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: Março de 2014

Previsão da Defesa: Novembro de 2016

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:

1. Desenvolver uma abordagem para o agrupamento hierárquico aglomerativo de séries espaço-temporais incluindo uma etapa de pré-processamento baseada em redução de dados, especificamente por meio da geração de amostragens aleatória uniforme

2. Avaliar experimentalmente a nova abordagem 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

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".

1.3 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.

2.2 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.

3 Métodos de Pesquisa

4 Agrupamento hierárquico aglomerativo de séries espaço-temporais

descrever aqui a tecnica proposta!!!

5 Resultados Preliminares

descrever aqui os resultados preliminares!!!

6 Cronograma de Execução

Uberlândia, 14 de Dezembro de 2015.

Assinatura do Orientador:

Assinatura do Aluno:

Referências

[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.

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