Uso do Método SKATER Para Agrupamento de Dados sem Informações Espaciais

Mateus Costa Soares

22 fevereiro 2018

Capa (Título, Identificação do(s) Aluno(s), do Orientador, da Disciplina e do Cur

Folha de Rosto

Sumário

Introdução (contendo breve revisão da literatura)

Em estatística espacial, um problema frequente dos pesquisadores é uma base de dados espaciais muito granular (com muitos objetos), o que pode ocasionar a detecção de tendências locais e a não detecção de tendências globais. Para contornar esse problema é comum utilizar técnicas de regionalização. Atualmente há quatro tipos de técnicas de regionalização: agrupamento não espacial seguida de processamento espacial, agrupamento não espacial com uma medida de dissimilaridade ponderada espacialmente, busca de tentativa e erro com otimização e agrupamento restringida espacialmente e particionamento (GUO, 2008). Neste último grupo estão métodos como o SKATER e o REDCAP, os quais foram testados e se mostraram significativamente superiores aos outros. Não poderiam, então, esses métodos serem utilizados para agrupamento de dados não espaciais? Se sim, seriam eles eficientes? Para responder essa pergunta serão utilizadas bases de dados simuladas para comparar a classificação realizada por esses métodos com métodos usuais de classificação de dados não espaciais, como k-means, agrupamento hierárquico, entre outros.

Objetivos gerais e específicos

Objetivos

Testar, por meio de dados simulados a qualidade da classificação dos métodos SKATER e REDCAP para dados não espaciais, em comparação com métodos usuais (k-means, agrupamento hierárquico, entre outros)

Material e Métodos

Introdução aos Algorítmos

Agrupamento é um problema de aprendizagem não supervisionada, onde busca-se encontrar grupos com indivíduos similares entre si e dissimilares à indivíduos em outros grupos. Para entendermos melhor os algoritmos a serem testados para a classificação, temos nessa seção uma breve introdução aos mesmos. Definiremos x_ij como o objeto i do grupo j, c_j o centroide do grupo j e (x_ij-c_j)^2 uma medida de distância entre um ponto do grupo j e seu respectivo centroide.

K-Means

K-means é um método que classifica os dados em k grupos, sendo k um número arbitrário, fornecido pelo usuário. O método distribui aleatoriamente os dados nos grupos, então são formados k centroides. Após

esse passo inicial, iterativamente, os dados são realocados aos centroides mais próximos, então os centroides são recalculados. Esse procedimento continua até os centroides ficarem estabilizados. A função objetiva do método é: J= ???*(j=1)ˆk??????*(i=1)ˆn??????(x_ij-c_j)²???

Agrupamento Hierárquico

Esse conjunto de métodos inicia com n grupos, com um dado em cada grupo e é calculada uma matriz de tamanho n*n contendo as distâncias entre os grupos. Iterativamente, os grupos com menor distância são unidos, resultando em um novo grupo, é recalculada a matriz de distâncias. Esse procedimento é realizado até todos os dados estarem em um grupo só. É feita então a representação dos passos em uma árvore hierárquica, conforme exemplo abaixo:

A partir da árvore é possível escolher o melhor número de agrupamentos. Para calcular a matriz de distâncias são utilizadas três técnicas: . Ligação simples: A distância entre dois grupos é igual a menor distância entre um membro do primeiro grupo e um membro do segundo grupo. . Ligação completa: A distância entre dois grupos é igual a maior distância entre um membro do primeiro grupo e um membro do segundo grupo. . Ligação média: A distância entre dois grupos é igual a distância média entre os membros do primeiro grupo e os membros do segundo grupo.

SKATER

O SKATER é um método que consiste em duas etapas: primeiramente é produzida uma árvore de custo mínimo (MST - Minimum Spanning Tree) em seguida a mesma é particionada k-1 vezes, gerando k grupos. Para obter-se a MST, é necessário primeiramente obter um grafo de conectividade. Em problemas de agrupamento espacial, o grafo liga o centroide de uma região aos centroides de todas as regiões fronteiriças através de arestas. Neste trabalho proporemos a sua utilização em problemas não espaciais, então uma sugestão é iniciar com o grafo completo, no qual cada nó está conectado com todos os demais. Cada aresta do grafo tem um custo proporcional à dissimilaridade dos objetos que estão sendo ligados. A MST consiste em um grafo com as n-1 arestas de menor custo, onde todos os dados estão ligados e não há circuitos. Temos então a MST. Para construir a MST, parte-se de um ponto inicial qualquer, este ponto é incluído na árvore T_1. Encontra-se a aresta de menor custo partindo da arvore T_1 e esse ponto é então inserido na árvore, produzindo assim a árvore??? T???*2. Esse procedimento é repetido até todos os pontos estarem ligados. Para medirmos a homogeneidade, temos a seguinte função: SSD= ???*(j=1)ˆm??????_(i=1)ˆ(n_j)??????(x_ij- x ??_j)²??? Para reduzirmos a MST, cortamos a aresta na qual a SSD seja a menor possível, ou seja, com a menor função objetivo: f_1 (S_iˆT)= ???SSD???_T-(???SSD???_Ta+ ???SSD???_Tb) Onde T é a árvore anterior e Ta e Tb são as árvores resultantes da remoção de uma aresta. É realizado o procedimento de remoção de arestas até obter-se o número de grupos desejado.

Redcap

Similar ao SKATER, esse método parte de um grafo de conectividade e contém seis diferentes abordagens para a construção da MST: . Agrupamento com ligação simples e com restrição de primeira ordem (First-Order-SLK): Parte do grafo restringido à objetos espacialmente ligados, cada objeto começa como um grupo, todos os custos das arestas são listados em ordem crescente e os menores são ligados, unindo os grupos, esse procedimento é realizado até todos os grupos estarem ligados; . Agrupamento com ligação média e com restrição de primeira ordem (First-Order-ALK): É semelhante a abordagem anterior, porém, após cada união de grupos, os custos entre o novo grupo e os demais são recalculados, considerando a distância como a média entre as distâncias de todos os dados dos grupos; . Agrupamento com ligação completa e com restrição de primeira ordem (First-Order-CLK): Igual ao First-Order-SLK, porém a distância entre os grupos é definida pela maior distância entre os indivíduos do grupo. . Agrupamento com ligação simples e com restrição de ordem total (Full-Order-SLK): Parte do grafo total e da matriz de contiguidade C, todos os custos das arestas são listados em ordem crescente e os menores são ligados, caso dois grupos contíguos forem ligados, a menor aresta contígua será adicionada ao grupo. A matriz de contiguidade é então atualizada, sendo as fronteiras do grupo consideradas para todos os objetos que o formam. . Agrupamento com ligação média e com restrição

de ordem total (Full-Order-ALK): Semelhante ao anterior, mas considerando a distância dos grupos como a média entre as distâncias de todos os indivíduos dos grupos. . Agrupamento com ligação completa e com restrição de ordem total (Full-Order-CLK): Igual ao Full-Order-SLK, porém a distância entre os grupos é definida pela maior distância entre os indivíduos do grupo. Para realizar a partição da árvore é adotado o mesmo procedimento do algoritmo SKATER.

Bases de Dados

Serão simuladas várias bases de dados, com distribuições diferentes para os testes de comparação dos métodos;

Recursos Computacionais

O software R, versão 3.4.2 (R CORE TEAM, 2017) será utilizado para testar os métodos de agrupamento. Os algoritmos estão contidos em pacotes auxiliares, com exceção do REDCAP, que deverá ser implementado com base na implementação do SKATER.

Cronograma das atividades

Referências (usar normas da ABNT ou da UFPR)