PLANO DE TRABALHO

Programa de Mestrado em Ciência da Computação
Universidade Federal de Uberlândia

Aluno: Rodolfo Martignon Sevilhano Mendes

Orientador: Prof. Dr. Humberto Luiz Razente

Título do Trabalho: Uma nova abordagem para aumento da escalabilidade na geração de agrupamentos hierárquicos de séries espaço-temporais

Título do Trabalho: Efeitos da Amostragem no Agrupamento Hierárquico de Séries Temporais

Data de Início como Aluno Regular: 11/08/2014

Previsão da Defesa: Fevereiro de 2017

1 Introdução

A análise de séries temporais de imagens de satélite tem se mostrado uma importante ferramenta no estudo dos impactos das mudanças climáticas globais. A partir destas séries temporais, é possível analisar o uso e a cobertura do solo em uma determinada região, e como este uso varia ao longo do tempo.

Uma das possíveis técnicas para extrair conhecimento destas imagens é o agrupamento de séries temporais. A partir das imagens de satélite, métricas como temperatura da superfície, índice de reflexão e índice de vegetação são extraídos para cada pixel da imagem, para cada instante no tempo. O agrupamento das séries temporais obtidas, combinado com a visualização em mapa a quais grupos cada série pertence, permite ao tomador de decisão entender como o uso do solo está distribuído em uma região.

Por sua vez, o avanço das tecnologias de sensoreamento têm permitido que satélites coletem imagens com resoluções cada vez maiores. Isso significa que os conjuntos de dados são representados por um número cada vez maior de séries temporais, na ordem de milhões e até mesmo bilhões de séries para serem agrupadas.

Por meio do algoritmo k-médias é possível agrupar dados com complexidade de tempo linear, $O(n)$. Porém, esse algoritmo possui algumas desvantagens com relação ao significado dos agrupamentos gerados. Uma destas desvantagens é a necessidade do usuário fornecer antecipadamente o número de agrupamentos desejados através do parâmetro k. Ou seja, é necessário que o usuário tenha algum conhecimento prévio sobre o domínio do problema e que ele tenha uma estimativa do número de grupos existente no conjunto

de dados. Outra limitação está relacionada ao formato dos agrupamentos no espaço dos atributos. No k-médias, cada instância do conjunto de dados é atribuída ao grupo cujo centro esteja mais próximo, que por sua vez é recalculado a cada iteração. Assim, os agrupamentos tendem a ter formatos esféricos ou regiões densas são divididas por hiperplanos separadores projetados, o que nem sempre corresponde à melhor descrição dos dados.

Uma abordagem alternativa são os 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-médias, 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 o algoritmo capaz de detectar agrupamento com formatos arbitrários. Apesar dessas vantagens, os algoritmos hierárquicos tem alto custo computacional, de ordem $O(n^2)$ (quadrática) ou $O(n^3)$ (cúbica) para tempo e espaço, tornando sua aplicação impraticável para grandes conjuntos de dados como séries temporais de imagens de satélites de alta resolução.

Neste trabalho, propomos uma nova abordagem para que se possa aproveitar as vantagens dos algoritmos hierárquicos mesmo em grandes conjuntos de dados. Esta abordagem consiste em, inicialmente, reduzir o conjunto de dados, e em seguida, aplicar o agrupamento hierárquico neste conjunto de dados reduzido. Por fim, as instâncias restantes são atribuídas ao seu vizinho mais próximo (1-NN) completando o agrupamento. Espera-se que, com esta abordagem, seja possível aplicar o agrupamento hierárquico em tempo consideravelmente menor, mas sem perder a qualidade dos agrupamentos resultantes.

1.1 Objetivos e Desafios de Pesquisa

O objetivo geral deste trabalho é desenvolver uma abordagem escalável para o agrupamento hierárquico de séries espaço-temporais de imagens de satélite que permita reduzir significativamente a complexidade de tempo de execução e espaço do algoritmo, sem no entanto reduzir a qualidade dos agrupamentos produzidos. Especificamente, deseja-se obter:

1. Desenvolvimento de uma abordagem para o agrupamento hierárquico aglomerativo de séries espaço-temporais incluindo uma etapa de pré-processamento baseada em redução de dados;

2. Desenvolvimento de uma abordagem para a geração de amostras de séries espaço-temporais por meio da análise de auto-similaridade;

3. Avaliar experimentalmente as novas abordagens em termos de tempo de execução, consumo de memória e qualidade dos agrupamentos gerados visando o agrupamento dos diversos tipos de vegetação para identificação de áreas de plantio de culturas como a cana-de-açúcar.

A etapa 1 encontra-se concluída e os resultados experimentais iniciais encontram-se descritos na Seção ???. A implementação da etapa 2 está em andamento espera-se concluí-la em ....

1.2 Hipótese

A hipótese deste projeto de pesquisa consiste em: "A redução do conjunto dos dados de entrada permite executar o agrupamento hierárquico aglomerativo em tempo menor, mas mantendo a qualidade dos agrupamentos produzidos".

1.2.1 Amostragem baseada em fractais

A hipótese de uso da teoria dos fractais para a geração de amostras significativas baseadas na análise da auto-similaridade vem do fato de que técnicas de amostragem ingênuas (*naive*), como amostragens aleatórias e amostragens estratificadas, não são adequadas para conjuntos de dados reais com presença de dados ruidosos e de exceções, uma vez que o comportamento desses algoritmos é imprevisível. Como não há atributos de classe disponíveis, pretende-se fazer uma amostragem baseada na densidade da estrutura *box count*. O objetivo é obter um conjunto mínimo de instâncias que represente as características do conjunto original.

1.3 Contribuição

A contribuição esperada é uma estratégia para o agrupamento hierárquico aglomerativo que produza respostas mais rápidas para o usuário.

2 Revisão da Literatura Correlata

Nesta seção apresentaremos os principais conceitos teóricos relacionados ao trabalho desenvolvido. Na subseção 2.1 apresentaremos o processo de descoberta de conhecimento em banco de dados e suas principais etapas.

2.1 Descoberta de Conhecimento em Bancos de Dados

A Descoberta de Conhecimento em Bancos de Dados, também conhecida pela sigla KDD (*Knowledge Discovery in Databases*) é o processo pelo qual dados brutos, coletados a partir das mais variadas fontes, são processados e transformados em informações úteis. Por sua vez, estas informações permitem o aprimoramento da tomada de decisão e até mesmo ampliação do conhecimento científico sobre um determinado fenômeno [Tan et al. 2009].

O processo de KDD envolve desde a aquisição dos dados até a disponibilização do conhecimento para o usuário final. De acordo com [Tan et al. 2009], este processo pode ser descrito pelas seguintes etapas:

1. Pré-processamento
2. Mineração de dados
3. Pós-processamento

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. Na tarefa de *limpeza dos dados*, são tratados atributos sem valor definido e ruídos. A tarefa de *integração de dados* consiste em consolidar fontes de dados de diversos tipos (arquivos de texto, planilhas, *web-services*, arquivos XML, bancos de dados) em uma única fonte de dados consolidada, usualmente um *data-warehouse*. A *redução da dimensionalidade* consiste em diminuir o número de atributos que serão considerados na mineração de dados. Dentre as principais técnicas podemos citar PCA (*Principal Component Analysis*) e DWT (*Discrete Wavelets Transforms*). Por fim, a *redução da numerosidade* busca representar o conjunto de dados através de um número reduzido de instâncias [Han et al. 2011].

O reconhecimento de padrões é efetivamente realizado na etapa de mineração de dados. As tarefas desta etapa são categorizadas de acordo com o conhecimento que se deseja extrair da base de dados analisada. Na tarefa de mineração de itens frequentes, deseja-se extrair de um banco de transações quais itens ocorrem conjuntamente com maior frequência. Na tarefa de classificação, o objetivo é inferir um modelo a partir do qual seja possível prever à qual classe uma determinada instância de dados pertence. Por fim, na análise de agrupamentos deseja-se descobrir a existência de grupos (*clusters*) de dados. Assim, é preciso que se estabeleça uma *medida de similaridade* entre as instâncias do banco de dados, de forma que se maximize a similaridade entre instâncias do mesmo grupo e se minimize a similaridade entre instâncias de grupos diferentes.

Por fim, na etapa de pós-processamento avalia-se se os padrões descobertos de fato representam um *conhecimento* novo sobre os dados. Para cada tipo de padrão descoberto, pode-se estabelecer uma *medida objetiva* sobre a qualidade do padrão [Han et al. 2011]. No caso dos agrupamentos, por exemplo, a qualidade destes pode ser medida em termos de *coesão* e *separação* [Tan et al. 2009].

Neste trabalho, será enfatizada a tarefa de agrupamento de dados, com atenção especial aos algoritmos hierárquicos de agrupamento. Também será discutido como as técnicas de redução de numerosidade influenciam o tempo de execução dos algoritmos hierárquicos e a qualidade dos agrupamentos produzidos, como etapa de pré-processamento.

2.2 Técnicas de Amostragem de dados

O objetivo das técnicas de amostragem de dados é reduzir o número de instâncias submetidas aos algoritmos de mineração de dados. Entre os desafios da amostragem de dados estão o balanceamento das instâncias com relação à ocorrência de instâncias raras ou de exceções. Conside um conjunto T com cardinalidade $|T| = N$. Entre as técnicas propostas na literatura destacam-se [García et al. 2015]:

- Amostragem aleatória de tamanho s sem substituição: criada pela escolha de s instâncias de T ($s < N$), onde a probabilidade de um exemplo ser escolhido é de $1/N$, de modo que todas as instâncias têm a mesma chance de serem escolhidas;

- Amostragem aleatória de tamanho s com substituição: semelhante à anterior, exceto pelo fato que a cada vez que uma instância é escolhida, permanece no conjunto e pode ser escolhida novamente;

- Amostragem balanceada: criada levando-se em consideração um conjunto de critérios pré-definidos, por exemplo, para manter a proporcionalidade de instâncias entre classes conhecidas;

- Amostragem de agrupamentos: escolha de grupos específicos resultantes de técnicas de agrupamento;

- Amostragem estratificada: obtida por meio da divisão de um conjunto T em partes mutualmente disjunta seguida da escolha de uma amostragem aleatória em cada divisão.

2.3 Análise de Agrupamentos

A análise de agrupamentos é uma tarefa de mineração de dados cujo objetivo é, automaticamente, particionar o conjunto de dados em subconjuntos chamados grupos. Os objetos reunidos em um mesmo grupo devem ser similares entre si, enquanto que objetos de grupos separados devem ser diferentes. Ao conjunto dos grupos resultantes da análise dá-se o nome de *agrupamento*.

A análise de agrupamentos pode ser usada como uma ferramenta para extração de conhecimento sobre um conjunto de dados ou então, como um etapa de pré-processamento para outras tarefas de mineração de dados. Por exemplo, em [Gonçalves et al. 2014], a análise de agrupamentos foi utilizada para identificar o uso do terreno em diferentes regiões do estado de São Paulo, Brasil. Já em [Petitjean et al. 2014], a análise de agrupamentos foi utilizada para eleger protótipos que posteriormente seriam utilizados como dados de treinamento para a tarefa de classificação 1-NN.

Existem diversas abordagens para o agrupamento de dados. No agrupamento por *particionamento* o conjunto de dados é dividido em k grupos, com cada grupo contendo pelo menos um objeto do conjunto. De maneira geral, estes algoritmos consistem em: a partir de um agrupamento inicial, iterativamente realocar os objetos em grupos mais significativos até que um critério de parada seja atingido. Podemos incluir nesta categoria os algoritmos *k-médias* e *k-medoids*.

Uma abordagem alternativa é o agrupamento *hierárquico*. Nesta abordagem, os objetos são organizados em uma hierarquia de grupos. Por sua vez, esta hierarquia pode ser construída por duas maneiras diferentes: *aglomerativa* e *divisiva*.

Na abordagem aglomerativa cada objeto de dados é inicialmente incluído em seu próprio grupo. Em seguida, cada grupo é aglomerado com o seu grupo mais próximo, formando uma relação "pai-filho"entre o grupo resultante e os grupos menores. Esse processo se repete até que um único grupo, que contenha todos os dados do conjunto, seja obtido. Já na abordagem divisiva o processo se inverte. Todos os objetos de dados são agrupados em um único grupo inicial, que será a raiz da hierarquia. Por sua vez, este grupo inicial é sucessivamente dividido em grupos menores, até que cada objeto esteja em seu próprio grupo.

Na subseção 2.4 a abordagem hierárquica será explorada com mais detalhes.

2.4 Abordagem Hierárquica para Agrupamentos

Na abordagem hierárquica de agrupamento, os grupos são organizados em árvore, de forma que cada nodo desta árvore representa um grupo. Desta maneira, estabelece-se uma relação pai-filho entre os grupos, tal que, dado um grupo pai C_p que tenha n filhos $\{C_1, C_2, ..., C_n\}$, então $C_i \subset C_p$ para todo $1 \leq i \leq n$, como mostra a Figura 1(b). Nas folhas desta árvore, encontram-se as instâncias de dados, cada uma incluída em seu próprio grupo. Por sua vez, na raiz encontra-se o grupo que abrange todo o conjunto de dados. A Figura 1(a) mostra a hierarquia entre os grupos através de um diagrama conhecido como *dendrograma*.

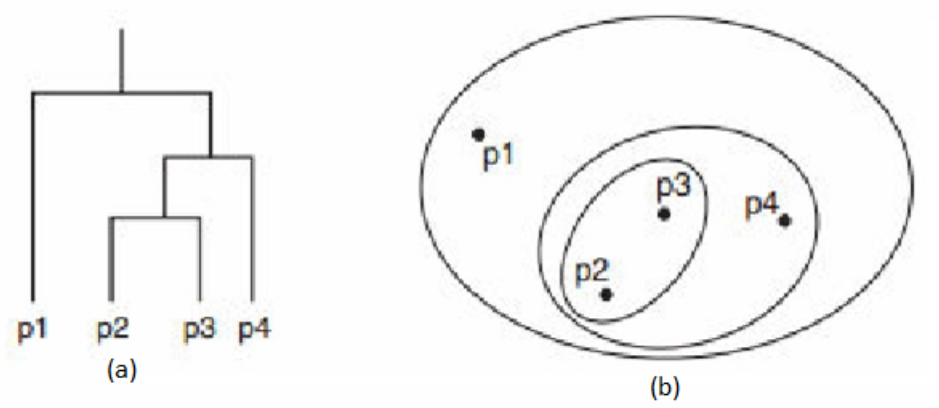

Figura 1: Exemplo de agrupamento hierárquico [Tan et al. 2009].

A organização de grupos em hierarquias tem aplicações em diversas áreas. Em recuperação da informação por exemplo, documentos podem ser agrupados de maneira hierárquica de acordo com o assunto, revelando uma estrutura de tópicos e sub-tópicos. Em Biologia, o agrupamento de espécies de acordo com suas características pode ajudar na compreensão sobre como essas espécies evoluíram ao longo do tempo.

Uma das principais vantagens da abordagem hierárquica de agrupamento é que não é necessário que o usuário informe o número k de grupos desejado previamente. Porém, nem sempre é interessante para o usuário analisar todos os grupos obtidos pelo agrupamento hierárquico. Nestes casos, a hierarquia original pode ser convertida em um particionamento através do procedimento de poda (*prunning*).

Nesta subseção serão discutidos os principais aspectos relativos ao agrupamento hierárquicos. Serão abordados: algoritmo AGNES para agrupamento aglomerativo, algoritmo DIANA para agrupamento hierárquico divisivo e as principais medidas de distância entre clusters: single linkage, complete linkage, average linkage, mean distance e Ward's distance. Por fim serão analisados vantagens e desvantagens da abordagem hierárquica.

2.4.1 AGNES: algoritmo aglomerativo

O algoritmo AGNES (*AGglomerative NESting*) é o algoritmo elementar para executar o agrupamento hierárquico aglomerativo. 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, como mostra o Algoritmo 1.

Algoritmo 1: Algoritmo aglomerativo para agrupamento hierárquico

Entrada: conjunto de dados X
Saída: agrupamento hierárquico
$n \leftarrow |X|$
$\mathcal{C} \leftarrow \{C_i = \{x_i\} \,|x_i \in X\}$
$M \leftarrow (m_{ij})_{n \times n} \mid m_{ij} = dist\,(C_i, C_j)\,, C_i, C_j \in \mathcal{C}$
repita
  encontra o par de grupos mais próximos C_i e C_j
  $C \leftarrow C_i \cup C_j$
  $\mathcal{C} \leftarrow (\mathcal{C} \setminus \{C_i, C_j\}) \cup \{C\}$
  recalcula M
até $C \equiv X$

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.

Dado o algoritmo básico do agrupamento hierárquico aglomerativo, a principal diferença entre as diferentes abordagens nesta categoria são as medidas de similaridade entre grupos usadas. Por sua vez, estas podem ser baseadas em grafos ou baseadas em protótipos [Tan et al. 2009]. São medidas de distância baseadas em grafos:

- **Ligação simples**: A ligação simples, ou *single linkage*, toma como similaridade entre dois grupos a distância entre seus elementos mais próximos, e é calculada pela Equação 1:

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

  Ao tomar-se as instâncias de dados como vértices de um grafo, e as ligações entre grupos como vértices ponderados, então o agrupamento gerado é correspondente a uma *árvore geradora mínima* [Han et al. 2011], de forma que os grupos formados tendem a ser contíguos no espaço dos atributos [Tan et al. 2009].

- **Ligação completa**: Por sua vez, a ligação completa, ou *complete linkage* é a medida oposta à ligação simples, pois toma como similaridade entre dois grupos a distância entre seus elementos mais distantes, sendo calculada pela Equação 2:

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

- **Distância Média**: Por fim, a similaridade entre dois grupos pode ser medida através da distância média entre os pares dos grupos (*group average*). Essa medida é um balanceamento entre a ligação simples e a ligação completa, e é obtida pela média das distâncias entre cada um dos pares ordenados (x_i, x_i), com $x_i \in C_i$ e $x_i \in C_j$. É obtida pela Equação 3:

$$dist\,(C_i, C_j) = \frac{1}{n_i n_j} \sum_{x_i \in C_i} \sum_{x_j \in C_j} |x_i - x_j| \tag{3}$$

2.4.2 DIANA: algoritmo divisivo

2.5 Fractais e a propriedade de auto-similaridade

Um fractal pode ser definido pelo conceito de auto similaridade, no qual partes de qualquer tamanho de um fractal são similares (exata ou estatisticamente) ao conjunto todo. Um exemplo clássico de um fractal criado por meio da construção repetitiva é o triângulo de Sierpinski, construído por meio de um processo iterativo, onde se retira de um triângulo o triângulo central e para cada triângulo resultante realiza-se o mesmo processo, recursivamente, conforme apresentado na Figura 2. O triângulo de Sierpinski apresenta características interessantes, como o fato de cada triângulo interior ser uma miniatura do triângulo em que está inserido, perímetro tendendo ao infinito e área tendendo a zero quando o número de iterações tende ao infinito [Schroeder 1991].

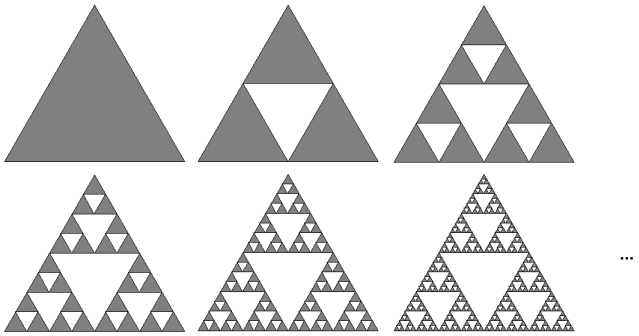

Figura 2: Construção do triângulo de Sierpinski [Schroeder 1991].

O conceito de auto similaridade está relacionado com periodicidade. Muitos fenômenos naturais e humanos acontecem com periodicidade. São exemplos o uso do solo pela agricultura, o valor das moedas e das ações, o comportamento de redes de comunicação, as filas de caixa de supermercado, o uso das estradas, as músicas que tocam nas rádios e os efeitos da economia na vida das pessoas. Uma razão para a universalidade dessas movimentações harmônicas é a linearidade aproximada de muitos sistemas e a sua invariância com deslocamento no espaço e tempo. Experimentos realizados com alguns conjuntos de dados sintéticos e reais mostram que os dados referentes aos fenômenos humanos caracterizam-se por apresentarem uma distribuição fractal [Traina-Jr. et al. 2010].

Para a análise da auto similaridade de conjuntos contendo fenômenos naturais e humanos, chamados de fractais estatisticamente auto-similares, utiliza-se o método *Box-Counting*. Para encontrar o *Box-Counting* de um conjunto de dados imerso em um espaço

E-dimensional, deve-se dividir esse espaço em células de um hipercubo de lado r, recursivamente, até encontrar um elemento por célula [Traina-Jr. et al. 2010]. O método foi proposto para cálculo da dimensão fractal, que corresponde ao número mínimo de dimensões para representação de um conjunto (dimensionalidade intrínseca).

Ainda preciso escrever sobre box plot com base em [Traina-Jr. et al. 2010].

3 Métodos de Pesquisa

4 Agrupamento hierárquico aglomerativo de séries espaço-temporais

descrever a tecnica proposta!!!

5 Resultados Preliminares

descrever os resultados preliminares!!!

6 Cronograma de Execução

Uberlândia, 14 de Dezembro de 2015.

Assinatura do Orientador:

Assinatura do Aluno:

Referências

[García et al. 2015] García, S., Luengo, J., and Herrera, F. (2015). *Data Preprocessing in Data Mining*, chapter Data Reduction, pages 147–162. Springer International Publishing.

[Gonçalves et al. 2014] Gonçalves, R., Zullo, J., Amaral, B. F. d., Coltri, P. P., Sousa, E. P. M. d., and Romani, L. A. S. (2014). Land use temporal analysis through clustering techniques on satellite image time series. In *Geoscience and Remote Sensing Symposium (IGARSS), 2014 IEEE International*, pages 2173–2176. IEEE.

[Han et al. 2011] Han, J., Pei, J., and Kamber, M. (2011). *Data Mining: Concepts and Techniques*. The Morgan Kaufmann Series in Data Management Systems. Elsevier Science.

[Petitjean et al. 2014] Petitjean, F., Forestier, G., Webb, G. I., Nicholson, A. E., Chen, Y., and Keogh, E. (2014). Dynamic time warping averaging of time series allows faster and more accurate classification. In *Data Mining (ICDM), 2014 IEEE International Conference on*, pages 470–479. IEEE.

[Schroeder 1991] Schroeder, M. (1991). *Fractals, Chaos, Power Laws: Minutes From an Infinite Paradise*. W. H. Freeman.

[Tan et al. 2009] Tan, P., Steinbach, M., Kumar, V., and FERNANDES, A. (2009). *Introdução ao datamining: mineração de dados*. Ciencia Moderna.

[Traina-Jr. et al. 2010] Traina-Jr., C., Traina, A. J. M., Wu, L., and Faloutsos, C. (2010). Fast feature selection using fractal dimension. *Journal of Information and Data Management (JIDM)*, 1(1):3–16.