Análise de Tendência aplicada a Estruturas Multivariadas em Tráfego de Fluxos de Redes

Arnoldo N. da Silva, Paulo Roberto Freire Cunha
Centro de Informática, UFPE
50732-970, Recife, PE
E-mail: ans2@cin.ufpe.br, prfc@cin.ufpe.br

Jorge L. Castro e Silva, José Everardo B. Maia
Universidade Estadual do Ceará - Departamento de Estatítica e Computação
60740-903, Campus do Itaperi, Fortaleza, CE
E-mail: jlcs@larces.uece.br, jmaia@larces.uece.br

Resumo: *A projeção de comportamento futuro auxilia na tomada de decisão de processos representados por séries no tempo. A automatização deste procedimento pode ser feita aplicando métodos de análise de tendência. Em estruturas multivariadas, encontradas em modelos de tráfego de redes, a alta dimensionalidade sugere a aplicação de métodos para a sua redução sem perda das propriedades, permitindo a análise de um conjunto de variáveis com um comportamento comum. A análise de componentes principais (PCA) permite representar estas estruturas através de um pequeno número de componentes. Este trabalho aplica análise de tendência e estatística multivariada em fluxos de tráfego de redes. A aplicação do PCA permite que um conjunto de variáveis tenha suas tendências de comportamento avaliadas simultaneamente. Estas variáveis representam pares de fluxos entre pontos de origem e destino na rede. Programas que implementam os métodos são desenvolvidos através do MATLAB e aplicados em dados reais da rede européia GEANT mostrando resultados satisfatórios.*

Introdução

A análise de tráfego de rede em relação a sua tendência de comportamento necessita de métodos matemáticos que permitam diagnóstico simultâneo dos diversos fluxos que trafegam nos pares de pontos da rede.

Métodos convencionais de análise de tráfego de redes são baseados em informações coletadas pelo SNMP (Sample Network Management Protocol)[8]. Esta ferramenta fornece dados referentes a todo o tráfego na interface do roteador, podendo parte destes dados estar apenas transitando e não sendo gerado ou destinado a este roteador.

Existe um conjunto de problemas importantes que requer uma modelagem e análise do tráfego na rede como um todo, ou seja, em todos os enlaces da rede simultaneamente. Porém as dificuldades de análise crescem proporcionalmente ao tamanho da rede, exigindo maior custo de processamento. Uma forma de abordar este tipo de análise é reconhecer que o tráfego observado em diferentes enlaces não é independente, e sim determinado por um conjunto de fluxos Origem-Destino (OD). Um fluxo OD é constituído por um conjunto de tráfego que é gerado por um ponto de entrada da rede e destinado a um ponto de saída comum.

A solução investigada neste trabalho busca aplicar a analise de tendência em um conjunto de fluxos OD que apresentam o mesmo padrão de comportamento. A estatística multivariada fornece meios que conduzem à solução proposta. Através da redução da dimensionalidade sem perda das propriedades dos elementos da amostra, são geradas estruturas que apresentam características comuns entre um conjunto de fluxos OD. A técnica de Análise dos Componentes Principais, PCA (do inglês "Principal Component Analysis") [4] é usada neste trabalho para obtenção desta redução.

Trabalhos Relacionados

O estudo inicial que mostra a viabilidade da aplicação de estatística multivariada em análise de fluxos OD é

apresentado em [5], através do uso de PCA, o autor conclui que em uma rede, os fluxos OD podem ser modelados usando um pequeno numero de componentes independentes. Alguns trabalhos [1] [2] [7] aplicam métodos de estatística multivariada em análise de tráfego baseadas em técnicas de agrupamento.

Com relação à previsão de tráfego, Pesquisas iniciais podem ser encontradas em [3], onde é usado um modelo de série temporal linear, mostrando que a série obtida pode ser modelada com ARIMA. Em [10], os autores focalizaram a projeção do tráfego na Internet sobre pequenas escalas de tempo, como segundos ou minutos, que são relevantes para a alocação dinâmica de recursos. [11] propõe um mecanismo para prever tendências congestionamentos após a série ser suavizada com a aplicação da transformada de wavelet discreta. Em [9], os autores utilizam a análise de multiresolução por wavelets para isolar as tendências de longo-termo e analisar a variabilidade em múltiplas escalas de tempo, sendo aplicado em amostras coletadas pelo SNMP.

Trabalhos citados acima apresentam soluções voltadas para análise de enlaces isoladamente. No entanto, não são considerados problemas que ocorrem devido à alta dimensionalidade da rede, que podem elevar o custo da gerência. Este trabalho diferencia-se por propor uma solução para tratar tais problemas.

Análise Estrutural

A amostra de dados coletada é organizada em uma matriz de tráfego (MT) que deve passar por um processo de pré-processamento, onde os dados atípicos (outliers) deverão ser tratados. A Análise de Componentes Principais pode ser utilizada para a identificação destas anomalias, através do gráfico da dispersão dos pontos entre a primeira e a segunda componentes principais. Os pontos que ficarem fora do intervalo onde há concentração de pontos serão candidatos à *outliers.* A estrutura gerada pela MT pode apresentar altas dimensões, pois a quantidade de pares de fluxos OD equivale ao quadrado do número de nós e a quantidade de instantes de tempo depende do período total de coleta e é inversamente proporcional ao intervalo entre duas coletas sucessivas.

Quando surge a necessidade de analisar uma estrutura de alta dimensão, é preciso buscar uma alternativa de aproximação para uma estrutura com baixa dimensionalidade, desde que preserve suas propriedades mais importantes. A técnica de estatística multivariada chamada Análise dos Componentes Principais (PCA) tem por finalidade básica a redução da dimensionalidade a partir de combinações lineares das variáveis originais.

Análise de Componentes Principais

PCA é um método de transformação de coordenadas que mapeia os dados medidos em novo sistema de eixos, chamados de componentes principais (CP). Cada CP tem a propriedade de apontar para a direção de variância máxima restante nos dados, considerando a energia já representada nos componentes anteriores. Os eixos principais são ordenados pela quantidade de energia nos dados por eles capturados.

Considerando uma matriz X, calcular os CPs é equivalente a resolver o problema de autovalores simétricos para a matriz X^TX. A matriz X^TX é a medida de covariância entre os fluxos. Cada componente principal v_i é o i-ésimo autovetor calculado a partir da decomposição espectral de X^TX.

$$X^T X v_i = \lambda_i v_i \qquad i = 1,...,n \tag{1}$$

onde λ_i é o autovalor correspondente a v_i. Devido ao fato X^TX ser definida positiva simétrica, seus autovetores são ortogonais e os autovalores correspondentes são reais não-negativos. Por convenção, os autovetores têm norma unitária e os autovalores são ordenados de forma decrescente, ou seja, $\lambda_1 \geq \lambda_1 \geq \cdots \geq \lambda_n$.

Uma vez que os dados foram mapeados em um espaço de componentes principais, os dados transformados em uma dimensão no tempo podem ser examinados. Considerando os dados mapeados em CPs, temos que a contribuição (equação 2) do eixo principal i em função do tempo é dada por Xv_i.

$$u_i = Xv_i \quad i = 1,\ldots,n \tag{2}$$

A equação acima mostra que todos os fluxos OD, quando pesados por v_i, produzem uma dimensão de dados transformados. Deste modo u_i captura a variação temporal comum a todos os fluxos por todo o eixo principal i.

Como os eixos principais estão em ordem de contribuição da energia total, u_1 captura a mais forte tendência comum temporal para todos os fluxos OD, u_2 captura o próximo mais forte, e assim por diante. Devido ao conjunto $\{u_i\}_{i=1}^{n}$ capturar a tendência comum de variação no tempo para os fluxos OD, este conjunto é referido como *autofluxo* de *X*. Assim a quantidade $r < n$ eixos principais que explicam a maior parte da variabilidade definio número de componentes.

O conjunto de CPs $\{v_i\}_{i=1}^{n}$ pode ser organizado na ordem de colunas de uma matriz principal *V*, cujo tamanho é *nxn*. Da mesma forma, podemos formar uma matriz *U* de dimensões *txn,* onde a coluna *i* é u_i. Assim temos:

$$X_i = U(V^T)_i \qquad i = 1,\ldots,n \qquad (3)$$

onde X_i é a série temporal do i-ésimo fluxo OD e $(V^T)_i$ é a i-ésima linha de *V*. A equação mostra que cada fluxo OD X_i está em torno de uma combinação linear de autofluxos, com pesos $(V^T)_i$ associados. Considerando as *r*'s primeiras componentes principais, podemos aproximar o *X* original como:

$$X' \approx \sum_{i=1}^{r} u_i v_i^T \qquad (4)$$

A aplicação de PCA em fluxos de tráfego desenvolvido por [5] classificou os autofluxos em três categorias: *d-autofluxo*; *s-autofluxo* e *n-autofluxo*. O grupo de autofluxos que apresentam fortes tendências e periodicidade é classificado como *d-autofluxo*, ou seja, autofluxos determinísticos, sendo estes indicados para a aplicação da análise de tendência. A segunda categoria, *s-autofluxo*, diz respeito aos autofluxos que exibem altos valores de curto período, ou seja, contém picos (spikes). Finalmente, os autofluxos que se apresentam aproximadamente estacionários são classificados como *n-autofluxo*, ou seja, representando ruídos (noises).

Cada um destes autofluxos carrega em sua estrutura características dos fluxos OD, como tendências de crescimento ou queda, períodos de estabilidade e comparações entre períodos de carga no tráfego.

Análise de Tendência

A análise de tendência abordada neste trabalho utiliza regressão linear e tem por objetivo encontrar uma função que represente a série temporal. O método dos mínimos quadrados é utilizado para calcular os coeficientes de uma curva que melhor representa um conjunto de dados observados. A função é obtida a partir dos pontos que minimizam a soma dos quadrados dos erros em relação aos pontos originais.

O polinômio de grau k, que compõe a função, representa a possível tendência. A curva estimada é calculada para o autofluxo, pois sua tendência de comportamento deve refletir nos fluxos OD que contém as características deste autofluxo. Com base neste cálculo, é possível verificar a tendência para os próximos períodos do tráfego em vários fluxos simultaneamente.

Estudo de Caso

A amostra de dados usada para o estudo de caso pertence à rede GEANT [12], um backbone da European National Research and Education Networks (NRENs) com 4 meses de coletas, sendo que neste caso, apenas um mês é avaliado. A topologia analisada é formada por 23 nós com 37 enlaces. O software Matlab[6] é usado para implementação e aplicação das operações de PCA e para estimar a curva média na análise de tendência.

A detecção de outliers é feita com a aplicação do PCA sobre os dados originais. O gráfico da figura 1 mostra a dispersão dos pontos entre a primeira e a segunda CP. O procedimento matemático utilizado para definir os pontos distantes da área da concentração é auxiliado pelo cálculo dos percentis. Os dados atípicos encontrados são substituídos nos fluxos por seus valores vizinho anteriores.

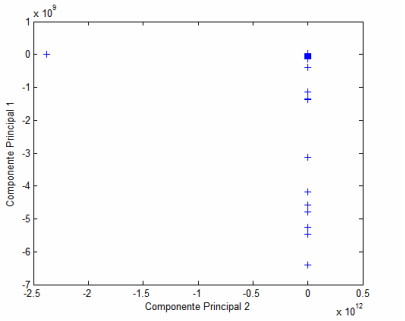


Figure 1: CP 1 x CP 2

A nova amostra gerada apresenta ausência de dados atípicos suficientes para a

aplicação eficiente da análise multivariada e posterior estudo de análise de tendência. A função *princomp* do MATLAB permite calcular o peso das CPs, os autofluxos e variância de cada componente. Com isso, é feito o cálculo do percentual da variância total. Os resultados, representado no gráfico da Figura 2 mostram que as 10 primeiras CPs explicam 84% da variabilidade total, permitindo obter uma boa representatividade dos dados de tráfego.

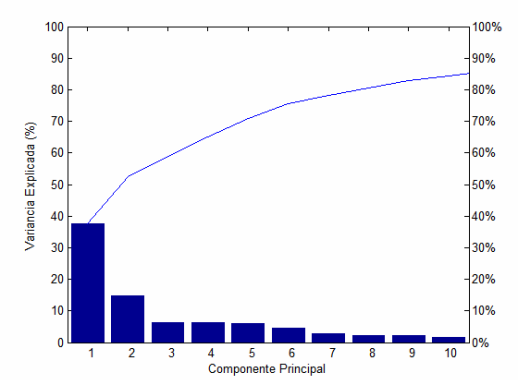
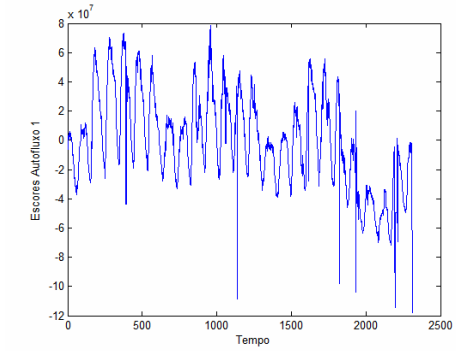
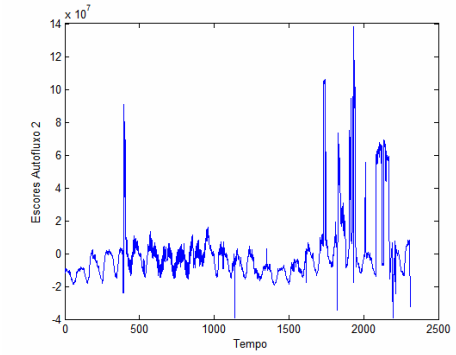


Figura 2: Gráfico do percentual de variabilidade da CPs

Os autofluxos gerados contêm as propriedades de um conjunto de variáveis (pares de fluxos OD) e a análise de tendência é aplicada nestas estruturas. A Figura 3 mostra, como exemplo, os autofluxos referentes às duas primeiras CPs.



(a)



(b)

Figura 3: (a) Escores do Autofluxo da CP 1; (b) Escores do Autofluxo da CP 2.

Neste estudo de caso, é mostrada a aplicação da análise de tendência no autofluxo da primeira CP. Um polinômio de grau 2 é usado para estimar a curva média na projeção a partir do tempo indexado por 1700 até o final da série no ponto 2311. A figura 4 mostra que, no autofluxo 1, a projeção da curva de ajuste segue uma tendência de queda acompanhando o comportamento da série original no intervalo de tempo analisado.

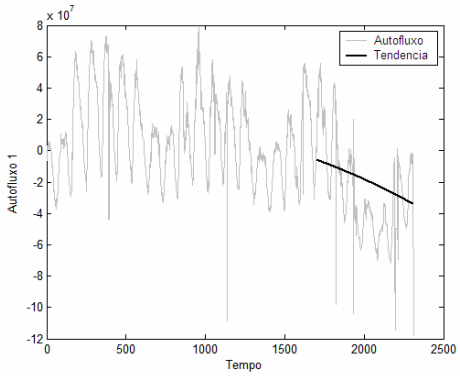


Figura 4: Gráfico com tendência no autofluxo 1.

O gráfico dos pesos mostrado na figura 5 permite identificar os fluxos que mais apresentam as características do autofluxo da CP 1.

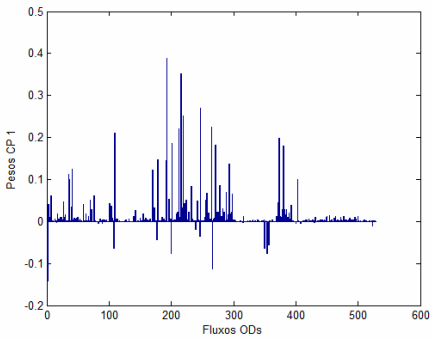


Figura 5: Gráfico dos pesos para a CP 1.

Um script é desenvolvido para automatizar esta busca, onde os fluxos identificados são aqueles cujos pesos são maiores em valores absolutos que um limiar definido em [5] por $1/\sqrt{n}$, onde n é o número de variáveis. A execução aplicada a esta componente aponta os fluxos mostrado na tabela 1.

Fluxos (índice)
7 27 35 36 41 70 76 101 109
170 178 192 193 196 201 212 215
216 219 221 224 232 242 247 255
257 265 270 271 278 288 293 298
372 373 380 403

Tabela 1: Índices dos fluxos com propriedades do Autofluxo 1.

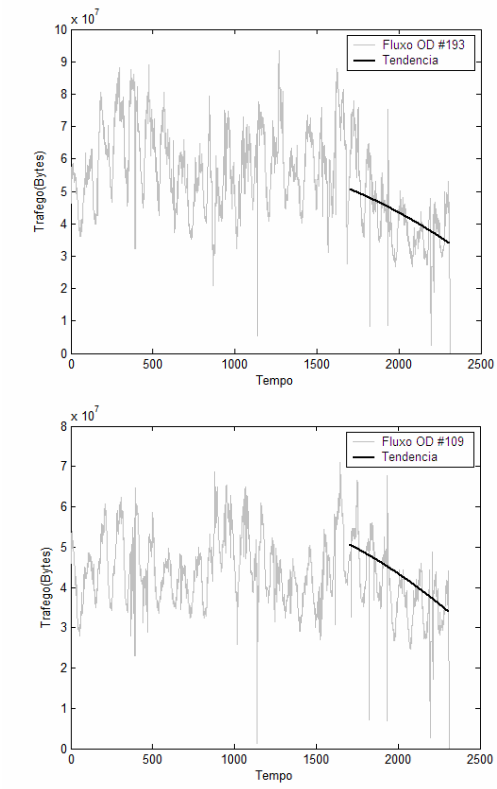


Figura 6: Gráficos com tendência nos fluxos 109 e 193.

Os fluxos identificados são, assim, analisados simultaneamente, pois a tendência de comportamento decrescente encontrada no autofluxo da CP 1 é refletida nestes fluxos. Como exemplo, é mostrada a análise de dois destes fluxos nos gráficos da figura 6, onde a mesma tendência é encontrada.

Conclusões

Este trabalho mostrou a viabilidade da análise de tendência do tráfego de redes a partir de estruturas multivariadas. A Análise de Componentes Principais atua na redução da dimensionalidade das séries. Os resultados mostraram que estudos de projeção de tráfego baseado nos autofluxos refletem em vários fluxos OD, auxiliando o administrador de grandes redes no planejamento da capacidade. Outras técnicas de estatística multivariada e de projeção serão experimentadas e avaliadas para investigar melhorias na precisão dos resultados.

Referências

[1] S. Fernandes, C. A. Kamienski, D. Mariz e D Sadok. Avaliação de Técnicas de Agrupamento na Amostragem de Tráfego na Internet. In: Simpósio Brasileiro de Redes de Computadores (SBRC 2006), 2006, Curitiba.

[2] R. H. Filho, Characterization and Compression of Internet Packet Header Traces using Cluster Analysis; Workshop em Desempenho de Sistemas Computacionais e de Comunicação (WPERFORMANCE): Anais do XXVI Congresso da Sociedade Brasileira de Computação. 2006.

[3] N. K. Groschwitz e G. C. Polyzos, A time serie model of long-term NSFNET backbone traffic. Proc. IEEE ICC'94 pp. 1400-1404, 1994.

[4] J. T. Jolliffe "Principal component analysis". Springer Verlag, New York, 1986

[5] A. Lakhina, , K. Papagiannaki, , M. Crovella, , C. Diot, , E. D. Kolaczyk and, N. Taft (2004). Structural Analysis of Network Traffic Flows. In: Proceedings of ACM SIGMETRICS, pp. 61-72, 2004.

[6] MATLAB for Windows User's Guide, The Math Works Inc., 1991.

[7] M. R. Oliveira, R. Valadas, A. Pacheco e P. Salvador "Cluster Analysis of Internet Users Based on Hourly Traffic utilization", IEICE Transactions on Communications, vol. E90–B, No.7, 2007.

[8] M. Schoffstall, M. Fedor, J. Davin e J. Case, "A Simple Network Management Protocol (SNMP), "RFC 1157, 1990.

[9] K. Papagiannaki, N. Taft, Z. Zhang, e C. Diot, Long-term forecasting of Internet backbone traffic. In IEEE Transactions on Neural Network, Vol. 16, No. 5, 2005.

[10] A. Sang e S. Li, A predictability analysis of network traffic. In Proc. of the IEEE INFOCOM'2000, Tel Aviv, Israel.

[11] J. L. C. Silva, ProCon – Prognóstico de Congestionamento de Tráfego de Redes usando Wavelets. Tese de Doutorado, (UFPE), 2004.

[12] S. Uhlig, B. Quoitin, S. Balon e J. Lepropre, Providing public intradomain traffic matrices to the research community. ACM SIGCOMM Computer Communication Review, 36(1), 2006.