Agrupamentos de Séries Temporais de Imagens de Satélite por *VNS* Básico com Busca Local e Restrições

Wanderson L. da Silva Francisco A. M. Neto

IMECC/Unicamp

2014

Sumário

1. Motivação
2. Formulação do problema
3. Proposta
   - Busca local
   - Estruturas de vizinhança
   - Aplicação em séries temporais
4. Conclusão

Plano

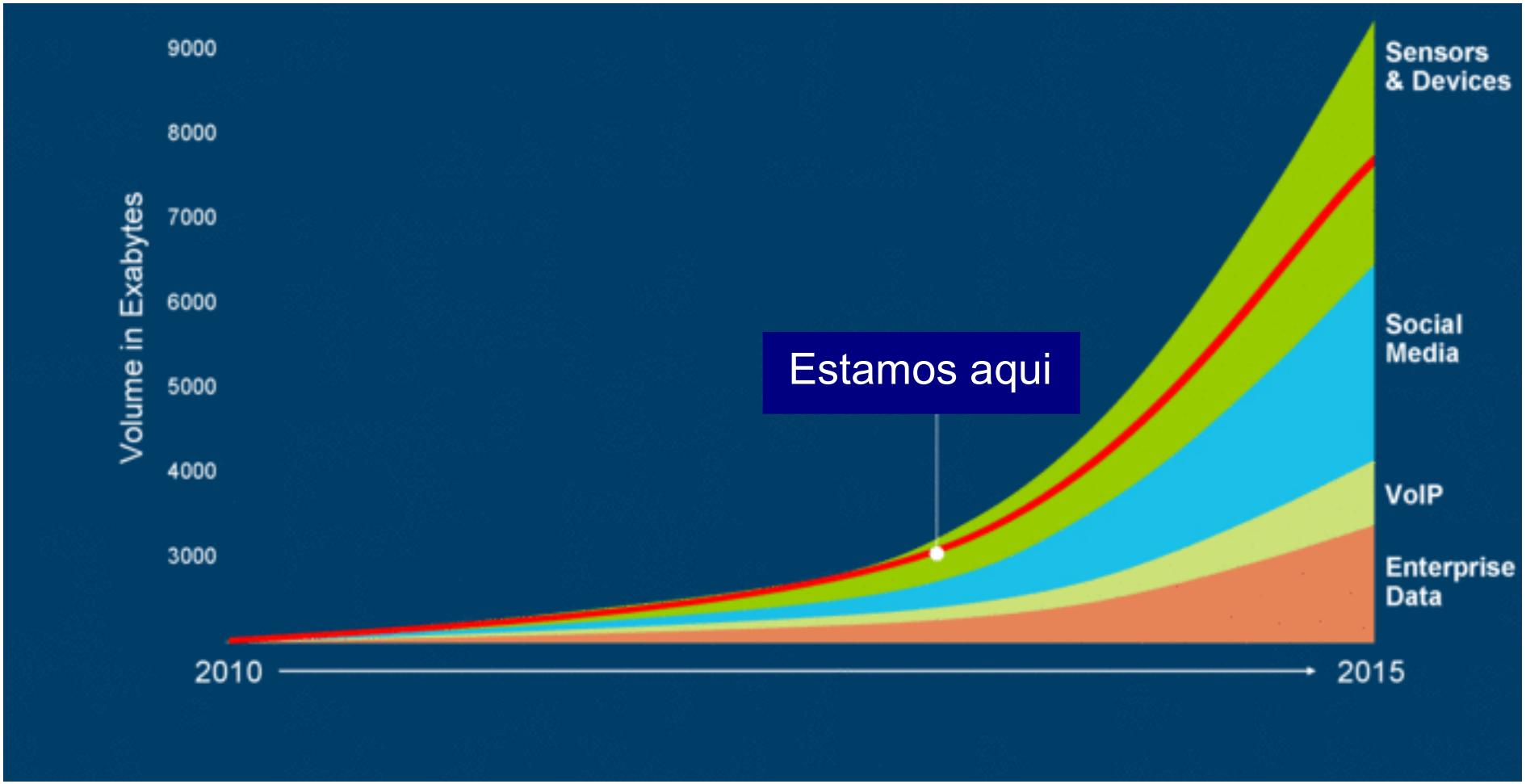
1 Motivação

2 Formulação do problema

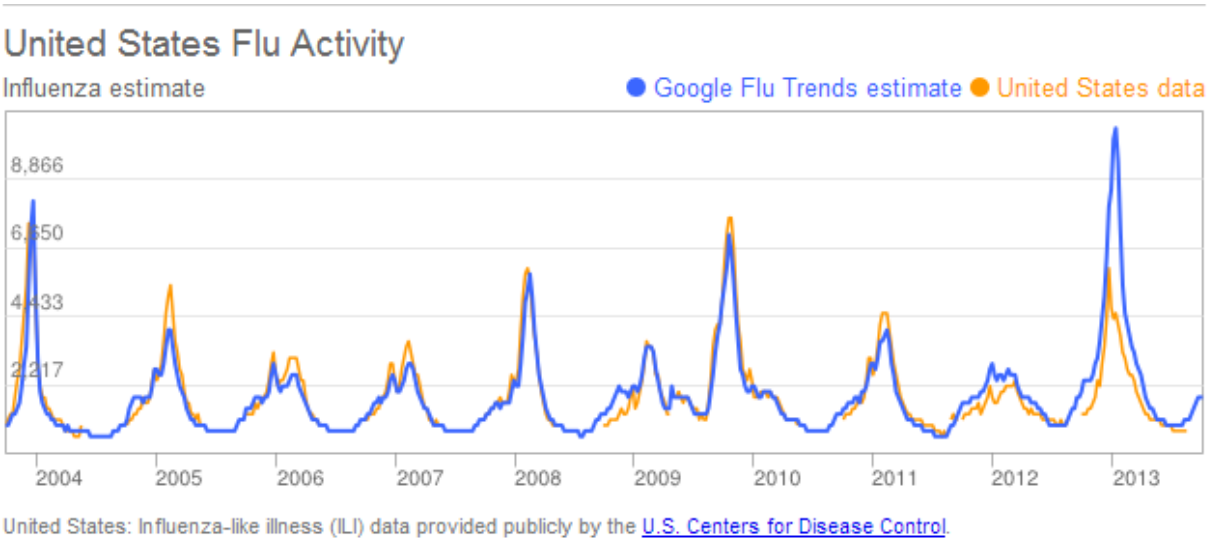
3 Proposta
- Busca local
- Estruturas de vizinhança
- Aplicação em séries temporais

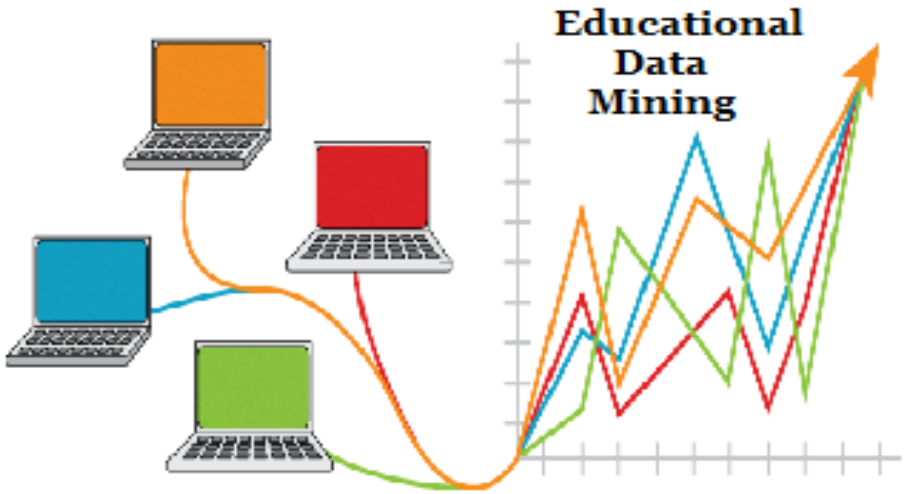
4 Conclusão

Pletora de dados



Aplicações


United States Flu Activity
Influenza estimate
Google Flu Trends estimate
United States data
8,866
6,650
4,433
2,217
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
United States: Influenza-like illness (ILI) data provided publicly by the U.S. Centers for Disease Control.


Educational
Data
Mining

Interdisciplinaridade

Otimização e mineração de dados

- *Linear and nonlinear separation of patterns by linear programming*
- *Integer programming and the theory of grouping*
- *A branch and bound algorithm for feature subset selection*
- *Evaluating alternative linear programming models to solve the two-group discriminant problem*
- *Improved linear programming models for discriminant analysis*
- *Misclassification minimization*
- *Support vector networks*
- *Mathematical programming in data mining*
- *Feature subset selection within a simulated annealing data mining algorithm*
- *Optimization-based data clustering using the nested partitions method*

Plano

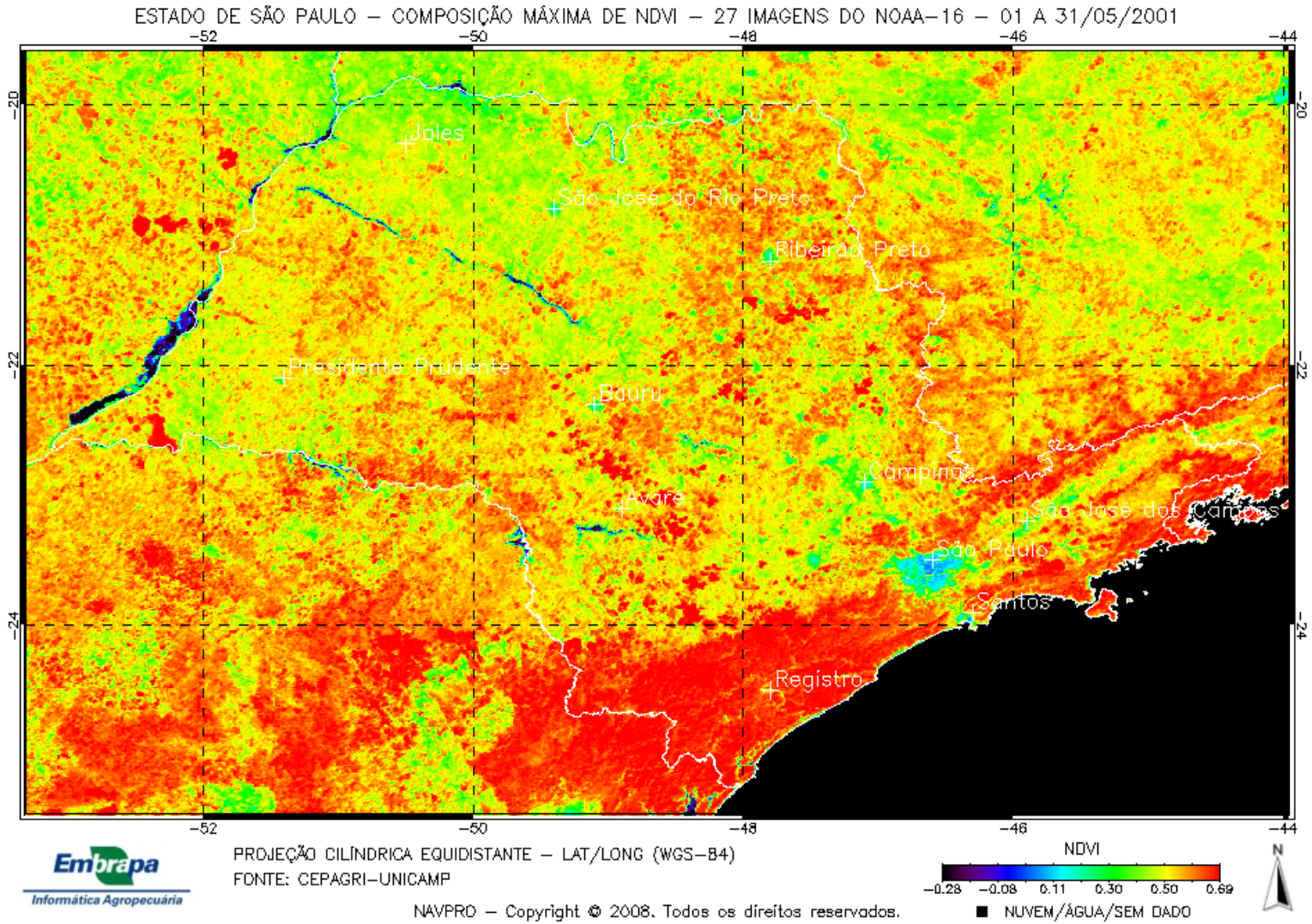
1 Motivação

2 Formulação do problema

3 Proposta
- Busca local
- Estruturas de vizinhança
- Aplicação em séries temporais

4 Conclusão

Problema original



Problema de agrupamento por partição

Seja $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$, onde $\mathbf{x}_i = (x_{i1}, \cdots, x_{id}) \in \mathbb{R}^d$. O problema de agrupamento consiste em obter uma partição $\mathbf{P}^*$ de $\mathbf{X}$ em k subconjuntos que atenda um determinado critério de qualidade Q, de forma que $Q(\mathbf{P}^*) \geq Q(\mathbf{P})$, $\forall \mathbf{P} \in \mathcal{P}$.

Problema de agrupamento por partição

Seja $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$, onde $\mathbf{x}_i = (x_{i1}, \cdots, x_{id}) \in \mathbb{R}^d$. O problema de agrupamento consiste em obter uma partição $\mathbf{P}^*$ de $\mathbf{X}$ em k subconjuntos que atenda um determinado critério de qualidade Q, de forma que $Q(\mathbf{P}^*) \geq Q(\mathbf{P})$, $\forall \mathbf{P} \in \mathcal{P}$.

Problema de agrupamento baseado em centroides

Determinar um conjunto $C^* = \{\mathbf{c}_1, \cdots, \mathbf{c}_k\}$ de centroides que formem k subconjuntos do tipo

$$C_j = \left\{ \forall x \in X \middle| \underset{1 \leq l \leq k}{\operatorname{argmin}}\, d(\mathbf{x}, \mathbf{c}_l) = j \right\},$$

de forma que C^* atenda um critério de qualidade Q tal que $Q(\mathbf{C}^*) \geq Q(\mathbf{C})$ para todo C.

Formalização do problema

$$\min \; f(C) = \sum_{i=1}^{n} \min_{1 \leq j \leq k} ||x_i - c_j||^2$$

$$\text{s.a} \quad C \in \Omega$$

tal que:

1. $C \in \Omega \Rightarrow \left(\forall \mathbf{c_j} \in C, \exists \mathbf{x} \in \mathbf{X} \text{ tal que } j = \underset{1 \leq l \leq k}{\operatorname{argmin}} ||x - c_l||_2^2 \right)$.
2. $\left(C_i, C_j \in \Omega, \text{com } C_i \neq C_j \right) \Rightarrow \mathbf{x} \in C_i \text{ e } \mathbf{y} \in C_j$, então $\exists \mathbf{w} \in \mathbb{R}^d$ tal que $(\mathbf{w}^t\mathbf{x}) \cdot (\mathbf{w}^t\mathbf{y}) < 0$

Plano

1 Motivação

2 Formulação do problema

3 Proposta
- Busca local
- Estruturas de vizinhança
- Aplicação em séries temporais

4 Conclusão

Objetivos

Uso combinado da *VNS*
com o *k*-médias projetado

Objetivos

Uso combinado da *VNS*
com o *k*-médias projetado

Uso do descritor $< \mathcal{F}, L_2 >$ para
aplicação de métodos de dados
estáticos em séries temporais.

Busca em vizinhança variável (*VNS*)

Busca em vizinhança variável (*VNS*)

1. Um mínimo local com relação a uma estrutura de vizinhança não é necessariamente um mínimo local com relação às outras estruturas de vinhança.

Busca em vizinhança variável (*VNS*)

1. Um mínimo local com relação a uma estrutura de vizinhança não é necessariamente um mínimo local com relação às outras estruturas de vinhança.

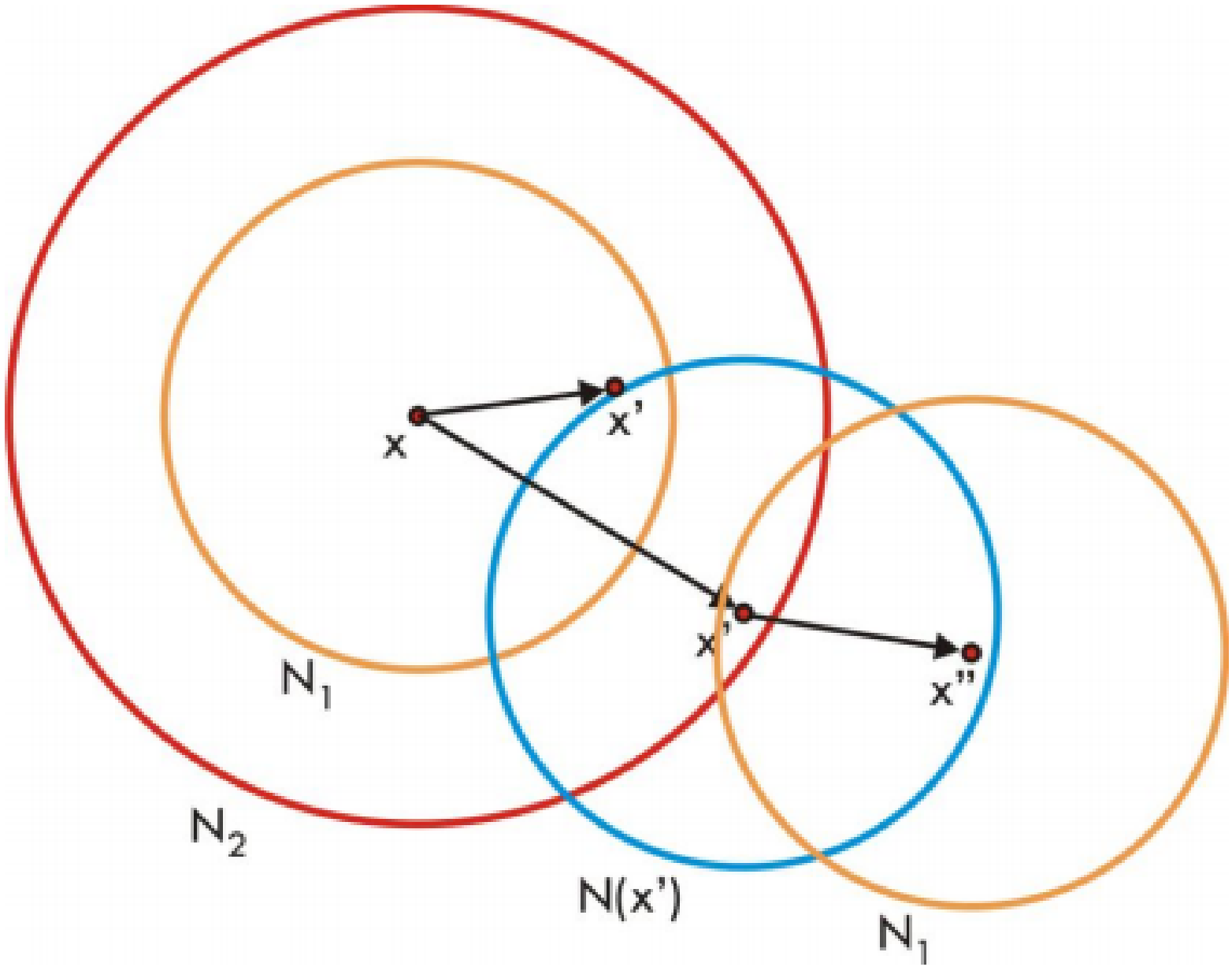
2. Um mínimo global é um mínimo local com relação a quaisquer estruturas de vizinhança.

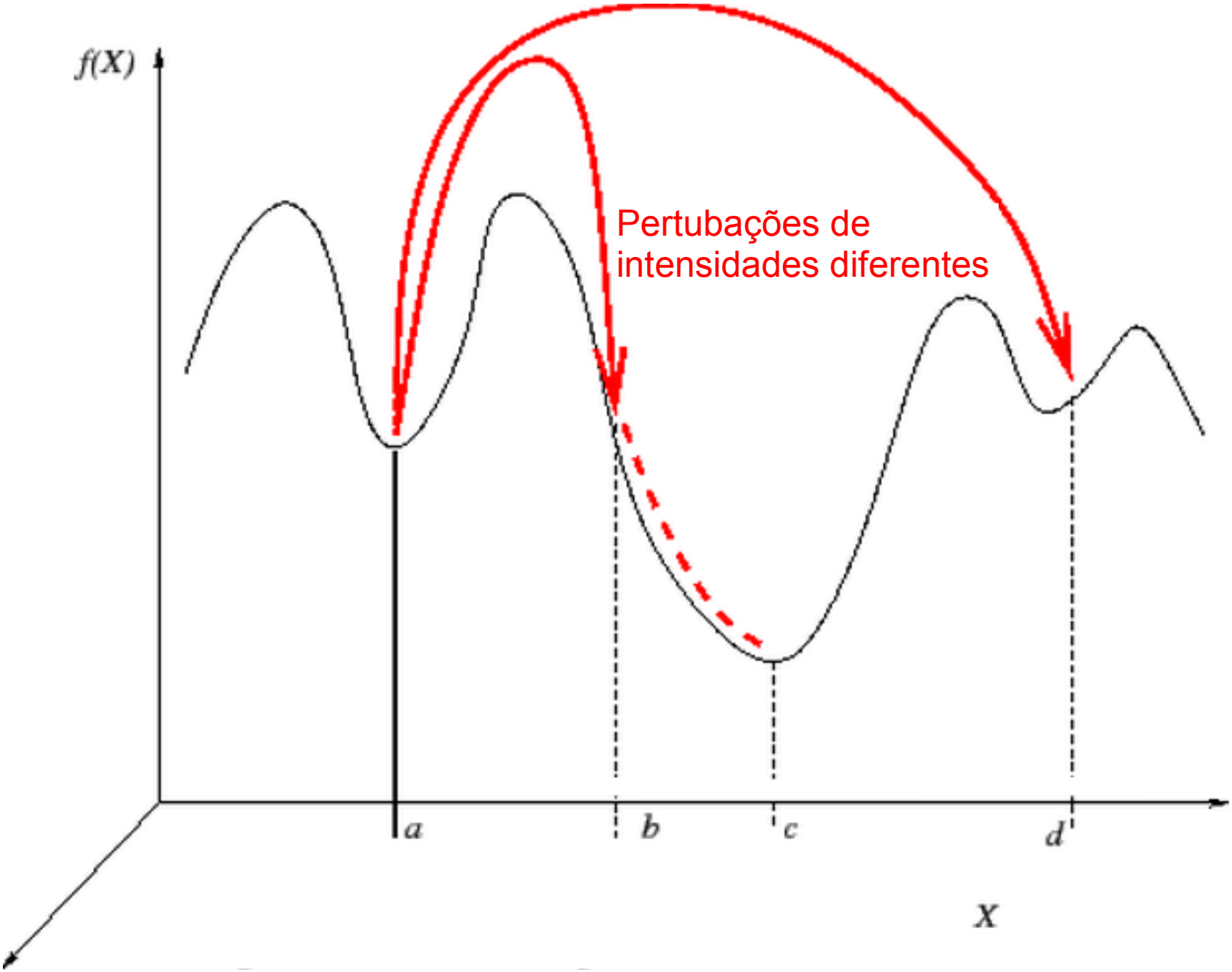
Busca em vizinhança variável (*VNS*)

1. Um mínimo local com relação a uma estrutura de vizinhança não é necessariamente um mínimo local com relação às outras estruturas de vinhança.

2. Um mínimo global é um mínimo local com relação a quaisquer estruturas de vizinhança.

3. Para um grande número de problemas, mínimos locais com relação a uma, ou a várias vizinhanças, são relativamente próximos.


x
x'
x''
N1
N2
N(x')
N1


f(X)
Pertubações de
intensidades diferentes
a
b
c
d
X

Algorithm 1: esquema geral do algoritmo proposto

input :

- $\mathbf{X}$ conjunto dos pontos $\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\}$ a serem agrupados
- $\mathbf{C}$ conjunto de centroides iniciais $\{\mathbf{c}_1, \mathbf{c}_2, \cdots, \mathbf{c}_k\}$
- Ω_c conjunto de restrições de caixa $\{\mathcal{H}_1, \mathcal{H}_2, \cdots, \mathcal{H}_{k'}\}$
- k número de grupos a serem gerados
- t_{max} tempo máximo de execução

output :

- $\mathbf{C}$ melhor solução (conjunto de centroides) obtida dentre as investigadas.

BasicVNS($\mathbf{X}, \mathbf{C}, \Omega_c, k, t_{max}$)
$i_{max} \longleftarrow \frac{n}{10}$
repeat
  $i \leftarrow 1$;
  repeat
    $\mathbf{C}' \leftarrow$ Shake($\mathbf{C}, i$);
    $\mathbf{C}'' \leftarrow$ LocalSearch($\mathbf{X}, k, \mathbf{C}', \Omega_c$);
    $\mathbf{C}, i \leftarrow$NeighbourhoodChange($\mathbf{C}, \mathbf{C}'', i$)
  until $i = i_{max}$;
  $t \leftarrow$ tempo de processamento
until $t > t_{max}$;

Incorporando conhecimento

formas de restrição

1. Relações de paridade,

Incorporando conhecimento

formas de restrição

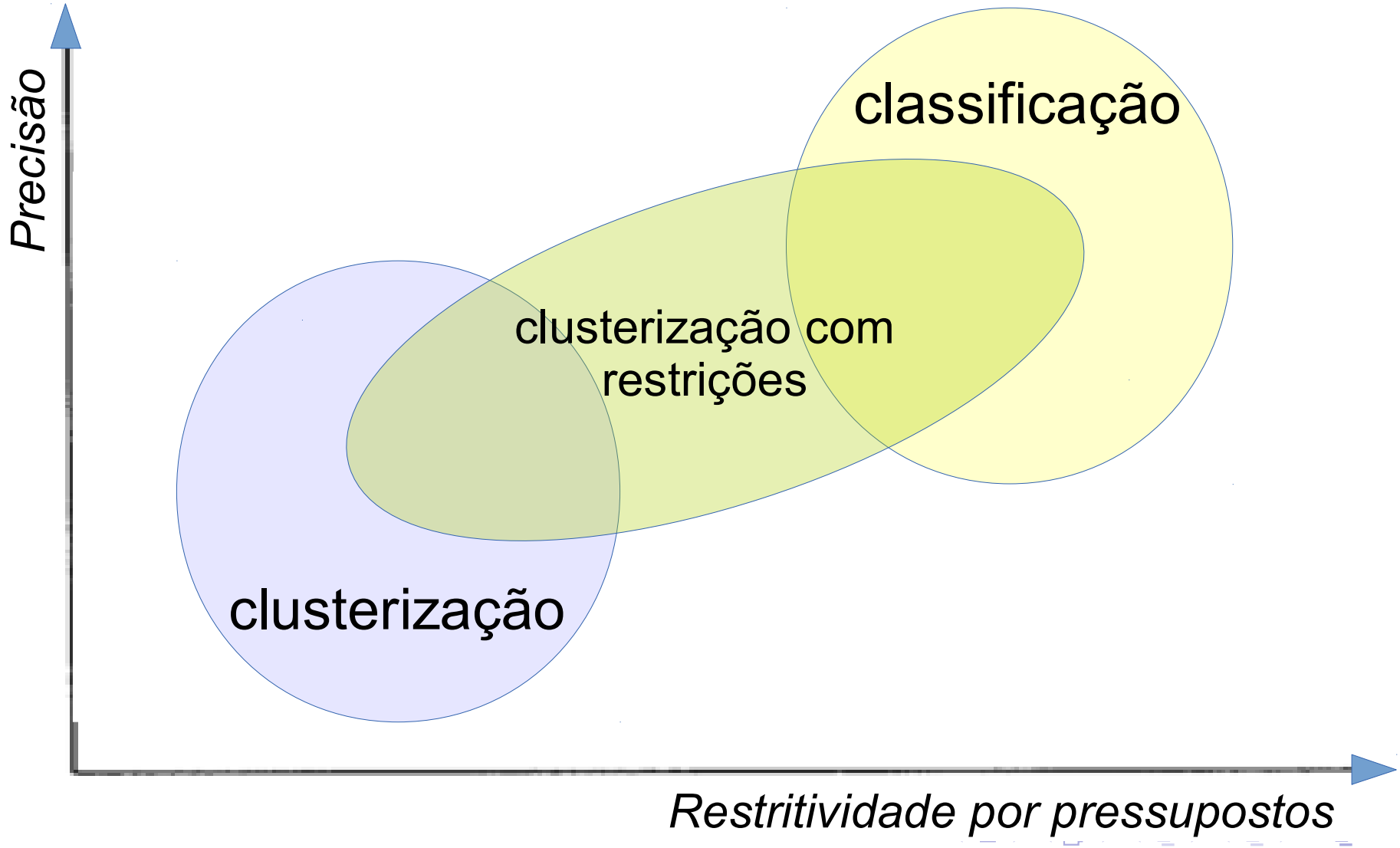
1. Relações de paridade,
2. Amostra de classes,

Incorporando conhecimento

formas de restrição

1. Relações de paridade,
2. Amostra de classes,
3. *Seeds* para inicialização.

Paradigmas supervisionados, não supervisionados e semissupervisionados



Função LocalSearch

Algorithm 2: k-médias projetado

$\mathbf{C} \leftarrow \text{LocalSearch}(\mathbf{X}, k, \mathbf{C}, \Omega_c)$

passo 1 (inicialização): usar C como solução inicial

passo 2 (atribuição): associar cada elemento x_j, com $j \in \{1, \cdots, n\}$, ao centroide mais próximo ($C \rightarrow P$)

passo 3 (teste de otimalidade local): se não houve alguma mudança no passo anterior, parar aqui

passo 4 (projeção): $C' \longleftarrow \text{Projection}(C', C, \Omega_c)$

passo 5 (atualização): trocar os centroides C da solução corrente por C' e voltar para o **passo 2**

Espaço factível

$\Omega_c = \{\mathcal{H}_1, \cdots, \mathcal{H}_{k'}\}$, com:

$$\mathcal{H}_i = \left\{ \mathbf{x} \in \mathbb{R}^d \middle| \mu_j - \frac{3\sigma_j}{\sqrt{n_i}} \leq x_j \leq \mu_j - \frac{3\sigma_j}{\sqrt{n_i}}, \forall j \in \{1, \cdots, d\} \right\},$$

onde μ_j e σ_j são, respectivamente, a média e o desvio padrão, na dimensão j, da amostra $A_i = \{\mathbf{x}_1^{(i)}, \cdots, \mathbf{x}_{n_i}^{(i)}\}$ de uma classe i.

Estrutura de vizinhança 1

realocação dos mais distantes

$$N_i^{(1)}(C_1) = \left\{ C \in \Omega \middle| (r(C_1, C) = i) \bigwedge (\mathbf{w}(C_1) \cdot \mathbf{e}_j \geq L) \right\},$$

com j varia entre 1 e i.

Medida de aderência

$$\eta(\mathbf{x}_j) = \frac{||\mathbf{x}_j - \mathbf{c}_l|| - \mu_l}{\sigma_l},$$

onde

$$\mu_l = \frac{\sum_{\mathbf{x} \in \mathbf{c}_l} ||\mathbf{x} - \mathbf{c}_l||}{\#\mathbf{c}_l},$$

$$\sigma_l = \left(\frac{1}{\#\mathbf{c}_l} \sum_{\mathbf{x} \in \mathbf{c}_l} \left| ||\mathbf{x}_j - \mathbf{c}_l|| - \mu_l \right|^2 \right)^{\frac{1}{2}}.$$

Estrutura de vizinhança 2

realocação do mais discrepante

$$N_i^{(2)}(C_1) = \left\{ C \in \Omega \middle| (r(C_1, C) = i) \bigwedge (\Theta(C_1) \cdot \mathbf{e}_j \geq L,) \right\}$$

onde:

- j varia entre 1 e i,
- L é o i-ésimo maior valor de η entre os elementos da solução C_1,
- $\Theta(C_1)$ é o vetor decrescente dos valores η dos elementos de C_1.

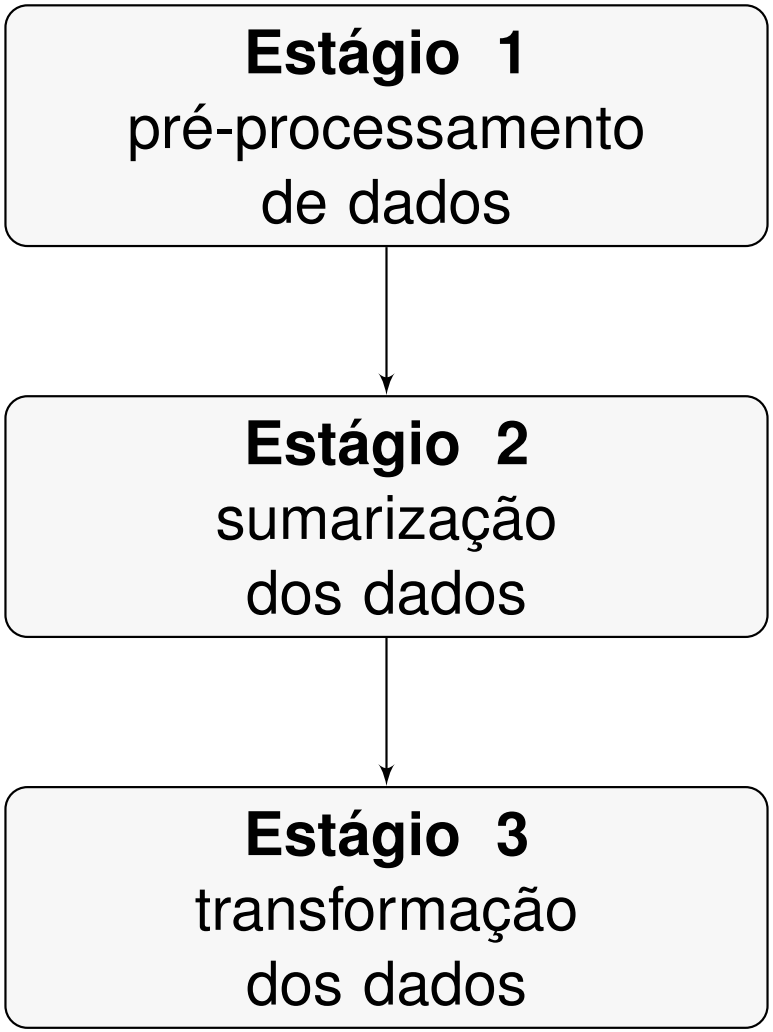
Estrutura de vizinhança 3

realocação intensa

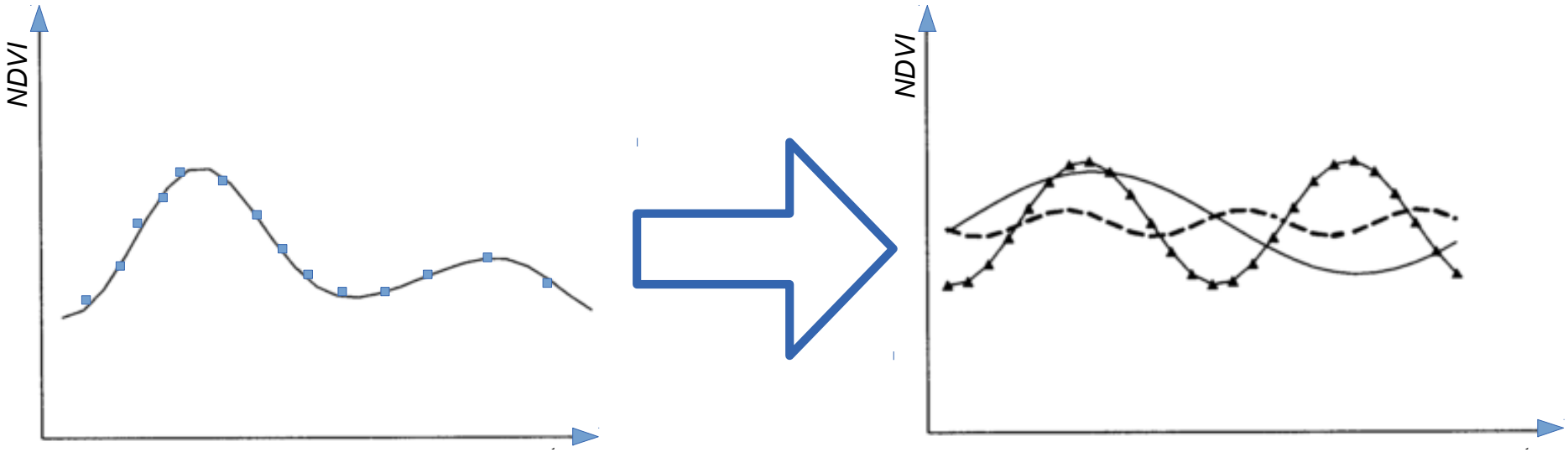
$$N_i^{(3)}(C_1) = \left\{ C \in \Omega \middle| \Theta(C_i) \cdot \mathbf{e}_j \geq L_i \right\},$$

com $L_i = 5 - 0,1i$.

transformação de dados



Análise funcional



Para $y(k)$, com $k \in \{1, 2, 3, \cdots, n\}$, têm-se:

$$y_t = \bar{y} + \sum_{k=1}^{\lfloor \frac{n}{2} \rfloor} \left[A_k \cos\left(\frac{2\pi kt}{n}\right) + B_k \operatorname{sen}\left(\frac{2\pi kt}{n}\right) \right],$$

onde $\bar{y}$ é a média aritmética dos dados e

$$A_k = \frac{2}{n} \sum_{t=1}^{n} y_t \cos\left(\frac{2\pi kt}{n}\right) \quad \text{e} \quad B_k = \frac{2}{n} \sum_{t=1}^{n} y_t \operatorname{sen}\left(\frac{2\pi kt}{n}\right).$$

Plano

1 Motivação

2 Formulação do problema

3 Proposta
- Busca local
- Estruturas de vizinhança
- Aplicação em séries temporais

4 Conclusão

Trabalhos futuros

benckmarking

Avaliar comparativamente o método proposto em dados sintéticos, reais e nas séries temporais de *NDVI*.

algoritmos	*B-cubed recall*	*B-cubed precision*	MSSC
k-médias			
Multi-start k-médias			
COP k-médias			
k-médias projetado			
VNS + $N^{(1)(2)(3)}$ + *k*-médias			
VNS + $N^{(1)(2)(3)}$ + COP *k*-médias			
VNS + $N^{(1)(2)(3)}$ + *k*-médias projetado			