20 e 21 de outubro
Instituto Nacional de Pesquisas Espaciais - INPE
São José dos Campos - SP

Análise de Séries Temporais para Caracterizar o Tráfego de Rede utilizando Mineração de Dados por Clusterização

Adriana C. Ferrari dos Santos [1], Jose Demísio Simões da Silva [2], Lília de Sá Silva [3], Milena P. da Costa Sene [3]

[1]Programa de Doutorado em Computação Aplicada – CAP, INPE – SJCampos-SP

[2]Laboratório de Computação Aplicada – INPE – SJCampos-SP

[3]Divisão de Desenvolvimento de Sistemas Solo – INPE – SJCampos-SP -

aferrarisantos@gmail.com, demisio@lac.inpe.br, lilia@dss.inpe.br, milena@dss.inpe.br

Abstract. *This paper describes the partial results obtained using data mining through the Adaptive Kohonen Map tool to characterize the behavior pattern of network traffic over time. Different measures of similarity and variance of values per attribute between the classes are used in cluster analysis of the data. Besides the above mentioned technique, the foundation also will be discussed relevant to the work, including the clustering technique, modeling of network traffic, the concept of network session, description of the attributes analyzed, as well as the anomalies of network traffic.*

Resumo. E*ste artigo descreve os resultados parciais obtidos a partir de mineração de dados por clusterização para caracterização do comportamento padrão do tráfego de rede ao longo do tempo. Para este propósito, utilizou-se a abordagem "Mapa de Kohonen Adaptável", considerando diferentes parâmetros de similaridade e taxa de desvio para clusterização das sessões do tráfego. Conceitos, tais como, técnica de clusterização, modelagem do tráfego, sessões de rede e descrição dos atributos estudados compreendem o embasamento teórico deste trabalho.*

Palavras-chave: *Análise de séries temporais, mineração de dados, detecção de anomalias em redes, aplicação de inteligência computacional, segurança de redes.*

1. Introdução

O crescimento do uso de redes de computadores conectadas a Internet, possibilitou às empresas agregarem mais valor aos seus produtos e serviços a partir da interação on-line com seus clientes e fornecedores. Por outro lado, as organizações têm se tornado muito dependente das tecnologias de rede, sentindo, imediatamente, o impacto quando seus

recursos não estão disponíveis. Este fato tem gerado preocupações para os responsáveis pela segurança, disponibilidade e integridade dos dados, tornando-se evidente a necessidade de estabelecer monitoramento e controle sobre o comportamento do tráfego de rede, como forma de garantir a identificação de problemas em tempo hábil, evitando prejuízos.

Com o intuito de manter o funcionamento correto e confiável de redes de computadores, diversas medidas preventivas, tais como implantação de sistemas antivírus, anti-spyware, firewalls e sistemas de detecção de intrusos (SDIs) têm sido propostas ao longo dos últimos anos. Entretanto, nos SDIs, observa-se a necessidade do uso de técnicas eficientes que proporcionem análise de grandes volumes de dados de rede em intervalos de tempo regulares, a fim de mapear o comportamento normal e anômalo do tráfego das redes monitoradas de modo preciso e em tempo satisfatório.

O objetivo deste trabalho é apresentar os resultados parciais da pesquisa em desenvolvimento que aborda a caracterização do comportamento padrão do tráfego HTTP de redes de computadores através das técnicas de Análise de Séries Temporais e Inteligência Computacional. As séries temporais estudadas correspondem à grandes volumes de dados de sessões do tráfego Web, onde cada registro de sessão é descrito por nove atributos. Diferentes parâmetros de similaridade e taxa de desvio foram utilizados para clusterização das sessões do tráfego, no intuito de refinar o número de atributos empregados na análise.

Este estudo é uma continuação das pesquisas de aplicação de técnicas de redes neurais para detecção de anomalias no tráfego de redes de computadores, iniciada em 2004 no Instituto Nacional de Pesquisas Espaciais (INPE), cujos trabalhos encontram-se descritos em [6],[7],[9],[10],[11],[12].

O referencial teórico deste trabalho, incluindo o conceito de tráfego de redes de computadores, sessões de rede, bem como a definição de comportamento padrão e anômalo da rede, é apresentado na próxima seção. A rede neural Mapa de Kohonen Adaptável utilizada para a mineração de dados por clusterização é abordada na seção 3. Na seção 4, descreve-se a modelagem do tráfego de rede, os atributos de rede analisados e os parâmetros escolhidos para a clusterização. Os resultados parciais da caracterização do comportamento padrão do tráfego de rede são discutidos na seção 5. Finalmente, na seção 6, são apresentados a conclusão deste trabalho e os próximos desafios.

2. Caracterização do Tráfego de Rede

A caracterização do comportamento do tráfego de rede envolve a análise das sessões do tráfego por meio de seus atributos e visa encontrar padrões nos conjuntos de dados analisados, mapeá-los e, através disso, desenvolver uma metodologia para busca de anomalias no comportamento do tráfego de rede, de acordo com o período do dia. O esforço significativo da tarefa de mineração de dados deve-se ao fato de se trabalhar com um volume gigantesco de dados e à diversidade de valores dos atributos das sessões gerados durante as comunicações do serviço Web.

Em geral, durante o processo de comunicação entre os hosts na rede são gerados pacotes de dados que constituem o tráfego daquela rede de computadores. Neste trabalho, considera-se como sendo uma sessão de tráfego HTTP qualquer seqüência de pacotes que caracterize a troca de informações entre pares de endereços IP (do cliente e

do servidor) durante uma comunicação do cliente com o serviço Web, em um determinado intervalo de tempo.

Cada sessão do tráfego de uma rede pode ser unicamente modelada através de atributos contidos no cabeçalho dos pacotes. Estes atributos podem ser primitivos, tais como: IP de origem, IP de destino e protocolo de aplicação, ou derivados, por exemplo: quantidade de pacotes recebidos pela estação servidora em determinado intervalo de tempo ou quantidade de bytes recebidos pela estação cliente naquela sessão.

O tráfego de uma rede se comporta de modo dinâmico, dependendo de fatores tais como tipos de serviços fornecidos, quantidade de usuários e hosts e os períodos do dia em que os serviços são acessados. Portanto, modelar o comportamento do tráfego de rede requer a investigação de valores de diferentes atributos que compõem os conjuntos de dados do tráfego, os quais se alteram dependentemente dos fatores mencionados.

Caracterizar o tráfego padrão de uma rede é definir o perfil de comportamento normal do trafego esperado em cada período do dia, a partir de análise dos dados atuais do tráfego. Com base neste perfil modelado, ou seja, a partir do conjunto de dados histórico do tráfego, torna-se possível classificar o tráfego corrente como padrão ou anômalo.

Possíveis desvios no perfil de comportamento padrão do tráfego mapeado são denominados anomalias. Exemplos de anomalias no tráfego incluem: traços de ataques, abusos na rede, eventos de falha em sistemas e serviços, problemas de infra-estrutura de rede, expansão de usuários ou adição de serviços na rede, entre outros processos. Resumindo, nem toda anomalia na rede é um ataque ou ameaça, mas constitui uma informação suspeita a ser analisada pelo administrador de rede, logo que identificada.

3. Técnica de Clusterização Adotada

A técnica de mineração de dados por clusterização vem sendo aplicada na análise de conjuntos de dados de grande volume, de modo a encontrar as melhores transformações a serem aplicadas aos dados originais, para extrair relações existentes entre os dados que não são detectáveis diretamente pelo observador humano, e que, no entanto, são potencialmente úteis para o objetivo de caracterizar o comportamento padrão do conjunto, diferenciando as classes de dados.

O processo de clusterização consiste em, dada uma base de dados, agrupar (clusterizar) os elementos desta base de modo que os dados mais similares, ou seja, elementos que possuem características em comum, sejam alocados no mesmo cluster e dados menos similares sejam alocados em clusters distintos.

Neste trabalho, é utilizada rede neural “Mapa de Kohonen Adaptável” para clusterização das sessões do tráfego de rede. Trata-se de um clusterizador que simula a rede neural SOM (Self Organizing Map), cuja arquitetura é ilustrada na figura 1.

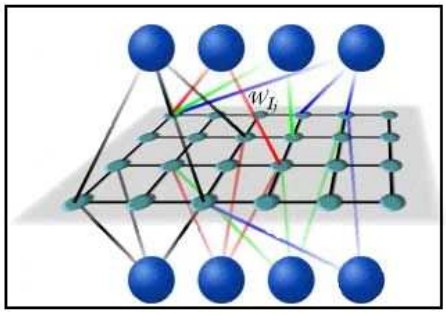


Figura 1. Arquitetura da rede neural SOM

A rede SOM é um classificador não-supervisionado, inspirada nos mapas corticais, onde os neurônios competem entre si para responder a um estímulo apresentado. Durante o aprendizado, formam-se agrupamentos de neurônios topologicamente organizados, sendo que cada grupo responde a uma classe de estímulos. A característica de auto-organização da rede SOM é proveniente do uso de regras de aprendizado não supervisionado, tal como regras de aprendizado competitivo. Neste tipo de aprendizado, a única informação apresentada à rede são os padrões de entrada. Sendo assim, as ligações sinápticas são definidas de forma a recompensarem o neurônio vencedor, sem requerer comparações com padrões desejados.

Enquanto a rede SOM busca o centro como representante de cada cluster, o "Mapa de Kohonen Adaptável" busca o representante de um novo cluster e agrupa os dados similares, alocando-os neste cluster representado. A principal diferença entre classificação e clusterização de dados é que na classificação os dados devem ser atribuídos a clusters previamente conhecidos, enquanto na clusterização deve-se, primeiramente, "descobrir" quem são os clusters.

O algoritmo "Mapa de Kohonen Adaptável" adotado neste trabalho analisa os dados processados a partir dos vetores de entrada (vetores de sessão do tráfego da rede contendo 9 valores de atributos) e os agrupa, testando a sua similaridade em relação aos clusters existentes. Durante o processo de clusterização, a matriz de pesos (w) da rede é construída da seguinte forma: as linhas desta matriz são atualizadas com os valores dos vetores de entrada que definem a existência de um novo cluster. Ao vetor que identifica a formação de um novo cluster dá-se o nome de "vetor representante do cluster". A matriz de pesos resultante, formada pelos vetores representantes de clusters, caracterizam o comportamento padrão do tráfego da rede num determinado período de tempo.

Neste trabalho, através do algoritmo "Mapa de Kohonen Adaptável", foram configurados dois parâmetros para a clusterização dos dados, os quais influenciam na qualidade dos resultados, sendo estes identificados como taxa de desvio e taxa de clusterização. A taxa de desvio refere-se à função de ativação de cada neurônio de entrada (atributo das sessões), sendo utilizada na seleção dos atributos da seção corrente a ser classificada e, neste trabalho, foram adotadas as taxas de desvio de 10% e 15%. A taxa de clusterização das sessões permite alocar a sessão ao cluster a que mais se assemelha. As taxas 70% e 100% foram consideradas para clusterização das sessões nos testes realizados.

Cada vetor de sessão de entrada X é alocado ao cluster mais semelhante, após os valores dos atributos serem comparados com os valores dos vetores contidos na matriz

de pesos W construída pelo algoritmo. Caso o vetor de entrada X não apresente similaridade com os vetores representantes de cluster existentes na matriz W, ele é incluído como novo membro da matriz W e passa a ser um vetor representante de novo cluster.

4. Descrição do Tráfego

Para os experimentos realizados neste trabalho foram coletados, entre os meses de março/2010 e maio/2010, dados provenientes de uma rede local de computadores do Instituto Nacional de Pesquisas Espaciais (INPE) em São José dos Campos, extraídos de pacotes de rede em quatro períodos do dia e em diferentes dias da semana, incluindo finais de semana. Para garantir a ausência de sessões maliciosas nas séries utilizadas para caracterizar o tráfego de comportamento normal, os conjuntos de dados analisados foram previamente sanitizados através da ferramenta Snort [1].

A tabela 1 apresenta uma amostra escolhida aleatoriamente dos dados analisados. P1 representa o período da madrugada e corresponde aos horários das 00h às 06h59min, P2 é o período da manhã no horário das 07h às12h59min, P3 corresponde ao período da tarde, compreendendo os horários das 13h às 18h59min e P4 é o período da noite no horário das 19h às 23h59min.

Tabela 1. Amostra das Séries Temporais Analisadas

Série Temporal	**Data/Dia da Semana**	**P1** (madrugada)	**P2** (manhã)	**P3** (tarde)	**P4** (noite)
		Total de Sessões	**Total de Sessões**	**Total de Sessões**	**Total de Sessões**
S09032010	09/03/2010 - Ter	4477	90813	85204	18435
S15032010	15/03/2010 - Seg	4033	96721	177265	5249
S28032010	28/03/2010 - Dom	20043	18852	17682	12328
S01042010	01/04/2010 - Qui	5541	89820	119535	2979
S14042010	14/04/2010 - Qua	19130	76061	103961	7591
S16042010	16/04/2010 - Sex	21420	68565	98741	18993
S17042010	17/04/2010 - Sab	20085	86932	76890	4237

Cada sessão analisada é reconstruída por um conjunto de 9 (nove) atributos incluindo: psizeCL (tamanho médio dos pacotes recebidos pelo cliente), psizeSV (tamanho médio dos pacotes recebidos pelo servidor), pnumCL (total de pacotes recebidos pelo cliente) e pnumSV (total de pacotes recebidos pelo servidor), smallpkt (porcentagem de pacotes pequenos – menores que 130 bytes), dataDIR (direção do trafego – cliente recebe 1 pacote: incrementa 1; servidor recebe 1 pacote: decrementa 1), brecvCL (total de bytes recebidos pelo cliente – número), brecvSV (total de bytes recebidos pelo servidor – número), duration (duração da sessão – timestamp do primeiro pacote menos o timestamp do último pacote da sessão em segundos).

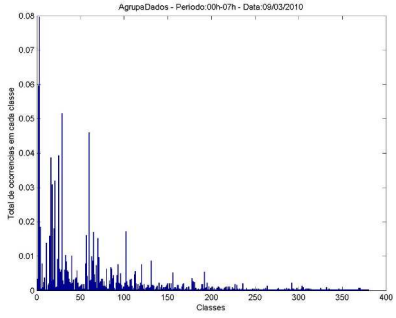
Os dados de pacotes da rede foram capturados através do sniffer de rede TCPdump e copiados da máquina de captura para HD externo por meio de scripts. As sessões foram reconstruídas através do sistema RECON – Sistema de Reconstrução de Sessões TCP/IP [2], e gravadas em base de dados MySQL, onde cada tabela da base contém os dados capturados em cada dia. Em seguida, foi executado um script para converter cada tabela da base em quatro arquivos de texto, um arquivo gerado para cada período de tempo acima descrito (P1 a P4), sendo cada um dos nove atributos de cada

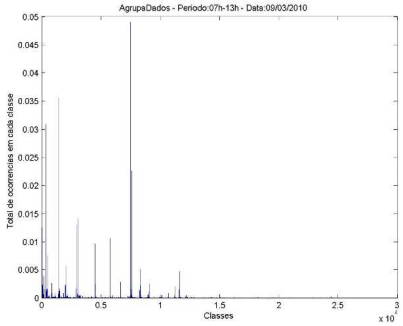
sessão separados por vírgula para uso posterior no Matlab, linguagem utilizada para o desenvolvimento da técnica de cluster.

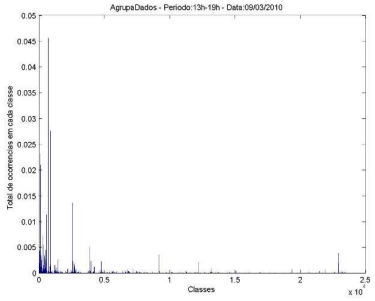
5. Análise dos Resultados Parciais

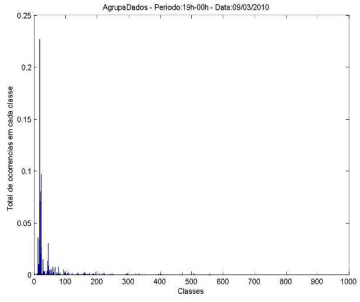
No contexto deste trabalho, o processo de análise do tráfego de rede consiste em investigar o valor dos atributos das sessões do tráfego referente à troca de pacotes de dados durante a comunicação do serviço Web entre os hosts na rede, em busca de informações de interesse, no caso, reconhecimento de padrões que caracterizam o perfil do comportamento "padrão" do tráfego.

Dos testes realizados, escolheu-se aleatoriamente a amostra do dia 09 de março, nos quatro períodos de tempo para apresentar os resultados da clusterização de modo a observar a quantidade de sessões agrupadas por cluster, com os dois parâmetros: desvio (ds) de 10% e similaridade de 100% (ver figura 2) e desvio (ds) de 15% e similaridade de 70% conforme figuras 2 e 3.

<table>
<tr><td colspan="2">Parâmetros: (ds)=desvio de 10% - (sum(ds)>=size(w,2))=elementos 100% similares
Dia 09/03/2010 (terça-feira) nos 4 períodos</td></tr>
<tr><td> 

(P1) – 380 clusters
Cluster 3 = 357 sessões = 7,97%</td><td> 

(P2)- 25245 clusters
Cluster 7498 = 4446 sessões= 4,9%</td></tr>
<tr><td> 

(P3) – 23867 clusters
Cluster 710 = 3878 sessões = 4,55%</td><td> 

(P4) (P4)– 967 clusters
Cluster 18 = 4180 sessões = 22,67%</td></tr>
</table>

Figure 2. Sessões do maior cluster – desvio=10% e similaridade=100%

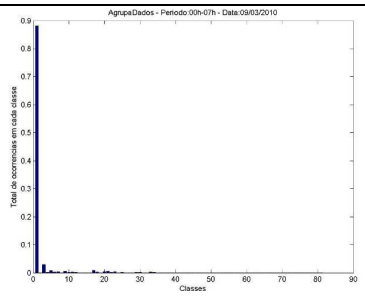
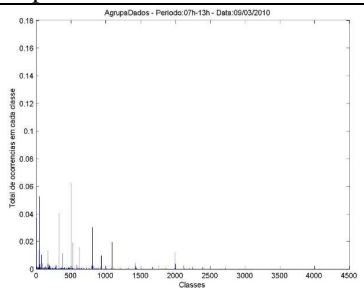
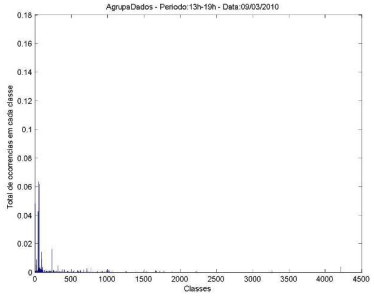
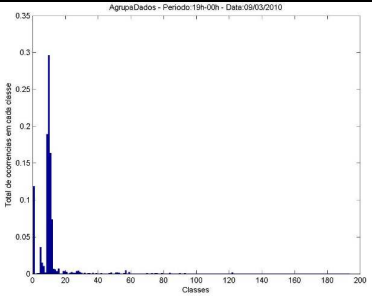
<table>
<tr><td colspan="2">Parâmetros: (ds)=desvio de 15% - (sum(ds)>=0.7*size(w,2))=elementos 70% similares
Dia 09/03/2010 (terça-feira) nos 4 períodos</td></tr>
<tr><td> 
(P1) – 81 clusters
Cluster 1 = 3949 sessões = 88,21%</td><td> 
(P2)- 4381 clusters
Cluster 1 = 14846 sessões= 16,35%</td></tr>
<tr><td> 
(P3) – 4337 clusters
Cluster 1 = 14348 sessões = 16,84%</td><td> 
(P4) – 193 clusters
Cluster 10 = 5458 sessões = 29,61%</td></tr>
</table>

Figure 3. Sessões do maior cluster – desvio=15% e similaridade=70%

A clusterização dos dados com parâmetro de desvio=15% e similaridade=70%, classifica maior quantidade de sessões nos primeiros clusters gerados na matriz de pesos W, sendo a maioria dos vetores de entrada agrupados nos clusters de posição 1. Este resultado permitiu observar que os registros (vetores de entrada de sessões com 9 atributos) contidos no início das séries temporais possuem características mais semelhantes, independente do período de tempo.

Diferente análise foi realizada a fim de extrair dados que caracterizassem o comportamento padrão da rede. Neste caso, considerou-se o total de clusters gerado para cada série. Os resultados obtidos a partir de uma amostra de séries temporais de diferentes dias da semana, nos quatro períodos do dia são apresentados nas tabelas 2, 3, 4 e 5. Os parâmetros selecionados para os testes foram respectivamente: ds=10% e similaridade=100%, ds=10% e similaridade=70%, ds=15% e similaridade=100% e ds=15% e similaridade=70%.

Tabela 2. Clusters gerados - Parâmetros ds= 10% e 100% de similaridade

Parâmetros: (ds)=desvio de 10% e (sum(ds)>=size(w,2))=elementos 100% similares					
Série Temporal	**Dia da Semana**	**Total de Clusters Gerados**			
		P1	**P2**	**P3**	**P4**
S09032010	Terça-feira	380	25245	23867	967
S15032010	Segunda-feira	260	27042	26789	777
S28032010	Domingo	295	292	334	276
S01042010	Quinta-feira	307	17830	16020	242
S14042010	Quarta-feira	451	18684	22858	2130
S16042010	Sexta-feira	414	20420	20546	861
S17042010	Sábado	395	495	399	378

Tabela 3. Clusters gerados - Parâmetros ds= 10% e 70% de similaridade

Parâmetros: (ds)=desvio de 10% e (sum(ds)>=0.7*size(w,2))=elementos 70% similares					
Série Temporal	**Dia da Semana**	**Total de Clusters Gerados**			
		P1	**P2**	**P3**	**P4**
S09032010	Terça-feira	98	8471	8262	254
S15032010	Segunda-feira	46	8710	8775	301
S28032010	Domingo	62	50	55	49
S01042010	Quinta-feira	46	6192	5708	24
S14042010	Quarta-feira	79	6620	7758	1020
S16042010	Sexta-feira	68	7317	7178	313
S17042010	Sábado	65	124	54	97

Tabela 4. Clusters gerados - Parâmetros ds= 15% e 100% de similaridade

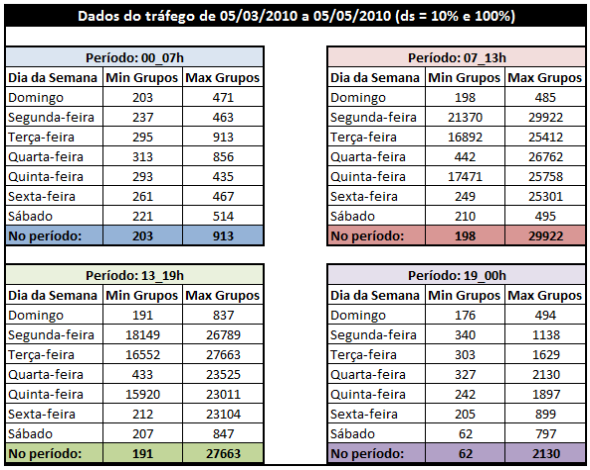
Parâmetros: (ds)=desvio de 15% e (sum(ds)>=size(w,2))=elementos 100% similares					
Série Temporal	**Dia da Semana**	**Total de Clusters Gerados**			
		P1	**P2**	**P3**	**P4**
S09032010	Terça-feira	266	20351	19477	697
S15032010	Segunda-feira	182	21393	21267	612
S28032010	Domingo	206	210	228	205
S01042010	Quinta-feira	205	14225	12785	161
S14042010	Quarta-feira	298	15121	18307	1812
S16042010	Sexta-feira	277	16417	16493	679
S17042010	Sábado	253	341	263	274

Tabela 5. Clusters gerados - Parâmetros ds= 15% e 70% de similaridade

Parâmetros: (ds)=desvio de 15% e (sum(ds)>=0.7*size(w,2))=elementos 70% similares					
Série Temporal	**Dia da Semana**	**Total de Clusters Gerados**			
		P1	**P2**	**P3**	**P4**
S09032010	Terça-feira	81	4381	4337	193
S15032010	Segunda-feira	41	4536	4432	228
S28032010	Domingo	59	43	52	45
S01042010	Quinta-feira	44	3275	3084	21
S14042010	Quarta-feira	73	3521	4144	700
S16042010	Sexta-feira	59	3835	3729	253
S17042010	Sábado	59	106	52	82

Como mostrado nas tabelas 3 e 5, a aplicação dos parâmetros de desvio 10% e 15% com similaridade de 70%, o total de clusters formado para cada série é consideravelmente menor, quando comparado com os resultados obtidos a partir dos mesmos parâmetros de desvio com similaridade de 100%, conforme ilustrado nas tabelas 2 e 4. Contudo, os parâmetros ds=15% e similaridade=70%, apresentaram resultado mais satisfatório para todas as séries temporais analisadas, independente do dia da semana e período de tempo.

Após o processo de clusterização das séries temporais coletadas, foram investigados os clusters mínimos e máximos gerados para todos os dias da semana (figuras 3, 4, 5 e 6), nos quatro períodos do dia e utilizando os mesmos pares de parâmetros (ds, similaridade) previamente escolhidos para agrupamento das sessões.



Dados do tráfego de 05/03/2010 a 05/05/2010 (ds = 10% e 100%)

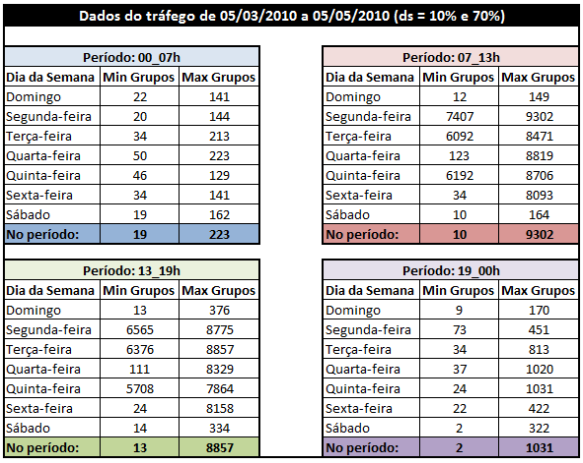
Período: 00_07h		
Dia da Semana	Min Grupos	Max Grupos
Domingo	203	471
Segunda-feira	237	463
Terça-feira	295	913
Quarta-feira	313	856
Quinta-feira	293	435
Sexta-feira	261	467
Sábado	221	514
No período:	203	913

Período: 07_13h		
Dia da Semana	Min Grupos	Max Grupos
Domingo	198	485
Segunda-feira	21370	29922
Terça-feira	16892	25412
Quarta-feira	442	26762
Quinta-feira	17471	25758
Sexta-feira	249	25301
Sábado	210	495
No período:	198	29922

Período: 13_19h		
Dia da Semana	Min Grupos	Max Grupos
Domingo	191	837
Segunda-feira	18149	26789
Terça-feira	16552	27663
Quarta-feira	433	23525
Quinta-feira	15920	23011
Sexta-feira	212	23104
Sábado	207	847
No período:	191	27663

Período: 19_00h		
Dia da Semana	Min Grupos	Max Grupos
Domingo	176	494
Segunda-feira	340	1138
Terça-feira	303	1629
Quarta-feira	327	2130
Quinta-feira	242	1897
Sexta-feira	205	899
Sábado	62	797
No período:	62	2130

Figure 3. Mínimos e Máximos de Clusters gerados com os parâmetros ds=10% e 100% de similaridade



Dados do tráfego de 05/03/2010 a 05/05/2010 (ds = 10% e 70%)

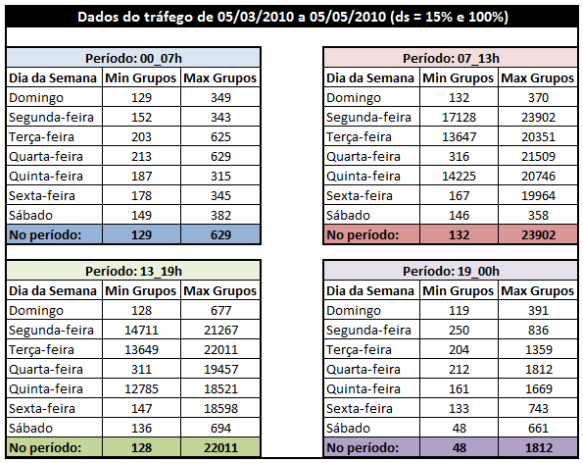
Período: 00_07h		
Dia da Semana	Min Grupos	Max Grupos
Domingo	22	141
Segunda-feira	20	144
Terça-feira	34	213
Quarta-feira	50	223
Quinta-feira	46	129
Sexta-feira	34	141
Sábado	19	162
No período:	19	223

Período: 07_13h		
Dia da Semana	Min Grupos	Max Grupos
Domingo	12	149
Segunda-feira	7407	9302
Terça-feira	6092	8471
Quarta-feira	123	8819
Quinta-feira	6192	8706
Sexta-feira	34	8093
Sábado	10	164
No período:	10	9302

Período: 13_19h		
Dia da Semana	Min Grupos	Max Grupos
Domingo	13	376
Segunda-feira	6565	8775
Terça-feira	6376	8857
Quarta-feira	111	8329
Quinta-feira	5708	7864
Sexta-feira	24	8158
Sábado	14	334
No período:	13	8857

Período: 19_00h		
Dia da Semana	Min Grupos	Max Grupos
Domingo	9	170
Segunda-feira	73	451
Terça-feira	34	813
Quarta-feira	37	1020
Quinta-feira	24	1031
Sexta-feira	22	422
Sábado	2	322
No período:	2	1031

Figure 4. Mínimos e Máximos de Clusters gerados com os parâmetros ds=10% e 70% de similaridade



Dados do tráfego de 05/03/2010 a 05/05/2010 (ds = 15% e 100%)

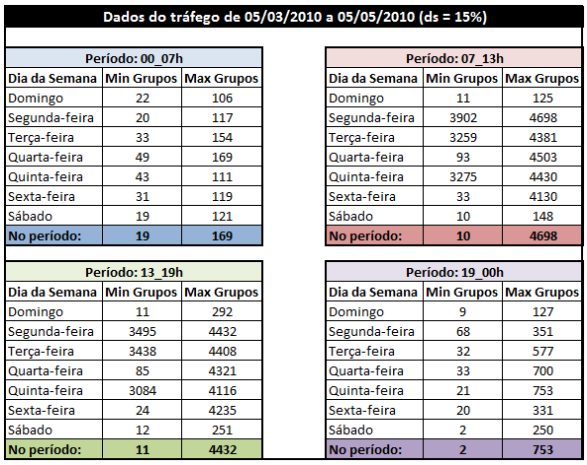
Período: 00_07h		
Dia da Semana	Min Grupos	Max Grupos
Domingo	129	349
Segunda-feira	152	343
Terça-feira	203	625
Quarta-feira	213	629
Quinta-feira	187	315
Sexta-feira	178	345
Sábado	149	382
No período:	129	629

Período: 07_13h		
Dia da Semana	Min Grupos	Max Grupos
Domingo	132	370
Segunda-feira	17128	23902
Terça-feira	13647	20351
Quarta-feira	316	21509
Quinta-feira	14225	20746
Sexta-feira	167	19964
Sábado	146	358
No período:	132	23902

Período: 13_19h		
Dia da Semana	Min Grupos	Max Grupos
Domingo	128	677
Segunda-feira	14711	21267
Terça-feira	13649	22011
Quarta-feira	311	19457
Quinta-feira	12785	18521
Sexta-feira	147	18598
Sábado	136	694
No período:	128	22011

Período: 19_00h		
Dia da Semana	Min Grupos	Max Grupos
Domingo	119	391
Segunda-feira	250	836
Terça-feira	204	1359
Quarta-feira	212	1812
Quinta-feira	161	1669
Sexta-feira	133	743
Sábado	48	661
No período:	48	1812

Figure 5. Mínimos e Máximos de Clusters gerados com os parâmetros ds=15% e 100% de similaridade



Dados do tráfego de 05/03/2010 a 05/05/2010 (ds = 15%)

Período: 00_07h

Dia da Semana	Min Grupos	Max Grupos
Domingo	22	106
Segunda-feira	20	117
Terça-feira	33	154
Quarta-feira	49	169
Quinta-feira	43	111
Sexta-feira	31	119
Sábado	19	121
No período:	**19**	**169**

Período: 07_13h

Dia da Semana	Min Grupos	Max Grupos
Domingo	11	125
Segunda-feira	3902	4698
Terça-feira	3259	4381
Quarta-feira	93	4503
Quinta-feira	3275	4430
Sexta-feira	33	4130
Sábado	10	148
No período:	**10**	**4698**

Período: 13_19h

Dia da Semana	Min Grupos	Max Grupos
Domingo	11	292
Segunda-feira	3495	4432
Terça-feira	3438	4408
Quarta-feira	85	4321
Quinta-feira	3084	4116
Sexta-feira	24	4235
Sábado	12	251
No período:	**11**	**4432**

Período: 19_00h

Dia da Semana	Min Grupos	Max Grupos
Domingo	9	127
Segunda-feira	68	351
Terça-feira	32	577
Quarta-feira	33	700
Quinta-feira	21	753
Sexta-feira	20	331
Sábado	2	250
No período:	**2**	**753**

Figure 6. Mínimos e Máximos de Clusters gerados com os parâmetros ds=15% e 70% de similaridade

A partir destes resultados, foi mapeada a faixa de clusters que caracteriza o comportamento padrão do tráfego da rede. Por exemplo, ao analisar o tráfego corrente coletado num domingo, utilizando parâmetros ds=15% e similaridade=70% (ver figura 6), se a quantidade de clusters obtido estiver contido no intervalo entre 22 e 106 no período P1, 11 e 125 no P2, 11 e 292 no P3 ou 9 e 127 no P4, o tráfego pode ser considerado como padrão, ou seja, não há traços de anomalia neste dia e período.

A etapa seguinte dos testes foi investigar qual dos nove atributos seria o responsável por gerar novos clusters com poucas sessões agrupadas, bem como a variação de seus valores, objetivando a remoção dos atributos não significativos para a caracterização do tráfego.

Para tanto, foi desenvolvido um script que analisa os resultados da clusterização e gera um relatório com os clusters que possuem menor número de sessões agrupadas e seus respectivos valores de atributos nas séries temporais. Ao analisar este relatório, observou-se características semelhantes nos dados dos períodos P1 (madrugada) e P4 (noite), bem como, nos dados de P2 (manhã) e P3 (tarde). Nos períodos P1 e P4, a maioria dos registros possuem valores diferentes de zero somente nos atributos smallpkt e duration, variando de 0 a 1 e de 0 a 601, respectivamente. Quando novos clusters com poucas sessões agrupadas foram gerados, observou-se que, nestas sessões, todos os atributos possuíam valores diferentes de zero. Este resultado permitiu concluir que não existe um único atributo responsável por gerar novos clusters nos dados avaliados.

Os conjuntos de dados referentes aos períodos P2 e P3 apresentaram registros na sua maioria com valores de atributos diferentes de zero, os quais variam entre: psizeCL= 0 a 1456; psizeSV= 0 a 1356; pnumCL= 0 a 595.205; pnumSV= 0 a 442.571; smallpkt= 0 a 1; dataDIR= -5441 a 51754; brecvCL= 0 a 860.404.736; brecvSV= 0 a 3.694.680; duration= 0 a 601. Nestes períodos, os novos clusters com menor número de sessões agrupadas são gerados pelos nove atributos (em conjunto ou separadamente), não sendo possível descartar quaisquer atributos do processo de clusterização.

6. Conclusões

Nos SDIs atuais observa-se a necessidade do uso de técnicas eficientes que proporcionem análise de grandes volumes de dados de rede em intervalos de tempo regulares, a fim de mapear o comportamento normal e anômalo do tráfego das redes monitoradas de modo preciso e em tempo satisfatório.

Neste trabalho, foi desenvolvida uma metodologia para caracterizar o comportamento padrão do tráfego de rede ao longo do tempo de forma automatizada, através da analise de atributos das sessões do tráfego, aplicando-se a técnica de mineração de dados por clusterização.

Através do perfil do tráfego padrão modelado é possível classificar os dados de tráfego corrente e identificar anomalias, permitindo ao administrador de rede aplicar contra medidas imediatas para restabelecer a segurança da rede.

Os melhores resultados de clusterização obtidos até o momento, aplicando-se a rede neural "Mapa de Kohonen Adaptável" nas séries temporais coletadas, foram gerados com os parâmetros de desvio 15% e similaridade 70%, reduzindo drasticamente o número de clusters formados. Através desta metodologia foram obtidos valores de limiar do numero de clusters que caracterizam o comportamento do tráfego nos diferentes dias da semana e períodos do dia.

Como meta para os trabalhos futuros, serão realizados testes em séries temporais de maior volume (4 meses de trafego) e séries contendo traços de ataques simulados em laboratório. A fim de analisar o custo computacional e a viabilidade da metodologia proposta para identificação de anomalias no tráfego de rede, serão pesquisados trabalhos recentes na área em busca de metodologias conhecidas na literatura que possam ser utilizadas para comparação.

7. Referências Bibliográficas

[1] Caswell, B., J. Beale, J. C. Foster, and J. Posluns, Snort 2 - Sistema de Detecção de Intruso Open Source, Editora Alta Books, Rio de Janeiro, 2003, chapter 1, pp. 23.

[2] Chaves, M. H. P. (2002) "Análise de Estado de Tráfego de Redes TCP/IP para Aplicação em Detecção de Intrusão". Dissertação de Mestrado em Computação Aplicada – INPE.

[3] Ertoz, L. et al.. (2003) "Detection and summarization of novel network attacks using data mining". Technical Report. Minneapolis, USA: University of Minnesota.

[4] Garcia, T.,P., Diaz, V. J., Fernandes, M. G. (2009) "Anomaly-basec network intrusion detection: Techniques, systems anda challenges". In: Computers & Security, volume 28, Issue 1-2, February, p. 18-28.

[5] Kayacik, H.G. et al. (2003) "On the capability of an SOM based intrusion detection system". In: IJCNN'2003 International Joint Conference on Neural Networks, 2003, Portland, Oregon, USA. Proceedings of… Piscataway, NJ, USA: IEEE,. v. 3, p. 1808-1813.

[6] Santos, A. C. F.; Silva, L.S.; Silva, J.D.S.; Rosa, R.R. (2009) "Aplicação de Técnicas de Análise de Séries Temporais em Dados de Tráfego de Rede". In:

Workshop dos Cursos de Computação Aplicada, 2009, INPE, São José dos Campos, SP.

[7] Silva, L.S.; Santos, A. C. F.; Mancilha, D. T.; Silva, J.D.S.; Montes, A. (2008) "Detecting attack signatures in the real network traffic with Annida". Expert Systems with Application: An International Journal, v. 34, issue 4, p.2326-2333. ISSN:0957-4174.

[8] Silva, L.S. (2007) "Uma Metodologia para Detecção de Ataques no Tráfego de Redes baseada em Redes Neurais". Dissertação (Doutorado em Computação Aplicada) – Instituto Nacional de Pesquisas Espaciais (INPE), São José dos Campos, SP, 2007, 254 p.

[9] Silva, L. S.; Santos, A. C. F.; Silva, J. D. S.; Montes, A. (2006) "Hamming net and LVQ neural networks for classification of computer network attacks: a comparative analysis". In: SBRN'2006 Brazilian Neural Networks Symposium, 9., 2006, Ribeirão Preto, São Paulo. Anais... [S.l.]: IEEE Explore Digital Library, p.13. ISBN 0769526802 http://doi.eeeecomputersociety.org/ 10.1109/SBRN.2006.21

[10] Silva, L.S.; Montes, A., Silva, J.D. S; Mancilha, T. D.; Santos, A. C. F. (2006) "A framework for analysis of anomalies in the network traffic". In: Workshop dos Cursos de Computação Aplicada, 6., 2006, INPE, São José dos Campos, SP. Anais... São José dos Campos. Disponível em: <eprint.sid.inpe.br/rep-/sid.inpe.br/ePrint@80/2006/12.20.23.21> Acesso em: 13 dez 2006.

[11] Silva, L. S.; Santos, A. C. F.; Silva, J. D. S.; Montes, A. (2005) "ANNIDA: Artificial Neural Network for Intrusion Detection Application - Aplicação da Hamming Net para detecção por assinatura". In: CBRN'2005 Congresso Brasileiro de Redes Neurais, 7., 2005, Natal, RN, Brasil. Anais... [S.l.]: [s.n.].

[12] Silva, L. S.; Santos, A. C. F.; Silva, J. D. S.; Montes, A. (2004) "Neural network application for attack detection in computer networks". In: IJCNN'2004 International Joint Conference on Neural Networks, 2004, Budapeste, Hungria. Proceedings… Piscataway, NJ, USA: IEEE,. (INPE-11626-PRE/7007).

[13] Tapiador, J.M. E. et al. (2004) "Measuring normality in HTTP traffic for anomaly-based intrusion detection". Computer Networks, v. 45, n. 2, p. 175-193.

[14] Zanero, S. (2005) "Improving Self Organizing Map performance for network intrusion detection". In: SDM'2005 SIAM Conference on Data Mining, 5., 2005, Newport Beach, CA.. Proceedings… [S.l.]: [s.n.], 2005. Disponível em: <http://citeseer.ist.psu.edu/zanero04improving.html>. Acesso em: 24 abr 2006.