Um Retrato da Aplicação de Recursos da Saúde e seu Impacto no IDH dos Municípios do Estado do Pará

Ádamo Lima de Santana[1], Carlos Renato L. Francês[1], Paulo de Tarso[2], João Crisóstomo W. A. Costa[1], Patrícia T. Endo[1], Aldebaro Klautau Jr.[1]

[1]Laboratório de Computação Aplicada (LACA) - Universidade Federal do Pará (UFPA)
Caixa Postal 479 - 66.075-110 - Belém - PA – Brasil

[2]Fundação Nacional de Saúde - Av. Visconde de Souza Franco, 616 - Reduto
CEP 66.053-000 - Belém- PA – Brasil

adamo@deec.ufpa.br, rfrances@ufpa.br, pptarso@funasa-pa.gov.br,
jweil@ufpa.br, endo@deec.ufpa.br, a.klautau@ieee.org

Abstract. *In the current state-of-art, it is not more satisfactory to get only the historical information from databases. It is important to make use of information a priori, that is, to make possible to carry out prospections from data of system. This work presents a strategy to investigate the application of public resources in the health area, verifying the impact caused in the index of human development (IDH), in the context of the cities of the State of Pará.*

Resumo. *No atual estado-da-arte, não é mais satisfatório obterem-se apenas as informações históricas de bases de dados. É importante dispor de informações a priori, isto é, que possibilitem realizarem-se prospecções do sistema. Este trabalho apresenta uma estratégia para investigar a aplicação de recursos públicos na área de saúde, verificando o impacto causado no índice de desenvolvimento humano (IDH), no contexto dos municípios do Estado do Pará.*

Palavras-chave: Saúde Pública, IDH, Data Mining, Redes Bayesianas, Observatório de Saúde da Amazônia.

Este trabalho foi parcialmente financiado pela CAPES.

1. Introdução

Pela exigüidade de recursos financeiros disponíveis nos países de terceiro-mundo, é imprescindível que esses poucos recursos sejam aplicados em projetos que efetivamente possam gerar impacto na melhoria das condições de vida da população. Desta forma, torna importante a capacidade de obter-se um conjunto de informações privilegiadas, através das quais seja possível traçar estratégias de políticas públicas e, eventualmente, redirecionando esforços, corrigindo rumos e realizando ações preventivas.

Uma abordagem possível é prover sistemas de suporte à decisão par os gestores do setor de saúde, combinando a capacidade estatísticas/probabilística com o know-how dos especialistas do domínio da saúde. Obviamente, a avaliação deverá ser realizada em função de um ou mais índices entendidos como plausíveis, para a verificação da eficácia das informações obtidas.

Este artigo apresenta uma estratégia implementada para uma instituição recente da Fundação Nacional de Saúde, chamada de "Observatório de Saúde da Amazônia - OSA" (Seção Pará). A idéia do Observatório é realizar a integração de diversas informações existentes em bases de dados isoladas, que contemplem informações sobre a Região Amazônica, utilizando-se a estrutura de comunicação e de geo-referenciamento do Sistema de Proteção da Amazônia (SIPAM), antigo Sistema de Vigilância da Amazônia (SIVAM).

Este trabalho apresenta uma parte do estudo pretendido pelo Observatório, abordando o impacto do bom/mau uso de recursos públicos nos índices/contas sociais, dentre os quais, destaca-se o Índice de Desenvolvimento Humano (IDH), dentro dos municípios do Estado do Pará.

Para realizar a tarefa, está-se utilizando a abordagem de "descoberta de conhecimento", baseada em redes bayesianas. Desta forma, este artigo apresenta a seguinte estrutura: a Seção 2 apresenta as particularidades da infra-estrutura do Observatório de Saúde da Amazônia. Na Seção 3 é apresentado o processo de extração de conhecimento, com ênfase em redes bayesianas. A Seção 4 mostra a aplicação de KDD (Knowledge Discovery in Database) na área de Saúde Pública, incluindo uma análise particular sobre as informações referentes ao estado do Pará. A Seção 5 faz referências sobre o andamento desse estudo e os seus passos futuros. A Seção 6 lista algumas considerações sobre o Observatório de Saúde da Amazônia e alguns passos à frente.

2. Infra-estrutura Utilizada

O Observatório de Saúde da Amazônia está utilizando duas abordagens para prover informações e extrair conhecimento sobre parâmetros de saúde da região amazônica:

- Utilização das bases de dados já disponíveis nas diversas instituições integrantes do projeto: ADA/SUDAM (Agência de Desenvolvimento da Amazônia/Superintendência de Desenvolvimento da Amazônia), SIPAM, Museu Emílio Goeldi, FUNASA (Fundação Nacional de Saúde), DATASUS (Departamento de Informática do Sistema Único de Saúde), SESPA (Secretaria Executiva de Saúde Pública do Pará), por exemplo. Nesta estratégia, atualmente, está sendo realizado o processo de convergência das diversas bases citadas para uma

base única, na qual será aplicado o processo de data mining;

- Aquisição de novos dados, utilizando-se a infra-estrutura de comunicação disponível nas diversas instituições participantes.

O SIPAM tem uma infra-estrutura comum e integrada de meios técnicos destinados à aquisição e tratamento de dados e para a visualização e difusão de imagens, mapas, previsões e outras informações. Esses meios abrangem o sensoriamento remoto, a monitoração ambiental e meteorológica, a exploração de comunicações, a vigilância por radares, recursos computacionais e meios de telecomunicações.

O SIPAM conta, em sua atual estrutura, com os seguintes subsistemas: Subsistema de Aquisição de Dados; Subsistema de Tratamento e Visualização de Dados; Subsistema de Telecomunicações; Subsistema de Suporte de Transmissão; Subsistema de Auxílio à Navegação Aérea. Dentre os quais, o Observatório utiliza com maior intensidade os subsistemas Aquisição de Dados e de Telecomunicações. Sucintamente, o subsistema de Aquisição de Dados é composto pelos recursos técnicos necessários para obter os dados necessários à geração de informações atualizadas e confiáveis, utilizando os seguintes recursos: Sensoriamento Remoto por Satélite; Sensoriamento Aéreo; Sensores Meteorológicos; Plataformas de Coleta de Dados; Radiodeterminação; Rede de Exploração de Comunicações; Rede de Detecção Radar. Já o Subsistema de Telecomunicações do SIVAM é constituído por Redes e Serviços que permitem a veiculação de vários tipos de mensagens operacionais, técnicas e administrativas entre os diversos Órgãos Usuários, dentre os quais se inclui o OSA. Este subsistema utiliza os meios do Suporte de Transmissão para interligação dos usuários, compondo suas Redes e Serviços.

Além da infra-estrutura já instituída no SIPAM, há uma série de outras estruturas de coleta e comunicação de dados, disponíveis em localidades não assistidas pelo SIPAM ou atuando de forma complementar, que também estão sendo utilizadas no OSA, podendo-se citar: a rede de acesso do DATASUS, com diversos pontos de presença no Pará e na Amazônia; e a rede da SESPA, com dimensão estadual, mas que possui uma grande capilaridade.

3. Extração de Conhecimento de Bases de Dados

A transformação de dados em conhecimento tem utilizado métodos eminentemente manuais para análise e interpretação de dados, o que torna o processo de extração de padrões em bases de dados caro, lento e altamente subjetivo, além de inviável em se tratando de grandes volumes de dados (Rocha e Rezende, 1999).

O interesse em solucionar esse problema tem fomentado várias pesquisas em um campo emergente chamado Extração de Conhecimento de Bases de Dados (KDD - Knowledge Discovery in Database) (Fayyad et al., 1996). O processo KDD desponta como uma tecnologia capaz de cooperar amplamente na busca do conhecimento embutido nos dados. Desta forma, o principal objetivo deste processo é encontrar padrões válidos e potencialmente úteis nos dados. Como ressalta (Turban, 2001), as tecnologias computacionais como KDD estão sendo cada vez mais utilizadas para auxiliar analistas em seus trabalho de investigação e análise de enormes conjuntos de dados, a fim de se extrair conhecimento.

Na literatura são encontradas várias nomenclaturas para o processo de KDD, entre elas destacam-se: Data Mining, processamento de padrões de dados, arqueologia de dados, coleta de informações, garimpagem de dados e siftware (Mannila, 1997). Apesar de alguns pesquisadores tratarem KDD e Data Mining de modo indiscriminado, este trabalho, da mesma forma como salienta (Fayyad et al., 1996), Data Mining pode ser visto como parte de um processo maior de extração de conhecimento de bases de dados, ou seja, como uma etapa do KDD.

3.1. Etapas do Processo de Extração de Conhecimento de Bases de Dados

A extração de conhecimento a partir de dados pode ser entendida como um processo contendo, pelo menos, os seguintes passos:

- Compreensão do domínio da aplicação.
- Seleção e preparação dos dados.
- Data Mining.
- Avaliação do conhecimento extraído.
- Consolidação e utilização do conhecimento extraído.

Um esquema representativo contendo todos esses passos é ilustrado na Figura 1. O processo KDD inicia com o entendimento do domínio da aplicação, considerando aspectos como os objetivos dessa aplicação e as fontes de dados (bases de dados da qual se pretende extrair conhecimento). Em seguida, é realizada uma seleção de dados a partir dessas fontes, de acordo com os objetivos da aplicação do processo KDD. Os conjuntos de dados resultantes dessa seleção são, então, pré-processados (efetuando, por exemplo, o tratamento dos dados ausentes e de ruídos) e submetidos aos métodos e ferramentas da etapa de Data Mining com o objetivo de encontrar padrões/modelos (conhecimento) a partir dos dados. Após esta etapa, esse conhecimento é avaliado quanto a sua qualidade e/ou utilidade para que, em caso positivo, seja utilizado para apoio a algum processo de tomada de decisão.

Vale acentuar que, apesar das ferramentas de visualização serem mais utilizadas na etapa de avaliação do conhecimento extraído, elas são de grande relevância para facilitar o entendimento e avaliação, principalmente do usuário final, dos resultados de cada etapa, conforme pode ser observado na parte inferior da Figura 1.

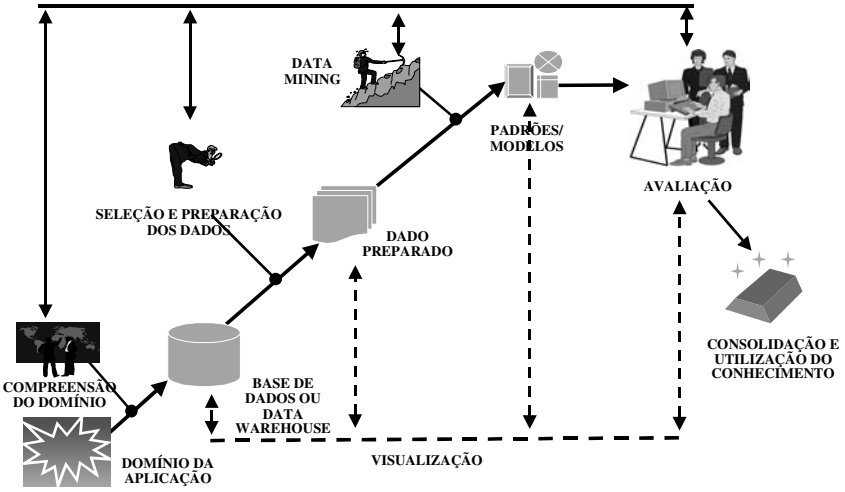

Figura 1. Etapas do Processo de Extração de Conhecimento de Bases de Dados.

Por representar o núcleo do processo de KDD, em razão do uso de métodos e algoritmos que possibilitem a extração de conhecimentos a partir de uma base de dados, a etapa de Data Mining será focalizada a seguir.

3.2. Data Mining

A etapa de Data Mining(DM) envolve a criação de modelos apropriados de representação dos padrões e relações identificados a partir dos dados. O resultado desses modelos, depois de avaliados pelo analista, especialista e/ou usuário final, são empregados para predizer os valores de atributos definidos pelo usuário final baseados em novos dados (Fayyad et al, 1996).

Para a Mineração de dados podem ser utilizados diversos algoritmos que fazem parte dos paradigmas de aprendizado disponíveis na literatura, dentre eles podem ser destacados os seguintes: simbólico, conexionista, genético e estatístico.

- Simbólico: aprendem construindo representações simbólicas de um conceito através da análise de exemplos e contra-exemplos desse conceito. Como exemplo, destacam-se as regras de produção do tipo “se-então” e Árvores de Decisão.
- Conexionista: formado pelas Redes Neurais, que são construções matemáticas que utilizam o mecanismo de paralelismo, no qual são conectados pequenas unidades de processamento (neurônios) ligadas em rede.
- Genético: caracterizado por um classificador genético, onde a partir de uma população, seus elementos competem entre si para realizar a predição, onde os elementos que possuem um desempenho fraco são descartados, enquanto os elementos mais fortes proliferam, produzindo variações de si mesmos.
- Estatístico: os classificadores estatísticos freqüentemente assumem que valores de atributos estão normalmente distribuídos, e então podem usar os dados fornecidos para determinar média, variância e co-variância da distribuição, entre outros. É neste paradigma que estão classificadas as Redes Bayesianas, que serão utilizadas neste trabalho.

3.3. Redes Bayesianas

As Redes Bayesianas podem ser entendidas como modelos que codificam os relacionamentos probabilísticos entre as variáveis que representam um determinado domínio (Russel e Norvig, 1995). Esses modelos possuem como componentes uma estrutura qualitativa, representando as dependências entre os nós, e quantitativa (tabelas de probabilidades condicionais (TPCs) desses nós), avaliando, em termos probabilísticos, essas dependências (Rezende e Rocha, 2000).

Na figura 2, é apresentado um esquema representativo desses componentes. Na qual, A, B, C, D e E são os nós que representam as variáveis do domínio de aplicação, a tabela representa a TPC do nó A, e P1, P2, P3 e P4 são as probabilidades de ocorrência dos valores (estados possíveis) de A, dados que os valores possíveis de B e E ocorrem. Em razão dos estados serem coletivamente exaustivos (a soma deve ser igual a 1), as probabilidades da não ocorrência de A, dados os valores de B e E são complementares (1-P1, 1-P2, etc.). Na Figura 2, é apresentada apenas uma tabela como exemplo, mas todos os nós possuem sua própria TPC.

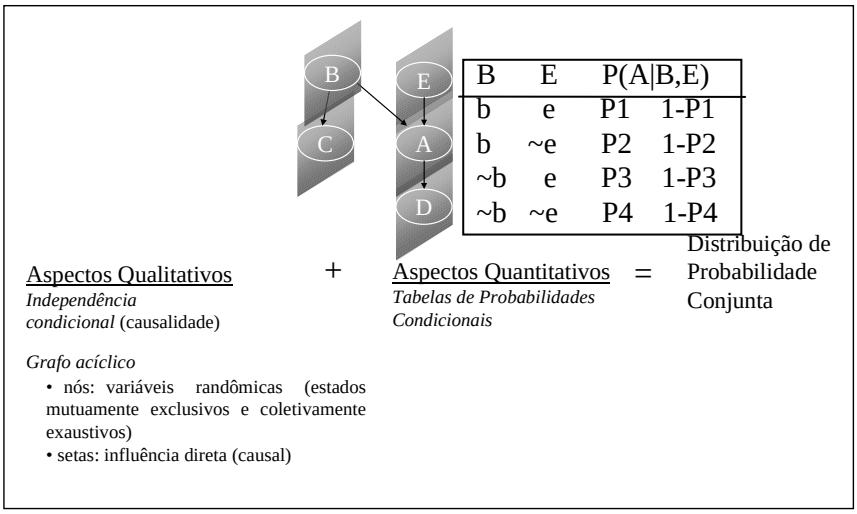

Figura 2. Aspectos qualitativo e quantitativo das redes Bayesianas.

Os componentes representados na Figura 2 propiciam uma representação eficiente da distribuição de probabilidade conjunta (DPC) do conjunto de variáveis X de um determinado domínio (Heckerman, 1997). A distribuição conjunta é dada pela seguinte equação:

$$P(X) = \prod_{i=n}^{n} P(X_i \mid Pa_i)$$

na qual Pai são os nós-pais do nó Xi. Essa representação acarreta uma redução substancial do número de probabilidades a serem manipuladas. Por exemplo, em uma rede com n nós (booleanos), no qual nenhum desses nós possui mais que k nós-pais, a ordem do número de valores de probabilidade a serem tratados pela rede Bayesiana seria O(2kn), ao invés de O(2n) da DPC.

É importante destacar que os componentes da semântica das redes Bayesianas facilitam, dada a inerente representação causal dessas redes, o entendimento e o processo de tomada de decisão, por parte dos usuários desses modelos. Isto se deve,

basicamente, ao fato das relações entre as variáveis do domínio poderem ser visualizadas graficamente, além da quantificação, em termos probabilísticos, dos efeitos dessas relações (Rocha e Rezende, 1999).

Dessa forma, é possível efetuar análises de causa e efeito a partir do cálculo da distribuição de probabilidade condicional de um conjunto de variáveis, dado o valor de um conjunto de variáveis de evidências. Por exemplo, a partir da rede Bayesiana representada na Figura 2, é possível calcular a probabilidade da ocorrência de D, dados os valore de B e E, ou ainda, a distribuição probabilística para qualquer subconjunto de variáveis, dado os valores ou distribuição de qualquer outro subconjunto das variáveis restantes. O cálculo de probabilidades de interesse em uma rede Bayesiana é geralmente referenciada como inferência probabilística.

A construção (ou aprendizado) das redes Bayesianas pode ser feita usando apenas o conhecimento prévio do domínio, os dados ou a combinação de ambos.

Os métodos de aprendizado Bayesianos a partir dos dados consideram, basicamente, dois aspectos. Primeiro, que a estrutura da rede Bayesiana (com o conhecimento prévio do domínio) pode ou não ser fornecida a priori e, segundo, que os valores dos atributos da base de dados podem ser completos ou com valores ausentes. Neste trabalho, é enfatizado o aprendizado de redes Bayesianas a partir de dados completos.

4. Extração de Conhecimento da Base de Dados

Nesta seção, serão aplicadas as etapas do processo KDD sobre a base de dados.

4.1. Seleção e Preparação dos Dados

Infelizmente não havia disponível uma base de dados completa, com todos os valores e atributos específicos e relevantes para a realização das análises, dessa forma foi necessário através de um longo pré-processamento formar uma.

Inicialmente verificados quais os dados relevantes para o problema, é hora de buscá-los e adaptá-los caso necessário.

Neste caso, os dados estavam dispersos em várias bases de dados diferentes. Partindo do atributo principal da análise, o IDH Municipal, os outros foram buscados, resultando num total de 8 atributos, como apresentados a seguir:

- IDH-M, atributo correspondendo no caso ao Índice de Desenvolvimento Humano Municipal de cada município do estado do Pará. Juntamente com o Idh-M foram pegos outros atributos correlatos: IDHM-L, IDHM-E, IDHM-R;
- IDHM-L, atributo correspondente ao Índice de longevidade existente em cada município do estado do Pará;
- IDHM-E, atributo correspondente ao Índice de educação existente em cada município do estado do Pará;
- IDHM-R, atributo correspondente ao Índice de renda existente em cada município do estado do Pará;

- Renda per capita, correspondente à Renda per capita verificada em cada município do estado do Pará;
- População, contendo o número de habitantes de cada município do estado do Pará;
- Investimentos, contendo a identificação de cada investimento aplicado em cada município do estado do Pará na área da saúde. Os investimentos existentes foram discretizados em 6 faixas de acordo com o tipo de aplicação destinado ao gasto:
    1. capacitação e cursos: referente à gastos com treinamento, e capacitação e especialização de pessoal na área da saúde;
    2. apoio e assistência: referente à gastos com suporte de comunidades e projetos vigentes;
    3. programas, projetos e eventos: gastos com implantação de novos programas, projetos e eventos relacionados com a área da saúde;
    4. reforma, ampliação e aquisições de equipamento: gastos com manutenção e melhoria de instalações existentes, assim como compra de novos equipamentos;
    5. construções de centros e postos de saúde: gastos com a construção de novos centros, postos e unidades de saúde;
    6. none: utilizado para referenciar os municípios que não foram beneficiados por investimento algum;
- Valor, contendo o valor gasto com cada um dos investimentos feitos.

Ao término desta etapa, a base de dados fora consolidada com 490 registros e os 8 atributos descritos.

De posse dos bancos de dados compostos apenas dos dados relevantes à análise, iniciou-se a etapa de Data Mining.

4.2. Data Mining

Analisando-se os atributos (verificados como sendo todos contínuos menos o atributo Investimentos, que é discreto) e suas dependências com os demais foi montada a seguinte rede bayesiana:

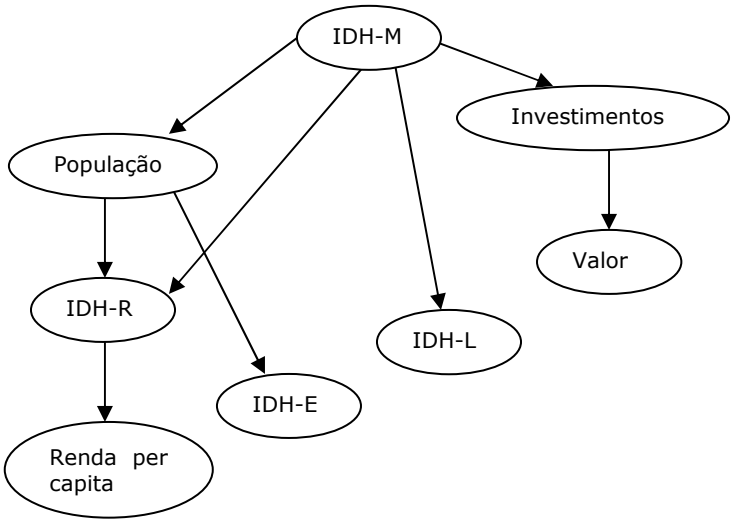

Figura 3. Redes Bayesiana criada a partir da base de dados.

Particularmente nesta etapa, foi utilizada para extração de padrões a ferramenta *Bayesware Discoverer* (http://bayesware.com), que constrói redes bayesianas a partir de uma base de dados, de modo a verificar mais facilmente de uma forma visual as probabilidades de cada nó e a propagação ou mudanças ocorridas na rede ao se realizar inferências sobre as relações entre as variáveis do domínio.

A rede bayesiana montada pela aplicação, juntamente com as probabilidades de alguns de seus nós pode ser vista a seguir:

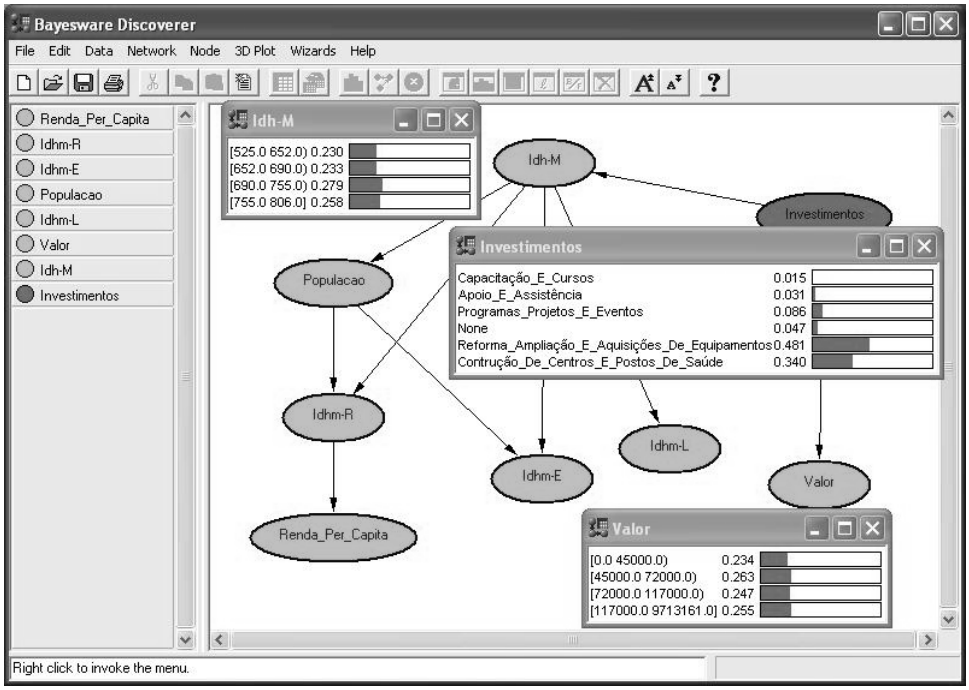

Figura 4. Rede bayesiana gerada pelo programa Bayesware Discoverer.

Face às características exploratórias das redes bayesianas, que permitem que sejam realizadas quaisquer inferências, torna-se possível efetuar uma série de análises de dependência entre os atributos.

4.3. Avaliação do Conhecimento Extraído

De posse da rede bayesiana criada, foram iniciadas as análises e inferências sobre a mesma; essa seção aponta alguns dos resultados obtidos.

Inicialmente pode ser verificado que a maior parte dos recursos investidos destinam-se à reforma, *ampliação e aquisições de equipamento*, seguido pelas *construções de centros e postos de saúde*, deixando as demais faixas com uma quantidade de recursos irregular.

Verificou-se também que um aumento no capital investido nos municípios impulsionaria o seu IDH Municipal em pelo menos 15%, em especial no que diz respeito à longevidade.

Foi também verificado que ao se evidenciar uma contínua maior aplicação dos investimentos na *ampliação e aquisições de equipamento*, ou nas *construções de centros e postos de saúde* (atualmente os pontos mais investidos), não causariam um impacto muito grande na melhoria do IDH Municipal, muito embora o investimento em *construções de centros e postos de saúde* apresentasse uma maior melhoria quando comparado com o primeiro. No entanto, um maior investimento nos demais pontos, como uma maior assistência para com as comunidades e projetos em funcionamento, assim como incentivo e implantação de novos programas e projetos, até uma melhor capacitação do pessoal e funcionários existentes impulsionaria o IDH Municipal em pelo menos 50%. Demonstrando que o investimento em áreas que agem de certa forma diretamente com a população viriam a gerar um benefício muito maior.

5. Estado Atual do Estudo e Passos Futuros

Atualmente este trabalho está sendo continuado com a implantação e utilização de outras técnicas de data mining, como Redes Neurais e SVM (Support Vector Machine), de modo a se averiguar qual técnica vem a apresentar melhores desempenhos e resultados para esta aplicação.

Além de técnicas alternativas, também estão sendo trabalhados métodos de paralelização para essas técnicas de inteligência computacional, de tal forma que possamos atingir o máximo de desempenho possível, não afetando na qualidade dos resultados, para bases de dados grandes e com enorme prospectiva de expansão.

6. Considerações Finais

O Observatório de Saúde da Amazônia está sendo atualmente implantado, dentro de uma ação maior do Plano de Desenvolvimento da Amazônia do Governo Federal. De forma quase consensual, ficou estabelecido que a prioridade seria estabelecer um mecanismo de fornecimento de informações privilegiadas para os gestores (em um nível de visão) e para a população em geral (em outro nível de visão), sendo que o segundo também visa à “prestação de contas” para a população.

O Sistema de Suporte à Decisão do OSA está sendo finalizado, em sua versão inicial. Entretanto, a idéia é criar-se um Observatório Sócio-Econômico para a Amazônia, integrando diversos Órgãos de desenvolvimento regional, tais como SIPAM e ADA.

A aplicação de técnicas de inteligência computacional no processo de aquisição de conhecimento amplia as possibilidades tradicionais dos bancos de dados, incutindo a idéia de prospecção e apresentação de tendências, baseando-se nas informações históricas.

7. Referências Bibliográficas

Box, G., Tiao, G. *Bayesian Inference in Statistical Analysys*. John Wiley and Sons, Inc. 1992.

Congdon, P. *Applied Bayesian Modelling*. John Wiley and Sons, Inc. 2003.

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

Heckerman, D. *Bayesian Networks for Data Mining*. Data Mining and Knowledge Discovery, Kluwer Academic Publishers, Vol. 1, p. 79-119, 1997.

Mannila, H. *Data Mining: Machine Learning, Statistic and Databases*. Department of Computer Science, University of Helsinki, 1997. URL: http://www.cs.helsinki.fi/~mannila/

Turban, E.; J. E. Aronson. *Decision Support Systems and Intelligent Systems*. 6ª Ed. Prentice-Hall. 2001.

Rezende, S. O, Rocha, C. A. J. *Bayesian Networks for Knowledge Discovery in a Database from the Program for Genetic Improvement of the Nelore Breed.* In: 2nd International Conference on Data Mining, 2000, Londres. Proceedings of 2nd International Conference on Data Mining. Londres: WIT Press, 2000. p.15 - 24.

Rocha, C. A. J.; Rezende, S. O. *Redes Bayesianas para Extração de Conhecimento de Bases de Dados*. Anais do II Encontro Nacional de Inteligência Artificial - ENIA. SBC'99, Rio de Janeiro, RJ, julho de 1999, pp 149 - 161.

Russell, S.; Norvig, P. *Artificial Intelligence - A Modern Approach*. Prentice-Hall. 1995.