Implementação Paralela de Mineração de Dados Aplicada ao Estudo de Núcleos Convectivos

Jacques Politi
Dr. Stephan Stephany, Dra. Margarete Oliveira Domingues
{jpoliti, stephan}@lac.inpe.br; margaret@cptec.inpe.br

Resumo

Neste trabalho está sendo implementado um ambiente para mineração de dados utilizando ferramentas da inteligência computacional e técnicas de processamento de alto desempenho. A mineração de dados tem sido utilizada para analisar grandes volumes de dados tentando identificar correlações, padrões freqüentes, anomalias, nos mais variados domínios de aplicações. Pretende-se utilizar a ferramenta desenvolvida ao estudo de um dos traçadores de núcleos convectivos, isto é, descargas do tipo nuvem-solo. Este estudo deverá buscar informações ocultas e potencialmente úteis em uma base de dados de natureza espaço-temporal. Devido a grande quantidade de dados disponíveis, a utilização de processamento paralelo torna-se necessária para identificar os padrões em um tempo aceitável. Assim, pretende-se utilizar uma máquina paralela de memória distribuída, paralelizando o código por meio da biblioteca de comunicação por troca de mensagens MPI (Message Passing Interface).

1. Introdução

Nas últimas duas décadas houve um crescimento significativo na quantidade de informação armazenada em formatos eletrônicos. Estima-se que a quantidade de informação no mundo dobra a cada 20 meses[8]. Isso foi proporcionado basicamente pela queda de preços dos equipamentos de armazenamento e processamento, e aos avanços nos mecanismos de captura e geração de dados, tais como, leitores de código de barras, sensores remotos e satélites espaciais.

Estes dados, produzidos e armazenados em larga escala, são inviáveis de serem lidos ou analisados por especialistas por meio de métodos manuais tradicionais segundo Piatetsky-Shapiro[6], tais como planilhas de cálculos e relatórios informativos operacionais. Por outro lado, sabe-se que grandes quantidades de dados equivalem a um maior potencial de informação.

Analisar essa crescente quantidade de informação, não é uma tarefa trivial, e necessita utilizar técnicas computacionais avançadas para descobrir padrões ocultos e potencialmente úteis entre os dados. Esse é o objetivo da mineração de dados, também conhecida como extração de conhecimento, arqueologia de dados ou colheita de informações.

2. Mineração de Dados

O processo de mineração ou *KDD* (*Knowledge Discovery in Database*s – Descoberta de Conhecimento em Banco de Dados) consiste basicamente de seis fases e cada fase pode possuir uma interseção com as demais. A figura 1 ilustra todo o processo.

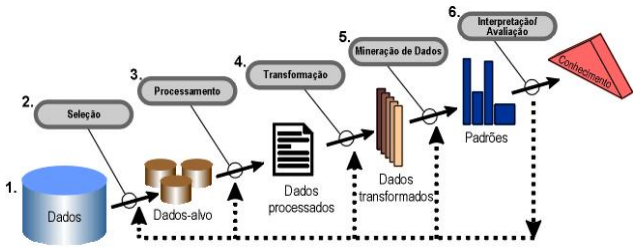

Figura 1. **Etapas do ciclo de KDD** [3]

1. Definir o problema*:* Inclui descrever cuidadosamente o problema, determinar se o uso da mineração de dados é apropriado, decidir a forma de entrada e saída dos dados, decidir relações custo/benefício etc.
2. Coletar e Selecionar os dados*:* Decidir como e quais dados serão coletados. Verificar se existe a necessidade de coletar dados de outros bancos, se existe alguma informação estatística sobre os dados, etc.
3. Pré-Processamento*:* Eliminação de ruídos e erros, estabelecimento de procedimentos para verificação da falta de dados; estabelecimento de convenções para nomeação e outros passos demorados para a construção de uma base de dados consistente.

4. Transformação*:* Alguns passos opcionais podem ser utilizados para auxiliar nas etapas seguintes, e são altamente recomendados, dentre eles temos a redução de dados e a compressão de dados.
5. Mineração de Dados*:* Aplicação dos algoritmos para descoberta de padrões nos dados; envolve a seleção de métodos/técnicas/modelos que são mais adequados para realizar a análise desejada.
6. Interpretação/Avaliação*:* Consiste na visualização dos resultados obtidos pelo processo de mineração de dados. Os padrões obtidos serão utilizados como ferramenta de suporte a decisão por parte do usuário. Este deverá avaliar a adequação dos padrões identificados pelo processo no tocante à extração de conhecimento desejado. Caso o resultado não seja satisfatório, o usuário poderá repetir um ou mais passos para refinar o processo.

As características principais do processo de mineração de dados são:

- O conhecimento descoberto é representado em uma linguagem de alto nível que pode ser entendido por usuários humanos.
- As descobertas retratam o conteúdo do banco de dados.
- O conhecimento descoberto é interessante de acordo com os usuários.
- O processo de descoberta é eficiente.

Os métodos de mineração de dados são formados pela interseção de diferentes áreas. As áreas mais relacionadas são: aprendizagem de máquinas [4][7], inteligência computacional, processamento de alto desempenho, estatística [2] e banco de dados.

3. Funcionalidades e Objetivos

As funcionalidades da mineração de dados e os tipos de conhecimento que podem ser descobertos são apresentados resumidamente abaixo [10]:

Caracterização – A caracterização de dados é um resumo geral das características dos objetos em uma classe alvo. Por exemplo, podemos querer caracterizar os consumidores de uma vídeo-locadora de vídeos que regularmente alugam mais de 30 filmes por ano.

Discriminação – É basicamente uma comparação das características gerais dos objetos entre duas classes referidas como classe alvo e classe oposta. Por exemplo, podemos comparar as características gerais dos consumidores que alugaram mais que 30 filmes no último ano com aqueles que alugaram menos de 5 filmes.

Associação – Estuda a freqüência de itens que ocorrem juntos em bancos de dados, e utiliza como critério de freqüência um limite chamado suporte, que identifica os conjuntos de itens freqüentes. Outro limite utilizado é a confiança, que é uma probabilidade condicional que um item aparece em uma transação quando outro item aparece, é usado como ponto pivô das regras de associação. Regras de associação são freqüentemente utilizadas em análise de mercados (*market basket analysis*). Por exemplo, poderia ser útil para o gerente da vídeo-locadora conhecer quais filmes sempre são alugados juntos ou se existe alguma relação entre alugar um certo tipo de filme e comprar pipoca ou refrigerante.

Classificação – Utiliza uma determinada classe rotulada para ordenar os objetos em uma coleção de dados. Normalmente utiliza um conjunto de treinamento onde todos os objetos são associados com as classes rotuladas conhecidas. O algoritmo de classificação aprende a partir do conjunto de treinamento e constrói um modelo. O modelo é utilizado para classificar novos objetos. Por exemplo, depois de começar uma política de crédito o gerente da vídeo-locadora pode analisar o comportamento dos consumidores, e rotulá-los de acordo com três possíveis valores "seguro", "risco" e "muito risco". Essa análise geraria um modelo que poderia ser utilizado para aceitar ou rejeitar pedidos de crédito no futuro.

Segmentação (*Clustering*) – Similar à classificação, segmentação é a organização de dados em classes. Entretanto, diferente da classificação, as classes rotuladas são desconhecidas e o algoritmo de segmentação deve descobrir classes aceitáveis.

Outlier Analysis – Outliers são elementos de dados que não podem ser agrupados em uma dada classe.

Análise de Evolução e Desvios – Fazem parte da análise de dados temporais. Na análise de evolução, os modelos extraem tendências nos dados, caracterizando, comparando, classificando ou agrupando os dados temporais. Em analises de desvio, por outro lado, considera as diferenças entre valores medidos e valores esperados, e tenta encontrar a causa para os desvios a partir dos valores antecipados.

É comum que os usuários não tenham uma idéia clara dos tipos de padrões que podem descobrir ou necessitam descobrir a partir dos dados que tem em mãos. Por isso é importante ter um sistema de mineração de dados versátil que permite descobrir diferentes tipos de conhecimento e em diferentes níveis de abstração. Isso torna a interatividade uma importante característica de um sistema de mineração de dados.

4. Mineração de Alto Desempenho

A mineração de dados necessita ser um processo eficiente, pois estamos lidando com grandes quantidades de informação e com algoritmos de complexidade computacional elevada. Existem basicamente três formas de acelerar esse processo:

Reduzir a Quantidade de Dados – Existem diversas formas de redução de dados, dentre elas temos a redução de dimensões, redução de valores e redução de casos [1]. Na redução de dimensões, o objetivo é identificar e remover atributos redundantes e irrelevantes. Na redução de valores, reduzimos o domínio de valores para um determinado atributo. A redução de casos consiste em selecionar subconjuntos de registros na base de dados.

Otimizar Algoritmos – Um algoritmo de mineração de dados baseado em indução de regras, procura por padrões (regras) em um espaço de busca muito grande. O tamanho desse espaço de busca faz com que uma busca exaustiva torne-se impraticável do ponto de vista computacional, necessitando utilizar alguma heurística para procurar apenas em algumas partes do espaço. Isto pode ser feito projetando novos algoritmos ou escolhendo um conjunto de parâmetros adequados para o algoritmo existente.

Processamento de Paralelo – Para que o processamento paralelo seja aplicado com sucesso, devemos identificar e paralelizar as rotinas que são críticas em relação ao tempo de processamento por meio de técnicas de temporização e *profiling*. Em principio um algoritmo de mineração de dados paralela descobre exatamente o mesmo conhecimento de sua versão seqüencial.

5. Núcleos Convectivos

Em geral, os núcleos convectivos estão associados a um ou mais aglomerados de nuvens Cumulonimbus (Cb). Essas nuvens são caracterizadas pelo forte movimento vertical e grande extensão, cerca de 16Km a 18Km de altura nos trópicos. O processo de formação destas nuvens depende da instabilidade atmosférica e das condições dinâmicas predominantes [5].

O ciclo de vida dessas nuvens Cb divide-se em três estágios:

- inicial (ou Cumulus);
- maduro e
- dissipativo

Estes estágios caracterizam-se em função do sentido do movimento vertical predominante das correntes de ar em seu interior. O ciclo de vida de uma Cb em geral é de uma a três horas [5].

Nesse contexto essas nuvens, em seu estado maduro, geram descargas elétricas. Elas são conseqüências das cargas elétricas que se acumulam a partir da colisão entre diferentes tipos de partículas como os cristais de gelo e granizo, atingindo às vezes a carga elétrica total de até centenas de coulombs. Admitem-se algumas variações para este processo de carregamento, que são os processos microscópicos e macroscópicos com variações, denominados processo indutivo e processo termoelétrico, respectivamente[9].

Essas descargas são classificadas em nuvem-solo, solo-nuvem, intranuvens, entre-nuvens, horizontais e para ionosfera. Os relâmpagos são constituídos por uma ou mais dessas descargas elétricas atmosféricas, de caráter transiente, portando uma alta corrente elétrica (em geral, superior a várias dezenas de quilo-ampéres). Eles são conseqüências das cargas elétricas que se acumulam em nuvens cumulonimbus (10-100C) e ocorrem quando o campo elétrico excede localmente a capacidade isolante do ar(>400 kV/m).

Neste trabalho focaliza-se uma estimativa da localização espaço/temporal dessas nuvens por meio de dados de descargas elétricas nuvem-solo (NS), tomados em intervalos de tempo muito menores que os dados convencionais e de satélite. Com isso pretende-se gerar uma ferramenta de auxílio ao acompanhamento de sistemas convectivos.

6. Metodologia de Mineração de Dados

Como já foi discutido na seção 1, devem-se seguir as etapas do processo de descoberta de conhecimento em banco de dados. A seguir descreve-se detalhadamente cada uma das etapas aplicadas ao problema:

Definição do Problema – Caracterizar núcleos convectivos utilizando dados de descargas atmosféricas do tipo nuvem-solo. Nessa caracterização pretende-se encontrar diversos tipos de correlações entre os parâmetros analisados.

Coleta e Seleção dos Dados – Os dados que estão sendo analisados são provenientes de vários instrumentos de medida e detecção de descargas elétricas atmosféricas. Os dados são de natureza espaço-temporal e se encontram disponíveis na forma de arquivos textos. Outras fontes de dados serão utilizadas, tais como, dados convencionais e de satélites. Os dados de descargas elétricas são identificados por meio de descargas de retorno. Quando uma descarga de retorno ocorre cada estação registra o momento exato da detecção, a localização(longitude e latitude), a polaridade e a intensidade de corrente das descargas. Com isso monta-se uma base de dados das ocorrências de descargas elétricas, em um formato padrão conhecido como UALF (*Universal ASCII Lighting Format*).

Pré-Processamento – Nesta etapa está sendo feita a seleção dos parâmetros que serão submetidos ao processo de mineração de dados, pois nem todos os parâmetros pertencentes ao formato UALF serão necessários para o nosso estudo. Os parâmetros iniciais analisados são: latitude e longitude, intensidade e polaridade. Outros parâmetros de outras fontes serão acrescentados no decorrer do projeto, como índices de precipitação, composição do solo etc.

Transformação – Devido à taxa de aquisição de dados ser relativamente alta, da ordem de nanosegundos, o

volume de dados total ultrapassa a ordem de Gigabytes. Analisar dados dessa magnitude exigiria grande capacidade computacional para obtermos resultados em um tempo aceitável. Uma tarefa crucial para obter um bom desempenho seria a redução de dados. Atualmente está sendo testado diversos algoritmos de segmentação (*clustering*) para reduzir o número de casos da base de dados. Nessa abordagem estamos agrupando espacialmente todas as descargas atmosféricas dentro de um intervalo de tempo, em um número reduzido de entidades, denominadas *clusters* ou centros de atividade. O método para encontrar esses agrupamentos é não-supervisionado, ou seja, não necessitamos definir o número de agrupamentos desejados, bastando apenas definir um diâmetro máximo para cada agrupamento.

Desses agrupamentos extraem-se outros parâmetros como a área, densidade e intensidade total, quanto à ocorrência do fenômeno em questão, que serão utilizados para montar uma tabela reduzida, pois representamos apenas o agrupamento e não as descargas.

Mineração de Dados – Nesta etapa aplicaremos os algoritmos seqüenciais e paralelos desenvolvidos, para encontrar padrões nos dados.

Interpretação/Avaliação – Consiste na visualização dos resultados obtidos pelo processo de mineração de dados. Caso o resultado inicial não seja satisfatório, será feito um refinamento do processo, a fim de atingir nosso objetivo. Será comparado também o desempenho dos algoritmos seqüenciais e paralelos.

7. Metodologia de otimização e paralelização

Inicialmente foram feitos um levantamento e análise sobre as ferramentas de mineração de dados disponíveis (da categoria *Open Source*), que melhor se adaptem aos tipos de dados disponíveis, isto é, dados espaço-temporais.

Constatou-se que nenhuma ferramenta se adequou perfeitamente às necessidades impostas pelo problema. Com isso desenvolveram-se algoritmos para as etapas de pré-processamento e transformação no ambiente MATLAB, devido à facilidade de manipulação de vetores e matrizes, bem como, à parte de visualização.

Esses algoritmos serão posteriormente convertidos na linguagem C ou Fortran 90 para que possam ser paralelizados utilizando a biblioteca de comunicação por troca de mensagens MPI (*Message Passing Interface*).

Quanto à implementação dos algoritmos da etapa de mineração de dados propriamente dita, pretende-se utilizar algoritmos baseados em árvores de decisão, devido a sua complexidade computacional reduzida.

Terminada a implementação, será feita a análise dos perfis de tempo de execução para identificarmos as rotinas que mais consomem tempo de processamento a fim de otimizá-las.

Depois de concluída fase de otimização e análise dos perfis de tempo de execução, será feita a paralelização das rotinas críticas por meio da biblioteca de comunicações por troca de mensagens MPI. Os algoritmos paralelizados serão validados por meio de casos de estudo mais simples, procedendo-se à análise do caso proposto.

8. Conclusão

Atualmente está sendo concluída a fase de transformação dos dados, que representa, tipicamente, de 50-80% do tempo total do processo de descoberta de conhecimento. Muitos desafios já foram superados, tais como a escolha da forma de representar os dados espaço-temporais em forma tabular. Existem diversas formas de se representar dados dessa natureza, e a análise dos parâmetros de entrada dos algoritmos das etapas posteriores permitiu concluir que não haveria necessidade de alterar a estrutura inicial dos dados para encontrarmos os padrões desconhecidos. Essa estrutura tabular consiste basicamente em duas colunas para a localização (latitude e longitude) e uma terceira com o instante de tempo considerado, a qual está agrupada em intervalos de tempo predeterminados. Na mineração de dados, esses padrões poderiam identificar correlações interessantes, como por exemplo, a ocorrência de uma maior polaridade das descargas nuvem solo e as condições atmosféricas e de superfície predominantes na região.

Entretanto outros desafios ainda não foram superados, por exemplo, a escolha do algoritmo de segmentação (*clustering*) mais adequado. Como se trata da análise de eventos discretos (ocorrência de descargas elétricas), técnicas tradicionais de *clustering* geram estruturas poligonais que, analisadas no decorrer do tempo, variam bruscamente de posição, com trajetória instável. Essa característica é indesejável, uma vez que se pretende identificar os núcleos convectivos, cuja trajetória é estável no decorrer do tempo. Outras formas de identificar esses agrupamentos estão sendo investigadas correntemente.

Referências

[1] Chen, Z - *Data Mining and Uncertain Reasoning: na integrated approach*.- John Wiley & Sons, Inc. (2001)

[2] Elder IV, J. F.; Pregibon, D. A statistical perspective on knowledge discovery in data bases.In: U. M. Fayyad et al. (Ed.) *Advances in Knowledge Discovery and Data Mining*, 83-113. AAAI/MIT Press, 1996.

[3] Fayyad, U. M., Piatetsky Shapiro, G., Smyth - The KDD Process for Extracting Useful Knowledge from Volumes of Data - *Knowledge Discovery In Communications Of The Acm* November 1996/Vol. 39, No. 11

[4] Langley, P. *Elements of Machine Learning*. Morgan Kaufmann, 1996.

[5] MacGorman, D. R.; Rust, W. D. *The electrical nature of storms*. Oxford: Oxford University, 1998. 422 p.

[6] Piatetsky-Shapiro, G. Knowledge discovery in real databases: *A report on the IJCAI-89 Workshop. AI Magazine,* Vol. 11, No. 5, Jan. 1991, Special issue, 68-70.

[7] Shavlik, J. W.; Diettrich, T. G.; (Eds.) *Readings in Machine Learning.* San Mateo, CA:Morgan Kaufmann, 1990.

[8] Szalay,A.; Kunszt,P. Z.; Thakar,A.; Gray,J.; Slut, D. R. Designing and mining multi-terabyte astronomy archives: The sloan digital sky survey. *Proceedings of the ACM SIGMOD,* pages 451-462. ACM Press, 2000.

[9] Uman, M. A. *The lightning discharge*. Florida: Academic Press, 1987. 377 p.

[10] Zaïane, O. R. - *Principles of Knowledge Discovery in Databases* – Chapter 1 - Department of Computing Science University of Alberta (1999) - http://www.cs.ualberta.ca /~zaiane/courses/cmput690/

Agradecimentos

Os autores agradecem ao Met. Cesar A. A. Beneti e ao RIDAT pelos dados de descargas elétricas atmosféricas utilizados neste trabalho.