Aplicação de *Clustering* em bancos de dados de Gravimetria

Jorge Luiz de Lima Matias (aluno) & Eder Cassola Molina (orientador) – IAG/USP

Resumo

Data Mining é utilizada em diversas áreas para extrair, de forma automatizada, padrões, regras ou modelos de bancos de dados, que não são perceptíveis de forma trivial. Este trabalho consiste em estudar a aplicabilidade de uma destas técnicas, o *Clustering*, em bancos de dados de gravimetria. As técnicas de agrupamento de dados ou *Clustering* tendem a serem úteis na análise de conjuntos de dados de naturezas distintas, casos em que análises triviais podem ser mais complicadas. Dentro dessa técnica foi dado enfoque no algoritmo do tipo hierárquico aglomerativo probabilístico GHBC, com o qual foram obtidos resultados preliminares bastante animadores com dados gravimétricos de uma região próxima da costa do nordeste brasileiro.

Introdução

Considerando a importância da análise de dados e a crescente quantidade de informações nos estudos geofísicos, o estudo de novas ferramentas e novas abordagens com potencial de ajudar nestas análises é fundamental. Este trabalho estuda a potencialidade do uso das técnicas de *Data Mining*, em principal o Clustering, em dados geofísicos. As técnicas de *Data Mining* têm por objetivo a identificação de padrões válidos, novos, potencialmente úteis, que estejam embutidos nos dados, e já foram aplicadas com sucesso em diversas áreas. Essas técnicas procuram ser uma ferramenta alternativa às formas mais triviais de análise de dados, automatizando o processo e procurando resultados difíceis de serem encontrados por um analista humano, facilitando a descoberta de conhecimento embutido em grandes bancos de dados. As tarefas de *Data Mining* são muitas, mas podem ser classificadas em dois tipos principais (detalhes em Mitchell, 1997): atividades preditivas (como Classificação e regressão) e atividades descritivas (como *Clustering* e Associação de Regras).

No presente trabalho foi dado enfoque em uma tarefa do segundo tipo, o *Clustering*, que tem sido freqüentemente utilizado em trabalhos de exploração de dados e extração de padrões, detectando grupos (*clusters*) com características comuns dentro de um banco de dados. O resultado da aplicação desta técnica é um conjunto de agrupamentos de dados, ou seja, uma descrição do dado. Esta técnica é muito útil para descrever os dados, além de outras utilidades, por exemplo utilizar os *clusters* obtidos como preditores em técnicas preditivas, ou também como parâmetros para preencher dados incompletos. O *Clustering* pode ser particional, baseado em grade, baseado em densidade, hierárquico, entre outros (Jain et. al., 1999). O *Clustering* Hierárquico foi o mais estudado neste trabalho, pois possui o diferencial que seu resultado não é apenas uma partição formada a partir do conjunto de dados iniciais, como ocorre nos demais métodos, e sim uma hierarquia de partições, que descreve um agrupamento diferente em cada nível, o que ajuda na análise de anomalias presentes em dados geofísicos, pois estas podem apresentar-se com distintas densidades e intensidades no conjunto de dados. Há diversos algoritmos de *Clustering* Hierárquico, e estes seguem duas estratégias principais: divisiva ou aglomerativa. Na maioria dos trabalhos da literatura relacionados a *Clustering* hierárquico os algoritmos aglomerativos são os mais comumente usados, e isto se deve à complexidade dos algoritmos divisivos, cujo tempo de cálculo cresce exponencialmente com o número de exemplos, o que os torna pouco aplicáveis, sendo de aplicação praticamente proibitiva para um conjunto de dados com mais de algumas centenas de exemplos; já os métodos aglomerativos são mais aplicáveis, apresentando cálculos com relação quadrática em relação ao número de exemplos do conjunto de dados. Dentro deste tipo de tarefa foi dada preferência ao algoritmo hierárquico aglomerativo GHBC (*Gaussian Hierarchical Bayesian Clustering*), onde se determina a similaridade entre os *clusters* de forma probabilística.

Metodologia

Com o objetivo de estudar anomalias presentes em conjuntos de dados geofísicos, utilizou-se a técnica de *Clustering* Hierárquico aglomerativo. Há diversos algoritmos deste tipo, e os tradicionais em geral seguem uma estrutura semelhante e podem ser divididos em três métodos (Metz, 2006): de ligação, de centróides, e de ligação por variância. Todos estes métodos possuem desvantagens e vantagens de acordo com a disposição dos dados e aos ruídos, mas o maior problema é a dificuldade de escolha métrica de distância utilizada, que pode ser de vários tipos (Shen, 2005). Pensando nestes problemas, os métodos aglomerativos probabilísticos possuem uma vantagem clara sobre todos estes métodos, pois modelam os grupos com distribuições de probabilidades e realizam o processo de aglomeração dos clusters baseados em critérios probabilísticos; assim, não é tão sensível a problemas que ocorrem nos métodos tradicionais devido à disposição dos dados e aos ruídos. Além disto, os métodos probabilísticos fornecem parâmetros para estudo de quão prováveis são

as partições, ou seja, com estes métodos tem-se uma medida para auxiliar na busca da partição ótima. Dos métodos aglomerativos probabilísticos, escolheu-se um baseado em máxima verossimilhança entre os *clusters*, o GHBC, que é uma modificação do algoritmo HBC (*Hierarchical Bayesian Clustering*) exposto em Iwayama e Tokunaga (1995), que o utilizaram com bons resultados para classificação de textos. Pela figura 1 observa-se que para calcular o parâmetro *U* que é máximo na melhor união são necessários os valores de verossimilhança (*SC*) dos dois grupos que serão unidos e do grupo formado à *posteriori*, onde a verossimilhança de um grupo é dada através do produtório das probabilidades elementares de pertinência dos exemplos do grupo (Iwayama e Tokunaga, 1995), e o cálculo deste termo é o que diferencia o HBC do GHBC, pois no GHBC as probabilidades elementares de pertinência são determinadas assumindo que os elementos dentro de cada grupo foram produzidos por uma distribuição normal multivariada ou Gaussiana (Christ et. al., 2007 e Everitt et. al., 2001). Este modelo gaussiano foi escolhido pela simplicidade da estimativa dos parâmetros necessários e pelo sucesso em diversas aplicações, inclusive em dados não necessariamente gaussianos (Murtagh & Raftery, 1984; Banfield & Raftery, 1993; Dasgupta & Raftery, 1998; Villanueva, E. R., 2007). Devido a este modelo os *clusters* da partição inicial não podem ser unitários, então para formá-la foi utilizado o *k-means*, um algoritmo particional dos mais tradicionais. Um esquema do algoritmo GHBC encontra-se na figura 1.

Entrada:
$D=\{d_1,d_2,...,d_n\}$ Conjunto de dados
$C_0=\{c_1,c_2,...,c_k\}$ Clusters iniciais

Saída:
Uma hierarquia de partições: $\{C_0,C_1,C_2,...,C_{k-1}\}$

Repetir:
$U=SC(c_i \cup c_j)/[SC(c_i)SC(c_j)]$, $c_i,c_j \in C_{k-1}$, $c_i \neq c_j$
$(c_x,c_y) \leftarrow argmax_{ci,cj}(U)$
$c_z \leftarrow c_x \cup c_y$
$C_k \leftarrow C_{k-1}+c_z-c_x-c_y$

Até:
n-1 vezes, quando todos os elementos estarão em um único grupo.

Figura 1: Resumo esquemático do algoritmo GHBC (modificado de Villanueva, 2007).

Resultados

Na figura 2 tem-se uma simulação do algoritmo GHBC em um conjunto de dados sintético, com uma grade de 51 posições em x, 51 posições em y e apenas um parâmetro z associado a cada coordenada, totalizando 2601 amostras de um parâmetro arbitrário. Este conjunto de dados é o mais simples possível, com uma única anomalia central, e esta simulação tem o objetivo de mostrar como o algoritmo GHBC funciona na prática. Começa-se com uma partição inicial formada por algum outro método, neste caso o *k-means*, e a cada iteração do GHBC forma-se uma nova partição onde estão unidos dois grupos (*clusters*) da partição anterior. Nesta figura tem-se 6 partições pertencentes ao dendograma ou hierarquia resultantes do GHBC, e pode-se observar a tendência de classificação da anomalia central.

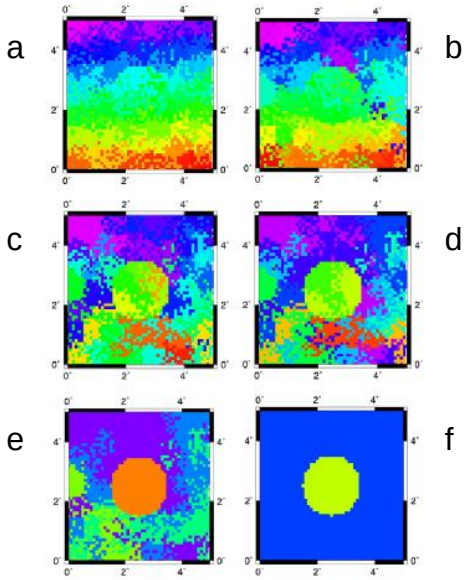

Figura 2 – Imagens de algumas partições, onde cada cor representa um cluster: a) partição inicial de 111 clusters gerada pelo k-means; b) partição gerada após 61 iterações do GHBC, ou seja, com 50 clusters; c) partição gerada após mais 30 iterações, ou seja, com 20 clusters; d) partição de 10 clusters; e) partição com 5 clusters; f) partição com 2 clusters.

Foi então aplicado o GHBC em dados reais, pertencentes a uma região do oceano próxima a costa do nordeste brasileiro (entre as longitudes -40º e -30º, e as latitudes 10º e 0º), analisada utilizando três parâmetros: anomalia ar-livre, altura geoidal e topografia (figura 3). O resultado do algoritmo pode ser observado na figura 4, onde se tem alguns exemplos do dendograma nos quais estão classificadas as feições mais interessantes da região.

Na figura 4.a existe uma quantidade de *clusters* tão grande que fica complicado de representar devido à limitação de cores, mas nota-se que os mesmos são pequenos. Na figura 4.b o GHBC já agrupou as anomalias em torno das ilhas que seguem o paralelo de -4º (setas amarelas) referentes à anomalias ar-livre negativas, agrupou a estrutura linear que segue o paralelo de -1º (seta branca) e um lineamento logo abaixo (setas pretas) claramente visível na figura 3.a e como baixo de topografia na figura 3.c, além de outras estruturas (setas amarelas) visíveis na figura 3.a. Na figura 4.c observa-se a evolução da classificação após 200 iterações, além de uma anomalia bem definida classificada no sudeste da região (apontada com seta branca). Na figura 4.d tem-se classificadas junto com esta mesma anomalia uma seqüência de direção SE-NW apontadas com setas brancas, e nas figuras 4.e e 4.f

tem-se, além da evolução da classificação do fundo, de outras anomalias e das já citadas, a classificação de um grupo anômalo relacionado com a plataforma continental. Na figura 4.g aparece a classificação em um grupo de uma anomalia linear que segue a costa (setas brancas). Nas figuras seguintes (4.h e 4.i) observa-se a evolução da classificação, onde as estruturas maiores classificadas no início mantém-se coesas, e outras menores começam a ser englobadas por grupos maiores, o que mostra a importância da análise da hierarquia gerada pelo método de *Clustering* hierárquico.

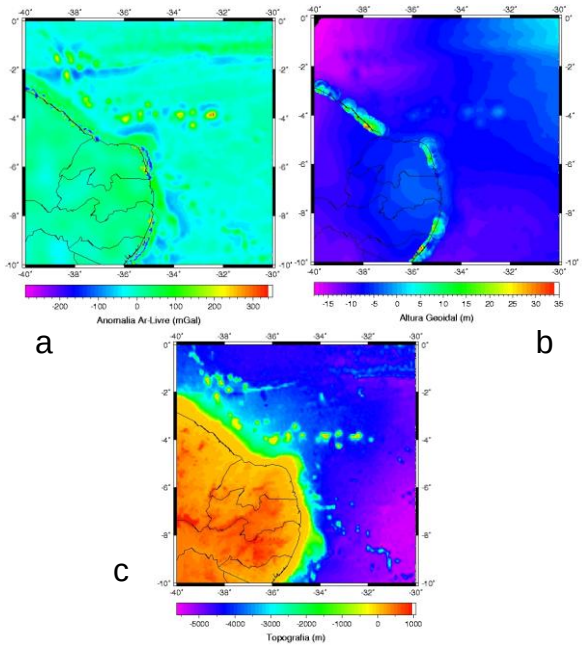

Figura 3: conjunto de dados da região: a) anomalia Ar-Livre (mGal); b) altura geoidal (m); c) topografia (m).

Discussão e Conclusões

O algoritmo GHBC gera bastante material para analisar, e são possíveis diversas abordagens de análise. Isso deixa claro que o pós-processamento de uma técnica de *Data Mining* é uma etapa muito importante, e percebe-se pela discussão apresentada que é uma etapa extensa, onde se deve analisar toda a hierarquia gerada, e provavelmente uma automatização maior nesta etapa faria com que alguma informação fosse perdida.

A técnica de *Clustering* hierárquico foi aplicada em dados gravimétricos e foram obtidos resultados animadores, onde a hierarquia gerada funcionou como um excelente guia para o analista perceber estruturas relevantes, sendo de grande auxílio na análise dos dados.

Como conclusão geral, tudo o que foi apresentado sinaliza que o *Clustering* e outras técnicas de *Data Mining* podem se constituir em uma importante ferramenta para a análise de dados geofísicos em estudos futuros.

Agradecimentos

Os autores agradecem ao apoio do FAPESP processo 2009/14944-5. Os mapas foram feitos usando o GMT (Wessel & Smith, 1998).

Referências

Banfield, J. D., Raftery, A. E., 1993. Model based Gaussian and non-gaussian Clustering. Biometrics, vol. 49: 803-821.

Christ, R. E., Villanueva, E. R., Maciel, C. D., 2007. Gaussian Hierarchical Bayesian Clustering algorithm. Proceedings of The seventh International Conference on Intelligent Systems Design and Applications (ISDA 2007), Rio de Janeiro(RJ), Brazil, 2007. In press.

Dasgupta, A.; Raftery, A. E., 1998. Detecting features in spatial point processes with clutter via model-based Clustering. American Statistical Association, vol. 93: 294-302.

Everitt, B. S., Landau, S. & Leese, M, 2001. Cluster Analysis. Oxford University Press Inc., New York, USA, 4 edition.

Iwayama, M., & Tokunaga, T., 1995. Hierarchical bayesian Clustering for automatic text classification. International Joint Conference on Artificial Intelligence, vol. 2:1322-1327.

Jain, A. K., Murty, M. N., & Flynn, P. J., 1999. Data Clustering: a review. ACM Computting Surveys, vol. 31(3):264-323.

Metz, J., 2006. Interpretação de clusters gerados por algoritmos de Clustering hierárquico. Dissertação de Mestrado, Instituto de Ciências Matemáticas e de Computação – USP São Carlos.

Mitchell, T. M., 1997. Machine Learning. WCB McGraw-Hill. Pyle, D., 1999. Data preparation for data mining. San Francisco, CA, USA: Morgan Kaufmann Publisher Inc.

Murtagh, F., Raftery, A. E. (1984). Fitting Straight Lines to Point Patterns. Pattern Recognition, vol.17(5): 479-483.

Shen, Z. Distance-Based Algorithms. Lectures Notes in Data Mining, 2005, p.73. Editado por M. W. Berry & M. Browne. University of Tennessee, USA.

Villanueva, E. R. (2007). Métodos Bayesianos aplicados em taxonomia molecular. Dissertação de Mestrado, Escola de Engenharia de São Carlos – USP São Carlos.

Wessel, P. & Smith, W.H.F., 1998. New, improved version of Generic Mapping Tools released, *EOS Trans. Amer. Geophys. U.*, vol. 79(47): 579.

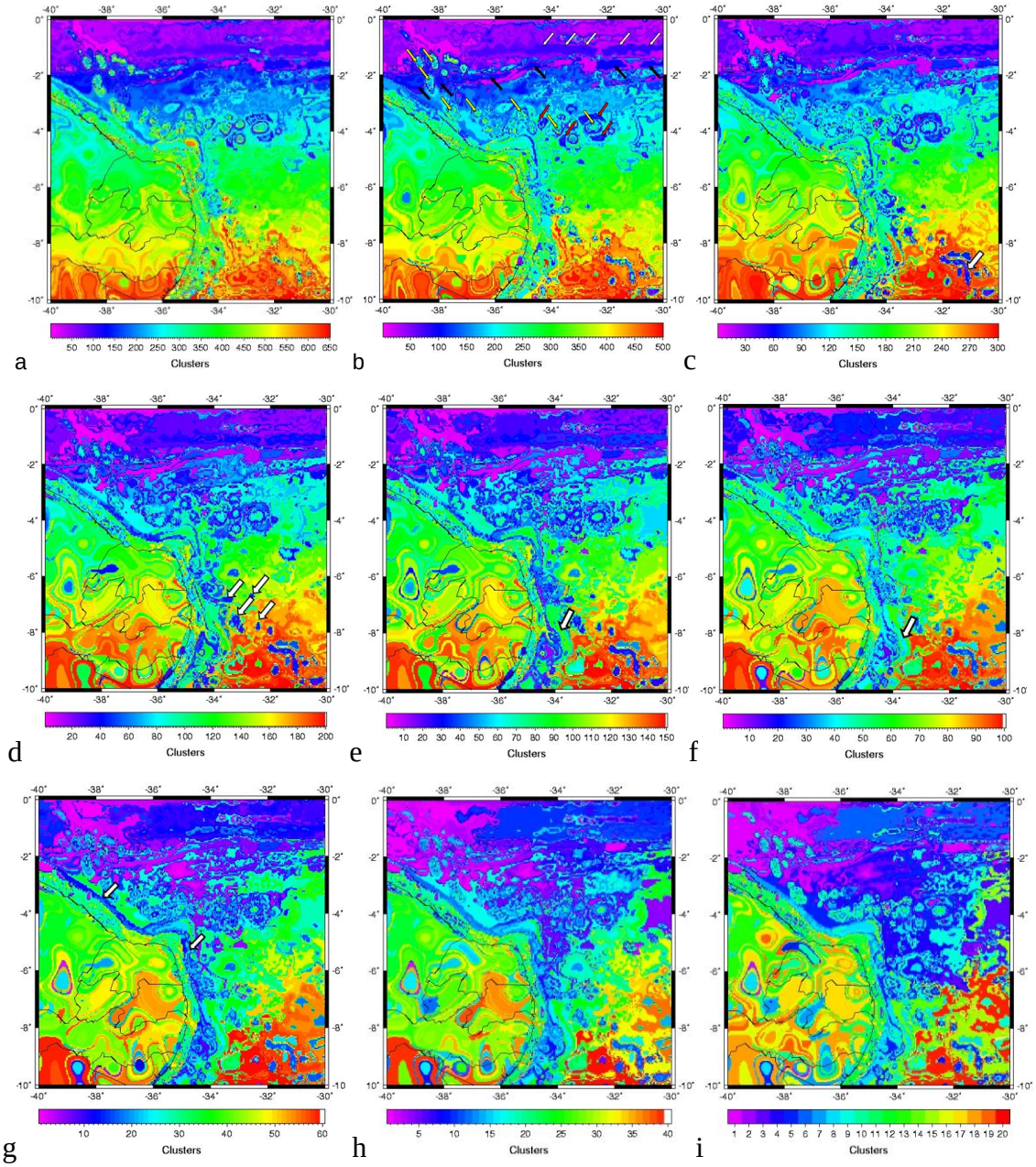

Figura 4: exemplos de partições da hierarquia gerada pelo algoritmo GHBC, com as feições principais indicadas por setas: a) partição com 650 clusters gerados pelo k-means; b) partição com 500 clusters gerados após 150 iterações do GHBC; c) partição com 300 clusters; d) partição com 200 clusters; e) partição com 150 clusters; f) partição com 100 clusters; g) partição com 60 clusters; h) partição com 40 clusters; i) partição com 20 clusters.