Evolutionary Design of Nearest Prototype Classifiers

Tiago José – tjs2

Tiago Neves - tnb

Centro de Informática U·F·P·E

Universidade Federal de Pernambuco

Motivação

- Base de dados muito grande
- Existência de parâmetros de aprendizado

Objetivo

- Criar protótipos para diminuir o tamanho da base
- Eliminar a existência de parâmetros de aprendizado

Conceitos

- Protótipo rotulado:

$$r_i = \langle p, s \rangle$$

p é o espaço do protótipo

s é a classe a que protótipo pertence

- Classificador:

$$C = \{r_1, \dots, r_n\}$$

Conceitos

- Padrão:

  v_r é cada exemplo usado para treinamento ou teste

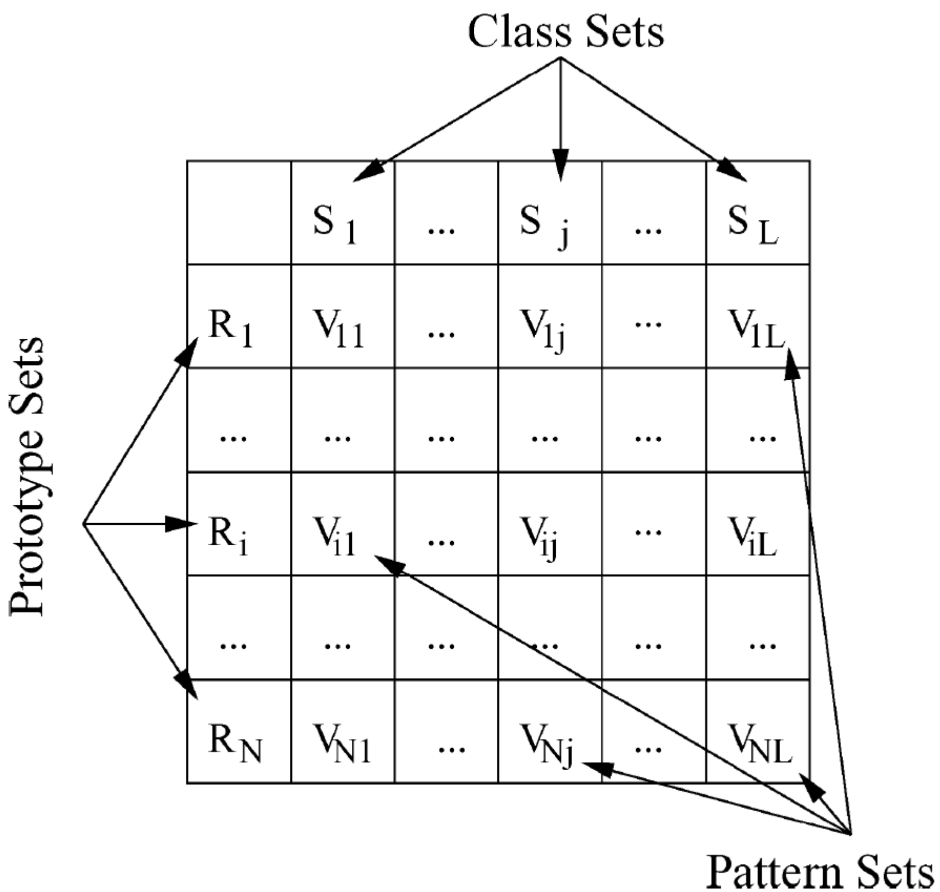
  $V = \{r_1, \dots, r_m\}$
- Classe:

  s_j

  $S = \{s_1, \dots, s_L\}$
- Qualidade de um protótipo:

  $quality(r_i)$

Estrutura


Class Sets
S_1 ... S_j ... S_L
R_1 V_{11} ... V_{1j} ... V_{1L}
... ...
R_i V_{i1} ... V_{ij} ... V_{iL}
... ...
R_N V_{N1} ... V_{Nj} ... V_{NL}
Prototype Sets
Pattern Sets

Estrutura – Conjunto de Classes

- $\forall v \in V,\ v = \langle p, s_j \rangle,\ s_j \in S, então\ v \in S_j$
- $s_j \in S$:
  - $regions(s_j) = \sum_{i=1}^{N} \delta(r_i, s_j), onde\ \delta = \begin{cases} 1, & sse\ r_i = \langle p, s_j \rangle \\ 0, & otherwise \end{cases}$
  - $expectation(s_j) = \frac{\|s_j\|}{regions(s_j)}$

Estrutura – Conjunto de Padrões

- $\forall v = \langle p, s_j \rangle \in V,\ s_j \in S,\ r_i = \langle p_i, s' \rangle \in C,\ então$
  $v \in V_{ij}\ \ sse\ \forall r_{i'} = \langle p_{i'}, s'' \rangle \in C, d(p_v, p_i) \leq d(p_v, p_{i'})$

- $d(x, y) = \sum_{i=0}^{i<K} (x[i] - y[i])^2$

Estrutura – Conjunto de Protótipos

- $\forall v = \langle p_v, s_j \rangle \in V,\ r_i = \langle p_i, s' \rangle \in C,\ então$
  $v \in R_i\ \ sse\ \forall r_{i'} = \langle p_{i'}, s'' \rangle \in C, d(p_v, p_i) \leq d(p_v, p_{i'})$
- $r_i = \langle p, s_j \rangle$:
  - $accuracy(r_i) = \frac{\|V_{ij}\|}{\|R_i\|}$
  - $apportation(r_i) = \frac{2\|V_{ij}\|}{expectation(s_j)}$
  - $quality(r_i) = min\left(1, accuracy(r_i) * apportation(r_i)\right)$

	S1	S2
$R_1 = \langle p_1, s_2 \rangle$	$\|V_{11}\| = 7$	$\|V_{12}\| = 9$
$R_2 = \langle p_2, s_1 \rangle$	$\|V_{21}\| = 10$	$\|V_{22}\| = 2$
$R_3 = \langle p_3, s_2 \rangle$	$\|V_{31}\| = 3$	$\|V_{32}\| = 9$

Estrutura – Exemplo

- Número de protótipos: 3
- Número de classes: 2
- Número de instâncias: 40

- $regions(s_1)$: 1 ($r_2 = \langle p_2, s_1 \rangle$)
- $\|s_1\|$: 20 ($\|V_{11}\| + \|V_{21}\| + \|V_{31}\| = 7 + 10 + 3$)
- $expectation(s_1)$: 20 $\left(\frac{\|s_1\|}{regions(s_1)} = \frac{20}{1}\right)$

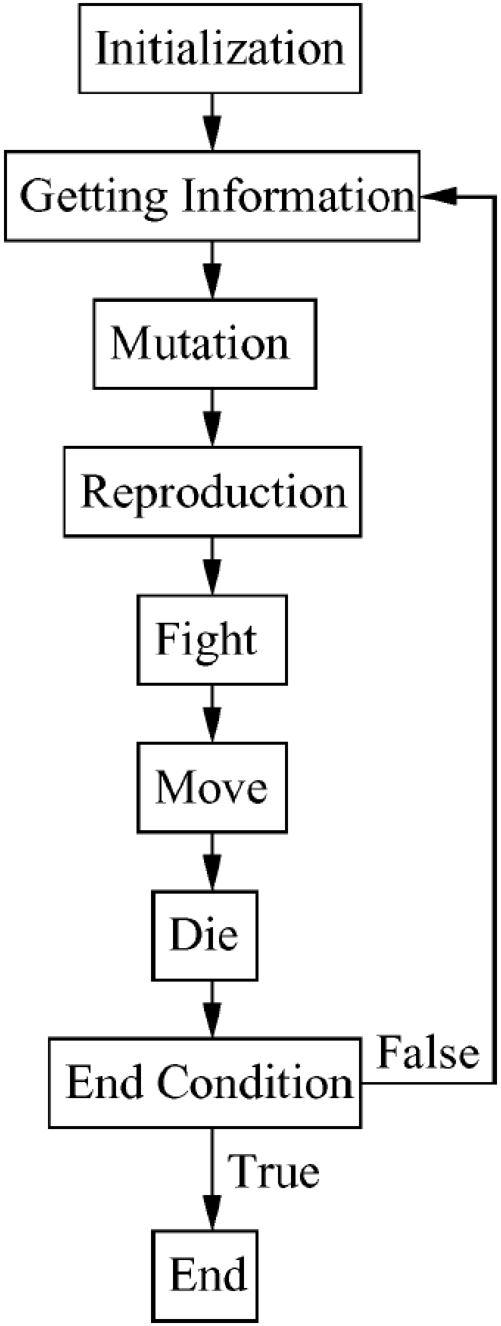
	S1	S2
$R_1 = \langle p_1, s_2 \rangle$	$\|V_{11}\| = 7$	$\|V_{12}\| = 9$
$R_2 = \langle p_2, s_1 \rangle$	$\|V_{21}\| = 10$	$\|V_{22}\| = 2$
$R_3 = \langle p_3, s_2 \rangle$	$\|V_{31}\| = 3$	$\|V_{32}\| = 9$

Estrutura – Exemplo

- $accuracy(r_2)$: 0.83 $\left(\frac{\|V_{ij}\|}{\|R_i\|} = \frac{10}{12}\right)$
- $apportation(r_2)$: 1 $\left(\frac{2\|V_{ij}\|}{expectation(s_j)} = \frac{2*10}{20}\right)$
- $quality(r_2)$: 0.83 $\left(\begin{matrix} min(1, accuracy(r_i) * apportation(r_i)) \\ \min(1,\ 0.83 * 1) \end{matrix}\right)$

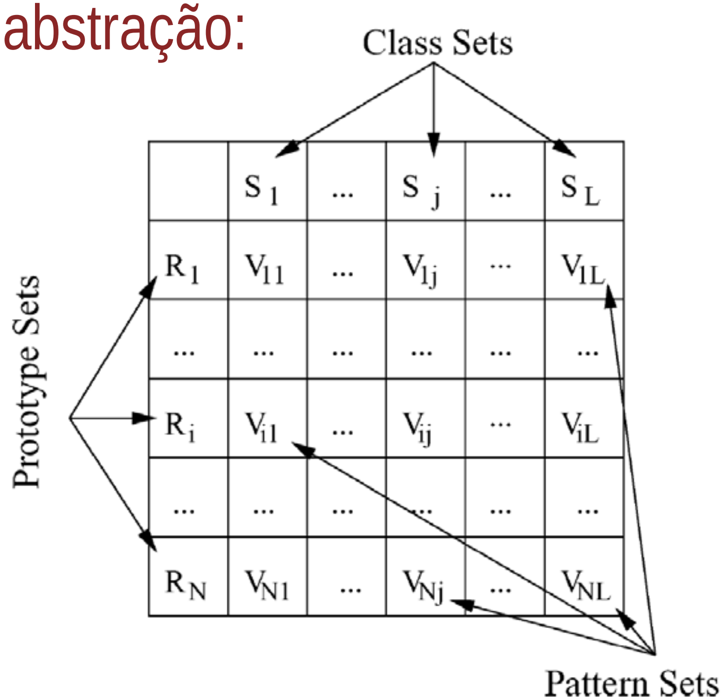
Algoritmo - Inicialização

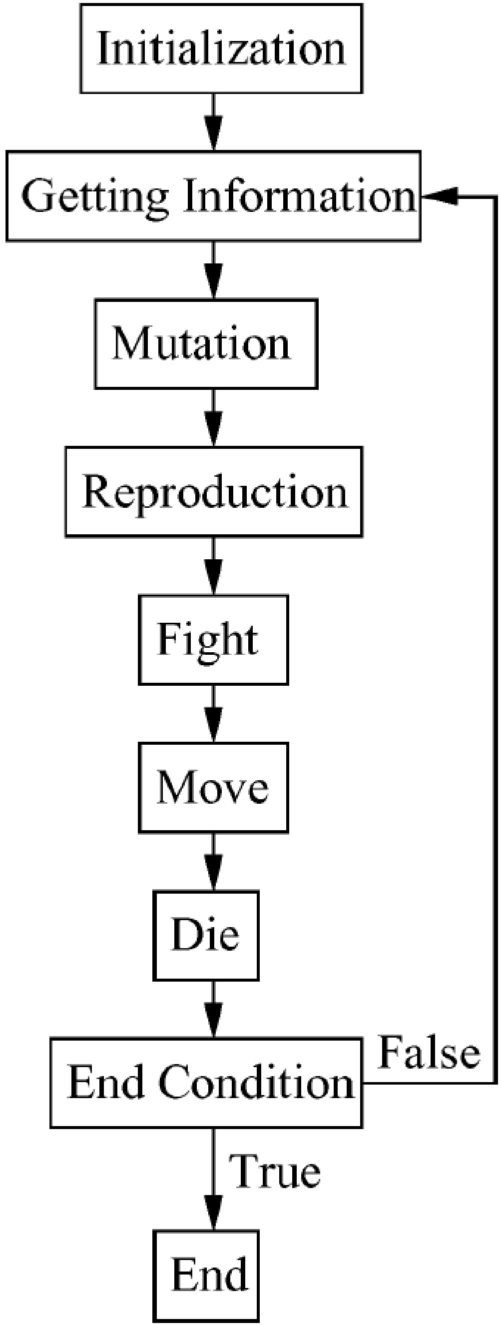
- O número inicial de protótipos é sempre um
- A localização desse protótipo inicial é irrelevante
- Não há parâmetros de aprendizado



Algoritmo – Obtenção de Informação

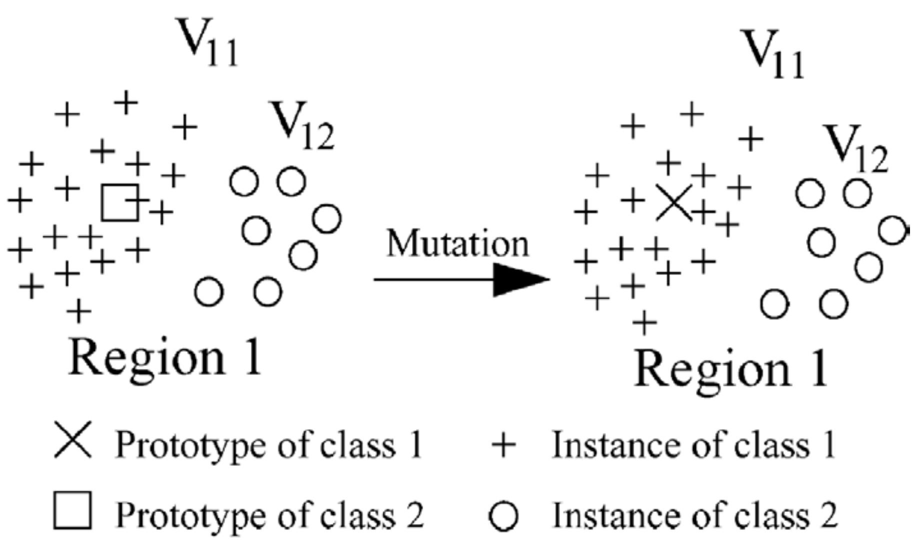
- Gera-se a abstração:

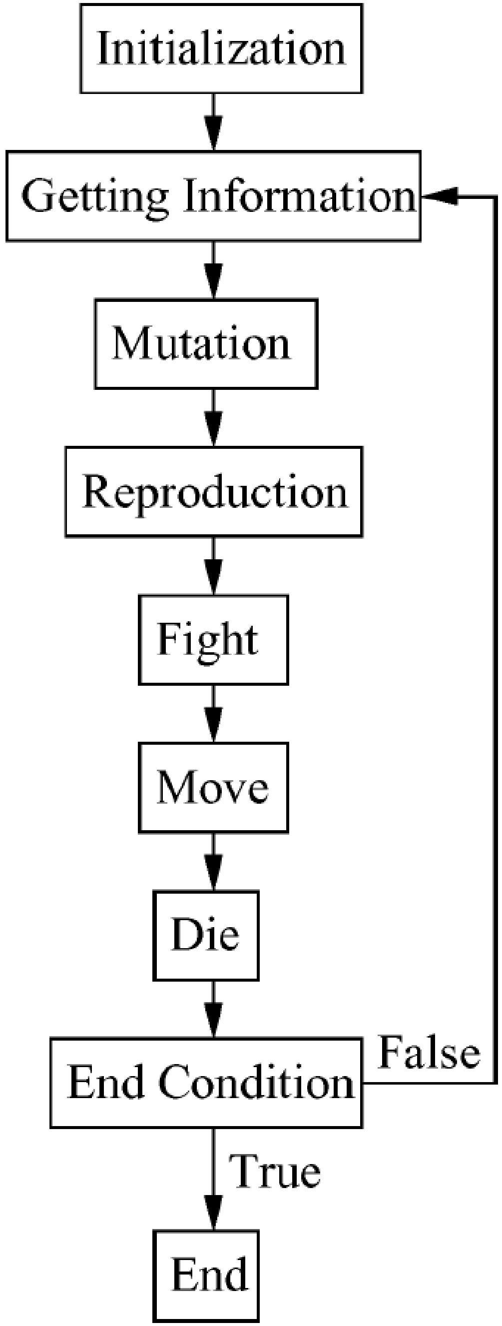




Algoritmo – Mutação

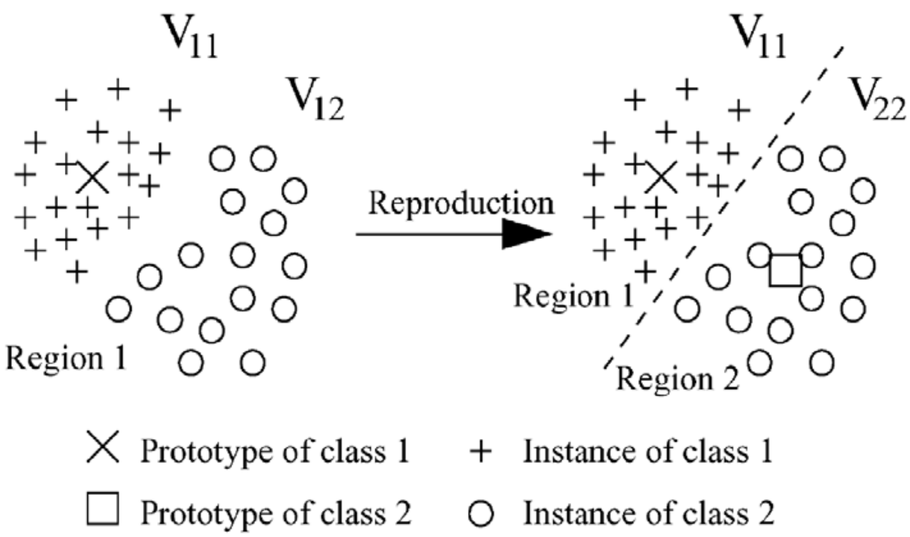
- $\forall r_i = \langle p, s \rangle \in C,\ \ s = \arg\max_j \|V_{ij}\|$

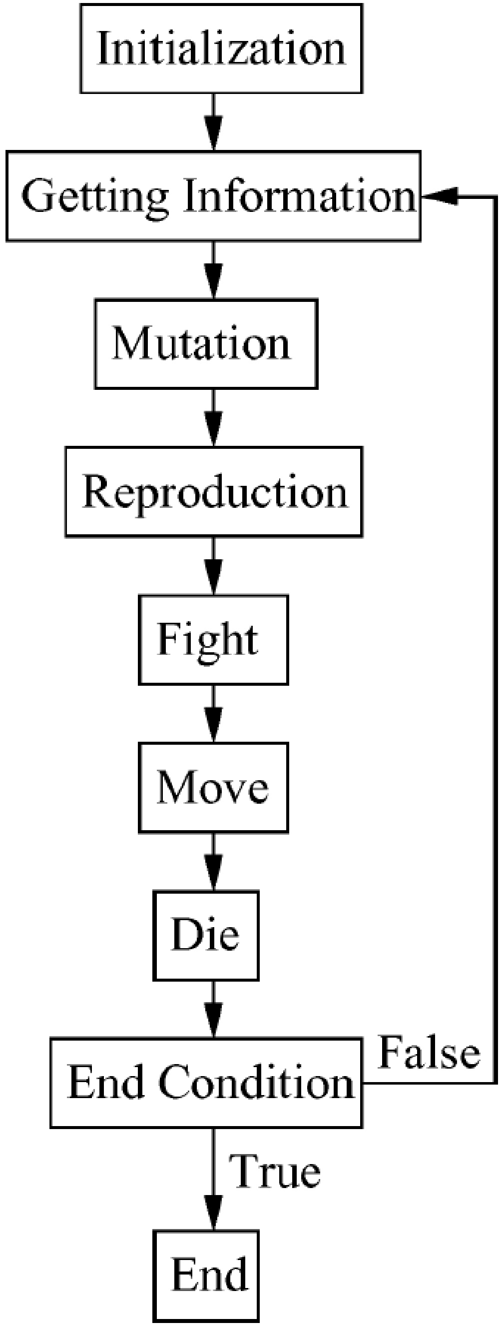




Algoritmo – Reprodução

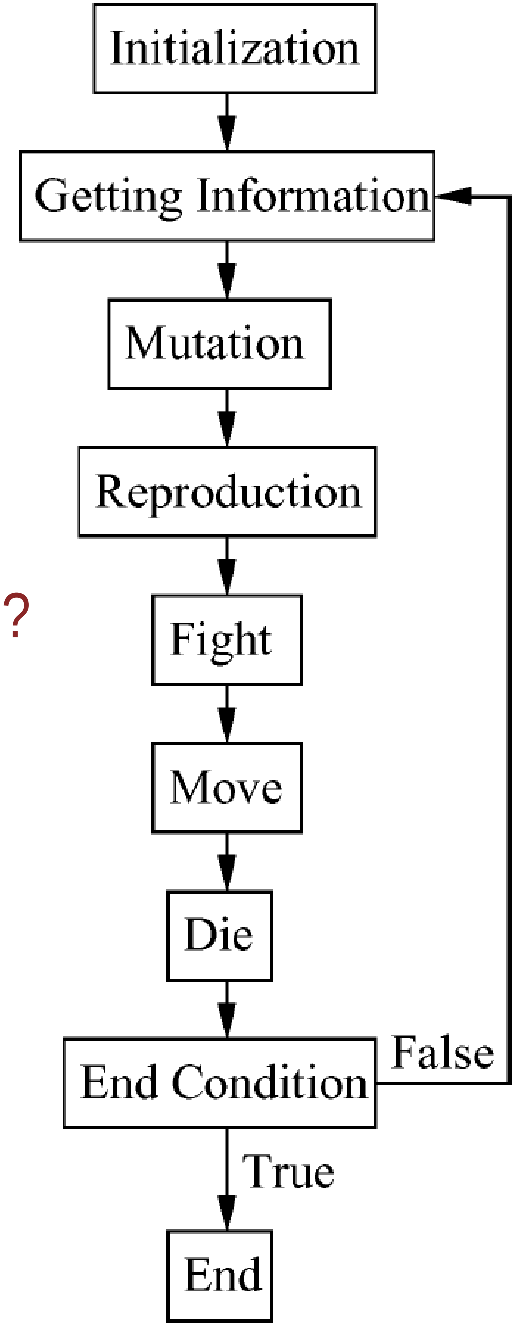
- A escolha do S_j é baseado no método da roleta
- O tamanho da fatia é proporcional a $\|V_{ij}\|$





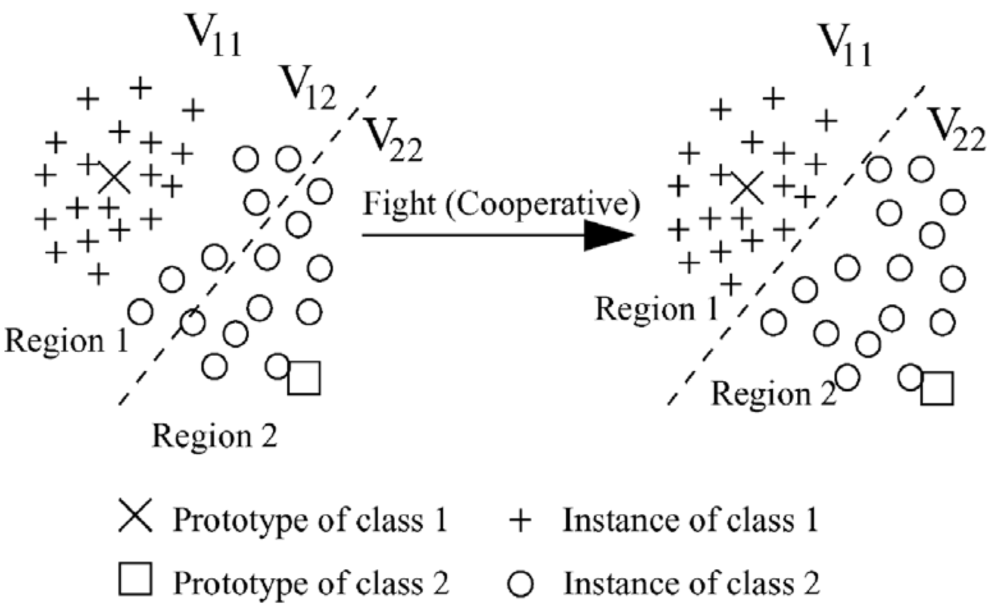
Algoritmo – Disputa

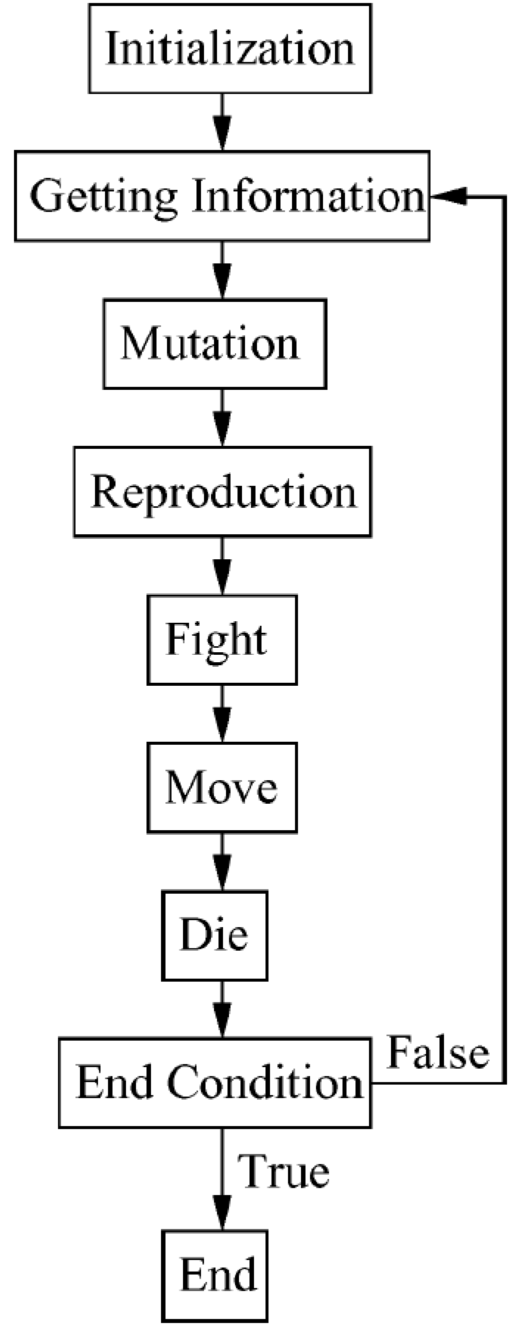
- Quais protótipos $r_{i'}$, tentarão desafiar o protótipo r_i?
  - Os protótipos $r_{i'} \in neighbours(r_i)$
- Qual protótipo $r_{i'}$ desafiará o protótipo r_i?
  - Método da roleta
  - O tamanho da fatia é proporcional a:
    - $|quality(r_i) - quality(r_{i'})|$
- O que determina se r_i aceita o desafio?
  - $P_{fight}(r_i, r_{i'}) = |quality(r_i) - quality(r_{i'})|$



Algoritmo – Disputa

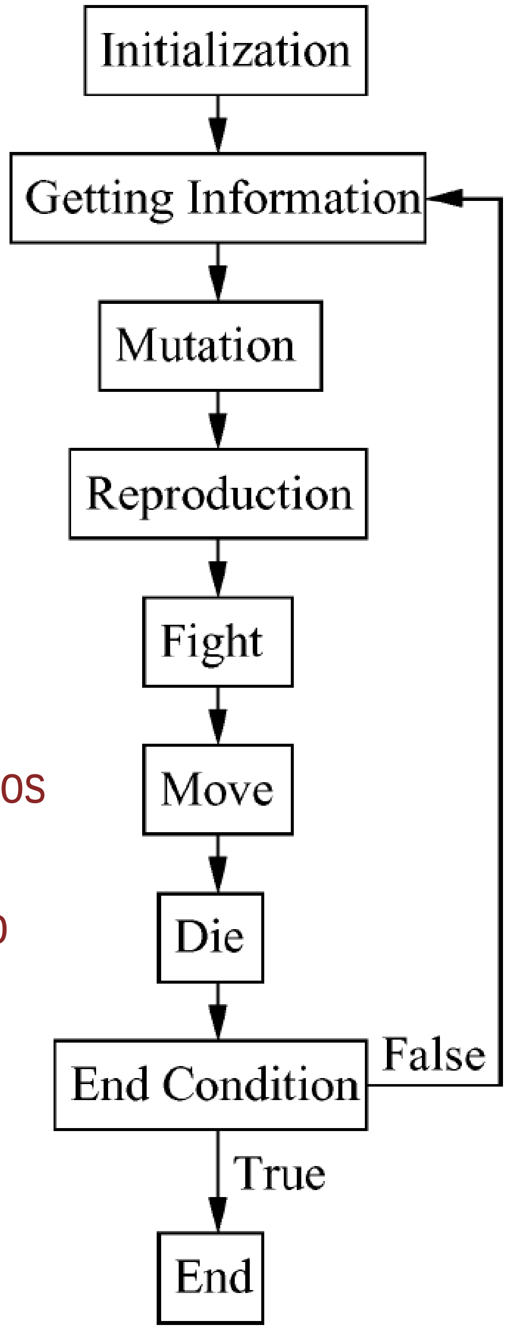
- $r_{i'} = \langle p_{i'}, s_{i'} \rangle$, $r_i = \langle p_i, s' \rangle$ e $s_i \neq s_j$:





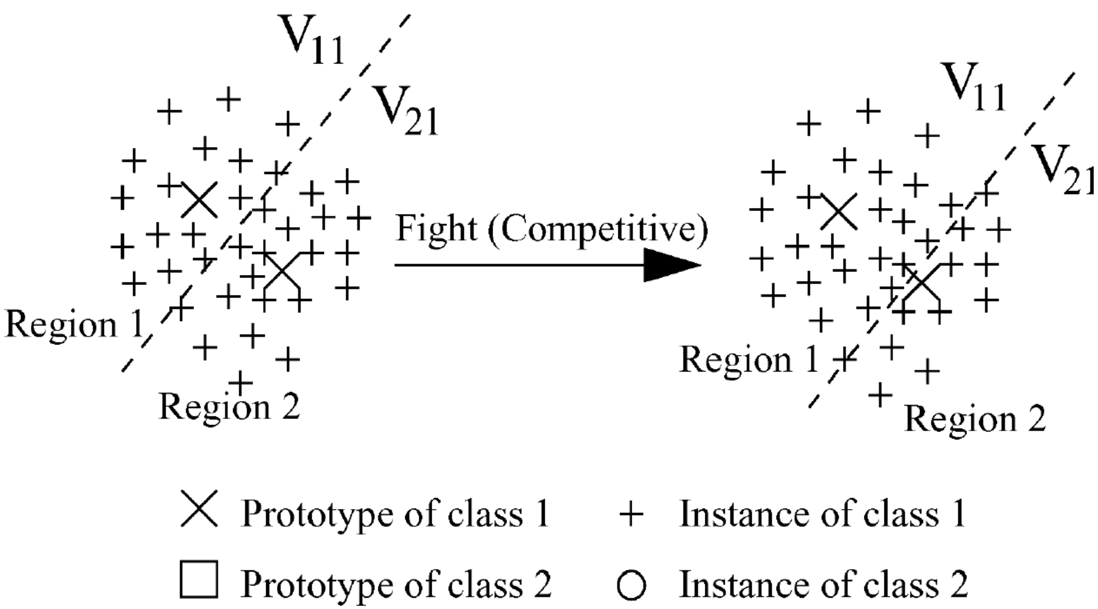
Algoritmo – Disputa

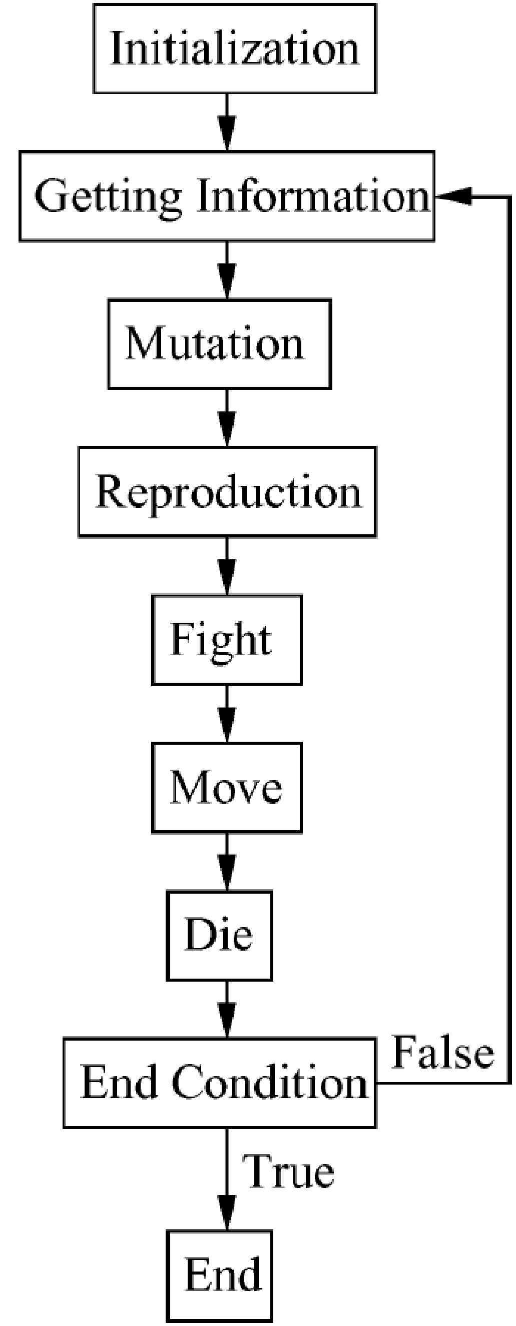
- $r_{i'} = \langle p_{i'}, s_{i'} \rangle$, $r_i = \langle p_i, s' \rangle$ e $s_i = s_j$:
  - Método da roleta para decidir o vencedor
  - O tamanho da fatia é proporcional a qualidade dos protótipos
  - A quantidade de padrões que são transferidos depende da probabilidade proporcional as qualidades de cada protótipo



Algoritmo – Disputa

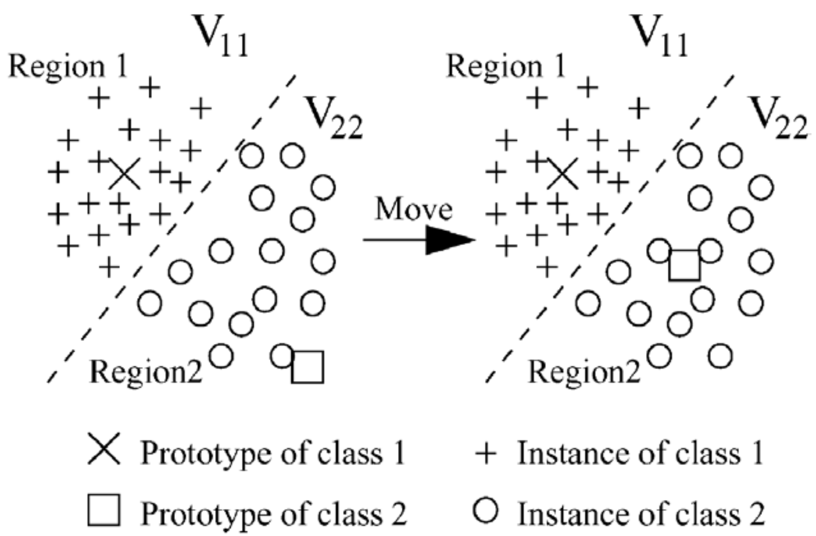
- $r_{i'} = \langle p_{i'}, s_{i'} \rangle$, $r_i = \langle p_i, s' \rangle$ e $s_i = s_j$:

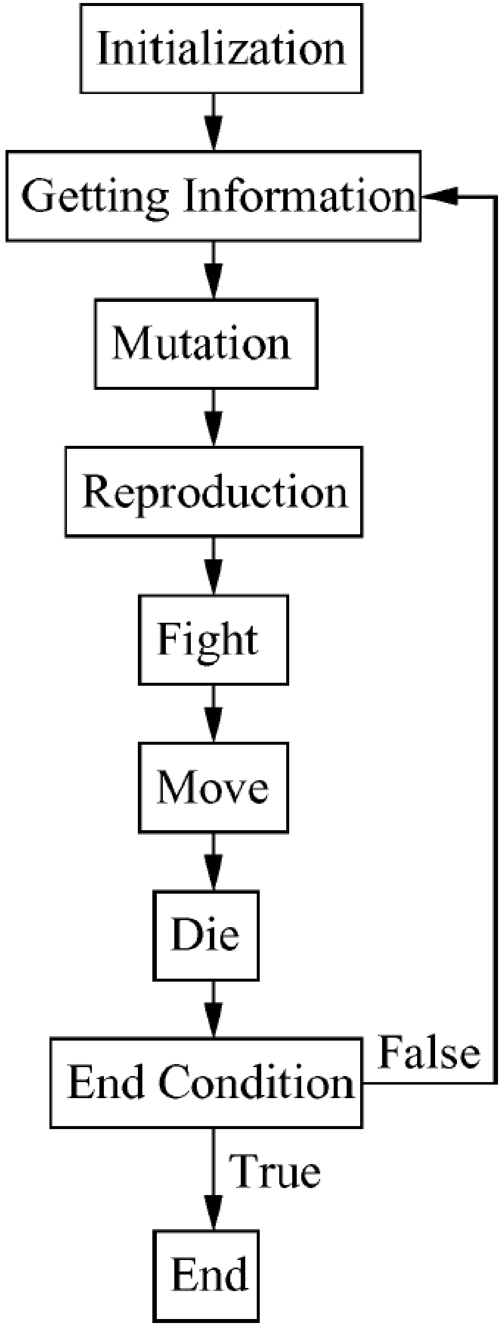




Algoritmo – Deslocamento

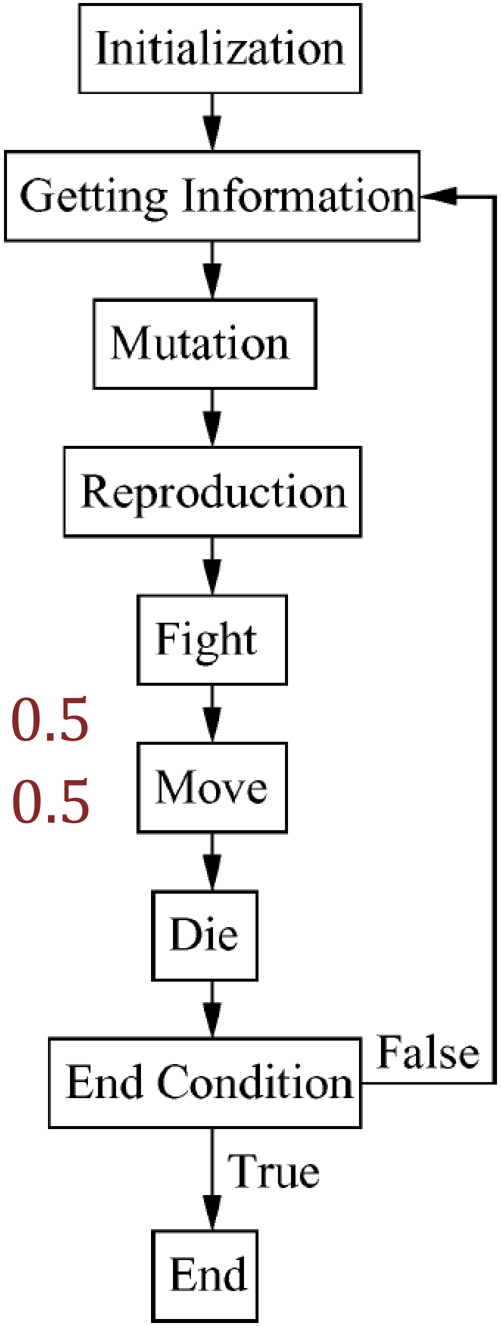
- $r_i = \langle p_i, s_j \rangle \rightarrow \langle centroide(V_{ij}), s_j \rangle$





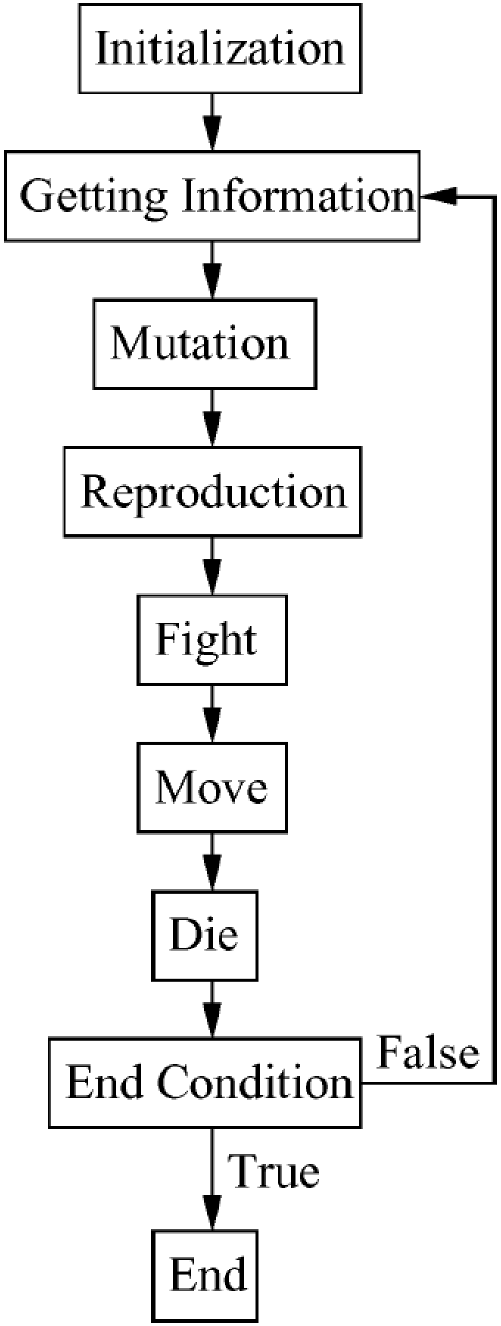
Algoritmo – Exclusão de Protótipos

- $$P_{die}(r_i) = \begin{cases} 0 & , quality(r_i) > 0.5 \\ 1 - 2 * quality(r_i), & quality(r_i) \leq 0.5 \end{cases}$$



Algoritmo – Condição de Parada

- Número máximo de iterações
- Acurácia desejada
- Convergência do número de protótipos
- Convergência da acurácia
- Combinação das abordagens citadas acima



Avaliação – ENPC – Bases Desbalanceadas

Dataset	Iterações	Gen. Accuracy		Maj. Accuracy		Min. Accuracy		AUC. Accuracy		Data Reduction	
glass1	100	0.80	0.03	0.85	0.04	0.71	0.04	0.78	0.03	0.89	0.04
	200	0.81	0.05	0.86	0.03	0.71	0.09	0.79	0.05	0.87	0.03
ecoli-0_vs_1	100	0.98	0.02	0.95	0.05	1.00	0.00	0.97	0.02	0.94	0.03
	200	0.96	0.02	0.91	0.07	0.99	0.01	0.95	0.03	0.95	0.02
iris0	100	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	0.94	0.02
	200	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	0.93	0.03
glass0	100	0.79	0.07	0.81	0.06	0.74	0.12	0.78	0.08	0.89	0.02
	200	0.81	0.08	0.83	0.04	0.76	0.25	0.80	0.12	0.84	0.06
ecoli1	100	0.91	0.04	0.95	0.05	0.78	0.13	0.87	0.06	0.94	0.03
	200	0.91	0.03	0.95	0.04	0.80	0.01	0.88	0.01	0.96	0.01
new-thyroid2	100	0.95	0.04	0.98	0.02	0.83	0.17	0.90	0.09	0.96	0.02
	200	0.97	0.02	0.98	0.02	0.94	0.11	0.96	0.05	0.97	0.01
new-thyroid1	100	0.98	0.02	0.98	0.02	1.00	0.00	0.99	0.01	0.95	0.03
	200	0.99	0.02	0.98	0.02	1.00	0.00	0.99	0.01	0.96	0.02
ecoli2	100	0.95	0.02	0.96	0.03	0.91	0.06	0.93	0.03	0.97	0.01
	200	0.96	0.02	0.97	0.02	0.91	0.06	0.94	0.03	0.96	0.01
glass6	100	0.90	0.06	0.91	0.07	0.83	0.15	0.87	0.08	0.99	0.00
	200	0.90	0.05	0.91	0.06	0.83	0.15	0.87	0.07	0.98	0.01
glass2	100	0.87	0.07	0.93	0.10	0.15	0.30	0.54	0.10	0.95	0.04
	200	0.89	0.04	0.97	0.05	0.00	0.00	0.48	0.02	0.97	0.01
shuttle-c2-vs-c4	100	0.99	0.02	1.00	0.00	0.90	0.20	0.95	0.10	0.92	0.01
	200	0.99	0.02	1.00	0.00	0.90	0.20	0.95	0.10	0.95	0.02
glass-0-1-6_vs_5	100	0.95	0.01	0.99	0.01	0.10	0.20	0.55	0.09	0.98	0.02
	200	0.94	0.03	0.98	0.03	0.10	0.20	0.54	0.08	0.98	0.02

Avaliação – ENPC x KNN – Bases Desbalanceadas

Dataset	Algoritmo	Gen. Accuracy		Maj. Accuracy		Min. Accuracy		AUC. Accuracy		Data Reduction	
glass1	ENPC - 200	0.81	0.05	0.86	0.03	0.71	0.09	0.79	0.05	0.87	0.03
	KNN	0.81	0.05	0.89	0.02	0.67	0.12	0.78	0.06	-	-
ecoli-0_vs_1	ENPC - 100	0.98	0.02	0.95	0.05	1.00	0.00	0.97	0.02	0.94	0.03
	KNN	0.97	0.02	0.95	0.07	0.98	0.03	0.96	0.03	-	-
iris0	ENPC - 200	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	0.93	0.03
	KNN	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	-	-
glass0	ENPC - 200	0.81	0.08	0.83	0.04	0.76	0.25	0.80	0.12	0.84	0.06
	KNN	0.84	0.06	0.86	0.08	0.80	0.08	0.83	0.06	-	-
ecoli1	ENPC - 200	0.91	0.03	0.95	0.04	0.80	0.01	0.88	0.01	0.96	0.01
	KNN	0.86	0.04	0.92	0.04	0.67	0.12	0.80	0.06	-	-
new-thyroid2	ENPC - 200	0.97	0.02	0.98	0.02	0.94	0.11	0.96	0.05	0.97	0.01
	KNN	0.99	0.01	0.99	0.01	0.97	0.06	0.98	0.03	-	-
new-thyroid1	ENPC - 200	0.99	0.02	0.98	0.02	1.00	0.00	0.99	0.01	0.96	0.02
	KNN	0.98	0.01	0.98	0.01	0.97	0.06	0.98	0.02	-	-
ecoli2	ENPC - 200	0.96	0.02	0.97	0.02	0.91	0.06	0.94	0.03	0.96	0.01
	KNN	0.95	0.04	0.96	0.04	0.87	0.12	0.92	0.07	-	-
glass6	ENPC - 200	0.90	0.05	0.91	0.06	0.83	0.15	0.87	0.07	0.98	0.01
	KNN	0.96	0.02	0.99	0.01	0.79	0.14	0.89	0.07	-	-
glass2	ENPC - 200	0.89	0.04	0.97	0.05	0.00	0.00	0.48	0.02	0.97	0.01
	KNN	0.87	0.05	0.92	0.05	0.22	0.19	0.57	0.11	-	-
shuttle-c2-vs-c4	ENPC - 100	0.99	0.02	1.00	0.00	0.90	0.20	0.95	0.10	0.92	0.01
	KNN	0.99	0.02	1.00	0.00	0.90	0.20	0.95	0.10	-	-
glass-0-1-6_vs_5	ENPC - 100	0.95	0.01	0.99	0.01	0.10	0.20	0.55	0.09	0.98	0.02
	KNN	0.96	0.02	0.97	0.02	0.80	0.24	0.89	0.12	-	-

Avaliação – ENPC – Bases Balanceadas

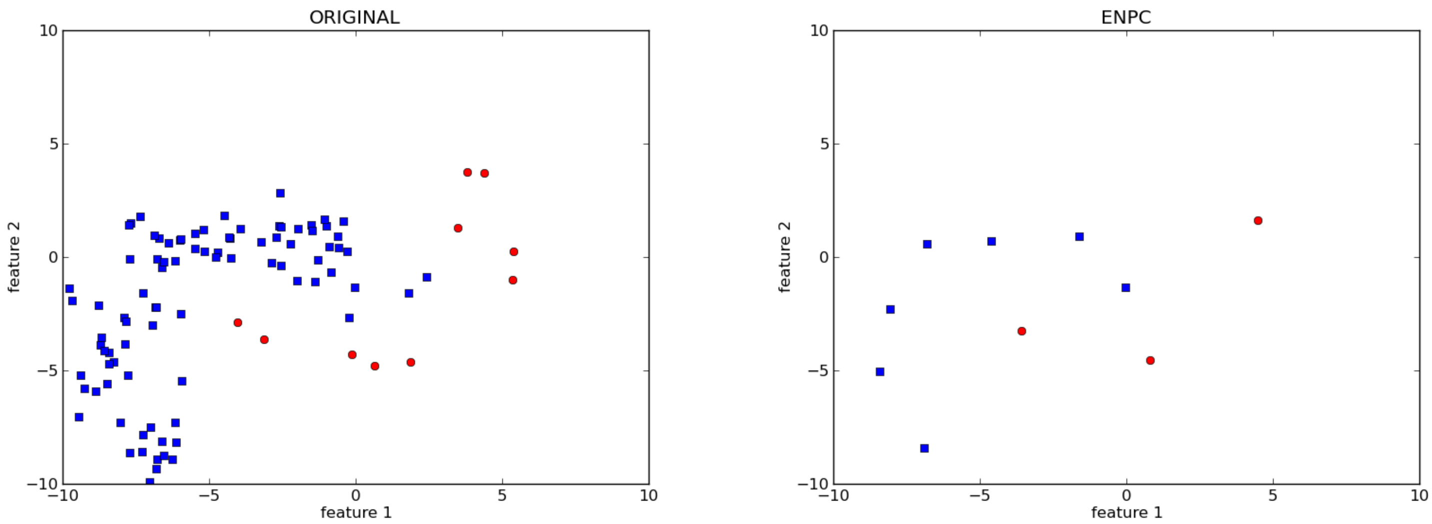
Dataset	Iterações	Gen. Accuracy		AUC. Accuracy		Data Reduction	
glass	100	0.71	0.09	0.72	0.09	0.68	0.01
	200	0.70	0.09	0.71	0.10	0.68	0.02
image_segmentation	100	0.93	0.01	0.96	0.01	0.98	0.00
	200	0.92	0.02	0.95	0.01	0.98	0.00
ionosphere	100	0.88	0.05	0.86	0.06	0.98	0.01
	200	0.89	0.03	0.87	0.03	0.97	0.01
iris	100	0.97	0.04	0.96	0.06	0.95	0.01
	200	0.97	0.03	0.96	0.04	0.94	0.02
liver	100	0.58	0.08	0.57	0.09	0.66	0.02
	200	0.62	0.08	0.61	0.08	0.66	0.01
pendigits	50	0.94	0.03	0.92	0.03	1.00	0.00
	100	0.95	0.01	0.92	0.03	1.00	0.00
pima_diabetes	100	0.71	0.06	0.68	0.07	0.79	0.03
	200	0.70	0.06	0.68	0.08	0.76	0.02
sonar	100	0.88	0.05	0.87	0.05	0.88	0.01
	200	0.86	0.06	0.85	0.06	0.88	0.02
spambase	100	0.82	0.03	0.81	0.02	1.00	0.00
	200	0.82	0.03	0.81	0.04	1.00	0.00
vehicle	100	0.65	0.03	0.60	0.05	0.73	0.01
	200	0.66	0.05	0.61	0.06	0.73	0.01
vowel	100	0.95	0.04	0.96	0.05	0.84	0.00
	200	0.96	0.02	0.97	0.03	0.84	0.00
wine	100	0.95	0.05	0.96	0.05	0.94	0.02
	200	0.97	0.04	0.96	0.05	0.95	0.01
yeast	100	0.48	0.03	0.47	0.04	0.57	0.07
	200	0.50	0.03	0.49	0.03	0.51	0.01

Avaliação – ENPC x KNN – Bases Balanceadas

Dataset	Iterações	Gen. Accuracy		AUC. Accuracy		Data Reduction	
glass	ENPC - 100	0.71	0.09	0.72	0.09	0.68	0.01
	KNN	0.70	0.05	0.71	0.06	-	-
image_segmentation	ENPC - 100	0.93	0.01	0.96	0.01	0.98	0.00
	KNN	0.97	0.01	0.98	0.00	-	-
ionosphere	ENPC - 200	0.89	0.03	0.87	0.03	0.97	0.01
	KNN	0.86	0.04	0.82	0.05	-	-
iris	ENPC - 200	0.97	0.03	0.96	0.04	0.94	0.02
	KNN	0.95	0.05	0.95	0.07	-	-
liver	ENPC - 200	0.62	0.08	0.61	0.08	0.66	0.01
	KNN	0.62	0.06	0.61	0.06	-	-
pendigits	ENPC - 100	0.95	0.01	0.92	0.03	1.00	0.00
	KNN	0.99	0.00	0.99	0.00	-	-
pima_diabetes	ENPC - 100	0.71	0.06	0.68	0.07	0.79	0.03
	KNN	0.70	0.05	0.66	0.06	-	-
sonar	ENPC - 100	0.88	0.05	0.87	0.05	0.88	0.01
	KNN	0.87	0.11	0.86	0.12	-	-
spambase	ENPC - 100	0.82	0.03	0.81	0.02	1.00	0.00
	KNN	0.91	0.01	0.91	0.01	-	-
vehicle	ENPC - 200	0.66	0.05	0.61	0.06	0.73	0.01
	KNN	0.70	0.05	0.63	0.07	-	-
vowel	ENPC - 200	0.96	0.02	0.97	0.03	0.84	0.00
	KNN	0.99	0.02	0.98	0.04	-	-
wine	ENPC - 200	0.97	0.04	0.96	0.05	0.95	0.01
	KNN	0.96	0.03	0.97	0.02	-	-
yeast	ENPC - 100	0.50	0.03	0.49	0.03	0.51	0.01
	KNN	0.52	0.04	0.51	0.05	-	-

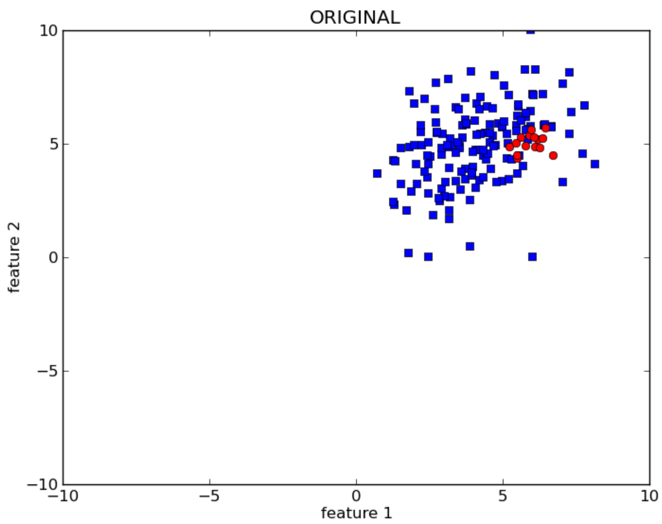
Avaliação – Visualização da redução - Banana

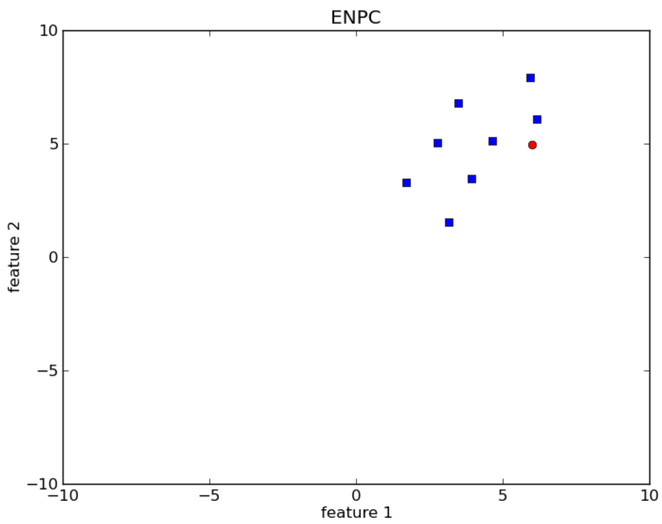
- 200 Iterações



Avaliação – Visualização da redução - Normal

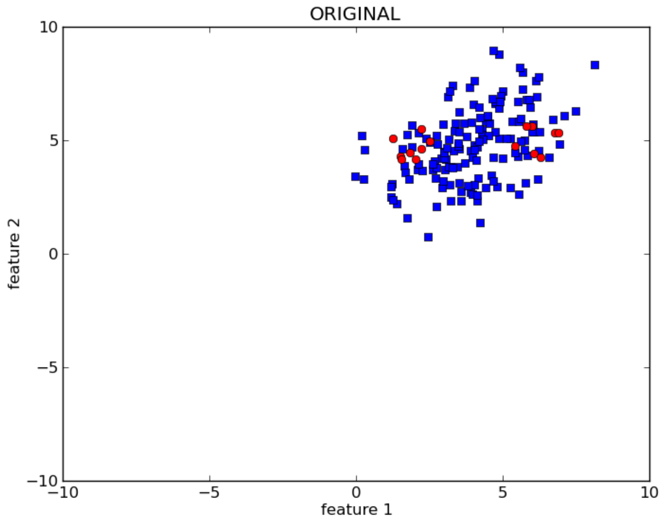
- 200 Iterações

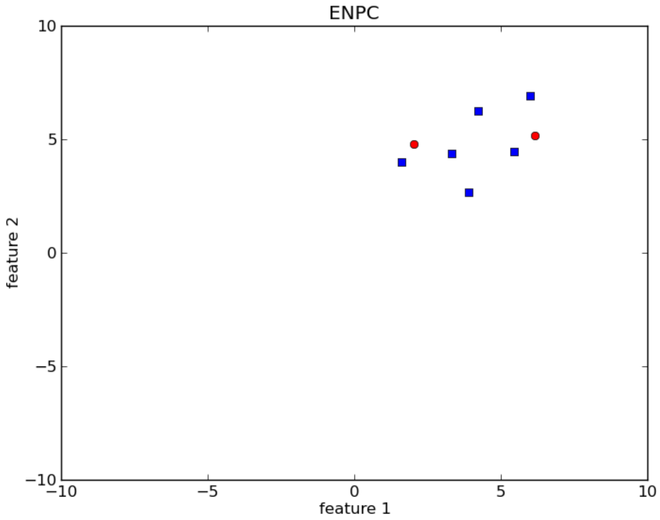




Avaliação – Visualização da redução – Norma Multimodal

- 200 Iterações





Conclusão

- Ótimo desempenho em bases desbalanceadas, mas dependendo da disposição dos dados, tende a não gerar protótipos da classe minoritária. Chega a ser melhor que o KNN;
- Bom desempenho em bases balanceadas chegando a ser tão bom quanto o KNN, perdendo, em alguns casos, por pouco.

Evolutionary Design of Nearest Prototype Classifiers

Tiago José – tjs2

Tiago Neves - tnb