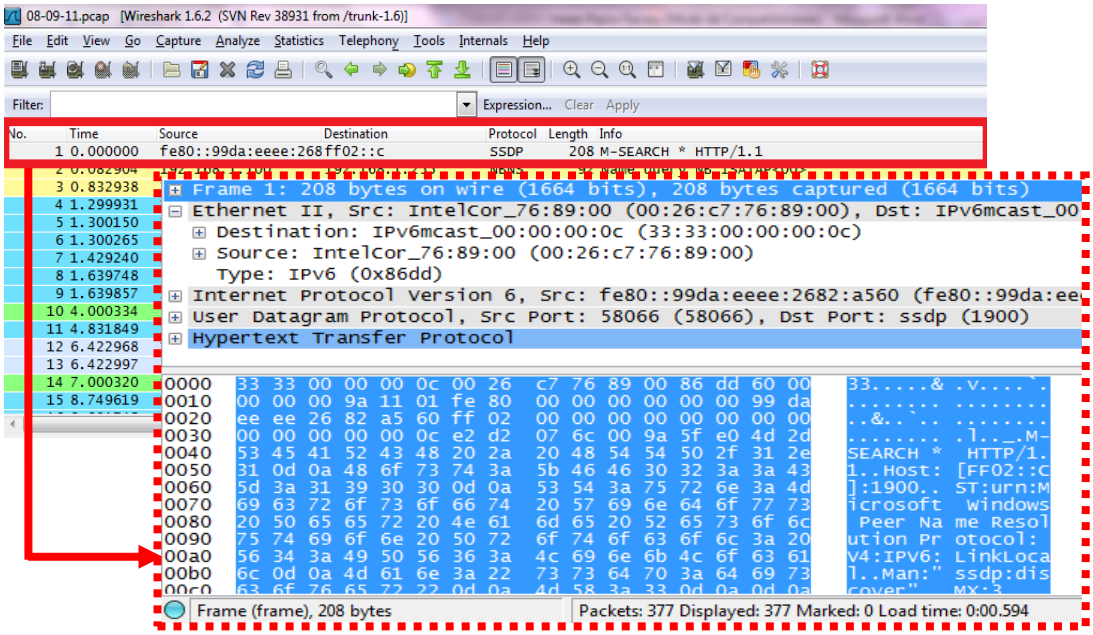


Figura 4.1. Filtragem e apresentação dos dados pelo *Wireshark*.

Com o *Wireshark* também é possível inspecionar cada bit em um pacote de dados, a aplicação realiza interpretação dos campos dos cabeçalhos, na maioria dos protocolos existentes, automaticamente.

Para filtrar os pacotes de dados por tipo, digite o nome no campo [Filter] ou acesse o *menu* [Analyze / Display Filters], após digitar ao menos uma letra, uma lista de nomes será exibida.

4.2.2 Identificação e Armazenamento dos Protocolos

O objetivo para obtenção sobre a identificação dos protocolos por camada, inicialmente, era desenvolver em linguagem JAVA um método no qual fosse lido o arquivo texto (gerado pelo *Wireshark*), extraído o pacote e assim identificar por meio de informações contidas no mesmo a qual camada este pertenceria. Todavia, a implementação deste método não foi realizada por não fazer parte do escopo para a proposta deste trabalho, portanto a classificação dos protocolos pertencentes à camada de Aplicação foi realizada de modo manual, baseando-se nas informações descritas pelo *sniffer Wireshark*.

Foram separados visualmente, 11 pacotes de dados coletados no dia 8 de Setembro de 2011, sendo do tipo HTTP, DNS e SSDP. Depois de selecionados, os pacotes de dados foram salvos como “package.txt” (arquivo de formato texto) dentro da pasta de projeto descrita como “ProjetoSOM”.

4.2.3 Leitura e Conversão dos Dados

Uma vez salvo o arquivo contendo os pacotes de dados com 11 protocolos selecionados, o *Wireshark* disponibiliza os pacotes de dados em hexadecimal e com separações por caracteres especiais (veja Figura 4.2). Antes que os dados fossem convertidos de hexadecimal para binário, faz-se necessário extrair as barras simples “|” e manter o delimitador de conteúdo do pacote de dados, representados pelos caracteres +---------+----------+----------+.

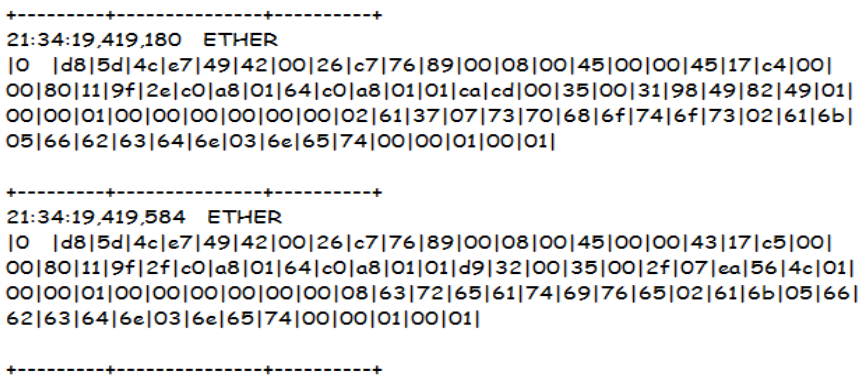


Figura 4.2. Amostra dos dados coletados gerado pelo *sniffer Wireshark*.

Para que seja efetuada a extração dos caracteres especiais, foi desenvolvido em linguagem JAVA um método que lê cada linha do arquivo texto e os retira do arquivo, obtendo deste modo somente os pacotes com os dados em hexadecimal. Após o método de extração, o arquivo texto é lido novamente de modo que cada bit é convertido de hexadecimal para binário, sendo 0 (zero) em hexadecimal representado como 0000 em binário , 1 (um) em hexadecimal como 0001 em binário, e assim consecutivamente até F em hexadecimal como 1111 em binário.

Após conversão dos dados, estes são armazenados em uma lista de *strings*, onde cada posição da lista é representada pelos dados brutos bit a bit do pacote de dados. Como dito anteriormente a rotulação foi realizada de forma manual, ou seja, o nome de cada protocolo ao qual o pacote é pertencente fica armazenado juntamente com os dados no campo denominado rótulo.

4.2.4 Apresentação dos Pacotes a entrada da RNA e Treinamento

Para que a RNA receba os dados e os organizem por semântica, a lista a qual os dados estão armazenados é apresentada a entrada da rede. A lista de pacotes não está ordenada por nomes, esta característica é essencial para que sejam verificados os resultados obtidos após o treinamento da rede. A apresentação ocorre de modo que cada neurônio de entrada da RNA lê cada posição da lista contendo os pacotes de dados, ou seja, o neurônio denominado 1 recebe os dados da posição 1 da lista, o neurônio 2 recebe os dados da posição 2 da lista e assim sucessivamente até o final da mesma.

O treinamento da rede neural proposta foi submetida há 9 passos da mesma forma como descrito no item 3.4.1 deste trabalho. Para melhor compreensão sobre a forma de apresentação dos dados à entrada de RNA, é apresentado um exemplo demonstrando-se uma parte do conteúdo do pacote de dado (após conversão e armazenado em um vetor de string). Sendo 1101100, uma parte da descrição do pacote de dados, para cada bit é associado um peso aleatório (variando-se de 0 a 1) e partindo do primeiro bit 1 (um) até o último, neste caso o 0, cada bit é submetido ao cálculo da equação 3.2 proposta como mostrado a seguir:

$d = (1 – 0.2)^2 + (1 – 0.3)^2 + (0 – 0.5)^2 + (1 – 0.8)^2 + (1 – 0.1)^2 + (0 – 0.3)^2 + (0 – 0.4)^2$.

O cálculo da distância é efetuado para todos os pacotes presentes no arquivo, neste caso para os 11 pacotes presentes na *string*. Após todos os cálculos é escolhido o neurônio no qual a entrada possui o menor valor da distância Euclidiana, e deste é determinado como o vencedor da disputa para ser ativo. O neurônio vencedor é retirado da disputa e um novo ciclo de cálculos é inicializado até que haja somente um único neurônio na disputa.

Após o treinamento foi obtido com sucesso uma lista ordenada dos pacotes de dados por tipos comuns de protocolos. Os resultados serão detalhados no item 4.3 que se segue.

4.3 Análise de Resultados

Após a captura dos pacotes de dados por meio do software verificador de tráfego de rede, gerado o arquivo de leitura em formato texto e conversão dos dados para binário. A rede proposta foi submetida ao treinamento proposto para uma RNA tipo SOM e obtido ordenação satisfatória quando a entrada apresentada a rede. A disposição dos dados de entrada e a ordenação obtida são apresentadas na tabela 2 a seguir.

Tabela 2. Demonstrativo dos resultados de treinamento.

Distribuição inicial dos dados por protocolo	Classificação final obtida
HTTP	SSDP
DNS	SSDP
DNS	DNS
HTTP	DNS
HTTP	DNS
DNS	DNS
DNS	DNS
DNS	DNS
DNS	HTTP
SSDP	HTTP
SSDP	HTTP

Como se pode observar, a tabela demonstra que a rede proposta alcançou o principal objetivo de classificação dos pacotes de dados em tipos comuns de protocolos.

Em uma segunda avaliação, foi alterado o cabeçalho de um dos pacotes que pertencia à identificação como rótulo DNS e novamente as entradas foram apresentadas à rede. Para identificar o pacote alterado foi aplicada a nomenclatura de rótulo ALTERADO. Após o treinamento foram obtidos os seguintes resultados:

package (1) = SSDP, package (2) = SSDP, package (3) = DNS, package (4) = DNS, package (5) = DNS, package (6) = DNS, package (7) = DNS, package (8) = HTTP, package (9) = HTTP, package (10) = ALTERADO e package (11) = HTTP.

Observou-se então que o pacote que anteriormente possuía características de um pacote DNS, agora se confunde entre os padrões de um pacote de dados HTTP. Portanto, um pacote que sofre alterações em seu cabeçalho, pode ser visualizado como anormalidade de padrão em uma rede neural tipo SOM.

4.4 Conclusão

Nesse capítulo foi proposta uma metodologia que inclui a implementação de uma rede neural capaz de classificar pacotes de dados em tipos comuns de protocolos. Os testes foram efetuados em aproximadamente cinco tipos distintos de coletas (quanto a horários) e em todas elas a obtenção de resultados satisfatórios.

Com base nas análises de resultados observadas no item 4.3, foi possível atingir a contribuição que busca auxiliar administradores de redes em detectar possíveis ataques a uma rede de computadores.

Capitulo 5

Considerações Finais e Trabalhos futuros

É crescente o número de usuários que acessam a Internet diariamente e juntamente com esse crescimento o número de tráfego de informações. Nasce aqui a necessidade por parte de administradores e pesquisadores o uso de relatórios referentes ao tráfego de rede para propor novas soluções ou análise de situações que podem colocar em risco a rede de empresas, universidades ou mesmo um bom funcionamento da Internet.

Diante desta situação, o trabalho apresentou o estudo de conceitos referentes a redes de computadores e redes neurais artificiais, propondo a metodologia que visa ordenar pacotes de dados por tipos comuns de protocolos e verificar anormalidades que podem estar presentes em uma amostra de tráfego de rede.

Com base em testes realizados e aplicação da metodologia proposta, conclui-se que pacotes não ordenados de forma padrão aos demais, podem ser considerados como ataques e necessitam de uma melhor análise pelo administrador de redes. Havendo o administrador verificado, por exemplo, a origem/destino do pacote que apresentou anormalidade de comportamento e detectado que realmente o pacote trás informações maliciosas, este pode disponibilizar a pesquisadores da área as conclusões das análises para que sejam desenvolvidas novas alternativas de segurança quanto à infraestrutura de uma rede.

Como trabalho futuro, uma possibilidade é o aprofundamento de estudos voltados à área de redes de computadores, especialmente quanto às características contidas em pacotes de dados como identificação da camada as quais pertencem. Obtendo os conhecimentos citados anteriormente, a classificação dos pacotes de dados seria de modo automático ao executar o código em JAVA proposto.

A metodologia prova que é possível classificar os pacotes de dados quanto aos padrões apresentados à rede neural. Em particular, o desenvolvimento de uma ferramenta completa que busca analisar tráfego de rede em tempo real é uma melhoria a proposta apresentada.

Referências

Arthur, D. & Panigrahy, R. (2006). Analyzing bittorrent and related peer-to-peer networks. In SODA '06: Proceedings of the seventeenth annual ACM-SIAM symposium on Discrete algorithm, pp. 961--969, New York, NY, USA. ACM Press.

Azevedo, Fernando Mendes de; Brasil, Lourdes Mattos; Oliveira, Roberto Célio L. de (2000). REDES NEURAIS com aplicações em Controles e em Sistemas Especialistas.

Barreto, Prof. Dr. Guilherme de Alencar (2007). Pertencente ao Departamento de Engenharia de Teleinformática Programa de Pós-Graduação em Engenharia Elétrica (PPGEE) da Universidade Federal do Ceará (UFC). Notas intituladas: Resumo dos Algoritmos WTA e SOM. Disponível em <http://www.deti.ufc.br/~guilherme/TIP705/wta_som.pdf> acessado em 15/08/2011.

Bishop, M.; Crawford, R.; Bhumiratana, B.; Clark, L.; Levitt, K. (2006). Some problems in sanitizing network data. In Society, I. C., editor, Proceedings of the 15th IEEE International Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises, pp. 307_312.

Braga, Antônio de P.; Ludermir, Teresa B.; Carvalho, André Carlos P. L. F. (2007). REDES NEURAIS ARTIFICIAIS, Teoria e aplicações.

Cardoso, Silvia Helena (2000). Artigo intitulado: Comunicação entre as células nervosas. Disponível em <http://www.cerebromente.org.br/n12/fundamentos/ neurotransmissores/neurotransmitters2_p.html> acessado em 30/05/2011.

Chappell, Laura (2011). Wireshark Network Analysis: The Official Wireshark Certified Network Analyst Study Guide. Disponível em <http://www.wiresharkbook.com/downloads.html>. Acessado em 30/08/2011.

Cantu, Evandro (2003). Apostila REDES DE COMPUTADORES E INTERNET apresentada como material de apoio ao curso de Telecomunicações na CEFET/SC – Unidade de São José.

Comer, Douglas E (2007). REDES DE COMPUTADORES E INTERNET. 4ª edição (ed. Bookman). ISBN 9788560031368.

Hruschka, Prof. Dr. Estevam Rafael Júnior (2011). Aprendizado de Máquina Sem Fim e a Leitura da Web, proposta apresentada no VIII ENIA (Encontro Nacional de Inteligência Artificial), ocorrido juntamente com o XXXI CSBC (Congresso da Sociedade Brasileira de Computação), entre os dias19 a 22 de Julho de 2011, no Centro de Convenções da cidade de Natal-RN.

Ibope Nielsen (2011). 73,9 milhões de pessoas têm acesso à internet no Brasil. Disponível em <http://www.ibope.com.br/calandraWeb/servlet/CalandraRedirect?temp=6&proj=PortalIBOPE&pub=T&nome=home_materia&db=caldb&docid=EA0526673CE1740D832578570054B23B> acessado em 10/04/11.

Ivcon (2011). Disponível em: <http://www.natcomp.com.br/lvcon/tema?tema=2> acessado em 20/10/11

Kohonen, Teuvo (1990). The Self-Organizing Map. Proceedings of the IEEE, v. 78, n.9, p. 1464-1479.

Kohonen, Teuvo (1997). Self-Organizing Maps, Springer Verlag (Berlim), 2ª edição.

Louredo, Paula (2011). Graduada em Biologia. Disponível em: <http://escolakids.uol.com.br/sistema-nervoso.htm > acessado em 20/10/11.

Melo, Marco Aurélio Vilaça de (2009). ASPECTOS TÉCNICOS E LEGAIS DA COLETA ANONIMIZAÇÃO DE TRÁFEGO DE REDES IP. Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais.

Mousavi, Mirrasoul J.; Butler-Purry, Karen L. (2005). Respective student member and senior member, from IEEE. Artigo intitulado: A Characterization Methodology for Distribution System Abnormalities Using Wavelet Packets and Self-Organizing Map Neural Networks.

Portal São Francisco (2011). Disponível em: <http://www.portalsaofrancisco.com.br/alfa/corpo-humano-sistema-nervoso/cerebro.php> acessado em 20/10/11.

Roque, Antonio Carlos (2008). Disponível em <http://acroque.blogspot.com/2008/10/o-modelo-de-mcculloch-e-pitts-parte-1.html> acessado em 20/10/11.

Russell, Stuart; Norvig, Peter (2003). Inteligência Artificial, tradução da 2ª edição.

Schubert, Vagner (2010). Instrutor de Informática e Técnico de Informática da Prefeitura Municipal de Rorainópolis. Disponível em: <http://schubertweb.blogspot.com/2010/11/o-modelo-de-referencia-osi_12.html> acessado em 20/10/11.

Searle, J. R. Minds (1991). Artigo intitulado: Brains and Science.

Segatto, Ênio Carlos; Coury, Denis Vinicius (2011). Artigo: Redes neurais aplicadas a relés diferenciais para transformadores de potência. Estudantes do Depto de Engenharia Elétrica – EESC - Escola de Engenharia de São Carlos – USP. Disponível em: <http://www.scielo.br/scielo.php?id=S0103-17592008000100009&script=sci_arttext> acessado em 20/10/11.

Senado Federal, B. (2008). Projeto de lei número 494. Disciplina a forma, os prazos e os meios de preservação e transferência de dados informáticos mantidos por fornecedores de serviço a autoridades públicas, para fins de investigação de crimes praticados contra criança e adolescentes, e dá outras providências.

Souza, Tyago (2008). ARTIGO PUBLICADO POR TYAGHO SOUZA EM 28/04/2008 | CATEGORIAS: DICAS, MUNDO, TECNOLOGIA. Disponível em: <http://www.noticiaki.com/rede-de-computadores-anuncia-possvel.html> acessado em 20/10/11.

Steding-Jessen, K.; Vijaykumar, N. L. & Montes, A. (2008). Uso de Honeypots de baixa interatividade para o estudo do abuso de Proxies abertos para o envio de Spam. INFOCOMP Journal of Computer Science.

Tanenbaum, Andrew S. (2003). Redes de Computadores. Tradução da 4ª Edição Americana.

Tafner, Malcon Anderson (1996). RECONHECIMENTO DE PALAVRAS FALADAS ISOLADAS USANDO REDES NEURAIS ARTIFICIAIS. Dissertação submetida à Universidade Federal de Santa Catarina para obtenção do Grau de Mestre em Engenharia.