Agrupamento de sismos com técnicas de *clustering*

Gabriel Silva[1] Marcelo Ladeira[1] Claus Aranha[2]

[1]Departamento de Ciência da Computação
Universidade de Brasília

[2]Department of Computer Science
University of Tsukuba

8 de julho, 2016

Conteúdo

Conteúdo

Definição do Problema

Dados acerca de terremotos que ocorreram previamente ficam registrados em catálogos

A partir dos catálogos, pode-se separar os terremotos em *mainshocks* (terremotos independentes) e *aftershocks* (terremotos ao menos parcialmente dependentes dos principais)

Objetivos

Aplicar métodos tradicionais de *declustering* ao catálogo de terremotos da Agência Meteorológica Japonesa (JMA, do inglês *Japan Meteorological Agency*) para:

- Analisar se *mainshocks* seguem propriedades previstas em teoria
- Verificar se a utilização de métodos já consolidados de *declustering* pode melhorar a performance do GAModel, um modelo de previsão de riscos de sismos

Tentar propor um método que utilize GA para a realização de *declustering*, e analisar a sua viabilidade

Conteúdo

Clustering e *Declustering*

Clustering

Agrupar uma coleção de dados, de modo que dados pertencentes a um mesmo grupo sejam o mais similar possível (tal similaridade depende da aplicação em questão), e sejam diferentes dos dados pertencentes a outro grupo.

Declustering

Corresponde a técnica de, a partir de um cátalogo de sismos, separar os terremotos em duas classes: *mainshocks* e *aftershocks*. Assim, o nome *declustering* vem da possibilidade de, após separarmos os *mainshocks* dos *aftershocks*, removermos os *aftershocks* e usarmos apenas os *mainshocks* em um determinado problema, evitando informações redundantes.

Métodos tradicionais de *Declustering*

- Método da janela
- Método SLC

Método da Janela

Descrição

Para um determinado terremoto do catálogo, os outros terremotos são identificados como *aftershocks* caso eles tenham magnitude menor e ocorram próximos do primeiro terremoto, tanto em termos de tempo quanto em termos de distância física

Caso um segundo terremoto esteja próximo temporalmente do primeiro terremoto, diz-se que o segundo está na janela do tempo do primeiro e, caso o segundo terremoto tenha ocorrido próximo geograficamente do primeiro, diz-se que o segundo está na janela da distância do primeiro

Método da Janela

Algoritmo

Fórmula para o cálculo das janelas:

$$d(M) = 10^{0.1238*M+0.983}[km] \tag{1}$$

$$t(M) = \begin{cases} 10^{0.032*M+2.7389}, & \text{se } M \geq 6.5 \\ 10^{0.5409*M-0.547}, & \text{caso contrário} \end{cases} [dias] \tag{2}$$

Método SLC

Descrição

Método Single-Link Clustering, também existente em outras áreas do conhecimento

Hierárquico e aglomerativo: é construída uma hierarquia de *clusters*, começando com todos os terremotos em *clusters* separados e unindo *clusters* a cada etapa.

Método SLC

Algoritmo

1. Inicialmente cada terremoto encontra-se em um *cluster* próprio.
2. Calcule uma distância máxima D, tal que *clusters* que estejam a uma distância maior do que D não possam ser unidos em uma único *cluster*.
3. Una cada *cluster* ao *cluster* mais próximo dele, obtendo assim um único *cluster*. Não realize essa ação apenas caso a distância entre esses dois *clusters* exceda D.
4. Caso tenha ocorrido alguma união entre *clusters* no passo 3, volte ao passo 3. Caso nenhuma união tenha ocorrido ou o número de *clusters* seja igual a 1 o algoritmo termina.

Distância entre dois *clusters*: menor distância entre dois terremotos que pertençam a eles

Fórmula para calcular a distância entre dois terremotos:

$$d_{st} = \sqrt{d^2 + C^2 * t^2} \tag{3}$$

C é uma constante que permite a comparação entre tempo e distância (tendo portanto $km * dia^{-1}$ como unidade de medida)

tendo sido usado o valor $C = 1$.

Método SLC

Cálculo da distância máxima

Para o cálculo da distância máxima utilizou-se a fórmula:

$$D = 9.4 * \sqrt{S_1} - 25.2 \quad (4)$$

S_1: mediana de todas as distâncias entre os terremotos

Método SLC

Separação de *mainshocks* e *aftershocks*

Aplicando-se o algoritmo SLC, obtêm-se um agrupamento de terremotos de modo que para finalizar a tarefa de *declustering*, basta selecionar de cada *cluster* o terremoto mais representativo

Selecionou-se o mais próximo do centróide do *cluster*

A distribuição de Poisson como métrica para avaliar a qualidade de *declustering*

De modo geral, a distribuição de terremotos em um determinado catálogo não segue uma distribuição de Poisson no tempo

Apesar disso, para regiões tectônicas grandes o suficiente, os terremotos principais com magnitude maior que um dado limite (comumente em torno de 3.8 ou 4.0) são homogêneos no tempo, ou seja, seguem uma distribuição de Poisson

Logo, verificar se um catálogo de terremotos contendo apenas os *mainshocks* com magnitude maior que o limite segue uma distribuição de Poisson pode ser uma boa medida para avaliar a qualidade do método de *declustering*

GADec

Algoritmo genético para a tarefa de *declustering* proposto nessa pesquisa

Principais aspectos da modelagem:

- Indivíduo
- População Inicial
- Função de *Fitness*
- Demais operadores

Indivíduo é representado como um *array*, e a quantidade de elementos do vetor é igual ao número de terremotos sendo analisado

Cada elemento do *array* pode assumir apenas dois valores: 0 ou 1

- 0 - Terremoto correspondente à posição não é centróide do seu *cluster*
- 1 - Terremoto correspondente à posição é centróide do seu *cluster*

Como determinar a qual *cluster* um terremoto com valor 0 associado pertence?
Calcula-se a distância do terremoto a todos os centróides e escolhe-se o centróide com a menor distância

Métrica de distância utilizada é a mesma do método SLC

A partir do *array* do indivíduo pode-se obter todo o agrupamento de terremotos que o indivíduo representa

Custo computacional: terremotos que não são centróides devem ser associados ao centróide mais próximo

Um indivíduo inicial é gerado através de um processo de dois passos:

- Escolhe-se o número de centróides do indivíduo de modo aleatório, com a restrição de ser uma quantidade entre o número mínimo de *clusters* encontrado nos métodos anteriores e o número máximo de *clusters* encontrado nos métodos anteriores.
- Escolhe-se quais terremotos serão centróides para aquele indivíduo. Tal escolha é aleatória, respeitando a restrição de que o número de centróides escolhidos deve ser correspondente ao calculado no primeiro passo.

A população inicial é então obtida simplesmente repetindo o processo descrito anteriormente, até que a quantidade de indivíduos seja igual ao estabelecido previamente.

Utilizou-se a fórmula:

$$\sum_{1 \leq i \leq k} D_{inter}(C_i)w - D_{intra}(C_i) \tag{5}$$

$D_{intra}(C_i)$ é a distância intracluster para o *cluster* (C_i)

$D_{inter}(C_i)$ é a distância de C_i aos demais *clusters*

k é o número de *clusters* e w é um parâmetro definido pelo usuário

Para o GADec, considerou-se w como sendo igual a $\frac{1}{2}$

Objetivo: maximizar a função de *fitness*

Crossover

One-point crossover com probabilidade de aplicação de 0.9

Mutação

flip-bit: *bits* do *array* do indivíduo são modificados (se valem 1 passam a valer 0, se valem 0 passam a valer 1)
Probabilidade de aplicação: 0.1

Seleção

Deterministic Tournament Selection, em que torneios com 5 indivíduos foram realizados e os 2 indivíduos mais aptos foram selecionados para *crossover*.

Elitismo

Os dois indivíduos mais aptos são colocados como indivíduos da geração seguinte.

Conteúdo

Catálogo da JMA

Dados acerca de terremotos coletados pela JMA, começando no início do ano 2000 e terminando no fim do ano de 2012

Para cada um dos terremotos, é conhecida sua latitude, longitude, magnitude, profundidade, data e momento de ocorrência

Divisão em regiões

Regiões usadas na pesquisa do GAModel:

- **Kanto**: região com latitude entre 34.8 e 37.05 e com longitude entre 138.8 e 141.05.
- **Kansai**: região com latitude entre 34.0 e 36.0 e com longitude entre 134.5 e 136.5.
- **Tohoku**: região com latitude entre 37.0 e 41.0 e com longitude entre 139.8 e 141.8.
- **Leste do Japão**: região com latitude entre 37.0 e 41.0 e com longitude entre 140.0 e 144.0.

Imagem da área abrangida pelo catálogo

Quantidade de terremotos por região

Região Geográfica	**Quantidade de terremotos**
Todo o Catálogo	220 195
Kanto	15 694
Kansai	1 869
Tohoku	14 072
Leste do Japão	43 561

Conteúdo

Testes de chi-quadrado

Para cada um dos métodos e cada uma das regiões, avaliou-se caso os *mainshocks* obtidos seguiam a distribuição de Poisson ou não

Utilização do teste de chi-quadrado para avaliar a hipótese nula H_0 = "Os terremotos seguem uma distribuição de Poisson no tempo"

Foram feitos 2 testes:

- No primeiro, consideram-se apenas *clusters* com terremotos com magnitude maior ou igual a 3.8
- No segundo, consideram-se apenas *clusters* com terremotos com magnitude maior ou igual a 4.0

Em ambos os testes são considerados intervalos de 10 dias

Como avaliar se o uso de técnicas de *declustering* melhora a performance do GAModel?

O GAModel é um modelo de previsão de risco de sismos que, a partir de um registro contendo dados de terremotos de anos anteriores, tenta prever onde ocorrerão os terremotos de um determinado ano

Avaliação do impacto dos métodos de *declustering* na performance do GAModel

1. Sendo GA estocásticos, executam-se 10 simulações do GAModel quando este recebe como entrada o catálogo original. Para este caso, o GAModel faz um pequeno filtro, mantendo apenas terremotos com magnitude maior do que 3.0 e com profundidade menor que 25km.
2. Executam-se 10 simulações do GAModel quando este recebe como entrada o catálogo declusterizado, ou seja, contendo apenas os dados relativos aos *mainshocks*.
3. Compara-se os desempenhos do GAModel com e sem o uso de *declustering* através de um teste de Kruskal-Wallis, permitindo assim verificar se é possível rejeitar a hipótese nula $H_0 =$ "A média de performance dos dois grupos é igual".

Problemas encontrados nos métodos

- Método SLC
- GADec

Problemas da distância no método SLC

Cálculo da distância máxima D conforme a fórmula 4 não foi possível (fez-se tal cálculo e obteve-se um valor negativo)

Escolheu-se então D como 12.0 (valor já trabalhado em estudos anteriores)

Problemas da eficiência no método SLC

Tempo de execução do algoritmo lento demais para que fosse executado para todo o catálogo

Simulações realizadas apenas para região de Kansai

10 simulações, com 50 indivíduos por geração e 100 gerações

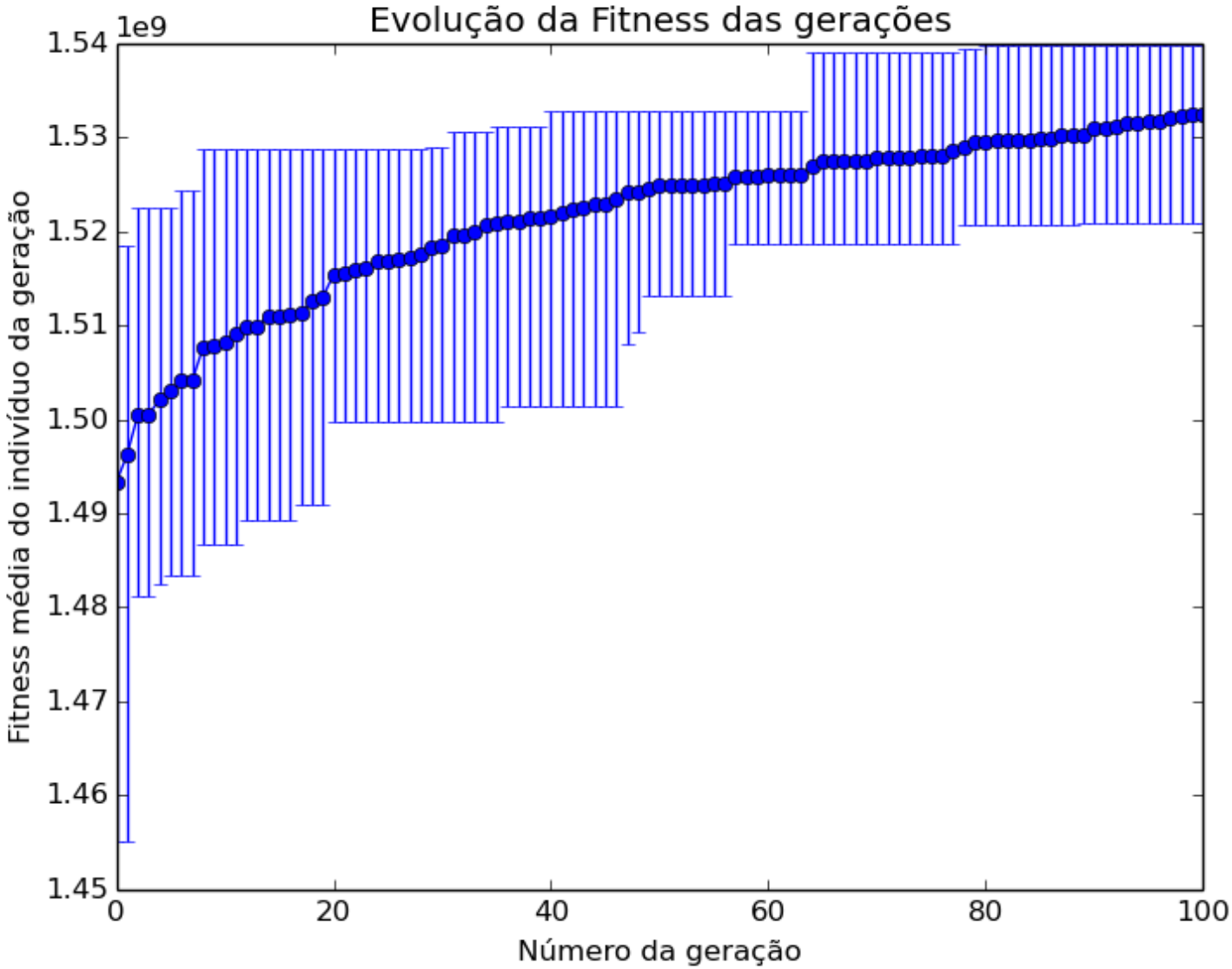
Análise preliminar

Feita análise preliminar da performance do GADec e de sua capacidade de convergência

Devido a resultados preliminares ruins, o restante da análise não foi feito

Conteúdo

Convergência do GADec



Análise do GADec

As barras de erro mostram que a convergência não foi significativa
Melhora da *fitness* média com o passar das gerações foi muito baixa (abaixo de 3%).

Resultado do teste de chi-quadrado

Valor de p-value no teste de chi-quadrado e possibilidade de rejeitar a hipótese nula de que terremotos seguem uma distribuição de Poisson no tempo com confiabilidade de 95%, para uma magnitude limite de 3.8

Região	**Sem *declustering***	**Janela**	**SLC**
Todo o catálogo	(rejeita-se, 0.0)	(rejeita-se,3e-37)	*
Kanto	(rejeita-se, 0.0)	(não rejeita-se, 0.11)	(não rejeita-se, 0.62)
Kansai	(rejeita-se, 1e-54)	(rejeita-se, 4e-6)	(não rejeita-se, 0.66)
Tohoku	(rejeita-se, 0.0)	(rejeita-se, 1.43e-6)	(não rejeita-se, 0.30)
Leste do Japão	(rejeita-se, 0.0)	(rejeita-se, 0.01)	(rejeita-se, 2e-28)

Resultado do teste de chi-quadrado

Valor de p-value no teste de chi-quadrado e possibilidade de rejeitar a hipótese nula de que terremotos seguem uma distribuição de Poisson no tempo com confiabilidade de 95%, para uma magnitude limite de 4.0

Região	**Sem *declustering***	**Janela**	**SLC**
Todo o catálogo	(rejeita-se, 0.0)	(rejeita-se, 2e-32)	*
Kanto	(rejeita-se, 0.0)	(não rejeita-se, 0.10)	(não rejeita-se, 0.56)
Kansai	(rejeita-se, 4e-18)	(rejeita-se, 2e-4)	(não rejeita-se, 0.77)
Tohoku	(rejeita-se, 0.0)	(rejeita-se, 9e-9)	(não rejeita-se, 0.19)
Leste do Japão	(rejeita-se, 4e-6)	(rejeita-se, 0.01)	(rejeita-se, 7e-12)

Análise dos testes de chi-quadrado

Para o catálogo como um todo, i.e. sem a aplicação de nenhum método de *declustering*, foi possível rejeitar a hipótese nula, o que está de acordo com o previsto em teoria

Já para o método da janela, foi possível rejeitar a hipótese nula para quatro das 5 regiões analisadas, sendo Kanto a única exceção. Em teoria, após a aplicação de tal método, não deveria ser possível refutar a hipótese nula com o grau de confiabilidade estabelecido para nenhuma região.

Para o método SLC os resultados do teste de chi-quadrado também divergiram do previsto em teoria, já que foi possível rejeitar H_0 para a região do Leste do Japão.

Resultado Kruskal-Wallis

Para ambos os métodos, a média de performance do GAModel foi superior ao obtido sem o uso de *declustering*, para todas as regiões e todos os métodos

Teste de Kruskal-Wallis permitiu refutar a hipótese nula H_0="A média de performance dos dois grupos é igual" com confiabilidade de 95% para quase todas as combinações de região e ano (a única exceção foi o Leste do Japão, para o ano de 2008)

Resultado ratifica que pode ser desejável usar *declustering* e trabalhar apenas com os *mainshocks* na área de modelagem de riscos de sismos

Conteúdo

Conclusão

Métodos de *declustering* tradicionais implementados (método da janela e método SLC) melhoram a performance do GAModel, mesmo que o tempo de ocorrência entre os *mainshocks* não siga uma distribuição de Poisson

GADec não convergiu para um bom resultado, além de não ser viável para catálogos com tamanho razoável, devido ao tempo necessário para execução

Fim

Obrigado!
Alguma Dúvida?