Descoberta de Conhecimento em uma Base de Dados de Bilhetes de Tarifação: Estudo de Caso em Telefonia Celular

Elionai Sobrinho[1,3], Jasmine Araújo[1,3], Luiz A. Guedes[2], Renato Francês[1]

[1]Departamento de Engenharia Elétrica e da Computação – Universidade Federal do Pará (UFPA) Caixa Postal 479 - 66.075-110 – Belém – PA – Brasil; Fone: +55 91 31831302, Fax: +55 91 31831634

[2]Departamento de Engenharia da Computação – Universidade Federal do Rio Grande do Norte (UFRN) Caixa Postal 1524 - 59.072-970– Natal – RN – Brasil, Fone: +55 84 32153771, Fax: +55 84 32153738

[3]Coordenação de Engenharia da Computação – Instituto de Estudos Superiores da Amazônia (IESAM) CEP – 66.055-260 – Belém – PA – Brasil, Fone: +55 91 40055400, Fax: +55 91 40055407

elionai@deec.ufpa.br, jasmine@deec.ufpa.br, affonso@dca.ufrn.br, rfrances@ufpa.br

Resumo: Nas grandes operadoras de serviços telecomunicações, há a necessidade de criação de técnicas automáticas para descoberta de conhecimento na enorme e complexa quantidade de dados gerados e armazenados através de suas operações diárias. Esse artigo apresenta a utilização do Modelo Bayesiano para inferência sobre o comportamento de uma rede de telecomunicações usando uma grande base de dados de registros detalhados de chamadas (CDR - Call Detailed Record). **Palavras-Chaves**: Desempenho da Rede Celular, Mineração de Dados, Registro Detalhado de Chamadas, Modelo Bayesiano, Descoberta de Conhecimento em Base de Dados.

Introdução

O manuseio eficiente de grandes bases de dados sempre foram um dos grandes desafios da empresas que lidam com cadastros de seus clientes, uma vez que suas transações diárias geram milhares de registros em suas bases de dados formando, ao longo do tempo, uma vasta quantidade de informação que, manipulada de forma adequada, poderá prover valiosas informações. Uma das técnicas utilizadas para extrair tais informações destas bases é a Mineração de Dados, técnica que pode ser agregada ao processo de Extração de Conhecimento de Bases de Dados (KDD – Knowledge Discovery in Database).

O foco deste estudo de caso é utilizar o método de mineração de dados em uma base de dados de bilhetes de tarifação de uma operadora de telefonia móvel celular, dados este que são basicamente compostos de registros sobre as chamadas realizadas na rede as quais são registradas nos bilhetes de tarifação conhecidos como CDR: Call Detailed Record (Registro Detalhado de Chamada).

Estes dados são comumente utilizados para obtenção de conhecimento sobre o modo de funcionamento do tráfego e comportamento de assinantes a partir das inferências e predições realizadas, além de prover meios confiáveis para auxiliar os analistas de tráfego, marketing

e rede no melhor processo de tomada de decisão [4]

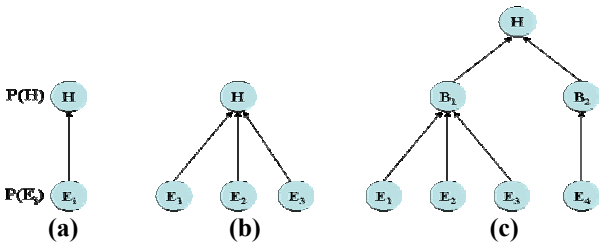
Uma Visão Geral Sobre Redes Bayesianas

As Redes Bayesianas são utilizadas para modelar situações (exemplo: diagnóstico médico) nas quais a causalidade tem sua função, no entanto, o entendimento do que está realmente acontecendo está incompleto[3].

As Redes Bayesianas são apontadas como classificadores estatísticos, que codificam os relacionamentos probabilísticos entre as variáveis que representam um determinado domínio, utilizando em seus cálculos as fórmulas de probabilidade condicional e condicional conjunta do Teorema de Bayes (1).

$$P(A|B)=\frac{P(B|A)P(A)}{P(B)}\ P(A|B)=\frac{P(A|B,E)P(A|E)}{P(B|E)}\ (1)$$

Uma Rede Bayesiana é caracterizada por um grafo, composto por vários nós, onde cada nó da rede irá representar uma variável aleatória, ou seja, um atributo da base de dados; setas ligando-os e cuja direção implicará na dependência que um possa ter sobre outro e tabelas de probabilidade para cada um dos nós.



P(H) – Probabilidade da Hipótese
$P(E_i)$ – Probabilidade da Evidência

Figura 1. Estruturas de Redes Bayesianas. (a)Evidência Simples, (b)Múltiplas Evidências e (c)Múltiplas Camadas.

A partir da observação das relações de dependência entre as variáveis aleatórias (os nós da rede), juntamente com alguns dados *a priori* dessas variáveis pode-se calcular eficientemente as probabilidades *a posteriori* de qualquer variável aleatória (através das chamadas Inferências Bayesianas) usando uma definição recursiva do Teorema de Bayes. Esses relacionamentos probabilísticos entre as variáveis podem ser de evidência simples, evidência múltipla e múltiplas camadas. A Figura 1 ilustra uma estrutura básica de uma Rede Bayesiana.

Em implementações práticas, o julgamento e análise de especialistas sobre as probabilidades dos nós da Rede Bayesiana tem se mostrado freqüentemente consistentes [1].

Neste trabalho, será utilizado o software Bayesware Discoverer para implementação da Rede Bayesiana.

A Base de Dados de Chamadas da Rede Móvel Celular

A cada chamada realizada pelos assinantes equivale a geração de um ou mais registros conhecidos por CDR (Call Datailed Record). Para o serviço de voz, alguns campos estão presentes nestes registros [7]:

- Número de Assinante Chamador;
- Número de Assinante Chamado;
- Rota de entrada da chamada;
- Rota de saída da chamada;
- Hora de início da chamada;
- Data de início da chamada;
- Duração da conversação;
- Categoria do Assinante A;
- Categoria do Assinante B e
- Fim de Seleção.

Para este trabalho, todos os tipos de chamadas foram armazenados, não somente as atendidas, mas também as mal sucedidas (não atendidas). A Tabela 1 fornece uma visão resumida dos

eventos associados a esses dois tipos de chamadas.

Tabela 1: Classificação de Chamadas Segundo o Evento Ocorrido.

Call Type	Descrição
Chamada Bem Sucedida	Chamada Atendida
	Queda de Chamada (após o atendimento)
Chamada Mal Sucedida	Assinante Não Responde
	Assinante Ocupado
	Assinante For a de Área ou Desligado
	Congestionamento
	Falha de Equipamento
	Desconexão
	Outros (Nenhum dos itens anteriores)

Preparação da Base de Dados de CDR

Uma das etapas para aplicação do processo de descoberta de conhecimento em base de dados é o passo de preparação desta base, o que equivale à eliminação de redundâncias, normalização e criação de faixas de valores para os campos que possuem valores numéricos. Vale ressaltar que para o referido estudo foram analisadas apenas as chamadas originadas na rede celular durante uma (01) hora, o que gerou após a preparação cerca de 48 mil registros. Os campos utilizados e as suas descrições, normalizações e criações de faixas numéricas, quando foi o caso, são descritos a seguir:

1-Status Call: Este campo fornece a informação necessária sobre o estado da chamada, isto é, se a chamada foi completada (houve atendimento) ou não, sendo que em caso negativo, ainda é possível saber quais os motivos que levaram ao não completamento. Para o estudo em questão, este campo basicamente assumirá dois valores:

- **Ok**: Chamada Completada e
- **Nok**: Chamada não Completada.

2-Call Type: Apresenta uma característica da ligação que poderá ser uma chamada normal ou uma ligação a cobrar. Este campo na base de estudo poderá ter os seguintes valores:

- **Normal**: Chamada Normal e
- **Collect**: Chamada a Cobrar.

3-Call Duration: Este campo possui informação sobre o tempo de conversação (em segundos) de cada chamada atendida. Como pode assumir muitos valores numéricos e esta diversidade poderá dificultar a análise, então foram atribuídas faixas de valores conforme a seguir:

- **Null**: 0 (zero);
- **Very_Short**: 1 a 10 segundos;
- **Short**: 10 a 60 segundos;
- **Normal**: 60 a 80 segundos;
- **Long**: 80 a 300 segundos e
- **Very_Long**: > 300 segundos.

4-Record Type: Este campo diz respeito ao tipo de registro da chamada, isto é, se é uma chamada do tipo Móvel-Móvel (MM) ou Móvel -Fixo (ML):

- **MM**: Chamada Móvel-Móvel e
- **ML**: Chamada Móvel-Fixo.

5-CodA e 6-CodB: Nomenclatura utilizada para identificar os números dos assinantes A e B respectivamente. Poderá assumir os valores:

- **Pre-paid:** Números relativos a assinantes Pré-pagos;
- **Pos-paid:** Números relativos a assinantes Pós-pagos e
- **Land:** Números relativos a assinantes da Rede Fixa.

Estes seis campos serão, portanto, as variáveis aleatórias da Rede Bayesiana.

Após a submissão da Base de Dados ao software Bayesware Discoverer, há um período de treinamento onde o software identifica o grau de depedência entre as variáveis (campos) fazendo a efetiva montagem da rede bayesiana. A Figura 2 mostra a rede resultante através de seus nós (campos) além dos relacionamentos ou grau de dependência identificados (setas):

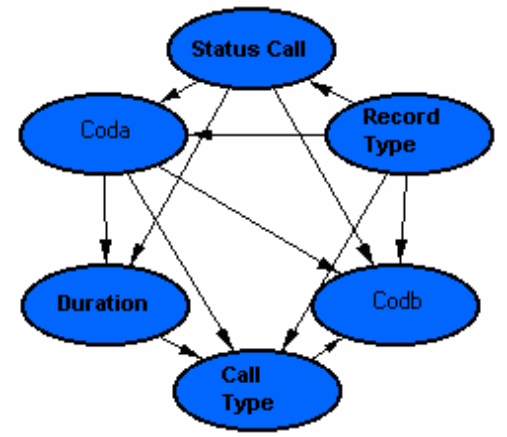


Figura 2: Rede Bayesiana criada para Base de Dados de Chamadas.

Resultados Obtidos

Após a criação da Rede Bayesiana, é gerada então uma tabela de probabilidades para cada nó. Estas probabilidades são obtidas a partir dos valores de cada campo da base de dados utilizada e são chamadas de probabilidades *a priori*. Neste ponto pode-se iniciar a verificação das inferências sobre a rede, criando-se situações (hipóteses) de modo a identificar o comportamento da mesma, gerando as probabilidades *a posteriori* para as situações criadas.

Algumas probabilidades *a priori* (probabilidades de evidências), encontradas são exibidas na Tabela 2:

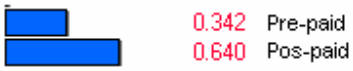
Tabela 2: Resultados *a priori*

Campos da Base de Dados	**Resultados *A Priori***
Call Type	 0.212 Collect 0.788 Normal 21,2 % das chamadas originadas são chamadas a cobrar e 78% são chamadas normais. Independentemente de serem chamadas atendidas ou não.
Call Duration	 0.517 Null 0.075 Very_Short 0.135 Short 0.169 Normal 0.067 Long 0.037 Very_Long Distribuição de probabilidade da duração das chamadas de acordo com a convenção utilizada
Status Call	 0.508 Nok 0.492 Ok 50,8% das chamadas originadas foram mal sucedidas, enquanto 49,2% foram bem sucedidas (atendidas)
Record Type	 0.187 MM 0.813 ML Das chamadas originadas, 18,7% são destinadas aos outros assinantes móveis, enquanto 81,3% são destinadas à rede de telefonia fixa.

A partir das inferências bayesianas realizadas sobre a rede da Figura 2, pôde-se obter diversos resultados conhecidos como resultados *a posteriori* ou probabilidades das hipóteses

inferidas sobre a rede. Como exemplo, a Tabela 3 mostra alguns resultados das probabilidades *a posteriori* produzidos:

Tabela 3: Inferências *a posteriori*

Hipótese	**Resultado *a posteriori***
Origem qualquer e chamadas completadas	**Distribuição da Originação**  **Distribuição do tipo de ligaçao**  Verifica-se que 34% das originações são efetuadas por pré-pagos e 64% por pós-pagos. A grande maioria (85%) são ligações normais e 15% são do tipo DDC
Origem pós-pago e chamadas completadas	**Distribuição do tipo de ligação**  Este resultado mostra que de todas as ligações originadas por pós-pagos, este é responsável por apenas 1% delas.

Conclusões

Como foi apresentado, as Redes Bayesianas, possuem suas particularidades e métodos para abordar e obter as devidas respostas e informações do sistema, demonstrando uma maior praticidade e flexibilidade para a realização das análises, não afetando no desempenho e integridade dos dados e resultados obtidos [5].

As etapas de preparação adequada das bases de dados foram de fundamental importância para realização coerente da análise dos resultados observados. Permitindo ao analista fazer, de forma rápida e clara, inferências acerca do comportamento dos parâmetros envolvidos a partir de hipóteses geradas artificialmente, possibilitando o isolamento de problemas ou falhas que não sejam tão importantes para o resultado final da análise em questão [2].

Referências:

[1] Chien, Chen-Fu et. All, (2002) "Using Bayesian Network for Fault Location on Distribuition Feeder".

[2] G. Box, G. Tiao, (1992) "Bayesian Inference in Statistical Analysis".

[3] Fayyad, U. Piatetsky-Shapiro, G.; Smyth, P., (1996) "The KDD Process for Extracting Useful Knowledge from Volumes of Data Communication of the ACM", vol. 39, nº 11, p. 27-34, November.

[4] Han, J.; Kamber. M. (2000) "Data Mining: Concepts and Techniques". Morgan Kaufmann Publ.

[5] Ramoni, M.; Sebastiani, P.(1997) "Discovering Bayesian Networks in Incomplete Databases". Knowledge Media Institute, The Open University, Technical Report, nº 46.

[6] Turban, E.; Aronson, J. E. (2001) "Decision Support Systems and Intelligent Systems". 6ª Ed. Prentice-Hall.

[7] ITUT Recommendation E.502. Traffic measurement requirements for digital telecommunication exchanges; International Telecommunication Union-Telecommunication Standardization Sector; 1992.