PENALIZAÇÃO POR NÃO-CONECTIVIDADE PONDERADA DE GRAFOS

Spencer Barbosa da Silva[1], Anderson Ribeiro Duarte[1]

Resumo: *O problema de detecção e inferência de clusters vem sendo recentemente tratado em muitos trabalhos através de técnicas de otimização. Recentes medidas de penalização são associadas à Estatístitica Scan Espacial para a detecção de clusters irregulares. Uma destas medidas é a de Não Conectividade, que se mostra bastante eficaz no auxílio para a detecção. Entretanto tal medida apresenta dificuldades para interpretar as diferenças existentes quanto a importância de cada conexão dentro de um possível cluster. Será proposta uma estratégia de ponderação para os termos associados à medida de Não conectividade visando aumentar a eficiência da medida anterior para detecção de clusters irregulares.*

Palavras-chave: Teoria de Grafos, Otimização Multi-objetivo, Penalização por Não-Conectividade, Penalização por Não-Conectividade Ponderada.

Introdução

Em geral, ao trabalhar com conglomerados espaciais, se tem por interesse mapear agrupamentos de regiões nos quais a incidência de um determinado fenômeno de interesse é discrepante, ou seja, muito acima ou abaixo dos valores esperados. São muito comuns estudos de fenômenos associados à doenças que em geral apresentam riscos para a saúde pública. Entretanto diversos outros fenômenos podem ser avaliados como: poluição, criminalidade, entre outros.

Diversos métodos de detecção e inferência para clusters espaciais se baseiam na maximização da estatística *Scan Espacial*. A referida estatística se encontra detalhada nos artigos de Kulldorff and Nagarwalla (1995)[6] e Kulldorff (1997)[7]. Ela é definida como uma razão de verossimilhança e busca identificar o cluster mais verossímil dentre algumas possíveis configurações de clusters no mapa em estudo.

Os métodos que utilizam a maximização da estatística Scan Espacial são bastante difundidos, mas apresentam algumas deficiências quando os clusters não apresentam formato regular (por exemplo, conjuntos não circulares) que são bastante comuns. Neste estudo, é bastante frequente a existência de clusters com formatos bastante irregulares.

Em diversos problemas os clusters não regulares podem ser observados como: problemas de tráfego, poluição, vigilância sindrômica. Em muitos destes casos, formatos não regulares se devem às características geográficas do mapa em estudo, tais como rios, regiões litorâneas, regiões montanhosas entre outras. Diversos métodos para a detecção de clusters com formatos irregulares já foram discutidos e pode-se encontrar uma interessante revisão bibliográfica deste assunto em Duczmal et al. (2009)[5].

O recorrente problema da irregularidade da forma dos possíveis clusters vem sendo discutido com profundidade nos últimos anos. Muitas estruturas para controlar a forma das possíveis soluções apresentadas pelos métodos de detecção já foram apresentadas. Cada uma destas

[1]Departamento de Matemática, ICEB, UFOP,
spencerbars@gmail.com, anderson@iceb.ufop.br

formas apresenta vantagens e desvantagens que, em geral, são inerentes à composição espacial do mapa completo, bem como à distribuição populacional ao longo do mapa.

Todos estes métodos são bastante importantes em estudos de vigilância sindrômica e também no estudo epidemiológico dentre outra possíveis áreas de aplicação.

Objetivos

Neste trabalho será utilizada uma proposta multi-objetivo para o procedimento de detecção de clusters. É importante salientar que neste momento não é objetivo principal apresentar detalhes profundos sobre a estratégia otimizadora.

Tem-se por objetivo principal, apresentar uma proposta de função objetivo que seja eficiente no processo de detecção de clusters. A função objetivo é denominada *Função de penalização por Não Conectividade ponderada de Grafos*. Esta função se baseia na função de penalização por Não Conectividade apresentada anteriormente por Yiannakoulias et al.(2007)[9].

Além disto observar que a função proposta tem aplicação não somente na área de detecção de clusters, mas pode ser utilizada em diversos problemas que envolvem a Teoria de Grafos.

Metodologia

A medida de penalização por Não Conectividade proposta por Yiannakoulias et al.(2007)[9] se baseia no subgrafo associado ao cluster candidato e se mostra bastante eficiente na detecção e inferência de clusters. Entretanto o formato desta penalização leva em conta apenas a contagem das arestas do subgrafo associado ao cluster candidato. Não existe uma consideração quanto ao grau de importância de uma aresta na conexidade do subgrafo. Em outras palavras, a pergunta interessante seria se existem arestas mais ou menos relevantes para a conexidade do grafo. Pensando apenas na análise do grafo, é fato que tal relevância não precisa ser considerada. Quando observamos que estamos trabalhando com subgrafos associados a conjuntos de regiões em um mapa, lembramos que as arestas são conexões de vizinhança entre regiões que podem ser muito ou pouco populosas. Neste contexto, observamos que existem sim arestas mais e menos importantes para a conexidade do subgrafo associado a um cluster candidato. A mesma análise pode ser realizada para o grau de importância de cada um dos vértices do subgrafo em estudo.

A medida de penalização por Não Conectividade é dada por:

$$y(z) = \frac{a(z)}{3\,(v(z) - 2)}$$

Ela se baseia em uma relação do número de vértices $v(z)$ e de arestas $a(z)$ do subgrafo associado ao cluster candidato z.

Para a medida ponderada, será estabelecida uma ponderação para os vértices e arestas do subgrafo associado a um cluster candidato. Tal ponderação será construída pensando na estrutura da distribuição populacional ao longo das regiões deste cluster candidato.

A ponderação das arestas do subgrafo associado ao cluster candidato z será definida pela média entre as populações das regiões cujos vértices são conectados pela aresta em questão. Portanto para uma aresta $a_{i,j}$ conectando os vértices v_i e v_j associados às regiões Ri e Rj com populações $pop(Ri)$ e $pop(Rj)$, teremos o seguinte peso ponderador:

$$P(a_{i,j}) = \frac{pop(Ri) + pop(Rj)}{2}$$

Já a ponderação dos vértices será dada pela população da região associada ao respectivo vértice, ou seja, para o vértice v_i associado à região Ri cuja população é $pop(Ri)$, teremos o seguinte peso ponderador:

$$P(v_i) = pop(Ri)$$

Se considerarmos a figura 1 com regiões definidas por $R1$, $R2$, $R3$ e $R4$, podemos construir diferentes cenários de distribuição populacional e então verificar a poderação das arestas e dos vértices para diferentes distribuições populacionais.

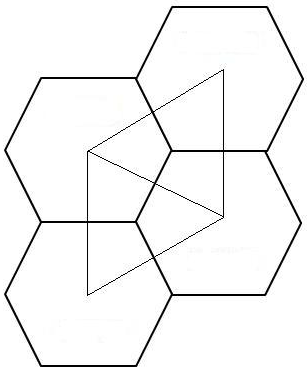

Figura 1: Possível cluster

Um destes possíveis cenários é ilustrado na figura 2 com as populações das regiões sendo: $pop(R1) = 1000$, $pop(R2) = 400$, $pop(R3) = 400$ e $pop(R4) = 800$.

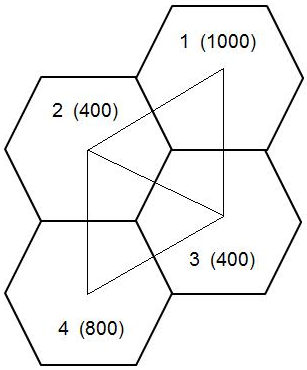

Figura 2: Ponderação de arestas e vértices

Teríamos então os seguintes pesos para as arestas:

$$P(a_{1,2}) = \frac{1000 + 400}{2} = 700, \quad P(a_{1,3}) = \frac{1000 + 400}{2} = 700, \quad P(a_{2,3}) = \frac{400 + 400}{2} = 400,$$
$$P(a_{2,4}) = \frac{400 + 800}{2} = 600, \quad P(a_{3,4}) = \frac{400 + 800}{2} = 600.$$

Já para os vértices os seguintes pesos:

$P(v_1) = 1000$, $P(v_2) = 400$, $P(v_3) = 400$, $P(v_4) = 800$.

Utilizando o mesmo formato de cluster, quando consideramos um outro cenário de distribuição populacional, os pesos associados às arestas ficam modificados.

Considere neste novo cenário as populações $pop(R1) = 400$, $pop(R2) = 1000$, $pop(R3) = 800$ e $pop(R4) = 400$, isto pode ser observado na figura 3.

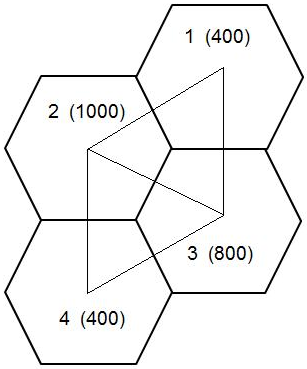

Figura 3: Ponderação de arestas e vértices

Teríamos então os seguintes pesos para as arestas:

$$P(a_{1,2}) = \frac{400+1000}{2} = 700, \quad P(a_{1,3}) = \frac{400+800}{2} = 600, \quad P(a_{2,3}) = \frac{1000+800}{2} = 900,$$
$$P(a_{2,4}) = \frac{1000+400}{2} = 700, \quad P(a_{3,4}) = \frac{800+400}{2} = 600.$$

Para os vértices os seguintes pesos:

$P(v_1) = 400$, $P(v_2) = 1000$, $P(v_3) = 800$, $P(v_4) = 400$.

Para reformular a função descrita, susbtituiremos as arestas e vértices por seus respectivos pesos ponderadores da seguinte forma:

$$yp(z) = \frac{\sum_{i=1}^{k-1} \sum_{j=i+1}^{k} P(a_{i,j})}{3\left[\sum_{i=1}^{k} P(v_i) - 2\left(\frac{\sum_{i=1}^{k} P(v_i)}{k}\right)\right]}$$

em que k é a quantidade de regiões no cluster candidato z.

Alguma dúvida pode pairar sobre o termo $\frac{\sum_{i=1}^{k} P(v_i)}{k}$ associado ao valor 2 no denominador, entretanto se pensarmos na suposição de todas as populações idênticas nas regiões da zona a ser avaliada, se faz necessário este termo para que tenhamos $y(z) = yp(z)$ para esta situação específica.

Com este novo formato estaremos levando em conta não somente a estrutura do subgrafo associado à zona z, mas também informações inerentes a estrutura da distribuição populacional dentro da zona z e o grau de relevância das vizinhanças entre regiões quanto às suas populações.

Voltando aos exemplos descritos através das Figuras 2 e 3 podemos observar o efetivo efeito da utilização deste formato de função de penalização. No exemplo apresentado através da Figura 2 as regiões mais populosas (R1 e R4) não estão conectadas, ou seja, não são vizinhas. Já no exemplo da Figura 3 as regiões mais populosas (R2 e R3) estão conectadas, ou seja, são vizinhas. A motivação deste formato de penalização fica evidênciada nesta situação. Em outras palavras, nossa suposição se baseia no fato de acreditarmos que vizinhanças de regiões mais populosas

devem gerar maior movimentação entre habitantes e portanto tendendo a reforçar a proliferação do fenômeno de interesse. Neste caso estaríamos reforçando o cluster em questão.

Note que utilizando a medida de Não Conectividade proposta anteriormente, os exemplos das Figuras 2 e 3 apresentariam a mesma medida de Não Conectividade que neste caso seria:

$$y(z) = \frac{5}{3\,(4-2)} = 0.833$$

Agora utilizando a medida de Não Conectividade Ponderada, as medidas seriam diferentes em cada um dos exemplos.

No exemplo da Figura 2 teríamos:

$$yp(z) = \frac{700 + 700 + 400 + 600 + 600}{3\left(1000 + 400 + 400 + 800 - 2\left(\frac{2600}{4}\right)\right)}$$

$$yp(z) = 0.769$$

No exemplo da Figura 3 teriamos:

$$yp(z) = \frac{700 + 600 + 900 + 700 + 600}{3\left(400 + 1000 + 800 + 400 - 2\left(\frac{2600}{4}\right)\right)}$$

$$yp(z) = 0.897$$

confirmando então o objetivo da proposição da nova função de Não Conectividade Ponderada.

Conclusões

Observamos que a Função de Não-Conectividade é eficiente na avaliação do grau de conexidade de um grafo, entretanto para problemas nos quais existem diferenças entre a importância das arestas e vértices, se faz necessário o auxílio das ponderações propostas através da Função de Não-Conectividade Ponderada.

Para o problema específico da detecção de clusters espaciais, já foi implementado um algoritmo genético como estratégia de otimização. No caso foi utilizado o algoritmo genético *NSGA-II* Deb et al.(2002)[1]. Quanto ao poder de detecção de clusters, foi utilizado um benchmark com dados reais de cancêr de mama no nordeste dos Estados Unidos. Neste mapa foram construídos clusters artificiais e a qualidade no processo de detecção foi comparada entre o algoritmo genético multi-objetivo em que um dos objetivos é a Função de Não-Conectividade e outro algoritmo genético multi-objetivo em que um dos objetivos é a Função de Não-Conectividade Ponderada.

Foi detectado em média um poder de detecção 4% superior ao método anterior. Dado que o método anterior já era bastante eficiente, esta melhoria aparentemente pequena já pode ser observada como significativa. Outrossim para outras abordagens de problemas, os efeitos da utilização da metodologia ponderada podem se tornar ainda mais evidentes.

Um fato relevante é que não ocorreu perda em tempo computacional na execução do algoritmo seja com a Função de Não-Conectividade ou com a Função de Não-Conectividade Ponderada.

Referências

[1] Deb, K., Pratap, A., Agrawal, S. and Meyarivan, T., A fast and elitist multiobjective genetic algorithm: NSGA-II, *IEEE Transactions on Evolutionary Computation*, **2(6)**, 182-197, 2002.

[2] Duarte, A.R., Duczmal, L., Ferreira, S.J. and Cançado, A.L.F., Internal cohesion and geometric shape of spatial clusters, *Environmental and Ecological Statistics*, **17**, 203-229, 2010.

[3] Duczmal, L., Cançado, A.L.F., Takahashi, R.H.C. and Bessegato,L.F., A genetic algorithm for irregularly shaped spatial scan statistics, *Computational Statistics & Data Analysis*, **52**, 43-52, 2007.

[4] Duczmal, L., Cançado, A.L.F. and Takahashi, R.H.C., Geographic Delineation of Disease Clusters through Multi-Objective Optimization, *Journal of Computational & Graphical Statistics*, **17**, 243-262, 2008.

[5] Duczmal, L., Duarte, A.R. and Tavares, R., Extensions of the scan statistic for the detection and inference of spatial clusters, em "Scan Statistics" (Glaz J., Pozydnyakov V. and Wallestein S., eds.) pp. 157-182, Birkhäuser, Boston, 2009.

[6] Kulldorff, M. and Nagarwalla, N., Spatial disease clusters: detection and inference, *Statistics in Medicine*,**14**, 799-810, 1995.

[7] Kulldorff, M., A Spatial Scan Statistic, *Communications in Statistics: Theory and Methods*, **26(6)**, 1481-1496, 1997.

[8] Silva, S. B. Dissertação de Mestrado: *Penalização por Não-Conectividade Ponderada de Grafos*, Departamento de Estatística-UFMG, 2010.

[9] Yiannakoulias, N., Rosychuk, R.J. and Hodgson, J., Adaptations for finding irregularly shaped disease clusters, *International Journal of Health Geographics*, **6**, 28, 2007.