Gerência de Tráfego de Redes Utilizando Baseline Bayesina

Cleverson Alessandro Veronez
veronez@lrg.ufsc.br

Sílvia Modesto Nassar
silvia@inf.ufsc.br

Carlos Becker Westphall
westphall@lrg.ufsc.br

Laboratório de Redes e Gerência (LRG)
Laboratório de Métodos de Trat. De Incerteza Comp. Evolucionária e Sist. Adaptativos (LISA)
Centro Tecnológico (CTC)
Universidade Federal de Santa Catarina (UFSC)
Caixa Postal 476, Florianópolis SC, Brasil. CEP 88040 970. +55(48)331-9739r235, Fax (48)331-9770

Resumo

A todo momento o administrador de rede de computadores, que tem suas decisões apoiadas basicamente em dados, depara-se com situações que exigem tratamento da incerteza. Apesar dos serviços oferecidos, as plataformas de gerência de redes atuais não são capazes de identificar problemas e sugerir ações corretivas, deixando ao administrador o encargo de interpretar gráficos e valores de variáveis. Neste contexto, este trabalho apresenta um estudo da aplicação do raciocínio probabilístico, através da implementação de uma rede bayesiana de conhecimento, a qual é capaz de reconhecer os relacionamentos entre os valores que representam o tráfego da rede. Em acréscimo, a rede bayesiana pode também ser utilizada como uma *baseline* dinâmica para sistemas especialistas proativos. O resultado deste trabalho consiste em um mecanismo que tem como objetivo oferecer suporte ao administrador no processo de tomada de decisão.

Palavras-chave

Inferência Bayesiana, Raciocínio Probabilístico, Sistema Especialista, *Baseline* Dinâmica, Gerência de Redes, *Knowledge Data Discovery*.

Abstract

At every moment the computer network administrator, which has his/her decisions supported basically on data, comes across situations that demand treatment of uncertainty. In spite of the offeredservices, the computer network management platforms are not capable of identifyin;problems and suggesting corrective actions, and as a consequence leaves to the administrator the burden of interpreting graphs and variable values. Within this context, this work presents a study on the application of probabilistic reasoning, through the implementation a bayesian belief network, which is capable of recognizing the relationships among the figures that represent traffic of the network. In addition, the bayesian network may also be used as a dynamic baseline for proactive expert systems. The outcome of this research resulted in a mechanism, which aims at supporting the administrator in the decision making process.

Key-Words

Bayesian Inference, Probabilistic Reasoning, Expert System, Dynamic Baseline, Network Management, Knowledge Data Discovery.

Através da expansão deste trabalho será implemento um sistema especialista bayesiano de apoio a gerência de redes que, além de identificar problemas, através das probabilidades das hipóteses diagnosticas, seja capaz de apresentar possíveis causas e soluções.

9. Referências Bibliográficas

[CHE 85] CHEESEMAN, P. "In defense of probability". Proceedings of 9th International Joint Conference on Artificial Intelligence. Los Angeles, pp 1002 -1009. 1985.

[CIS 99] CISCO. URL: http://cisco.com/warp/public/733/7000/, janeiro de 1999.

[COH 85] COHEN, P. R. "Heuristic reasoning about uncertainty: artificial intelligence approach". Boston: Pitman. 1985.

[FAY 96] FAYYAD, U.M.; PIATETSKY-SHAPIRO, G.; SMYTH, P.; UTHURUSAMY, R. "Advances in Knowledge Discovery and Data Mining". American Association for Artificial Intelligence, Menlo Park, Califórnia EUA. 1996.

[FRA 97] FRANCESCHI, A.S.M.; ROCHA, M.A.;WEBER, H.L.; WESTPHALL, C.B. "Employing Remote Monitoring and Artificial Intelligence Techniques to Develop the Proactive Network Management". Proceedings of the International Workchop on Applications of Neural Networks to Telecommunication 3. Laurence Erlbaum Associates, Publishers. Mahwah, (NJ), USA. 1997.

[FRA 91] FRAWLEY, W.J.; PIATETSKY-SHAPIRO, G.; AND MATTHEUS, C.J. 1991. "Knowledge Discovery in Datavases: Na Overview". In Knowledge Discovery in Databases, ed. G. Piatetsky-Shapiro and B. Frawley. Cambridge, Mass: AAAI/MIT Press, 1-27.

[IET 99] IETF Internet Engening Task Force URL: http://www.ietf.cnri.reeston.va.us/rfc/rfcl213.txt, janeiro de 1999.

[KNO 97] KNOBBE, A. J. "Data Mining for Adpative System Management". In proceedings of PADD. 1997.

[LAU 88] LAURITZEN, S. L. & SPIEGELHALTER, D. J. "Local computations with probabilities on graphical structures and their applications to expert systems". J. Royal Statist. Soc., B, 50(2): 154-227. 1988.

[LIN 96] LINDA, C. Van Der Gaag. "Bayesian Belief Networks: Odds and Ends". The Computer Journal. Volume 39. Número 02. 1996.

[LIN 82] LINDLEY, D. V. "Scoring rules and the inevitability of probability". International Statistical Review, (50):1-26. 1982.

[MEG 98] MEGAPUTER. URL: http://www.megaputer.com , setembro de 1998.

[MIC 99] Microsoft. URL: http://microsoft.com/office/,janeiro de 1999.

[NAS 98] NASSAR, S.M. "Estatística e Informática: Um Processo Iterativo Entre Duas Ciências". Trabalho apresentado no concurso para professor titular. Departamento de Informática e Estatística. Centro de Tecnologia. Universidade Federal de Santa Catarina. 128p. Abril, 1998.

[NET 99] NETICA. URL: http://www.norsvs.com. janeiro de 1999.

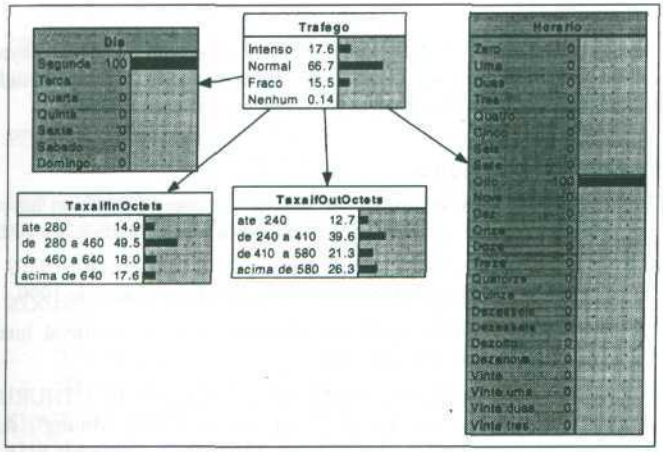

Figura 7.3: Rede Bayesiana *a posteori,* dado que é segunda-feira e oito horas

Veja abaixo a tabela *1.2,* apresentando a evolução das hipóteses diagnosticas dadas as evidências.

Hi	P(Hi)	**P(Hi/Seg)**	**P(Hi/ Seg e Oito_horas)**
Intenso	0,0900	0,1193	0,1762
Normal	0,5300	0,6322	0,6669
Fraco	0,3700	0,2452	0,1552
Nenhum	0,0100	0,0033	0,0014

Tabela 7.2: Evolução das Hipóteses Diagnosticas do Tráfego

Da forma como foi demonstrada, utilizando a rede bayesiana, com um simples clique do mouse podemos obter facilmente qualquer probabilidade sobre o tráfego da rede onde as probabilidades *a priori* vão se alterando através da aquisição de informação das evidências.

8. Conclusões e Perspectivas Futuras

Este trabalho contribuiu para o estudo de modelos probabilísticos. Com os experimentos realizados e através do protótipo implementado constatou-se a adequação do enfoque probabilístico no desenvolvimento de um sistema especialista de apoio à gerência de redes. O protótipo implementado ajuda também a compreender melhor o raciocínio em situações de incerteza, podendo ser usado para o treinamento de futuros administradores.

Em acréscimo, o presente trabalho apresentou um novo conceito na área de Gerência de Redes, o conceito de "*Baselines* Dinâmicas". Onde a rede bayesiana implementada é utilizada para expressar o comportamento da rede, atualizando-se com as mudanças da mesma. Uma das vantagens da utilização da *baseline* implementada é que ela reflete o comportamento da rede através de probabilidades, ou seja, um determinado comportamento pode estar X% dentro do esperado e não, como ocorre nas *baselines* convencionais, simplesmente estar ou não de acordo com o perfil da rede monitorada.

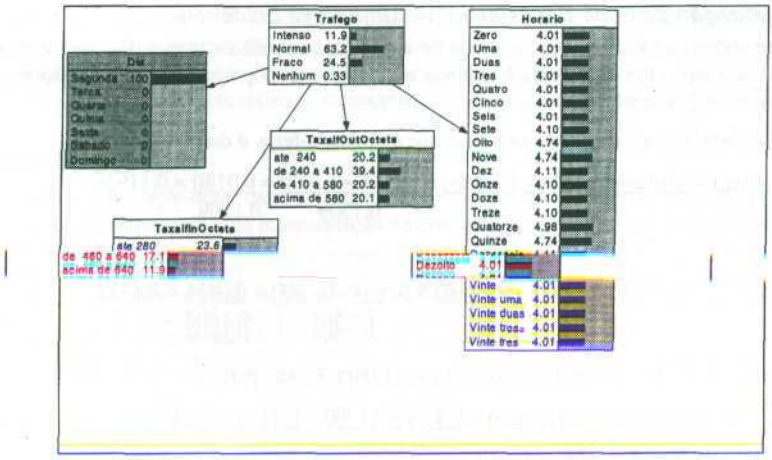

Figura 7.2: Rede Bayesiana *a posteori*, dado que é segunda-feira

Se além do evento Segunda-feira acrescentarmos que são Oito horas as probabilidades seriam

$$P(\text{Intenso/seg}\,/\text{Oito\_horas}) = \frac{P(\text{Intenso}_{/\text{Seg}}\ \text{n}\ \text{Oito\_horas})}{P(\text{Oito\_horas})} =$$

$$= \frac{P(\text{Intenso}_{/\text{Seg}}) \bullet P(\text{Seg/Intenso}_{/\text{Oito\_horas}})}{P(\text{Oito\_horas})} = \frac{0,1193 \bullet 0.074}{0,0474} = 0,1762$$

$$P(\text{Normal}_{/\text{Seg}}\ |\text{Oito\_horas}) = (0,6322 \bullet 0,0500) / \ 0,0474 = 0,6669$$
$$P(\text{Fraco}_{/\text{Seg}}\ |\text{Oito\_horas}) = (0,2452 \bullet 0,0300) / 0,0474 = 0,1552$$
$$P(\text{Nenhum}_{/\text{Seg}}\ |\text{Oito\_horas}) = (0,0033 \bullet 0,0200) / 0,0474 = 0,0014$$

Onde a probabilidade de haver tráfego às oito horas sabendo que é Segunda-feira é dada por:

$$P(\text{Oito\_horas}) = P(\text{Intenso}_{/\text{Seg}} \cap \text{Oitojioras}) + P(\text{Normal}_{/\text{Seg}}\ \text{n Oitojioras}) + P(\text{Fraco}_{/\text{Seg}}\ c$$
Oitojioras) + $P(\text{Nenhum}_{/\text{Seg}}$ n Oitojioras) = $0,0700 \bullet 0,1193 + 0,0500 \bullet 0,6322 +$
$0,0300 \bullet 0,2452 + 0,0200 \bullet 0,0033 = 0,0474$

De forma análoga são calculadas as probabilidades *a posteori* para todas as outras evidências resultando a rede *a posteori* apresentada na figura 7.3.

7.3. Atualização da Rede Bayesiana Para Uma Nova Evidência

Dada a ocorrência de uma evidência a rede deve atualizar as probabilidades. Por exemplo, se for constatado que o dia da semana é segunda feira, altera-se as probabilidades das Hipóteses Diagnosticas, veja a tabela 7.1.

A probabilidade do tráfego ser Intenso dado que é Segunda-feira é dada por:

$$P(Intenso/Seg) = \frac{P(Intenso) \cdot P(Seg/Intenso)}{P(Seg)} = \frac{0,0900 \cdot 0,2000}{0,1509} = 0.0180 = 0,1193$$

A probabilidade do tráfego ser Normal se é Segunda-feira é dada por:

$$P(Normal/Seg) = \frac{P(Normal) \cdot P(Seg/Normal)}{P(Seg)} = \frac{0.5300 \cdot 0.1800}{0,1509} = 0.0954 = 0,6322$$

A probabilidade do tráfego ser Fraco se é Segunda-feira é dada por:

$$P(Fraco/Seg) = \frac{P(Fraco) \cdot P(Seg/Fraco)}{P(Seg)} = \frac{0.3700 \cdot 0.1000}{0,1509} = 0.0370 = 0,2452$$

A probabilidade do tráfego ser Nenhum se é Segunda-feira é:

$$P(Nenhum/Seg) = \frac{P(Nenhum) \cdot P(Seg/Nenhum)}{P(Seg)} = \frac{0,0100 \cdot 0,0500}{0,1509} = 0.0005 = 0,0033$$

Hipóteses Diagnosticas	P(Hi)
Intenso	0,1193
Normal	0,6322
Fraco	0,2452
Nenhum	0,0033

Tabela 7.1: Probabilidades das hipóteses diagnosticas *a posteori*

Da mesma forma, as probabilidades das outras evidências também se alteram dada a certeza de ocorrência de uma das evidências.

A probabilidade da Taxa Média ifOutOctets ser menos do que 240 Kbps dado que é segunda:

$$P(TaxaifOutOctets\_até\_240) = P(Intenso_{/Seg} \cap TaxaifOutOctets\_até\_240) + P(Normal_{/Seg} \cap TaxaifOutOctets\_até\_240) + P(Fraco_{/Seg} \cap TaxaifOutOctets\_até\_240) + P(Nenhum_{/Seg} \cap TaxaifOutOctets\_até\_240) = P(Intenso_{/Seg}) \cdot P(TaxaifOutOctets\_até\_240 / Intenso_{/Seg}) + P(Normal_{/Seg}) \cdot P(TaxaifOutOctets\_até\_240/Normal_{/Seg}) + P(Fraco_{/Seg}) \cdot P(TaxaifOutOctets\_até\_240/Fraco_{/Seg}) + P(Nenhum_{/Seg}) \cdot P(TaxaifOutOctets\_até\_240 / Nenhum_{/Seg}) = P(TaxaifOutOctets\_até\_240) = 0,0000 \cdot 0,1193 + 0,0000 \cdot 0,6322 + 0,8100 \cdot 0,2452 + 1,0000 \cdot 0,0033 = 0,2019$$

De forma análoga são calculadas as probabilidades *a posteori* para todas as outras evidências, resultando a rede *a posteori* apresentada na figura 7.2.

Uma Rede Bayesiana de Conhecimento pode ser utilizada como *baseline*, apresentando um novo conceito na área de Sistemas Especialistas de Gerência de Redes, o conceito de *"Baselines* Dinâmicas". Dentre as vantagens de seu destacam-se que ela atualiza-se com as mudanças da rede e que ela reflete o comportamento da rede através de probabilidades.

7.1. A *Shell* Utilizada

As *Shells* são *softwares* que facilitam a construção de Sistemas Especialistas pelo fornecimento de esquemas de representação do conhecimento e de máquinas de inferência.

A *Shell* utilizada neste trabalho chama-se Netica [NET 99], foi desenvolvida pela *Norsys Software Corp.* em Vancouver, BC, Canadá e utiliza redes de crença para realizar vários tipos de inferência usando algoritmos modernos e rápidos. Dado um novo caso, pelo qual o usuário tem conhecimento limitado, Netica encontrará os valores ou probabilidades apropriadas para todas as variáveis desconhecidas.

7.2. A Rede Bayesiana a Priori

Utilizando a *Shell* Netica, dada as probabilidades das Hipóteses e as probabilidades das evidências, foi montada a rede bayesiana *a priori.*

Os cálculos das probabilidades foram realizados de acordo com a Probabilidade Bayesiana. Por exemplo, a probabilidade de ter tráfego na Segunda-feira, P(Seg), é dada por:

$$P(Seg) = P(Intenso \cap Seg) + P(Normal \cap Seg) + P(Fraco \cap Seg) + P(Nenhum \cap Seg)$$

$$P(Seg) = P(Intenso) \cdot P(Seg/Intenso) + P(Normal) \cdot P(Seg/Normal) + P(Fraco) \cdot P(Seg/Fraco) + P(Nenhum) \cdot P(Seg/Nenhum)$$

$$P(Seg) = 0{,}0900 \cdot 0{,}2000 + 0{,}5300 \cdot 0{,}1800 + 0{,}3700 \cdot 0{,}1000 + 0{,}0100 \cdot 0{,}0500 = 0{,}1509$$

De forma análoga foram efetuados os cálculos para os demais dias da semana, e para as demais evidências, resultando nas suas respectivas probabilidades *a priorí,* e consequentemente na rede bayesiana *a priorí* apresentada na figura 7.1.

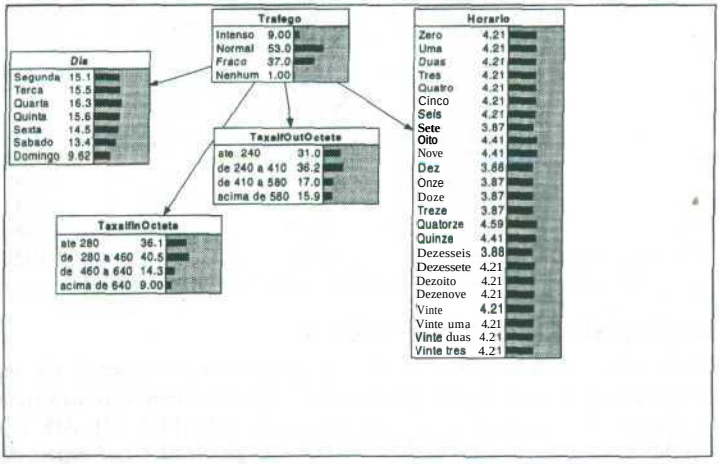

Figura 7.1: Rede Bayesiana *a priorí*

Taxa Média ifInOctets	P(ek/Intenso)	P(ek/Normal)	P(ek/Fraco)	P(ek/Nenhum)
Até 280 Kbps	0,0000	0,0000	0,9500	0,1000
De 280 a 460 Kbps	0,0000	0,7300	0,0500	0,0000
De 460 a 640 Kbps	0,0000	0,2700	0,0000	0,0000
Acima de 640 Kbps	1,0000	0,0000	0,0000	0,0000

Tabela 6.3: Probabilidades condicionais da evidência Taxa Média iflnOctets

Taxa Média ifInOctets	P(ek/Intenso)	P(ek/Normal)	P(ek/Fraco)	P(ek/Nenhum)
Até 240 Kbps	0,0000	0,0000	0,8100	1,0000
De 240 a 410 Kbps	0,0000	0,5500	0,1900	0,0000
De 410 a 580 Kbps	0,0000	0,3200	0,0000	0,0000
Acima de 580 Kbps	1,0000	0,1300	0,0000	0,0000

Tabela 6.4: Probabilidades condicionais da evidência Taxa Média ifOutOctets

Horária	P(ek/Intenso)	P(ek/Normal)	P(ek/Fraco)	P(ek/Nenhum)
Zero	0,0200	0,0400	0,0500	0,0600
Uma	0,0200	0,0400	0,0500	0,0600
Duas	0,0200	0,0400	0,0500	0,0600
Três	0,0200	0,0400	0,0500	0,0600
Quatro	0,0200	0,0400	0,0500	0,0600
Cinco	0,0200	0,0400	0,0500	0,0600
Seis	0,0200	0,0400	0,0500	0,0600
Sete	0,0700	0,0400	0,0300	0,0100
Oito	0,0700	0,0500	0,0300	0,0200
Nove	0,0700	0,0500	0,0300	0,0200
Dez	0,0700	0,0400	0,0300	0,0200
Onze	0,0700	0,0400	0,0300	0,0100
Doze	0,0700	0,0400	0,0300	0,0100
Treze	0,0700	0,0400	0,0300	0,0100
Quatorze	0,0900	0,0500	0,0300	0,0200
Quinze	0,0700	0,0500	0,0300	0,0200
Dezesseis	0,0700	0,0400	0,0300	0,0200
Dezessete	0,0200	0,0400	0,0500	0,0600
Dezoito	0,0200	0,0400	0,0500	0,0600
Dezenove	0,0200	0,0400	0,0500	0,0600
Vinte	0,0200	0,0400	0,0500	0,0600
Vinte e uma	0,0200	0,0400	0,0500	0,0600
Vinte e duas	0,0200	0,0400	0,0500	0,0600
Vinte e três	0,0200	0,0400	0,0500	0,0600

Tabela 6.5: Probabilidades condicionais da evidência Horário

7. A Rede Bayesiana de Conhecimento

A *baseline* é uma caracterização estatística do funcionamento normal do segmento monitorado da rede [NET 98]. Elas podem ser criadas de diversas formas, as mais comuns são através de técnicas de simulação ou de monitoração da rede [FRA 97]. Até o presente momento as *baselines* foram criadas de forma estática. O problema é que depois de algum tempo, estas *baselines* não refletem mais o comportamento da rede.

Ao final de uma coleta a aplicação "*stand by*" por um período de aproximadamente cinco minutos, até ser novamente executada.

6. O Processo de *Knowledge Data Discovery* (KDD)

6.1. Preparação dos Dados e Estimativa das Probabilidades A Priori

Foi implementada uma outra aplicação java para calcular os valores das probabilidades *a priori.* Esta aplicação utiliza o arquivo de dados realizando uma contagem para assim estimar as probabilidades da ocorrência das hipóteses diagnosticas e dos eventos associadas a elas. A figura 6.1 apresenta a interface de controle da aplicação de cálculo das probabilidades.

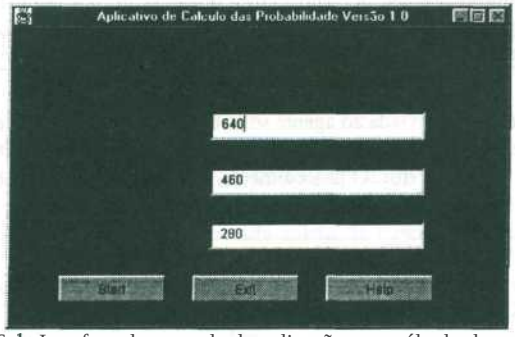

Figura 6.1: Interface de controle da aplicação para cálculo das probabilidades

Calculadas as probabilidades *a priori* **P(Hi)** e as probabilidades condicionais **P(e/Hi)**, a aplicação grava estas probabilidades em um arquivo .MBD, que contém as estruturas das tabelas 6.1, 6.2, 6.3, 6.4 e 6.5. Os valores apresentados nestas tabelas correspondem aos valores obtidos com os experimentos realizados.

Hipóteses Diagnosticas	P(Hi)
Intenso	0,0900
Normal	0,5300
Fraco	0,3700
Nenhum	0,0100

Tabela 6.1: Probabilidades das hipóteses diagnosticas

Dia da Semana	P(ek/Intenso)	P(ek/Normal)	P(ek/Fraco)	P(ek/Nenhum)
Segunda	0,2000	0,1800	0,1000	0,0500
Terça	0,2000	0,1800	0,1100	0,0500
Quarta	0,1700	0,1800	0,1400	0,0500
Quinta	0,1600	0,1900	**0,1100**	0,0500
Sexta	0,1500	0,1500	0,1400	0,0500
Sábado	0,1000	0,0900	0,2000	0,3000
Domingo	0,0200	0,0300	0,2000	0,4500

Tabela 6.2: Probabilidades condicionais da evidência Dia da Semana

Dia	Mês	Ano	Dia Semana	Hora	Minuto	ifOutOctets	ifInOctets
18	Jan	1999	Mon	16	15	802270	912464

Tabela 5.1: Exemplo da estrutura do arquivo de dados do programa de coleta

Também é utilizado um arquivo texto onde são gravadas informações sobre qualquer problema ocorrido. Como as variáveis monitoradas são contadores, elas podem retornar a zero, ocasionando erros nas taxas. Outro problema que podemos ter na coleta é a falhas de comunicação na rede. No caso de ocorrência de erros, os dados coletados não são gravados e a aplicação realiza imediatamente outra tentativa de coleta. No arquivo de erros é gravado a data, horário, tipo do erro e onde o erro ocorreu, conforme o quadro 5.1.

```
Fri Jan 22 18:37:02 GMT-02:00 1999 O Tempo decorrido negativo ou igual a
zero
Sat Jan 23 12:33:50 GMT-02:00 1999 timed out na comunicação para:
200.135.0.30 Erro no ifOutOctets
```

Quadro 5.1: Exemplo do arquivo de erros

A PDU *(Protocol Data Unit)* enviada ao agente solicitando o envio dos valores das variáveis coletadas é montada pelo próprio programa de coleta, dando liberdade para qualquer alteração futura e maior controle sobre os erros. O procedimento utilizado para compor a PDU pode ser visto no quadro 5.2.

```
Public String getID (String Xid, String sq) {
       SnmpAPI api;
       String val = "";
       String mib="Rfc1213-mib";
       api = new SnmpAPI();
       api.start();
       SnmpSession session = new SnmpSession(api);
       try{
          Thread.sleep(10);
          System.out.println("thread acordou...");
           } catch (Exception ie) {
           }
      session = new SnmpSession(api);
       session = new SnmpSession(api);
       session.peername = sq;
       session.community = "public";
       SnmpPDU pdu = new SnmpPDU(api);
       pdu.command = api.GET_REQ_MSG;
       SnmpOID oid;
       oid = new SnmpOID(Xid,api);
       pdu.addNull(oid);
       try(
            session.open();
            pdu = session.syncSend(pdu);
           SnmpVarBind varaux= (SnmpVarBind) pdu.variables.firstElement();
           val=varaux.variable.toString();
           } catch (SnmpException e) { val = "Erro";
               fileErro.println(tempoAtual + "SnmpException:   "+ e);
             } catch (RuntimeException e) { val = "Erro";
                 fileErro.println(tempoAtual + "RuntimeException:  "+ e);
               } catch (Exception e) { val = "Erro";
                 fileErro.println(tempoAtual + "Exception:   "+ e);
               }
      if (api.client == null) {
               val = "Erro";        } // fim do If api.client
   session.close();
   api.stop();
   System.out.println("taxa: " + val);
   return(val);
} // fim do método getID
```

Quadro 5.2: Procedimento que monta a PDU de coleta dos dados

- **ifInDiscards** - o número de pacotes recebidos em uma *interface* acima do limite. Estes pacotes são escolhidos para serem descartados mesmo que não tenham sido detectados erros. Uma razão para descartar pacotes é que ele pode ser maior do que o espaço livre no *buffer* do roteador. Nome: IF-MIB!ifInDiscards; Identificador: 1.3.6.1.2.1.2.2.1.13;
- **ifInErrors** - o número de unidades de transmissão recebidos com erros. Estes pacotes ou unidades de transmissão são descartados, impedindo que os erros se propaguem para o protocolo de nível mais alto. Nome: IF-MIB!ifInErrors; Identificador: 1.3.6.1.2.1.2.2.1.14.

Em uma fase inicial deste trabalho foram monitoradas todas as variáveis acima, porém com a implementação de um protótipo, percebeu-se que as variáveis ifOutDiscards, ifOutErrors, iflnDiscards e iflnErrors mantinham-se constantes (e seus valores próximos de zero), portanto a certeza da ocorrência destas evidências não altera a probabilidade das hipóteses diagnosticas, assim sendo, o programa de coleta, descrito abaixo, foi aprimorado para coletar apenas os valores das variáveis relevantes para este trabalho, explorando a esparcidade entre as variáveis, aumentando assim a performance do sistema.

5. A Coleta dos Dados

Foram coletados dados relativos à *interface* serial 4 do roteador citado anteriormente, utilizando para isto uma aplicação java. Os dados foram coletados com espaços de, aproximadamente, cinco minutos entre cada coleta. O tempo decorrido entre uma coleta e outra é variável; esta variação é decorrente do tempo de rede existente entre a comunicação do processo gerente (a aplicação de coleta) e o agente no roteador. Como este trabalho faz uso apenas dos valores das taxas médias dos dados coletados, a variação entre as coletas é irrelevante para o resultado final do trabalho.

A coleta tem como objetivo capturar dados para estimar as probabilidades das hipóteses diagnosticas de tráfego na rede, utilizando técnicas estatísticas de *Knowledge Data Discovery*.

5.1. O programa de coleta dos dados

Foi criada uma aplicação em Java para realizar a coleta dos dados. Esta aplicação ficou sendo executada em uma estação de trabalho IBM RS/6000, coletando os valores das variáveis monitoradas e os valores de suas taxas médias. A interface de controle da aplicação de coleta pode ser vista na figura 5.1.

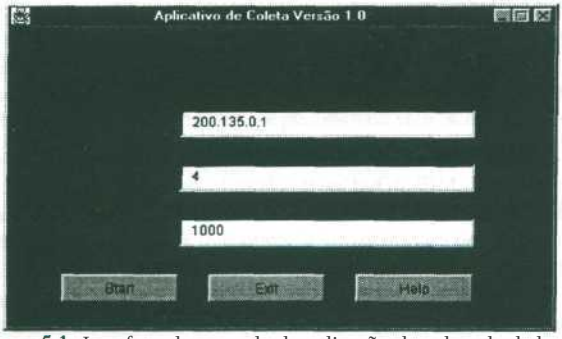

Figura 5.1: Interface de controle da aplicação de coleta de dados

A aplicação de coleta armazena os dados em um arquivo .MBD. Os dados são coletados conforme a estrutura da tabela 5.1.

Para o processo de KDD ser realizado com sucesso, são necessários seguir alguns passos fundamentais [MEG 98]: definição de objetivos, seleção, entendimento, limpeza, enriquecimento e preparação dos dados e; criação do modelo para DM.

Existem vários métodos que podem ser utilizados no processo de KDD e DM entre eles: sistemas analíticos orientados ao assunto, métodos estatísticos, redes neurais, programação evolucionária, raciocínio baseado em casos, árvores de decisão, algorítmos genéticos e métodos de regressão não-linear. Este trabalho faz uso de métodos estatísticos no processo de KDD.

4. Domínio da Aplicação

As variáveis monitoradas nos dão informações sobre o tráfego existente entre a RNP, Rede Nacional de Pesquisa [RNP 99] e a RCT, Rede Catarinense de Ciência e Tecnologia [RCT 99], pois são relativas ao roteador e a respectiva porta que realiza a comunicação entre estas duas redes, conforme a figura 4.1.

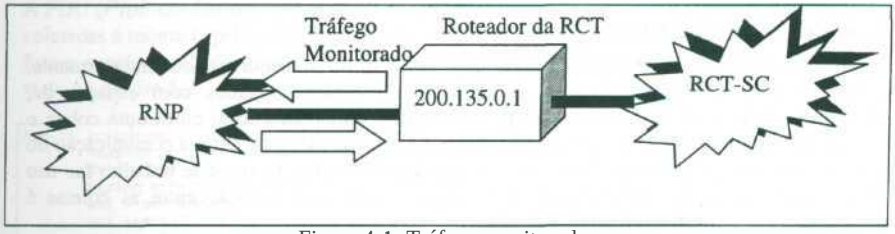

Figura 4.1: Tráfego monitorado

4.1. O Roteador Monitorado

O roteador monitorado é da marca Cisco, da série Cisco 7000 e localiza-se fisicamente nas instalações do Núcleo de Processamento de Dados da Universidade Federal de Santa Catarina em Florianópolis, SC. Este roteador é multiprotocolo. As *interfaces* de rede consistem em processadores de *interface* modulares, que oferecem uma conexão direta entre os barramentos de alta velocidade Cisco Extended Bus (CxBus) e a rede externa [CIS 99].

4.2. As Variáveis Monitoradas

As variáveis monitoradas pertencem ao grupo *interfaces* da MIB2 - Internet especificada na RFC 1213 [ET 99]. O Grupo *Interfaces* oferece dados sobre cada *interface* de um dispositivo gerenciável da rede. Essas informações são úteis para o gerenciamento de falhas, de configuração, de performance, e de contabilização. As variáveis monitoradas foram:

- **ifOutOctets** - o número total de bytes transmitidos por uma *interface*, incluindo caracteres de cabeçalho. Nome: IF-MIB!ifOutOctets; Identificador: 1.3.6.1.2.1.2.2.1.16;
- **ifOutDiscards** - o número de pacotes a serem transmitidos por uma *interface* acima do limite. Estes pacotes são escolhidos para serem descartados mesmo que não tenham sido detectados erros. Nome: IF-MIB!ifOutDiscards; Identificador: 1.3.6.1.2.1.2.2.1.19;
- **ifOutErrors** - o número de unidades de transmissão que contiveram erros. Estes pacotes ou unidades de transmissão são descartados, impedindo que os mesmos se propaguem pela rede. Nome: IF-MIB!ifOutErrors; Identificador: 1.3.6.1.2.1.2.2.1.20;
- **ifInOctets** - o número total de bytes recebidos em uma *interface*, incluindo caracteres de cabeçalho. Nome: IF-MIB!ifInOctets; Identificador: 1.3.6.1.2.1.2.2.1.10;

Nas aplicações dos sistemas especialistas probabilísticos os Hj's são as hipóteses diagnosticas concorrentes. O evento e pode ser visto como uma evidência. O conhecimento da ocorrência desta evidência leva à mudanças na probabilidade *a priori* P(Hj) para a probabilidade condicional P(H_i|e), que por sua vez considera a evidência e.

Sejam os eventos *e*\, e_2 ç E, se $P(e_1 \wedge e_2) = P(e_1) \cdot P(e_2)$ então os eventos e_1 e e_2 são independentes. Genericamente, para qualquer subconjunto E = { $e_{i1}, e_{i2}, ..., e_{1ik}$} de {$e_1, e_2, ..., e_n$} se $P(e_{i1} \wedge e_{i2}$ A ... $\wedge e_{ik}$ |H) = P(e_{i1} l H). P(ei2 l H) ... P(e_{ik} l H) então pode-se dizer que os eventos e_i's são eventos mutuamente independentes dado a hipótese H. A idéia básica do conceito probabilístico de independência é que o conhecimento de certa informação não traz informação adicional sobre outra informação. Isto é, se e somente se, saber que o evento *e*\ ocorreu não trouxer informação sobre o evento e_2, e saber que o evento e_2 ocorreu não trouxer informação sobre o evento e_1 então diz-se que ocorre a independência entre estes eventos.

Seja H uma hipótese e $e^n = e_1, e_2, ..., e_n$ uma seqüência de eventos independentes, observados no passado, e seja e um novo fato. A probabilidade condicional para a nova evidência é:

$$P(H|e^n \wedge e) = P(H \wedge e^n \wedge e) / P(e^n \wedge e) = (P(e^n) \cdot P(H|e^n) \cdot P(e|e^n \wedge H)) / ((P(e^n) \cdot P(e|e^n))$$

resultando em: P (H l e^n A e) = P(HI e^n).((P(el e^n A H)/P(el e^n))

O resultado acima mostra que uma vez calculada a probabilidade condicional da hipótese H dado o conjunto e^n de evidências, isto é o valor P(H|e^n), os dados passados e^n podem ser desprezados e assim pode ser obtido o impacto da nova evidência. A crença velha (H|e^n) assume o papel de crença *a priori* no cálculo do impacto da nova evidência; a probabilidade P(H|e^n) sumariza completamente a experiência passada e para sua atualização necessita somente ser multiplicada pela ***LIKELIHOOD ratio*** P(el e^n A H). Esta razão mede a probabilidade do novo evento e considerando a hipótese H e os dados passados e^n.

O comportamento de uma rede de computadores pode ser considerado como um processo estocástico, seus estados e sua evolução podem ser modelados utilizando a teoria da probabilidade. Portanto, torna-se relevante investigar a adequação do enfoque bayesiano para desenvolver o sistema especialista de verificação [NAS 98].

3. *Knowledge Data Discovery* (KDD)

KDD não é urna nova técnica mas sim um conjunto de tecnologias que envolvem aprendizado de máquina, estatística, tecnologias de bancos de dados, sistemas especialistas e visualização de dados. No processo de KDD o *Data Mining* (DM)[FAY 96] representa 20 % os outros 80 % são atribuídos à preparação dos dados.

Utilizando técnicas de DM [KNO 97] mostrou que o conhecimento adquirido contribui para o facilitar o trabalho de gerência de redes.

Várias definições são aplicadas para KDD: KDD é o processo não trivial de identificar válidos, novos, potencialmente úteis, e ultimamente compreensíveis, padrões nos dados, dada por [FRA 91]. Para [FAY 96], KDD é o processo de, usando métodos (algoritmos) de mineração de dados, extrair (identificar) de acordo com o que é julgado conhecimento pelas métricas e saídas esperadas, usando uma base de dados onde é requerido algum preprocessamento, subexemplificação e transformações.

práticas estão sendo desenvolvidas, por exemplo, para diagnóstico e prognóstico médico e para recuperação de informação probabilística [LIN 96].

Nos sistemas especialistas bayesianos os valores de probabilidade refletem a crença do especialista sobre o que espera que ocorra em situações similares às que tem experienciado e aprendido. A idéia de que as probabilidades se alteram com a mudança de conhecimento é crucial para estes sistemas. Eles têm em sua base de conhecimentos fatos e regras que representam o conhecimento do especialista num domínio de aplicação. Aos fatos e às regras são associadas às incertezas presentes no domínio, e é explicitada a crença em sua ocorrência através de valores de probabilidade. O raciocínio realizado pelo sistema deve considerar estas probabilidades para associar o vetor de probabilidades ao conjunto de hipóteses diagnosticas. A hipótese com maior probabilidade de ocorrência pode ser considerada a conclusão do sistema, note que esta conclusão está associada ao grau de certeza da resposta do sistema.

O teorema de Bayes é um método quantitativo para a revisão de probabilidades conhecidas, com base em nova informação amostrai. No processo de tomada de decisão significa calcular uma probabilidade pela aplicação de um teste diagnóstico (probabilidade *a posteriori),* considerando uma probabilidade já disponível (probabilidade *a priori).* O conceito de probabilidade condicional permite considerar as novas informações de forma a obter as novas probabilidades.

Exemplificando: sejam A e B eventos compostos de um espaço de probabilidades (e ,P). Suponha que um evento simples e ocorra. A probabilidade P(B) é a probabilidade de que e e B, dado o conhecimento inicial refletido por P. Intuitivamente, P(BIA) é a probabilidade que e ∈ B quando se tem a informação adicional de que e e A. Seja (e, P) um espaço de probabilidade e seja A ç e tal que P(A) ≠ O . Define-se o espaço de probabilidade (e, f) da seguinte forma: se e e A então (e, f) = P(e) / P(A); se e∉A então (e, f) = f(e) = 0. Para qualquer B c e a probabilidade condicional de B dado a ocorrência de A é igual a f(B). Observe que neste caso A é o novo espaço de probabilidade, onde B deve ser analisado. Se A = e então P(BIA) = P(B).

Sejam e, H_1, H_2, ... ,H_K c e eventos compostos, desde que nenhum desses eventos tenha probabilidade nula, então:

$$P(H_i|e) = \frac{P(e|H_i) \cdot P(H_i)}{P(e)}$$

Se P (Hj A e) ≠ O para todo i então:

$$\frac{P(H_i|e)}{P(Hj|e)} = \frac{P(H_i)}{P(H_j)} \cdot \frac{P(e|H_i)}{P(e|H_j)}$$

Se os eventos H_1 u fyu ...u H_k = e e $H_i \cap H_j = 0$ para todo i ≠ j então:

$$P(e) = P(H_1) \cdot P(e|H_1) + P(H_2) \cdot P(e|H_2) + \ldots + P(H_k) \cdot P(e|H_k)$$

Resultando: $$P(H_i|e) = \frac{P(e|H_i) \cdot P(H_i)}{\sum_{j=1}^{k} (P(H_j) \cdot P(e|H_j))}$$

1. Introdução

Uma rede de computadores pode existir sem mecanismos de gerenciamento, todavia seu uso pode encontrar dificuldades com congestionamento, segurança, roteamento, etc. [ROC 97].O gerenciamento é usado para controlar as atividades e monitorar os recursos da rede. Simplificando, o trabalho básico da gerência de rede é obter informação extraída de dados, para um possível diagnóstico e executar ações para resolver os problemas. Para alcançar estes objetivos, as funções de gerenciamento devem estar contidas em diversos componentes da rede permitindo o diagnóstico, a reação e a prevenção dos problemas [WES 91,WES 96].

Existem basicamente quatro linhas de estudos na representação do conhecimento incerto [LAU 88]. O modelo lógico, utiliza-se apenas de processamento simbólico [COH 85] enquanto o modelo lingüístico baseia-se no raciocínio *fuzzy* para interpretar sentenças imprecisas da linguagem natural [ZAD 83]. A teoria de Dempster-Shafer [SHA 76] representa o conhecimento incerto através das funções de crença e o modelo estatístico está baseado no cálculo de probabilidades, o que lhe dá consistência e confiabilidade [LIN 82] possuindo um forte apelo pragmático, já que este modelo possui flexibilidade e meios operacionais de avaliação crítica e aprendizado de dados [CHE 85], além de prover uma metodologia adequada à compreensão humana.

A estatística Bayesiana passou a ser aplicada em sistemas especialistas sendo uma teoria consistente e que permite a representação de conhecimentos certos e incertos. A maior dificuldade encontrada foi o esforço computacional exigido, pois no cálculo das distribuições de probabilidade há uma explosão combinatória. Mas, quando é explorada a esparsidade das relações entre as variáveis, este esforço computacional é reduzido [PEA 88].

Abordando a gerência de redes pela visão estatística, o presente trabalho implementa uma rede bayesiana de conhecimento, a qual é capaz de reconhecer os relacionamentos entre os valores que representam o tráfego da rede e é capaz de estimar o vetor de probabilidades de diferentes estados de comportamento da rede. Em acréscimo, a rede bayesiana pode também ser utilizada como uma *baseline* dinâmica para sistemas especialistas proativos. O resultado deste trabalho consiste em um mecanismo que tem como objetivo oferecer suporte ao administrador no processo de tomada de decisão.

A seção 2 deste trabalho aborda a Estatística Bayesiana, a seção 3 trata sobre *Knowledge Data Discovery*, a seção 4 apresenta o domínio da aplicação implementada, a seção 5 mostra como foram coletados os dados, a preparação dos dados e o processo de *Knowledge Data Discovery* (KDD) são vistos na seção 6, a rede bayesiana implementada é vista na seção 7, seguida das conclusões e das perspectivas futuras.

t

2. Probabilidade Bayesiana

Após a metade da década de 80, a pesquisa sobre raciocínio probabilístico em sistemas especialistas resultou na introdução das Redes Bayesianas, também chamadas de Redes Causais. Estas redes têm sua origem na teoria da probabilidade e são caracterizadas por um poderoso formalismo que representa o conhecimento no domínio e pelas incertezas associadas a este domínio. Mais especificamente, o formalismo proporciona uma representação concisa de uma distribuição conjunta de probabilidades em um grupo de variáveis. Associados a este formalismo estão os algoritmos para calcular eficientemente as probabilidades relevantes e para processar as evidências; estes algoritmos constituem os blocos básicos para o raciocínio com o conhecimento. Desde sua introdução, a estrutura de redes bayesianas vem ganhando popularidade e está começando a mostrar o seu valor em domínios complexos. Aplicações

[NET 98] NETO, F.W. "Aplicando a Técnica de Séries Temporais em Gerenciamento Pró-Ativo de Redes de Computadores". Anais do Simpósio Brasileiro de Redes de Computadores. Rio de Janeiro (RJ). Maio de 1998.

[PEA 88] PEARL, J. "Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference". San Mateo, Calif.: Morgan Kaufmann, 1988.

[RCT 99] RCT - Rede de Ciência e Tecnologia de Santa Catarina. URL: http://www.pop-ufsc.rct-sc.br, janeiro de 1999.

[RNP 99] RNP - Rede Nacional de Pesquisa. URL: http://www.rnp.br, janeiro de 1999.

[ROC 97] ROCHA, M.A.; WESTPHALL, C.B. "Proactive Management of Computer Networks using Artificial Intelligence Agents and Techniques". Proceedings of the Symposium on Integrated Network Management. San Diego (CA), USA. May, 1997.

[SHA 76] SHAFER, G. "A mathematical theory of evidence". Princeton, Princeton University Press. 1976.

[WES 96] WESTPHALL, C.B.; KORMANN, L.F. "Usage of the TMN Comcepts for Configuration Management of ATM Network". International Symposium on Advanced Imaging and NetWork Technologies. Berlim, Alemanha Out. 7-11, 1996.

[WES 91] WESTPHALL, C.B. "Conception et développement de l'architecture d'administration d'un réseau métropolitain". Thèse de doctorat nouveau regime. L'université Paul Sabatier. Toulouse, le 16 juillet 1991.

[ZAD 83] ZADEH, L. A. "The role of fuzzy logic in the management of uncertainty in expert systems". Fuzzy Sets and Systems. (11): 199-228. 1983.