Impacto da amostragem aleatória uniforme para o aumento da escalabilidade na geração de agrupamentos hierárquicos de séries espaço-temporais *

Rodolfo M. S. Mendes[1], **Humberto Razente**[1],
Maria Camila N. Barioni[1], **Luciana Alvim Santos Romani**[2]

[1]Faculdade de Computação – Universidade Federal de Uberlândia (UFU)
Campus Santa Mônica – Uberlândia – MG

[2]Embrapa Informática Agropecuária
Campinas – SP

rodolfomendes@mestrado.ufu.br, humberto.razente@ufu.br,
camila.barioni@ufu.br, luciana.romani@embrapa.br

Abstract. *This paper presents the results of a scalable approach to build hierarchical clustering from space-time series. The goal is to reduce the complexity in terms of space and time. The approach explores data sampling pre-processing techniques to reduce the numerosity of the data. The experiment indicates it is needed the development of more efficient strategies than the naive selection of samples (uniform sampling).*

Resumo. *Este trabalho apresenta resultados do emprego de uma abordagem escalável para o agrupamento hierárquico de séries espaço-temporais de imagens de satélite visando a redução da complexidade de tempo de execução e espaço do algoritmo. Para tanto são exploradas as técnicas de pré-processamento para redução da numerosidade, particularmente por meio de amostragem de dados. O experimento indica que é necessário o desenvolvimento de uma estratégia mais eficiente que a seleção ingênua de amostras (amostragem uniforme).*

1. Introdução

A análise de séries espaço-temporais [Dias et al. 2005] de imagens de satélite vem sendo utilizada como uma ferramenta complementar em estudos sobre os possíveis impactos decorrentes dos cenários de mudanças climáticas globais. Essas séries permitem, por exemplo, analisar o uso e a cobertura do solo em uma determinada região e sua variação ao longo do tempo.

Imagens de satélite de baixa e média resolução espacial possibilitam a extração de séries temporais como as de temperatura da superfície e índice de vegetação para cada pixel da imagem. Diferentes análises podem ser realizadas a partir das séries extraídas, como agrupamento ou classificação. Para facilitar o uso destas análises pelo tomador de decisão, pode-se georeferenciar os grupos/classes gerados em mapas bidimensionais permitindo visualizar o uso do solo em uma determinada região.

O avanço das tecnologias de sensoriamento remoto tem permitido que satélites coletem cada vez mais imagens com intervalos de tempo menores (a cada poucos minutos).

*Trabalho realizado com apoio financeiro da Capes, CNPq e Fapesp.

Desta forma, a quantidade de imagens a serem armazenadas aumenta significativamente passando de terabytes para petabyes de dados. Além disso, satélites com resolução espacial média geram milhões de séries que, para serem analisadas, dependem totalmente do uso de métodos computacionais eficientes, preferencialmente, de complexidade linear.

Os algoritmos de particionamento, por exemplo o *k-means* [Jain 2010], apresentam desvantagens com relação ao significado dos agrupamentos gerados, entre elas a necessidade de se fornecer antecipadamente o número de agrupamentos desejados e a tendência a encontrar agrupamentos esféricos ou regiões densas divididas por hiperplanos projetados, o que nem sempre corresponde à melhor descrição dos dados.

Uma abordagem alternativa é o uso de algoritmos hierárquicos aglomerativos. Nestes algoritmos cada instância do conjunto de dados é colocada em seu próprio grupo inicialmente. Em seguida, os agrupamentos são aglutinados em grupos maiores, formando uma hierarquia de grupos. Diferentemente do *k-means*, o usuário não precisa fornecer o número k de grupos previamente. Embora a saída do algoritmo hierárquico seja um dendograma (uma hierarquia de grupos), este pode ser convertido em uma partição de k grupos por meio de uma poda na árvore resultante. Por sua vez, as partições resultantes deste processo não estão limitadas a formatos esféricos, sendo possível detectar agrupamentos com formatos arbitrários. Apesar dessas vantagens, os algoritmos hierárquicos aglomerativos tem alto custo computacional para tempo e espaço, tornando sua aplicação impraticável para grandes conjuntos de dados como séries temporais de imagens de satélites de alta resolução temporal. Nesse sentido, podem ser aplicadas técnicas de redução de dados para acelerar a execução.

Neste trabalho empregou-se uma técnica de pré-processamento para que se possa aproveitar as vantagens dos algoritmos hierárquicos mesmo em grandes conjuntos de dados. Esta abordagem consiste em reduzir o número de instâncias do conjunto de dados, para em seguida, aplicar o agrupamento hierárquico neste conjunto de dados reduzido. Por fim, todas as instâncias do conjunto de dados foram atribuídas ao grupo mais próximo, utilizando a mesma medida de distância entre grupos utilizada no agrupamento inicial. Com esta abordagem é possível aplicar o agrupamento hierárquico em tempo consideravelmente menor, mantendo a qualidade dos agrupamentos resultantes. O objetivo é avaliar a redução na qualidade dos agrupamentos ao empregar uma estratégia de seleção ingênua de amostras (amostragem uniforme).

Este artigo está organizado da seguinte maneira. Na Seção 2 são apresentados os trabalhos correlatos. A Seção 3 apresenta a abordagem para agrupamento das séries. Na Seção 4 são discutidos os resultados preliminares e na Seção 5 as considerações finais e trabalhos futuros.

2. Trabalhos Correlatos

O objetivo da etapa de pré-processamento é preparar os dados que alimentarão a etapa de mineração de dados. Nesta etapa, podem ser realizadas uma série de tarefas que visam aumentar a qualidade dos dados fornecidos à mineração de dados. Por exemplo, a tarefa de *limpeza dos dados* trata, por exemplo, de atributos sem valor definido e ruídos. Já a *redução da dimensionalidade* consiste em diminuir o número de atributos que serão considerados na mineração de dados. Por fim, a *redução da numerosidade* busca representar o conjunto de dados por meio de um número reduzido de instâncias [Han et al. 2011].

As estratégias para redução do número de instâncias são baseadas em amostragem de dados. Entre os desafios da amostragem estão o balanceamento das instâncias com relação à ocorrência de instâncias raras ou de exceções [García et al. 2015]. As principais técnicas são: amostragem aleatória uniforme, amostragem balanceada, amostragem estratificada e amostragem de agrupamentos. Como em muitos conjuntos de séries temporais não há atributos de classe disponíveis, necessários nas abordagens clássicas de amostragens balanceadas e estratificadas, não é possível empregá-las neste trabalho.

Nas últimas décadas, vários trabalhos empregaram técnicas de amostragem como etapa de pré-processamento de mineração de dados, como em [Meek et al. 2002]. A redução do tamanho do conjunto de dados por meio de amostragem aleatória é uma técnica intrínseca de vários algoritmos de agrupamento, entre eles CURE [Guha et al. 1998], CLARA [Kaufman and Rousseeuw 1990], CLARANS [Ng and Han 2002] e YADING [Ding et al. 2015]. Nestes trabalhos, a amostragem aleatória é aplicada estimando-se o tamanho mínimo das amostras para que as características dos grupos sejam preservadas. Outros trabalhos correlatos abordaram a seleção de atributos para redução da dimensionalidade dos dados [Bones et al. 2016]. O BIRCH [Zhang et al. 1997] cria agrupamentos hierárquicos por meio de uma representação multidimensional compacta. Entretanto, nenhum desses trabalhos avaliou o uso de técnicas de amostragem para o agrupamento hierárquico de séries espaço-temporais.

3. Agrupamento hierárquico aglomerativo de séries espaço-temporais

A estratégia consiste em reduzir o tamanho do conjunto de dados por meio de técnicas de amostragem, aplicar um agrupamento aglomerativo, e finalmente, atribuir as instâncias restantes aos seus grupos mais próximos, como apresentado no Algoritmo 1.

Algoritmo 1: Agrupamento Hierárquico Aglomerativo com Amostragem

Entrada: Conjunto de dados D, número de grupos K
Saída: Agrupamento $\mathcal{C}$
1. Selecionar amostra $\mathcal{D}$, tal que $|\mathcal{D}| < |D|$;
2. Obter o agrupamento hierárquico $\mathcal{H} \leftarrow AGNES\,(\mathcal{D})$;
3. Aplicar o procedimento de poda em $\mathcal{H}$, obtendo o agrupamento $\mathcal{C}_0$, com K grupos ;
4. Atribuir os objetos restantes em $D \setminus \mathcal{D}$ aos grupos em $\mathcal{C}_0$, obtendo o agrupamento final $\mathcal{C}$;

O algoritmo AGNES (*AGglomerative NESting*) é o algoritmo elementar (clássico) para executar o agrupamento hierárquico aglomerativo [Han et al. 2011]. Basicamente, seu procedimento consiste alocar inicialmente cada instância de dados em seu próprio grupo e então, sucessivamente, fundir os grupos mais próximos entre si, até que todos os objetos sejam aglomerados em um único grupo. Após cada instância de dados ser atribuída ao seu próprio grupo, a matriz de distâncias M é calculada, armazenando a distância de cada par de grupos existente. Então, sucessivamente, o algoritmo localiza a menor entrada na matriz de distância M, que equivale a encontrar o par dos grupos mais próximos, aglomera os grupos encontrados e recalcula a matriz de distância M. A aplicação de diferentes medidas de similaridade entre grupos resulta, entre outros, nos

algoritmos: *single-linkage* calculado pela Equação 1 e *complete linkage* calculado pela Equação 2.

$$dist\left(C_i, C_j\right) = \min_{x_i \in C_i, x_j \in C_j} \left\{|x_i - x_j|\right\} \tag{1}$$

$$dist\left(C_i, C_j\right) = \max_{x_i \in C_i, x_j \in C_j} \left\{|x_i - x_j|\right\} \tag{2}$$

A redução de dados tem papel fundamental na obtenção da escalabilidade dos algoritmos de agrupamento. Assim, o verdadeiro desafio na aplicação das técnicas de redução de dados é manter as características do conjunto de dados original, para que seja possível descobrir a estrutura dos grupos presentes.

4. Resultados Preliminares

O conjunto de dados utilizado nos experimentos são valores de índice NDVI (*Normalized Difference Vegetation Index*) de uma área localizada entre as latitudes $-8{,}55$ e $-8{,}45$, e as longitudes $-38{,}25$ e $-37{,}25$, que corresponde a uma região do Estado de Pernambuco. Estes dados foram extraídos a partir de imagens coletadas pelo sensor MODIS (*Moderate Resolution Imaging Spetroradiometer*) durante o ano de 2003, em períodos de 16 dias. Assim, a cada *pixel* da imagem é associada uma série temporal composta pelos valores do índice NDVI coletados ao longo do ano. O conjunto de dados utilizado possui um total de 9812 séries temporais, cada uma com 23 valores do índice NDVI.

O experimento consistiu em comparar o algoritmo hierárquico aglomerativo por amostragem com o algoritmo AGNES, ambos utilizando a ligação simples (*single linkage*) como medida de distância entre grupos. Os algoritmos foram comparados em relação ao seu tempo de execução e à qualidade dos agrupamentos gerados. Para medir a qualidade dos agrupamentos, utilizou-se os índices Dunn [Dunn 1973] e Davies-Bouldin [Davies and Bouldin 1979].

Para cada um dos algoritmos comparados, foi realizada a poda da hierarquia gerada em grupos $k = 3, 5, 7$. Por ser determinístico, o algoritmo AGNES foi executado uma única vez para cada valor de k. Já o algoritmo por amostragem foi executado com amostras de tamanho $m = 10, 100, 1000$. Por sua natureza probabilística, o algoritmo por amostragem foi executado 10 vezes para cada combinação de m e k, e foram calculadas as médias do tempo de execução e das métricas de qualidade dos agrupamentos. O gráfico da Figura 1 apresenta a comparação dos tempos de execução em escala logarítmica (a amostra de tamanho 9812 corresponde à execução do algoritmo AGNES para a base completa).

Embora o algoritmo baseado em amostragem precise, ao final do algoritmo, atribuir as instâncias restantes ao grupo mais próximo, o tempo de execução dessa etapa é da ordem de $O\left(nm\right)$, onde n é o tamanho do conjunto de dados. Assim, a diminuição do número de instâncias processadas pelo algoritmo aglomerativo representou uma redução significativa no tempo de execução do agrupamento, como esperado.

Além do tamanho da amostragem, também foi avaliado se o número de grupos utilizados na poda influenciaria a qualidade dos agrupamentos produzidos. O índice Dunn é obtido pela razão entre a menor distância entre pares de grupos diferentes e a maior distância entre pares de um mesmo grupo. Assim, para agrupamentos de maior qualidade,

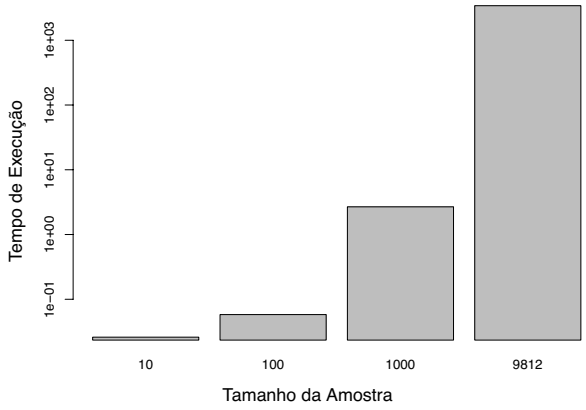

Figura 1. Tempo de execução por tamanho de amostra em escala logarítmica.

o valor desse índice é maior. Como pode ser observado no gráfico da Figura 2(a), o número de grupos utilizado na poda influenciou pouco a qualidade dos agrupamentos obtidos. Por sua vez, o fator determinante na qualidade foi o tamanho da amostra utilizada nos algoritmos.

Houve grande diferença entre a qualidade dos agrupamentos produzidos pelo algoritmo AGNES e os agrupamentos produzidos pelo algoritmo baseado em amostragem. Dado que o índice Dunn é influenciado pela menor distância entre pares de grupos diferentes, esse desempenho pode ser explicado pelo fato do algoritmo AGNES garantir que os pares de objetos mais próximos serão colocados no mesmo grupo. Por outro lado, no algoritmo baseado em amostragem, existe a possibilidade de que objetos muito próximos sejam colocados em grupos separados, afetando negativamente seu desempenho.

A qualidade inferior dos agrupamentos produzidos pelo algoritmo baseado em amostragem também foi refletida no índice Davies-Boulding. Este índice é obtido para cada par de grupos como a relação entre a soma dos desvios-padrão e a distância entre as médias dos grupos. Assim, quanto maior a qualidade dos agrupamentos, menor será o valor desse índice, pois menor será a dispersão dentro de um grupo e maior será a distância entre seus centros. O gráfico da Figura 2(b) apresenta os resultados obtidos.

Os dados apresentados no gráfico da Figura 2(b) também indicam que a qualidade dos agrupamentos produzidos pelo algoritmo baseado em amostragem foi inferior. Isso também pode ser explicado pela garantia que o algoritmo AGNES oferece de agrupar os pares mais próximos já no primeiro passo do algoritmo, pois diminui a dispersão dos dados. No gráfico da Figura 2(b) é possível verificar que o índice Dunn sofreu maior variação de acordo com o número de grupos e o tamanho das amostras.

5. Conclusões

A estratégia proposta permite a redução do tempo de execução dos agrupamentos hierárquicos aglomerativos de séries espaço-temporais. Porém, a amostragem aleatória impactou negativamente na qualidade dos agrupamentos obtidos. Entre os trabalhos futuros destaca-se o desenvolvimento de um novo método de amostragem baseado em fractais para gerar amostragens estratificadas com relação à densidade de objetos em volumes hiper-dimensionais analisados em várias escalas.

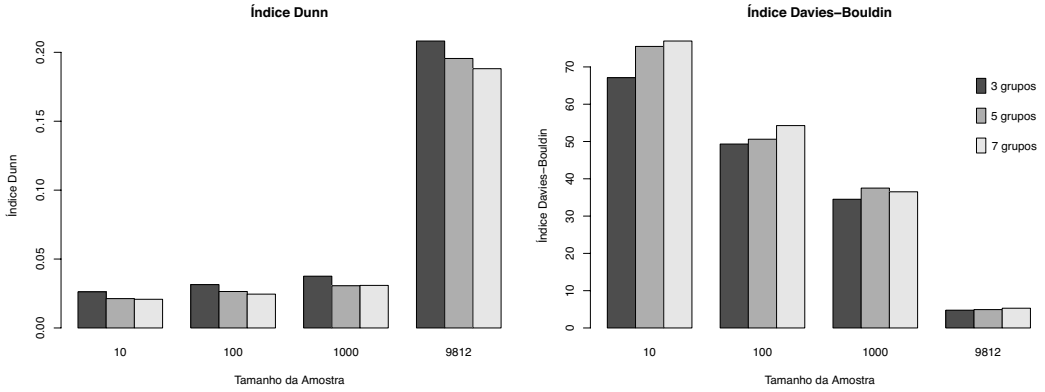

Figura 2. Tamanho da amostra por número de grupos: (a) Índice Dunn; (b) Índice Davies-Bouldin.

Referências

Bones, C., Romani, L., and Sousa, E. (2016). Improving multivariate data streams clustering. *Procedia Computer Science*, 80:461 – 471.

Davies, D. L. and Bouldin, D. W. (1979). A cluster separation measure. *IEEE Trans. Pattern Anal. Mach. Intell.*, 1(2):224–227.

Dias, T. L., Câmara, G., and A. Davis Jr., C. (2005). *Bancos de Dados Geográficos*, capítulo Modelos espaço-temporais, páginas 137–169. MundoGEO.

Ding, R., Wang, Q., Dang, Y., Fu, Q., Zhang, H., and Zhang, D. (2015). YADING: Fast Clustering of Large-Scale Time Series Data. *VLDB Endowment*, 8(5):473–484.

Dunn, J. C. (1973). A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters. *J. Cybern.*, (3):32–57.

García, S., Luengo, J., and Herrera, F. (2015). *Data Preprocessing in Data Mining*, capítulo Data Reduction, páginas 147–162. Springer.

Guha, S., Rastogi, R., and Shim, K. (1998). Cure: an efficient clustering algorithm for large databases. In *ACM SIGMOD Record*, volume 27, páginas 73–84. ACM.

Han, J., Kamber, M., and Pei, J. (2011). *Data Mining: Concepts and Techniques*. Morgan Kaufmann.

Jain, A. K. (2010). Data clustering: 50 years beyond k-means. *Pattern Recogn. Lett.*, 31(8):651–666.

Kaufman, L. and Rousseeuw, P. J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley Series in Probability and Statistics. Wiley.

Meek, C., Thiesson, B., and Heckerman, D. (2002). The learning-curve sampling method applied to model-based clustering. *Journal of Machine Learning Research*, 2:397–418.

Ng, R. T. and Han, J. (2002). Clarans: A method for clustering objects for spatial data mining. *IEEE Trans. Knowl. Data Eng. (TKDE)*, 14(5):1003–1016.

Zhang, T., Ramakrishnan, R., and Livny, M. (1997). Birch: A new data clustering algorithm and its applications. *Data Mining and Knowledge Discovery*, 1(2):141–182.