Trabalho apresentado no CNMAC, Gramado - RS, 2016.

Proceeding Series of the Brazilian Society of Computational and Applied Mathematics

Uma Comparação Entre o Algoritmo K-means e o Algoritmo Espectral Para Agrupamento de Dados com Curvatura Acentuada

Luciano G. Garcia[1]
Instituto de Matemática, Estatística e Física, FURG, Rio Grande, RS
Leonardo R. Emmendorfer[2]
Centro de Ciências Computacionais, FURG, Rio Grande, RS

1 Introdução

Diversos métodos de agrupamento de dados buscam encontrar agrupamentos ótimos, porém surgem certas limitações correspondentes a cada algoritmo utilizado. Com isto, existe uma grande necessidade de encontrar métodos eficientes que realizem esta tarefa de maneira bem geral, independente do formato do conjunto de dados. A principal tarefa do método espectral é obter informações da estrutura organizacional dos dados a partir da relação de similaridade entre eles, de modo que os agrupamentos acontecam de forma mais natural do que outros algoritmos como, por exemplo, o *k-means* [1].

2 O Método Espectral

De acordo com [2], o algoritmo pode ser resolvido eficientemente por um software de Álgebra Linear, e muitas vezes supera algoritmos de agrupamento tradicionais como o *k-means*. Para uma melhor compreensão do método, considere um conjunto com n pontos $x_1, ..., x_n$. Cada par de pontos é relacionado por um parâmetro de similaridade levando em consideração a Distância Euclidiana $f(d(x_i, x_j), \theta)$, e assim, é contruída uma matriz simétrica e positiva com as relações de similaridades entre cada par

$$W_{ij} = f(d(x_i, x_j), \theta). \tag{1}$$

Calcula-se a matriz Laplaciana L a partir da diferença da matriz diagonal de pesos dos vértices D e da matriz de similaridade W. Após, obtém-se os autovetores da matriz Laplaciana, fazendo um mapeamento nos autovetores a fim de obter um corte no grafo que representa os dados, sugerindo a criação de agrupamentos.

[1]lucianogarim@gmail.com
[2]leonardo.emmendorfer@gmail.com

O cálculo da matriz Laplaciana pode váriar dependendo de como os dados do conjunto estão posicionados. No caso em que os dados estão mais dispersos uns dos outros o algoritmo que utiliza a Laplaciana Normalizada Simétrica é mais recomendável [3]; assim define-se a matriz L de acordo com (2):

$$L = D^{-\frac{1}{2}}(D - W)D^{-\frac{1}{2}} \tag{2}$$

Outra possibilidade útil no caso de um conjunto de dados com pontos mais dispersos é o cálculo da Laplaciana Normalizada Assimétrica L_{ass} proposta por [4] onde o algoritmo utiliza os autovetores generalizados da matriz L os quais correspondem aos autovetores de L_{ass}. Quando quaisquer um dos algoritmos espectrais são aplicados a um conjunto de dados não convexos, os resultados são os melhores possíveis se comparados ao algoritmo k-means. Pode-se concluir isto a partir da Figura 1 .

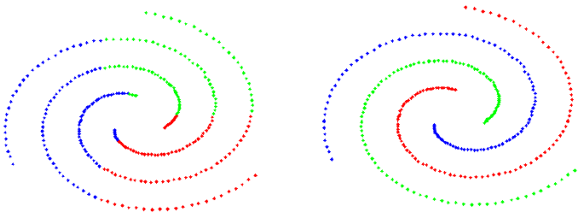

Figura 1: Banco de Dados Utilizado

3 Conclusões

Neste trabalho pode-se observar a eficiência do método espectral em agrupar, por exemplo, figuras que possuem curvatura acentuada conforme simulação feita utilizando o *software* Octave versão 4.0.0. Utilizando resultados obtidos da literatura com o método k-means não é possível obter o agrupamento esperado como pode-se ver na Figura 1 à esquerda, já à direita o algoritmo espectral resulta em um agrupamento ótimo de um ponto de vista prático.

Referências

[1] M. Jordan and Y. Weiss. On spectral clustering: Analysis and an algorithm, *In Advances in Neural Information Processing Systems*, volume 14, 2001.

[2] U. V. Luxburg. A Tutorial on Spectral Clustering, *Stat. and Comp.*, volume 17, 2007.

[3] M. W. Mahoney, L. Orecchia and N. K. Vishnoi. A Local Spectral Method for Graphs: With Applications to Improving Graph Partitions and Exploring Data Graphs Locally, *Journal of Machine Learning Research*, volume 13, 2012.

[4] J. Shi and J. Malik. Normalized cuts and image segmentation, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, volume 12, 2000.