Descoberta de Conhecimento em Bancos de Dados Espaço-Temporais Para Previsão de Risco Hidrológico

Petrônio C. de L. e Silva[1], Clodoveu A. Davis Jr[2], Elizabeth G. Davis[3]

[1]Instituto de Informática - Pontifícia Universidade Católica de Minas Gerais

petronio.candido@gmail.com

[2]Departamento de Ciência da Computação - UFMG

clodoveu@dcc.ufmg.br

[3]Serviço Geológico do Brasil - CPRM

elizabeth@bh.cprm.gov.br

Resumo. *Este artigo apresenta uma aplicação do processo de descoberta de conhecimento em bancos de dados na geração de alertas hidrológicos. Foi desenvolvida uma nova técnica de pré-processamento, baseada em grafos de vizinhança espaço-temporais entre estações fluviométricas em uma bacia. Os dados hidrológicos assim preparados podem ser usados para prever eventos de alerta e inundação em um ponto da bacia, utilizando redes bayesianas. O artigo apresenta resultados de um estudo de caso,baseado na experiência do Serviço Geológico do Brasil (CPRM) na operação do Sistema de Alerta contra Enchentes da Bacia do Rio Doce.*

Abstract. *This paper presents an application of the knowledge discovery in databases process for generating flood alerts. A novel preprocessing technique was developed, based on spatio-temporal neighborhood graphs involving measurement stations along a river basin. Hydrologic data prepared as such can be used to forecast flood events in a given point in the basin, using bayesian networks, The paper presents results from a case study, developed over the experience of the Brazilian geological survey (CPRM) in the operation of a flood warning for the Doce river basin in Southeastern Brazil.*

1. Introdução

Temos acompanhado com crescente frequência a ocorrência de eventos climáticos extremos, em todas as partes do planeta. Em nosso país, grande parte desses eventos está ligada à ocorrência de chuvas intensas ou de grande duração, que levam a inundações e deslizamentos de terra, com consideráveis perdas materiais e humanas. Ao longo das últimas décadas, algumas organizações têm se dedicado, entre outras coisas, ao registro sistemático de dados ambientais, como forma de compreender melhor os fenômenos naturais. Um dos principais focos de atenção de nossas autoridades é a coleta sistemática de dados fluviométricos, tendo em vista a proximidade de várias cidades brasileiras a rios de grande porte e o histórico de eventos de inundação.

Além disso, como o Brasil tem grande foco na geração de energia pelo aproveitamento hidrelétrico dos rios, certas bacias contam com abundância de dados, coletados

frequentemente e disponíveis em longas séries históricas. Apenas recentemente, no entanto, tornou-se possível utilizar parte desses dados tentar para antecipar a ocorrência de enchentes em cidades, de modo a reduzir o impacto dos eventos climáticos extremos sobre a população. Este artigo apresenta uma pesquisa realizada com o objetivo de verificar a adequação de técnicas de descoberta de conhecimento em bancos de dados (*knowledge discovery from databases*, KDD) para prever situações de risco hidrológico.

Segundo [Fayyad et al. 1996], KDD é um processo de extração de padrões não óbvios a partir dos dados, revelando informação útil. Existem inúmeras metodologias que guiam esse processo de descoberta e todas elas consistem em diversas etapas, que embora variem, comumente possuem as fases de pré-processamento, mineração de dados (ou fase de processamento) e pós-processamento (ou verificação). A mineração de dados espaciais é caracterizada por [Koperski and Han 1995] como a "extração de conhecimento implícito, relacionamentos espaciais e outros padrões não explicitamente armazenados em bancos de dados espaciais". A grande diferença existente entre o KDD convencional e o KDD Espaço-Temporal está no pré-processamento e processamento dos dados pelos algoritmos. Os dados convencionais e os dados georreferenciados e temporais diferem em inúmeros aspectos. As representações computacionais de tempo e espaço tendem a ser não-escalares, baseadas em vetores e matrizes e com formato próprio de armazenamento [Davis Jr. and Queiroz 2005].

As variações de um atributo no tempo ou no espaço, registradas em um banco de dados, devem ser analisadas respeitando-se os relacionamentos espaciais, como área, comprimento, proximidade, interseção e sobreposição, e os relacionamentos temporais, como inicialização, finalização, duração, sequencia, periodicidade e ocorrência. A maioria dos fenômenos naturais, por sua vez, é representada por dinâmicas espaço-temporais, onde há variação de um determinado parâmetro no tempo ou no espaço ou em ambos. O entrelaçamento dos relacionamentos espaciais origina o conceito de vida ou existência de geo-objetos, formando uma matriz de relacionamentos topológicos mista de tempo e espaço. Um estudo aprofundado sobre os relacionamentos topológicos entre objetos espaciais pode ser encontrado em [Egenhofer and Franzosa 1991] e [Egenhofer and Franzosa 1995]. Os diversos relacionamentos temporais, bem como representações temporais são estudados em [Allen 1983] e em [Allen 1991].

O pré-processamento de dados espaço-temporais visa transformar os objetos e relacionamentos geográficos armazenados em bancos de dados junto com marcas e intervalos temporais em representações intermediárias para uso por algoritmos convencionais de mineração de dados. Tais técnicas, quando aplicadas a dados hidrológicos, podem ser utilizadas na geração de modelos descritivos e preditivos capazes de auxiliar nos processos de previsão de enchentes.

O objetivo deste artigo é demonstrar a aplicação do processo de Descoberta de Conhecimento em Bancos de Dados, mais especificamente em sua etapa de mineração de dados, na previsão de enchentes na cidade de Governador Valadares através da utilização de de técnicas pré-processamento de dados espaço-temporais baseados em grafos de vizinhança direcionados espaço-temporais. O artigo está organizado como se segue. A Seção 2 apresenta trabalhos relacionados. A Seção 3 introduz conceitos de grafos de vizinhança espaço-temporais, usados no problema estudado, que é detalhado na Seção 4. Finalmente, a Seção 5 traz conclusões e lista trabalhos futuros.

2. Trabalhos Correlatos

Diversos autores têm trabalho na aplicação de métodos computacionais para a previsão de enchentes. Em [Alcoforado and Cirilo 2001] é apresentado um sistema de previsão de enchentes do Rio Capibaribe na região metropolitana de Recife(PE). Este sistema é um SIG acoplado a um módulo de redes neurais para previsão da vazão do rio em tempo real e a simulação da área inundada. A utilização dos SIGs na previsão das áreas inundadas também é o foco de [Costa et al. 2007]. [Andrade 2006] apresenta um modelo conceitual de previsão hidrometeorológica de precipitação baseado em equações termodinâmicas e modelo simplificado de física das nuvens seguido de um modelo chuva-vazão. A antecedência proporcionada pelo modelo hidrometeorológico aplicado é de 30 minutos para variáveis de entrada observadas.

Diversas pesquisas têm sido publicadas na aplicação direta de técnicas de mineração de dados para a previsão de desastres naturais, entre elas a previsão de enchentes. [Evsukoff et al. 2007] produziram modelos de previsão de vazões para a bacia do Rio Iguaçu, utilizando lógica fuzzy e técnicas de mineração de dados.

O presente trabalho diferencia-se dos mencionados pela utilização do pré-processamento das feições geográficas e atributos temporais baseado em grafos de vizinhança espaço-temporais. Esse grafo é estruturado a partir dos relacionamentos topológicos entre as feições geográficas contidas no banco e o relacionamento métrico de distância temporal (tempo decorrido) entre elas. Esse grafo é utilizado para realizar uma pivotação em dados hidrológicos espaço-temporais, a partir de sua representação usual como séries históricas de cotas ou vazões registradas em estações fluviométricas a fim de adaptá-los à geração de modelos preditivos como o de Redes Bayesianas.

3. Grafos de Vizinhança Espaço-Temporais

Os grafos de vizinhança espaciais (*neighborhood graphs*) são uma forma de modelar os relacionamentos espaciais, representados pelas arestas, entre entidades geográficas, representadas pelos vértices. Os relacionamentos espaciais podem ser topológicos, métricas e direcionais. [Ester et al. 2001] definem o *grafo de vizinhança espacial* (*spatial neighbourhood graph*) G_{neigh} para algum relacionamento espacial $neigh$ como sendo um grafo (N, E) em que o conjunto de vértices N, em que $n \in N$ representa um objeto espacial e dois vértices n_i e n_j são ligados por um vértice $e \in E \mid E \subseteq N \times N$ se e somente se entre eles houver um relacionamento espacial.

[Cao et al. 2005] definem *sequencia espaço-temporal* como uma lista S de eventos na forma $(x_1, y_1, t_1), (x_2, y_2, t_2) \ldots, (x_n, y_n, t_n)$, onde t_i representa o instante no local (x_i, y_i). Esse conceito é muito próximo ao de trajetória de um objeto móvel, onde se associa momentos no tempo a posições no espaço percorridas por esse objeto.Baseado na definição de sequencia espaço temporal [Cao et al. 2005], definimos *sequencia de eventos espaço-temporais*, como uma lista S de eventos na forma $(x_1, y_1, t_1, e_1), (x_2, y_2, t_2, e_2) \ldots, (x_n, y_n, t_n, e_n)$, onde t_i representa o instante de ocorrência do evento e_i no local (x_i, y_i).

Baseado no conceito de grafo de vizinhança pode-se definir um *caminho de vizinhança* (*neighbourhood path*) em algum grafo G como a lista de vértices $[n_1, n_2, \ldots, n_k]$, de tamanho k em G tal que uma aresta de G conecta pares de vértices sucessivos na lista.

Com os dados estruturados na forma de um grafo de vizinhança direcionado, pode-se traçar a sucessão de fatores precedentes que culminam em um evento a partir do caminho de vizinhança entre dois vértices do grafo. Esse caminho pode ser representado como uma transação contendo todos os fatores (isto é, eventos) responsáveis pelo acontecimento de um evento de interesse. Os eventos, originalmente no formato de um conjunto de tuplas, podem ser pivotados para uma conjunto de transações, cada transação contendo vários eventos. No caso de eventos hidrológicos, o grafo se limita a uma árvore, e os eventos são medições de cotas ou vazões ao longo do curso de um rio.O formato de transação, ou cesta de itens, é muito comum em banco de dados e diversos algoritmos de mineração de dados são capazes de lidar com este formato, entre eles os algoritmos de Regras de Associação, Clustering, Árvores de Decisão e Redes Bayesianas.

4. Estudo de Caso

O Sistema de Alerta Contra Enchentes da Bacia do Rio Doce (SAEBRD) é composto por uma rede de estações de coleta de dados espalhadas pela bacia do rio Doce e uma central de operações que funciona nas dependências da CPRM, em sua Superintendência Regional de Belo Horizonte. A operação do sistema é composta pelas seguintes etapas: coleta de dados, armazenamento, análise, elaboração da previsão hidrológica e meteorológica, transmissão das informações.

A rede de estações que compõem o sistema de coleta pertencem à empresas privadas e a orgãos públicos e é composta de estações hidrometereológicas, fluviométricas e pluviométricas, que coletam dados como a cota (nível do rio em cm), vazão defluente (em m^3/s) no caso das hidrelétricas e a precipitação diária (em mm).

A partir dos dados coletados e devidamente armazenados é realizada a análise da evolução dos níveis dos rios, bem como a previsão hidrológica com antecedência de 3 a 24 horas dependendo da localidade. A previsão hidrológica consiste na estimativa da vazão e da cota dos rios para algumas cidades integrantes do Sistema. As previsões do SAEBRD são definidas por meio de modelos hidrológicos, elaborados a partir de dados coletados nos períodos chuvosos anteriores.

Para algumas cidades consideradas estratégicas foram definidas cotas de alerta e cotas de enchente. Estas foram determinadas no campo, através de nivelamento topográfico da cota de início de enchente no ponto mais baixo da cidade. Já a cota de alerta foi definida de acordo com o tempo decorrido na subida dos hidrogramas da cheia de janeiro de 1997, discretizados a cada 12 horas. A cota de alerta definida é, no mínimo, 40 centímetros menor que a cota de enchente.

A cada momento, um desses locais, para efeito da previsão hidrológica, encontra-se em uma de três situações possíveis: a situação normal, alerta ou enchente. A situação normal caracteriza-se quando o nível nas estações encontra-se abaixo da cota de alerta. Até essa cota, a coleta dos dados é feita nas estações automáticas, de duas em duas horas.

A situação de alerta é atingida quando a cota de alerta é ultrapassada e está abaixo da cota de enchente. Neste caso, a coleta dos dados das estações automáticas e usinas hidrelétricas passa a ser feita de uma em uma hora.Regularmente, ao longo de todo o período de operação do SAEBRD, são elaborados boletins boletins informativos, que são enviados para a Defesa Civil dos locais em questão e divulgados na Web.

Atualmente o sistema usa modelos empíricos de previsão hidrológica, baseados nas curvas de vazão em diversos pontos do rio. A utilização de meios alternativos aos atuais implantados no CPRM busca explorar técnicas que visem aumentar a acurácia, a precedência ou ambos, explorando a massa de dados existente, acumulada pelas sucessivas operações do sistema. Essa massa de dados modela o funcionamento da bacia onde e quando ela é amostrada, nas estações de coleta e durante o período de operação do SAEBRD, correspondente ao período chuvoso na região.

Para elaborar um modelo preditivo de enchentes em uma bacia hidrográfica é necessário antes compreender a dinâmica espaço-temporal do fenômeno da enchente. A propagação da onda de cheia, que pode ser notada a partir de séries históricas de cotas em pontos consecutivos ao longo do curso de um rio, é a movimentação da água no sentido nascente para a foz dos rios dentro da área da bacia. Um montante extra de água, que atinge o curso do rio diretamente pela chuva ou por outros processos como o escoamento superficial, altera temporariamente o nível do rio elevando a sua cota.A elevação de cota propaga-se pela bacia em uma velocidade que depende de diversos fatores, como a morfologia do curso do rio e a declividade do terreno. Dessa forma, os bancos de dados com informações de cota, vazão e precipitação ao longo do tempo e em diversos pontos de monitoramento espalhados no curso de um rio podem ser utilizados para a composição de um modelo explicativo de como se forma uma onda de cheia. Discutimos a seguir como implementar esse modelo usando técnicas de mineração de dados.

4.1. Mineração de Dados

O processo de mineração de dados aplicado nos bancos de dados do SAEBRD teve como meta a extração de um modelo de previsão específico para a cidade de Governador Valadares (representada pela estação de coleta da ANA cuja sigla é GV e está localizada na própria cidade) utilizando as estações de coleta a montante da cidade. Doravante, todas as citações à estação de coleta localizada em Governador Valadares serão referenciadas pela sigla GV. Os dados do Sistema de Alerta contra Enchentes da Bacia do Rio Doce foram cedidos pela CPRM. A cada período chuvoso, que compreende os meses de novembro, dezembro, janeiro, fevereiro e março, é gerado um banco de dados independente com as séries históricas das medições de cota do rio, aferidos nas estações de coleta, precipitação chuvosa, colhido em estações meteorológicas e vazão defluente nas hidrelétricas da bacia do rio Doce.

A fase de pré-processamento contou com diversas etapas: a integração dos diversos bancos de dados individuais (períodos chuvosos a cada ano) em um único conjunto de dados, organizado com o objetivo montar um pequeno *data warehouse* orientado à fluviometria, com as séries históricas de cota e vazão por estação hidrológica, e à pluviometria, com séries históricas de precipitação por estação meteorológica. Em seguida, foram executadas a etapa de enriquecimento dos dados, com a inclusão de dados geográficos externos como os mapas temáticos da região da bacia do Rio Doce, a etapa de discretização dos dados contínuos de cota, vazão e precipitação e a etapa de balanceamento dos dados, ajustando a porcentagem de instâncias de cada classe[1]. Para essa fase foi utilizado o SGBD PostgreSQL[2] com a extensão PostGIS[3] que permite a utilização de

[1]Nas séries históricas do SAEBRD, existe grande desbalanceamento entre o número de instâncias de medições em que a cota dos rios está normal, em relação às instâncias de ocorrência de alerta ou de enchente

[2]http://www.postgresql.org/

[3]http://postgis.refractions.net/

feições espaciais no banco de dados.

Para estruturar o grafo de vizinhança é necessário definir quais os objetos e relacionamentos espaço-temporais que irão representar os vértices e arestas do mesmo. A modelagem conceitual do sistema SAEBRD mostra que há uma entidade continuante espaço-temporal que é imóvel mas cujos atributos são dinâmicos, o Rio Doce e seus afluentes.

Ligado a esse continuante existem ocorrentes (eventos) conforme Bittner (2001), em pontos específicos, as medições nas estações de coleta em intervalos regulares de tempo. Determinados eventos, quando relacionados, caracterizam um processo de cheia em uma determinada estação. O objetivo é estruturar um grafo de vizinhança em que os vértices sejam as estações de coleta e as arestas representem o relacionamento de vizinhança entre elas no curso do rio. Esse relacionamento é espacial e temporal, pois cada aresta é valorada com o tempo de propagação da onda de cheia de uma estação a outra. A partir das feições geográficas do Rio Doce e afluentes e os pontos das estações de coleta pode-se estruturar o grafo de vizinhança através de algoritmos de simplificação cartográfica [Davis Jr. and Queiroz 2005] ou percorrendo as feições diretamente.

A transformação de pivotação objetiva converter dados sequenciais em paralelos e vice-versa. Neste caso específico, deseja-se criar uma "cesta de eventos"relacionados com um evento de cheia ou alerta em uma determinada estação. Neste caso o objetivo é transformar as sequencias espaço-temporais contidas na tabela de medições em transações com eventos espaço-temporais. Cada evento na transação é representado como uma tupla $(e_i, t_i, c_i, p_i, v_i)$ onde e_i é a estação de coleta onde a medição aconteceu, t_i é a janela de tempo entre o horário da medição e o horário do evento em questão, em horas, c_i é a medida da cota na estação, em cm, p_i é a medida de precipitação na estação, em mm, e v_i é a vazão na estação em cm^3/s.O processo de pivotação consiste em achar caminhos de vizinhança dentro do grafo de vizinhança espaço-temporal entre o nó que representa a estação GV e todos os outros nós conectados a ele geograficamente e precedentes no tempo.

Dos atributos da tabela de transações apenas dois são relevantes para a extração de modelos, o atributo STATUS, que contém a situação em que se encontrava a estação GV no momento da medição, e os itens da transação.Após o pré-processamento, as massas de dados estavam prontas para serem remetidas aos algoritmos da fase de processamento, para a geração dos modelos descritivos e preditivos da base. Para a fase de processamento foi utilizada a ferramenta Weka [Witten and Frank 2005]. Foram formados dois modelos de redes bayesianas com estruturas distintas com o intuito de avaliar diferentes formatos de previsão. As redes bayesianas foram escolhidas pela sua capacidade de gerar previsões (isto é, saídas da rede) mesmo com dados ausentes, que no caso do SAEBRD significa que informações de determinadas estações de coleta estão disponíveis e outras não. Essas previsões estão associadas a uma probabilidade de ocorrência que significa o grau de crença da rede na ocorrência do evento inferido.

Dos modelos formados, o modelo I utiliza um paradigma sequencial de previsão (cada estação de coleta gera uma previsão isolada das demais) e o modelo II utiliza o paradigma transacional de previsão (uma previsão é composta pelos eventos de várias estações de coleta). O primeiro modelo é composto de 11 variáveis, sendo uma para cada estação de coleta a montante da estação GV e mais a variável STATUS, que representa

a classificação final (a previsão) e pode tomar um dos três valores: Normal, Alerta e Enchente. Cada variável que representa uma estação está ligada diretamente à variável STATUS. Nessa estrutura visa-se isolar os eventos de cada estação e medir o poder de previsão de cada estação isoladamente. Os eventos de cada variável que representa uma estação estão no formato (c_v_p), onde c o valor da cota em cm, v o valor da vazão em m^3/s e p o valor da precipitação em mm. No segundo modelo, são usadas as mesmas variáveis do primeiro modelo, porém encadeadas na forma de um grafo de vizinhança. Com esse modelo, pretende-se avaliar o poder de previsão da rede com as observações encadeadas ao longo do tempo. Os eventos da variável estão no formato c_v_p, onde c o valor da cota em cm, v o valor da vazão em m^3/s e p o valor da precipitação em mm.

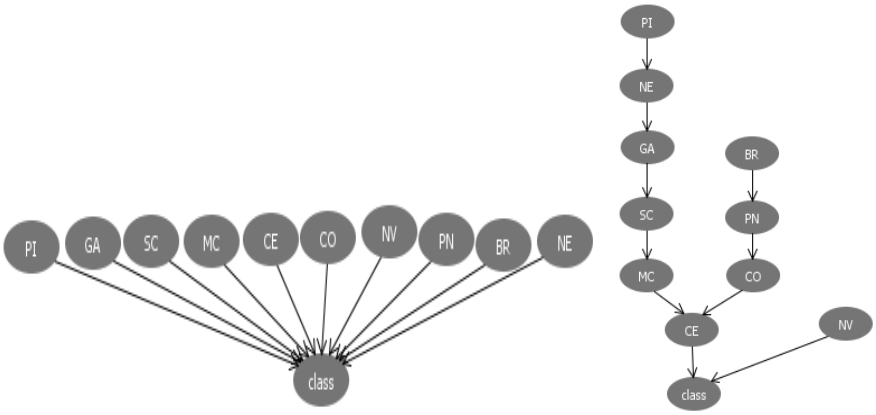

Figura 1. Modelos de Redes Bayesianas Gerados

Para aferir o poder preditivo dos modelos gerados foi implementado um motor de inferência capaz de receber como entrada o modelo da rede bayesiana no formato BIF (*Bayesian Networks Interchange Format*)[4] e um conjunto de dados para classificação. Esse motor de inferência foi desenvolvido na linguagem JAVA e utiliza-se das bibliotecas da ferramenta Weka, em especial do pacote de inferência bayesiana.

O conjunto de dados consiste nas leituras realizadas pelas estações de coleta em que cada registro contém o *timestamp* da coleta, a sigla da estação coletora e os dados coletados (cota, vazão, precipitação). A cada registro lido uma previsão é gerada para a estação GV. O horário dessa previsão e sua acurácia dependem da estação onde o dado foi coletado e, à medida que novos dados forem coletados outras previsões serão geradas para o mesmo horário.

4.2. Avaliação dos Modelos

A metodologia de testes baseou-se na comparação entre previsões geradas pelas redes bayesianas e as previsões geradas pelo SAEBRD, armazenadas no *data warehouse*. Para avaliar os modelos gerados foi utilizada a análise ROC (*Receiver Operating Characteristic*) e a métrica AUC(*Area Under Curve*). Segundo [Fawcett 2006], a análise ROC é utilizada para avaliar o desempenho de classificadores, baseado na quantidade de classificações corretas (verdadeiros-positivos e verdadeiros-falsos) e classificações incorretas (falsos-positivos e falsos-negativos) apresentados pelo classificador.

[4]http://www.cs.cmu.edu/ fgcozman/Research/InterchangeFormat/

A curva ROC é um gráfico que mostra a relação entre as taxas TP_R (taxa de acerto do classificador) e FP_R (taxa de erro do classificador). TP_R mede a porcentagem de verdadeiros-positivos em relação ao total de instâncias positivas, e FP_R mede a porcentagem de falsos-negativos em relação ao total de instâncias negativas. A comparação das curvas ROC de diversos classificadores permite escolher o melhor classificador, pela observação da curvas de melhor desempenho. A métrica AUC representa a área abaixo da curva ROC do classificador, calculado por algum processo de integração. O valor dessas áreas estará no intervalo $[0, 1]$ e será considerado o melhor classificador aquele que maximizar o valor de AUC.

A Figura 2 mostra o comportamento de cada modelo de rede bayesiana em relação às previsões geradas. Nota-se que o modelo II (transacional) tem um desempenho superior ao modelo I, em particular na previsão de eventos da classe Normal. Essa diferença se torna mais expressiva na Tabela 1, onde nota-se que o desempenho do modelo II fica acima dos 95% e o modelo I fica abaixo dos 90%.

Pela avaliação exposta a aplicação desse classificador no Sistema de Alerta Contra Enchentes da Bacia do Rio Doce é aceitável como método auxiliar, capaz de trabalhar com cenários de falha do sistema principal, como por exemplo se uma das estações de coleta utilizadas no modelo empírico falhar, deixando de informar os dados periódicos. O modelo de rede bayesiana é capaz de lidar com essas variações pois a previsão mantém sua eficácia mesmo na ausência de mais de uma variável. Com qualquer quantidade de dados de entrada na rede previsões serão geradas com uma margem de erro decrescente à medida que novos dados são observados. Com isso previsões podem ser geradas com até 44 horas de antecedência, que é o tempo que a onda de cheia leva para chegar da estação temporalmente mais distante até a estação localizada na cidade de Governador Valadares.

Classe	Modelo I	Modelo II
NORMAL	0.891	0.977
ALERTA	0.75	0.943
INUNDACAO	0.818	0.967
Média	0.825	0.964

Tabela 1. Valor do AUC médio por modelo de Rede Bayesiana

Como exemplo citamos os eventos: (CE_14_400), (MC_23_300) e (CE_14_450). O evento (CE_14_400) significa que a cota 400 mm foi lida na estação de coleta Cenibra e que essa onda de cheia demora 14 horas para chegar no seu local destino, isto é, Governador Valadares. Considerando a série histórica de dados e as regras geradas pela rede Bayesiana, esse evento isoladamente geraria uma previsão de Alerta em Governador Valadares, com antecedência de 14 horas. Se somado ao evento (MC_23_300) (estação Mário de Carvalho, 23 horas antes do horário alvo, cota de 300mm) será gerada uma previsão de Enchente. Já o evento (CE_14_450) isoladamente já geraria uma previsão de enchente.

Empiricamente o formato sequencial demonstrou-se capaz de fazer previsões a partir de eventos isolados, com menor precisão mas com maior antecedência. Já no formato transacional, todos os dados relacionados a uma previsão, i. e., todas as informações coletadas em todas as estações a montante da estação objetivo, são necessárias para compor a previsão. Isto torna a previsão mais confiável, embora a antecedência da previsão seja menor.

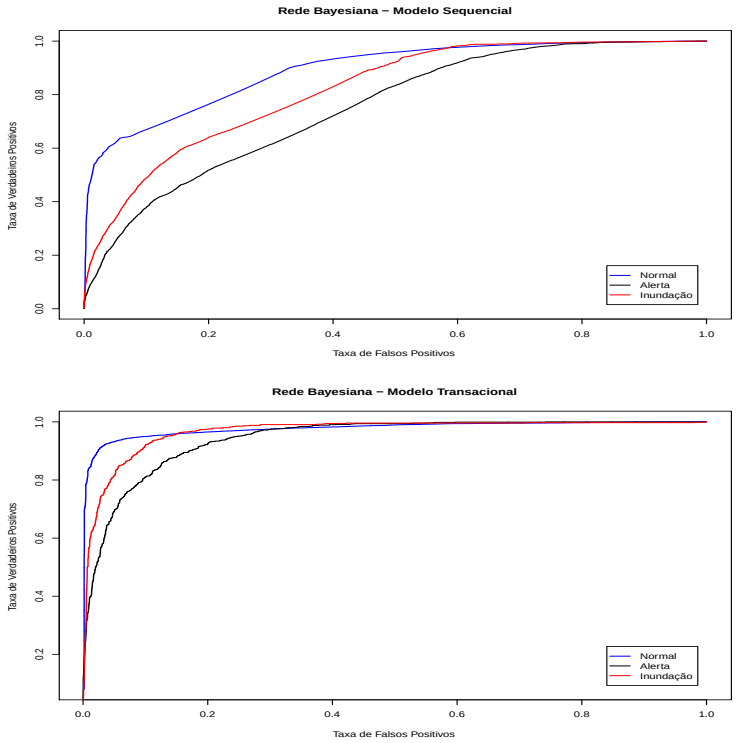

Figura 2. Curva ROC da Rede Bayesiana com dados sequenciais(a) e transacionais(b)

5. Conclusão e Trabalhos Futuros

O método de pivotação de dados através de grafos de vizinhança espaço-temporal, utilizado no pré-processamento dos dados do SAEBRD, mostrou-se efetivo para estruturação dos dados em um formato que possibilitou a utilização dos algoritmos convencionais de mineração de dados para tratar a dinâmica espaçotemporal dos eventos hidrológicos representados no banco de dados.

Os modelos de classificação gerados a partir do banco de dados devidamente pré-processado mostraram-se efetivos, alcançando taxas de 95% de acerto e com antecedência superior ao atual modelo em produção no SAEBRD.

Algumas questões ficaram fora do escopo desse artigo, como a utilização de outras técnicas de mineração de dados e estatística, tais como a análise de Séries Temporais, particularmente a análise de séries espaço-temporais, análise de características e outras técnicas, cujos resultados podem contribuir na geração de novos modelos explicativos ou preditivos, auxiliares no processo de predição de inundações.

6. Agradecimentos

Os autores agradecem pelo apoio da CPRM ao trabalho, especialmente pela cessão dos dados e por interações bastante positivas para o resultado final. Clodoveu Davis registra e agradece o apoio do CNPq (projetos 302090/2009-6, 474303/2009-8 e 551037/2007-5), e da Fapemig (PPM-00168-09) a seus projetos de pesquisa, e também ao suporte do Instituto Nacional de Ciência e Tecnologia para a Web (CNPq projeto 573871/2008-6).

Referências

Alcoforado, R. G. and Cirilo, J. A. (2001). Sistema de suporte à decisão para análise, previsão e controle de inundações. *RBRH - Revista Brasileira de Recursos Hídricos*, 6(4):133–153.

Allen, J. F. (1983). Maintaining knowledge about temporal intervals. In *Communications of ACM*, pages 832–843. ACM.

Allen, J. F. (1991). Time and time again: The many ways to represent time. In *International Journal of Intelligent Systems*, pages 341–356.

Andrade, J. P. M. d. (2006). Previsão hidrometeorológica visando sistema de alerta antecipado de cheias em bacias urbanas. Master's thesis, Universidade de Sãoi Paulo, São Carlos.

Cao, H., Mamoulis, N., and Cheung, D. W. (2005). Mining frequent spatio-temporal sequential patterns. *Data Mining, IEEE International Conference on*, 0:82–89.

Costa, M. D. G. A., Daré, R., Gomes, M. S. C. F., V., E. M., and Lani, J. L. (2007). Utilização de técnicas de geoprocessamento para a previsão de enchentes em atrativos turísticos em mariana - mg. pages 2479–2484.

Davis Jr., C. A. and Queiroz, G. R. (2005). *Bancos de Dados Geográficos*, chapter Algoritmos geométricos e representações topológicas, pages 53–92. MundoGEO.

Egenhofer, M. and Franzosa, R. (1995). On the equivalence of topological relations. In *International Journal of Geographical Information Systems, 9:133–152, 1995.*

Egenhofer, M. J. and Franzosa, R. D. (1991). Point set topological relations. *International Journal of Geographical Information Systems*, 5:161–174.

Ester, M., Kriegel, H.-P., and Sander, J. . (2001). *Algorithms and Applications for Spatial Data Mining*.

Evsukoff, A. G., Ebecken, N. F. F., Souza, F. T. d., Tavares, G. M., Alegre, M. P., Terra, G. S., and Hora, A. F. (2007). Uma abordagem de mineração de dados para a previsão de vazões com incorporação de previsão de precipitação da bacia do rio iguaçu. In *Simpósio Brasileiro de Recursos Hídricos*.

Fawcett, T. (2006). An introduction to roc analysis. *Pattern Recognition Letters*, (27):861–874.

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). From data mining to knowledge discovery in databases. *American Association for Artificial Intelligence*, pages 37–56.

Koperski, K. and Han, J. (1995). Discovery of spatial association rules in geographic information databases. In *Proc. 4th Int. Symp. on Large Spatial Databases*, pages 47–66, Portland, ME. IEEE.

Witten, I. and Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, 2 edition.