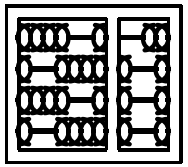


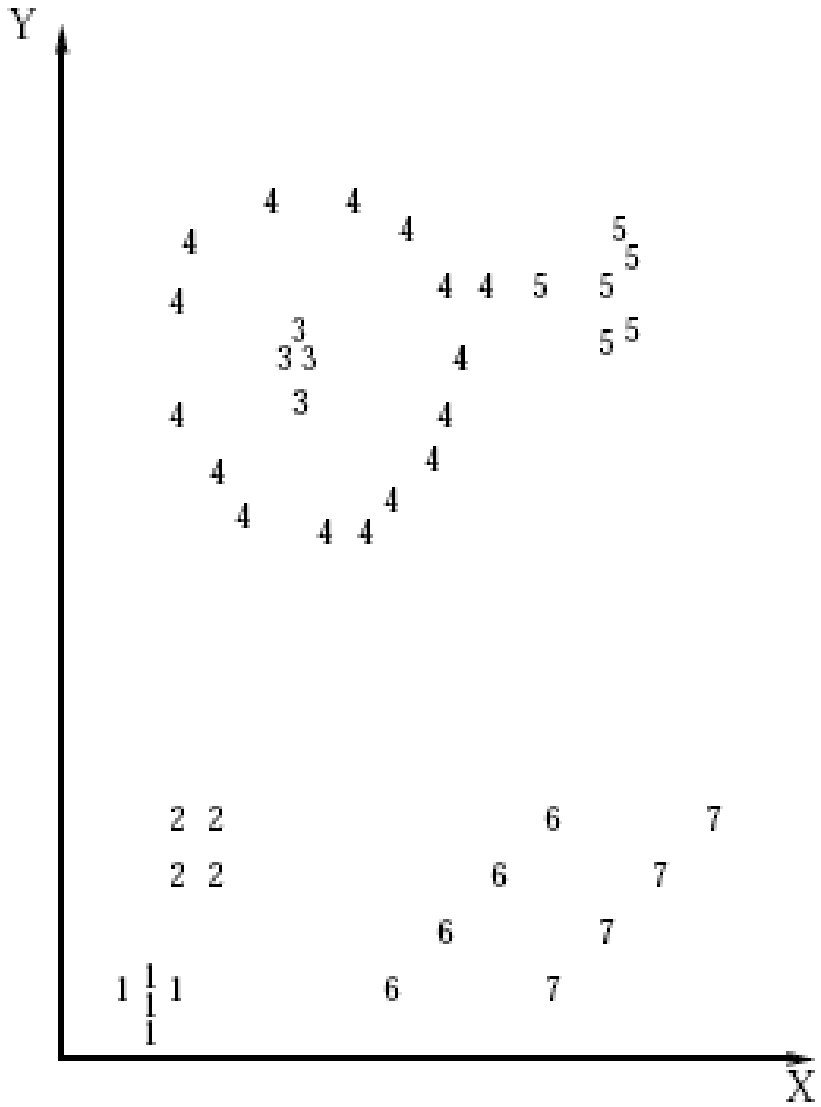
Universidade Estadual de Campinas - UNICAMP
Instituto de Computação - IC

Busca por Similaridade em Espaços Métricos utilizando Técnicas de Agrupamento de Dados

Jurandy Gomes de Almeida Junior

Campinas, 8 de outubro de 2008.

Problema



Soluções

- Busca seqüencial
- Métodos de acesso
  - Espaciais
  - Métricos

Soluções

- Busca seqüencial
- Métodos de acesso
  - Espaciais
  - Métricos
    - Estáticos
    - Dinâmicos

Soluções

- Busca seqüencial
- Métodos de acesso
  - Espaciais
  - Métricos
    - Estáticos
    - Dinâmicos

Soluções

- Busca seqüencial
- Métodos de acesso
  - Espaciais
  - Métricos
    - Estáticos
    - Dinâmicos

Soluções

- Busca seqüencial
- Métodos de acesso
  - Espaciais
  - Métricos
    - Estáticos
    - Dinâmicos

Soluções

- Busca seqüencial
- Métodos de acesso
  - Espaciais
  - Métricos
    - Estáticos
    - Dinâmicos

Conceitos básicos

- Espaço métrico
  - Positividade
  - Simetria
  - Identidade
  - Desigualdade triangular
- Funções de distância
  - Discreta
  - Contínua
- Consultas por similaridade
  - Abrangência
  - Vizinhos próximos
  - Vizinhos próximos reverso
  - Junção
  - Combinação

Conceitos básicos

- Espaço métrico
  - Positividade
  - Simetria
  - Identidade
  - Desigualdade triangular
- Funções de distância
  - Discreta
  - Contínua
- Consultas por similaridade
  - Abrangência
  - Vizinhos próximos
  - Vizinhos próximos reverso
  - Junção
  - Combinação

Conceitos básicos

- Espaço métrico
  - Positividade
  - Simetria
  - Identidade
  - Desigualdade triangular
- Funções de distância
  - Discreta
  - Contínua
- Consultas por similaridade
  - Abrangência
  - Vizinhos próximos
  - Vizinhos próximos reverso
  - Junção
  - Combinação

Conceitos básicos

- Espaço métrico
  - Positividade
  - Simetria
  - Identidade
  - Desigualdade triangular
- Funções de distância
  - Discreta
  - Contínua
- Consultas por similaridade
  - Abrangência
  - Vizinhos próximos
  - Vizinhos próximos reverso
  - Junção
  - Combinação

Conceitos básicos

- Espaço métrico
  - Positividade
  - Simetria
  - Identidade
  - Desigualdade triangular
- Funções de distância
  - Discreta
  - Contínua
- Consultas por similaridade
  - Abrangência
  - Vizinhos próximos
  - Vizinhos próximos reverso
  - Junção
  - Combinação

Conceitos básicos

- Espaço métrico
  - Positividade
  - Simetria
  - Identidade
  - Desigualdade triangular
- Funções de distância
  - Discreta
  - Contínua
- Consultas por similaridade
  - Abrangência
  - Vizinhos próximos
  - Vizinhos próximos reverso
  - Junção
  - Combinação

Conceitos básicos

- Espaço métrico
  - Positividade
  - Simetria
  - Identidade
  - Desigualdade triangular
- Funções de distância
  - Discreta
  - Contínua
- Consultas por similaridade
  - Abrangência
  - Vizinhos próximos
  - Vizinhos próximos reverso
  - Junção
  - Combinação

Conceitos básicos

- Espaço métrico
  - Positividade
  - Simetria
  - Identidade
  - Desigualdade triangular
- Funções de distância
  - Discreta
  - Contínua
- Consultas por similaridade
  - Abrangência
  - Vizinhos próximos
  - Vizinhos próximos reverso
  - Junção
  - Combinação

Conceitos básicos

- Espaço métrico
  - Positividade
  - Simetria
  - Identidade
  - Desigualdade triangular
- Funções de distância
  - Discreta
  - Contínua
- Consultas por similaridade
  - Abrangência
  - Vizinhos próximos
  - Vizinhos próximos reverso
  - Junção
  - Combinação

Conceitos básicos

- Espaço métrico
  - Positividade
  - Simetria
  - Identidade
  - Desigualdade triangular
- Funções de distância
  - Discreta
  - Contínua
- Consultas por similaridade
  - Abrangência
  - Vizinhos próximos
  - Vizinhos próximos reverso
  - Junção
  - Combinação

Conceitos básicos

- Espaço métrico
  - Positividade
  - Simetria
  - Identidade
  - Desigualdade triangular
- Funções de distância
  - Discreta
  - Contínua
- Consultas por similaridade
  - Abrangência
  - Vizinhos próximos
  - Vizinhos próximos reverso
  - Junção
  - Combinação

Conceitos básicos

- Espaço métrico
  - Positividade
  - Simetria
  - Identidade
  - Desigualdade triangular
- Funções de distância
  - Discreta
  - Contínua
- Consultas por similaridade
  - Abrangência
  - Vizinhos próximos
  - Vizinhos próximos reverso
  - Junção
  - Combinação

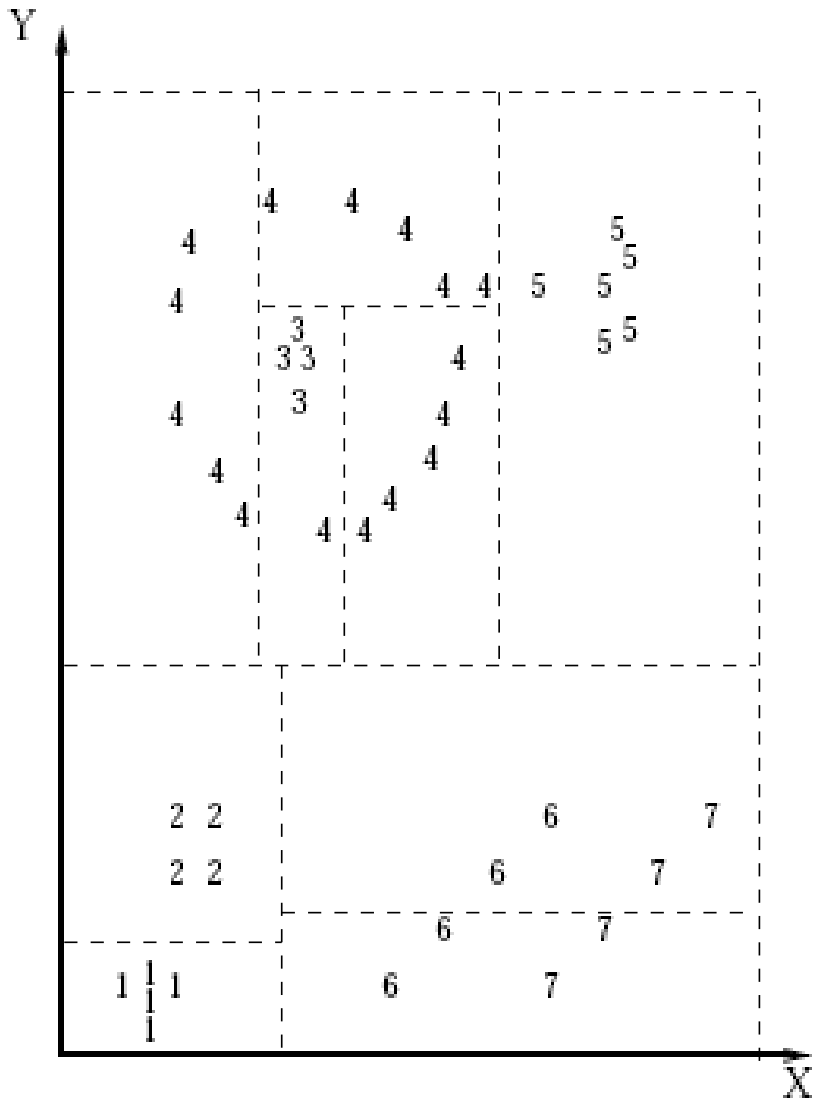
Conceitos básicos

- Espaço métrico
  - Positividade
  - Simetria
  - Identidade
  - Desigualdade triangular
- Funções de distância
  - Discreta
  - Contínua
- Consultas por similaridade
  - Abrangência
  - Vizinhos próximos
  - Vizinhos próximos reverso
  - Junção
  - Combinação

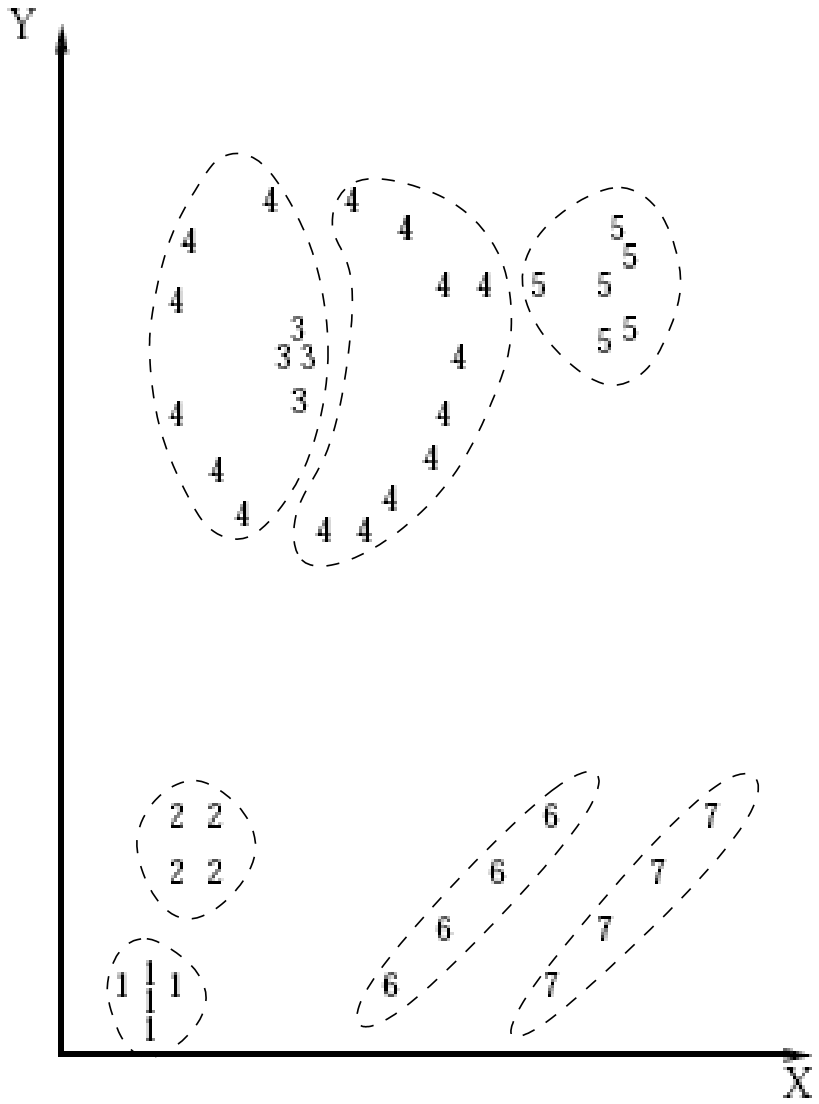
Conceitos básicos

- Espaço métrico
  - Positividade
  - Simetria
  - Identidade
  - Desigualdade triangular
- Funções de distância
  - Discreta
  - Contínua
- Consultas por similaridade
  - Abrangência
  - Vizinhos próximos
  - Vizinhos próximos reverso
  - Junção
  - Combinação

Estruturas de indexação



Agrupamento de dados



Paradigmas de agrupamento

- Particional
- Hierárquico
  - Aglomerativo
  - Divisivo

Paradigmas de agrupamento

- Particional
- Hierárquico
  - Aglomerativo
  - Divisivo

Paradigmas de agrupamento

- Particional
- Hierárquico
  - Aglomerativo
  - Divisivo

Paradigmas de agrupamento

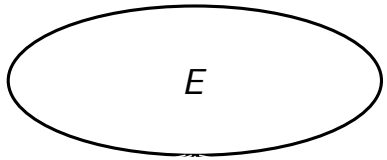
- Particional
- Hierárquico
  - Aglomerativo
  - Divisivo

DAHC

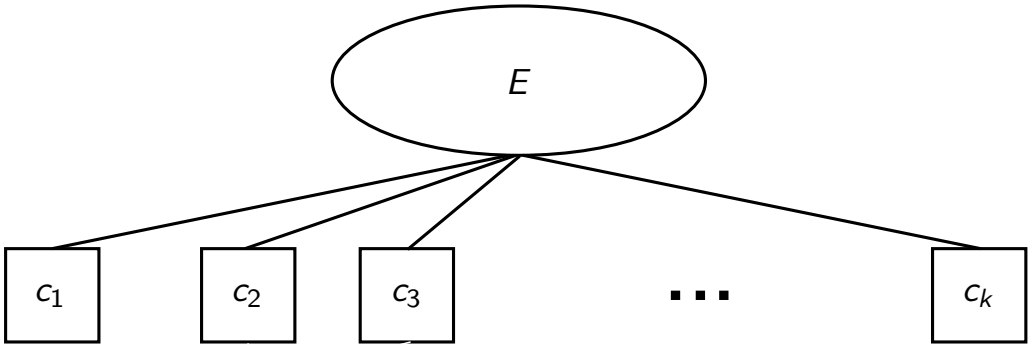
DAHC = Hierárquivo Divisivo +
Hierárquico Aglomerativo +
Fator de Reagrupamento

Símbolo	**Significado**
c, c_{rep}, c_{child}	Um grupo, o elemento representativo do grupo c e um apontador para o nível inferior de c na hierarquia de grupos.
C_i	O conjunto de grupos no i-ésimo nível.
k	O número de grupos em cada tarefa de agrupamento.
$f \in [0, 1)$	O fator de reagrupamento.
E_i	O conjunto de elementos sob análise no i-ésimo nível.
D	Uma medida para avaliar as dissimilaridades entre os elementos em E.

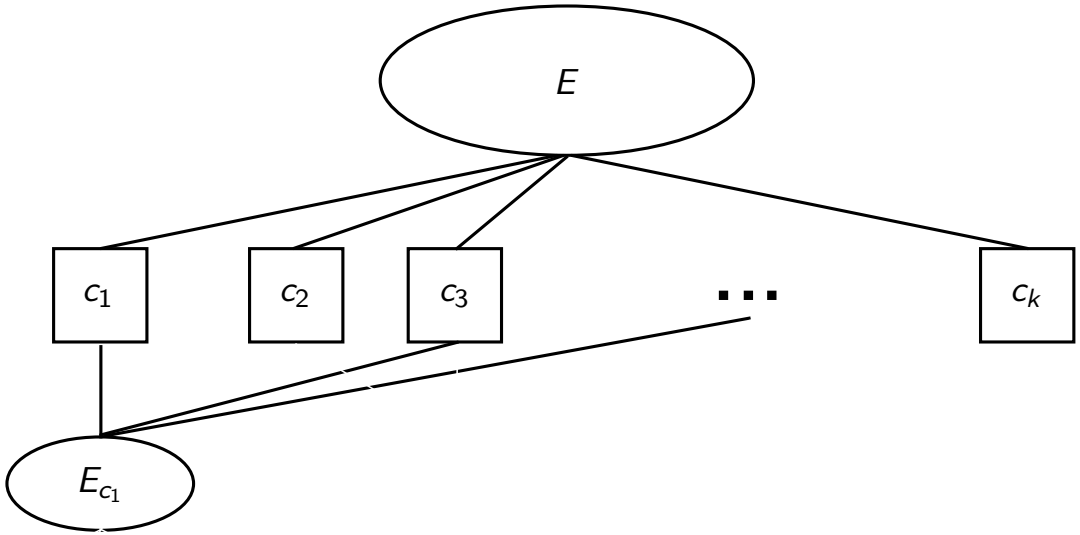
DAHC

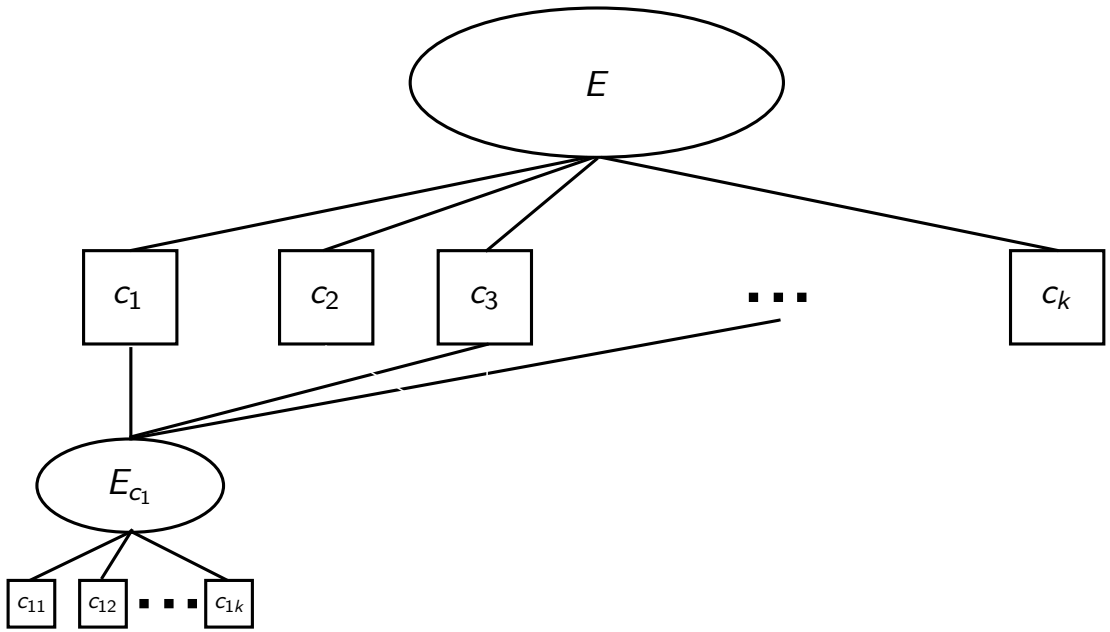


DAHC

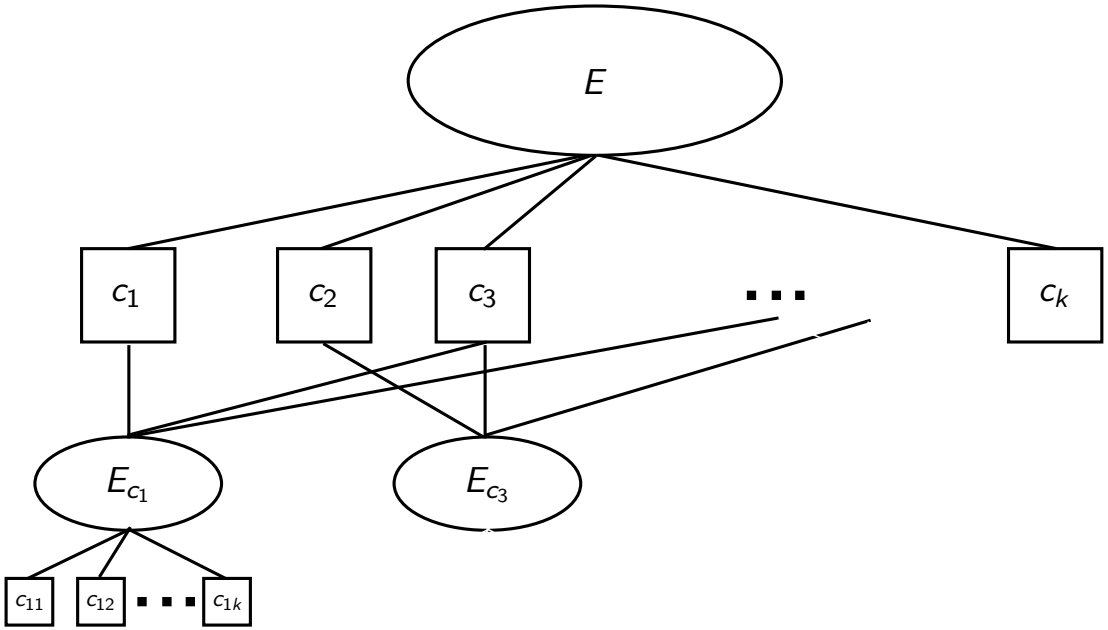


DAHC

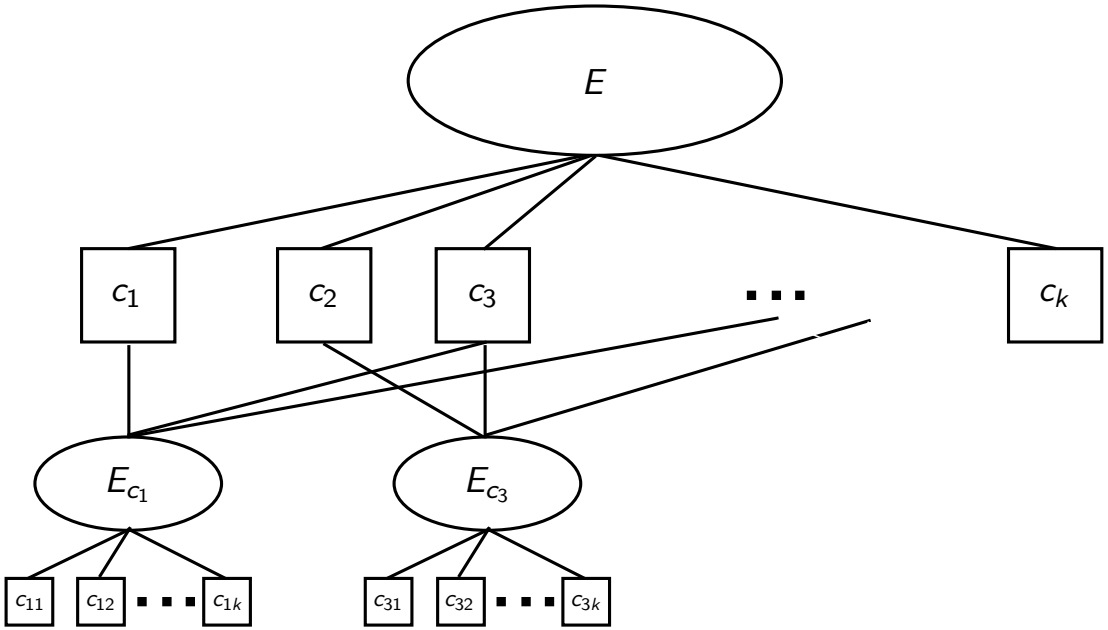



E
c_1
c_2
c_3
c_k
E_{c_1}
c_{11}
c_{12}
c_{1k}

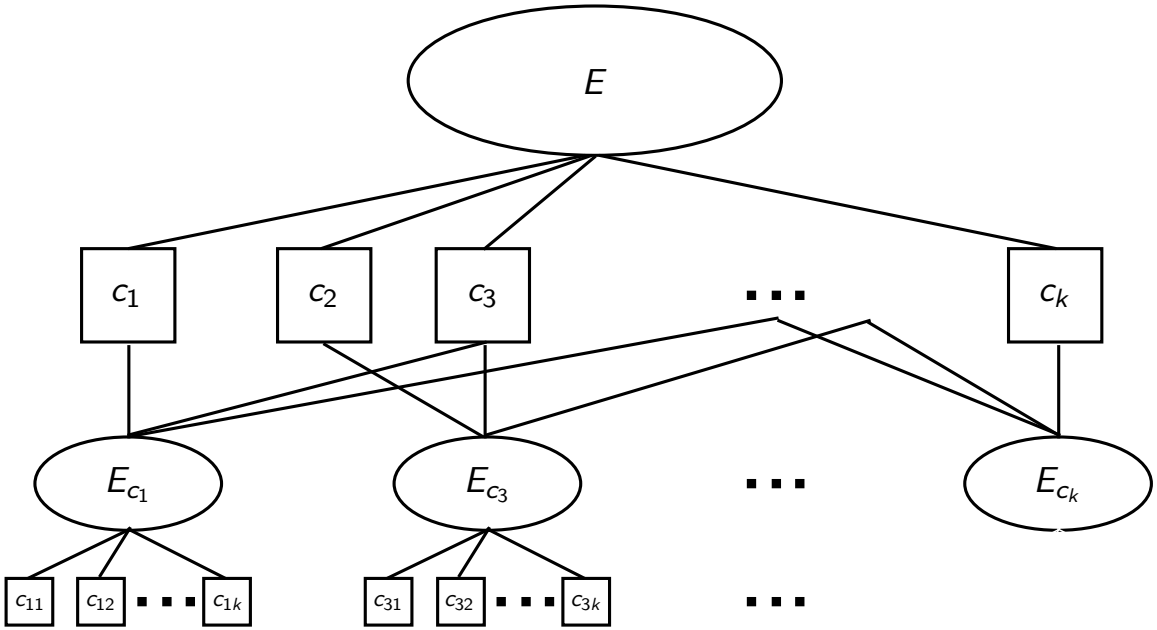
DAHC



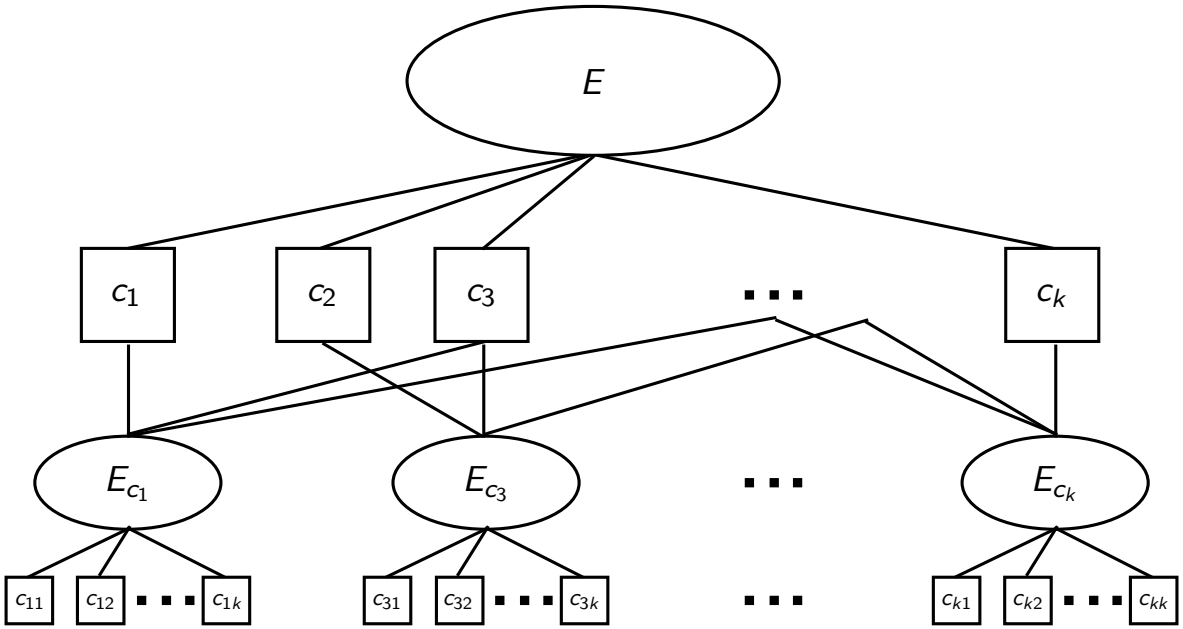
DAHC



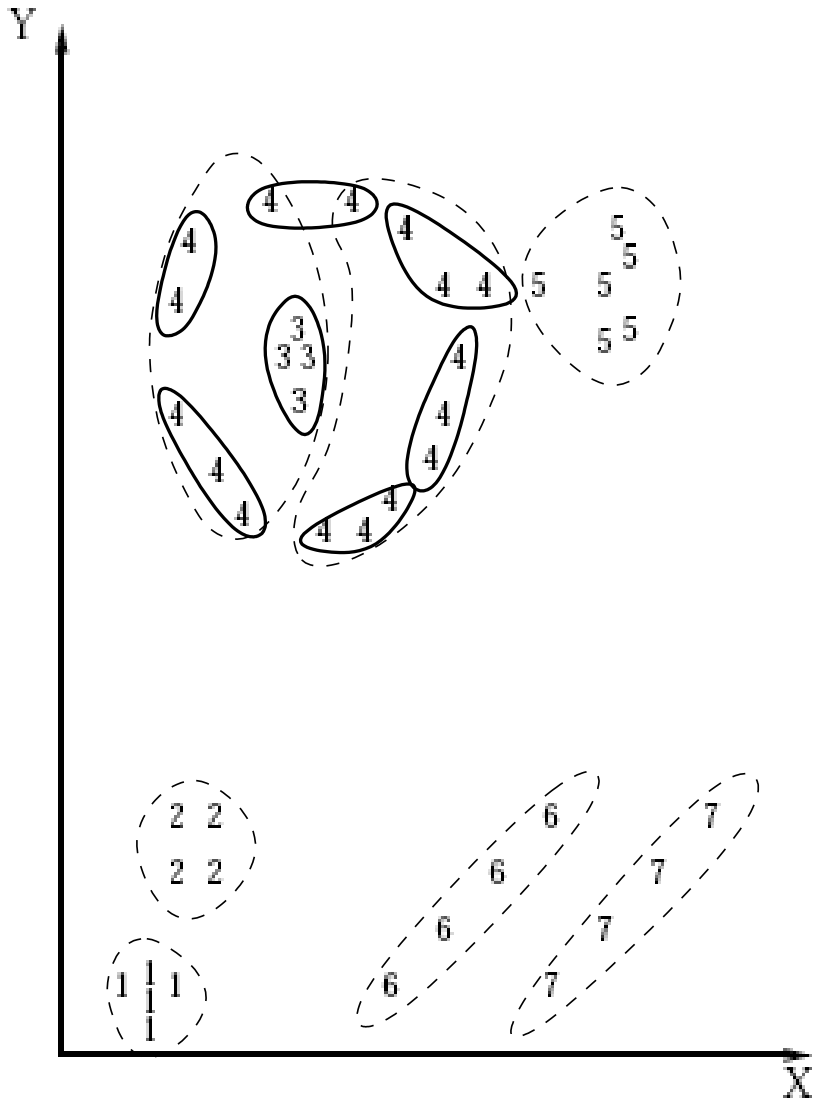
DAHC



DAHC



DAHC



Entrada: O número de grupos k, o fator de reagrupamento $f \in [0, 1)$, o conjunto de elementos E e a função de distância D.

1: **procedimento** DAHC(k, f, E, D)
2: $C \leftarrow \textsc{Cluster}(k, E, D)$ ▷ Etapa divisiva
3: **para** $c \in C$ **faça**
4: $C^* \leftarrow \lfloor f \times k \rfloor$ grupos mais próximos de $c \in C$
5: $E^* \leftarrow \{\}$
6: **para** $c^* \in C^*$ **faça** ▷ Etapa aglomerativa
7: $E^* \leftarrow E^* \cup c^*$
8: **fim para**
9: **se** *cardinalidade*$(E^*) > k$ **então** ▷ Aprofundamento
10: $c_{child} \leftarrow \mathrm{DAHC}(k, f, E^*, D)$
11: **fim se**
12: **fim para**
13: **fim procedimento**

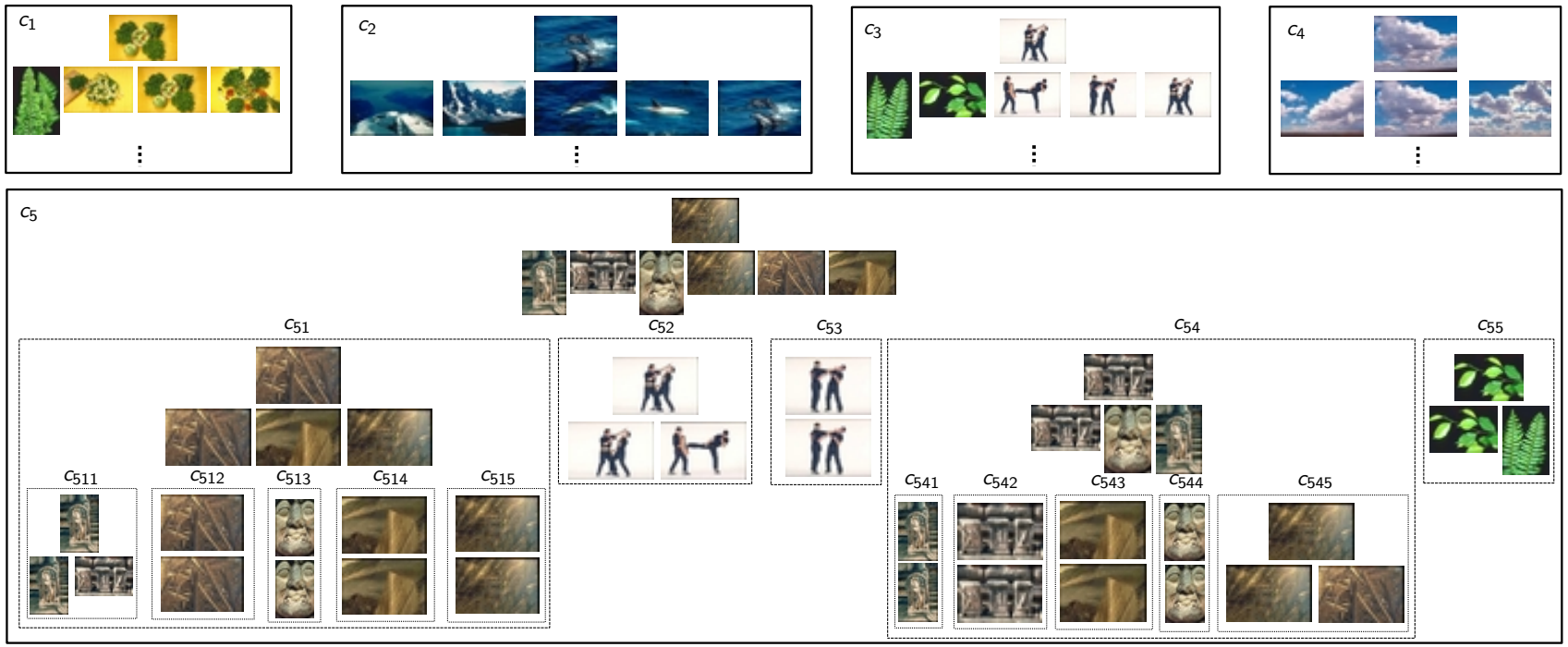


Figura: Exemplo da estrutura hierárquica criada pelo DAHC.

Experimentos

- Consultas por exemplo
- Bancos de dados
  - Corel RRSets
  - FreeFoto
- Validação cruzada
- Medidas
  - Precisão após 30 imagens
  - Número de comparações

Experimentos

- Consultas por exemplo
- Bancos de dados
  - Corel RRSets
  - FreeFoto
- Validação cruzada
- Medidas
  - Precisão após 30 imagens
  - Número de comparações

Experimentos

- Consultas por exemplo
- Bancos de dados
  - Corel RRSets
  - FreeFoto
- Validação cruzada
- Medidas
  - Precisão após 30 imagens
  - Número de comparações

Experimentos

- Consultas por exemplo
- Bancos de dados
  - Corel RRSets
  - FreeFoto
- Validação cruzada
- Medidas
  - Precisão após 30 imagens
  - Número de comparações

Experimentos

- Consultas por exemplo
- Bancos de dados
  - Corel RRSets
  - FreeFoto
- Validação cruzada
- Medidas
  - Precisão após 30 imagens
  - Número de comparações

Experimentos

- Consultas por exemplo
- Bancos de dados
  - Corel RRSets
  - FreeFoto
- Validação cruzada
- Medidas
  - Precisão após 30 imagens
  - Número de comparações

Experimentos

- Consultas por exemplo
- Bancos de dados
  - Corel RRSets
  - FreeFoto
- Validação cruzada
- Medidas
  - Precisão após 30 imagens
  - Número de comparações

Experimentos

- Consultas por exemplo
- Bancos de dados
  - Corel RRSets
  - FreeFoto
- Validação cruzada
- Medidas
  - Precisão após 30 imagens
  - Número de comparações

Experimentos

(a) Precisão após 30 imagens para cada descritor.

	BIC	GCH	CCV
Corel RRSets	53.0%	41.8%	40.7%
FreeFoto	68.1%	55.3%	59.9%

(b) Número de comparações.

Corel RRSets	*FreeFoto*
$\approx 21,000$	$\approx 48,000$

Tabela: Valores de referência utilizando uma busca seqüencial.

Resultados

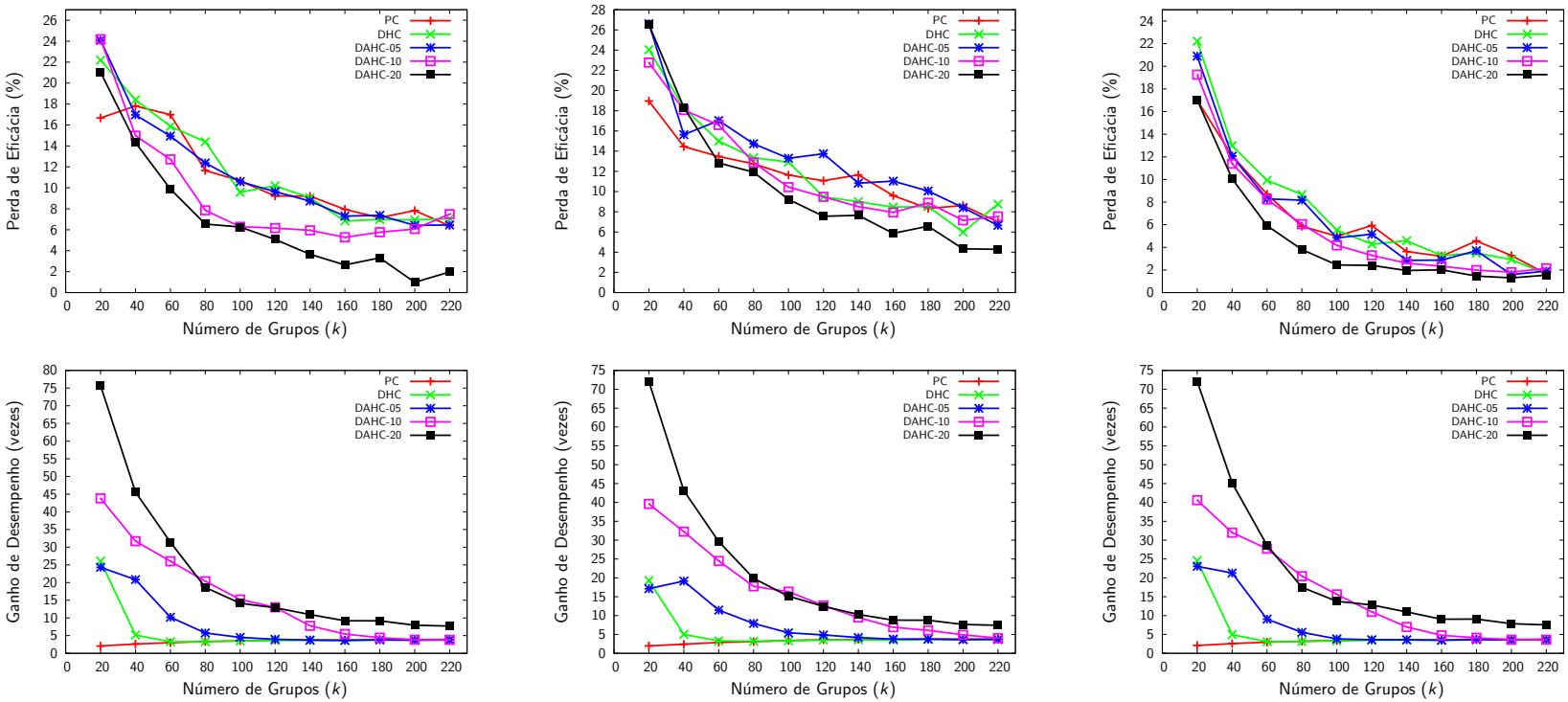


Figura: Perda de eficácia (acima) e ganho de desempenho (abaixo) para o banco *Corel RRSets*. Resultados dos descritores GCH, CCV e BIC, são mostrados, respectivamente, da esquerda para direita.

Resultados

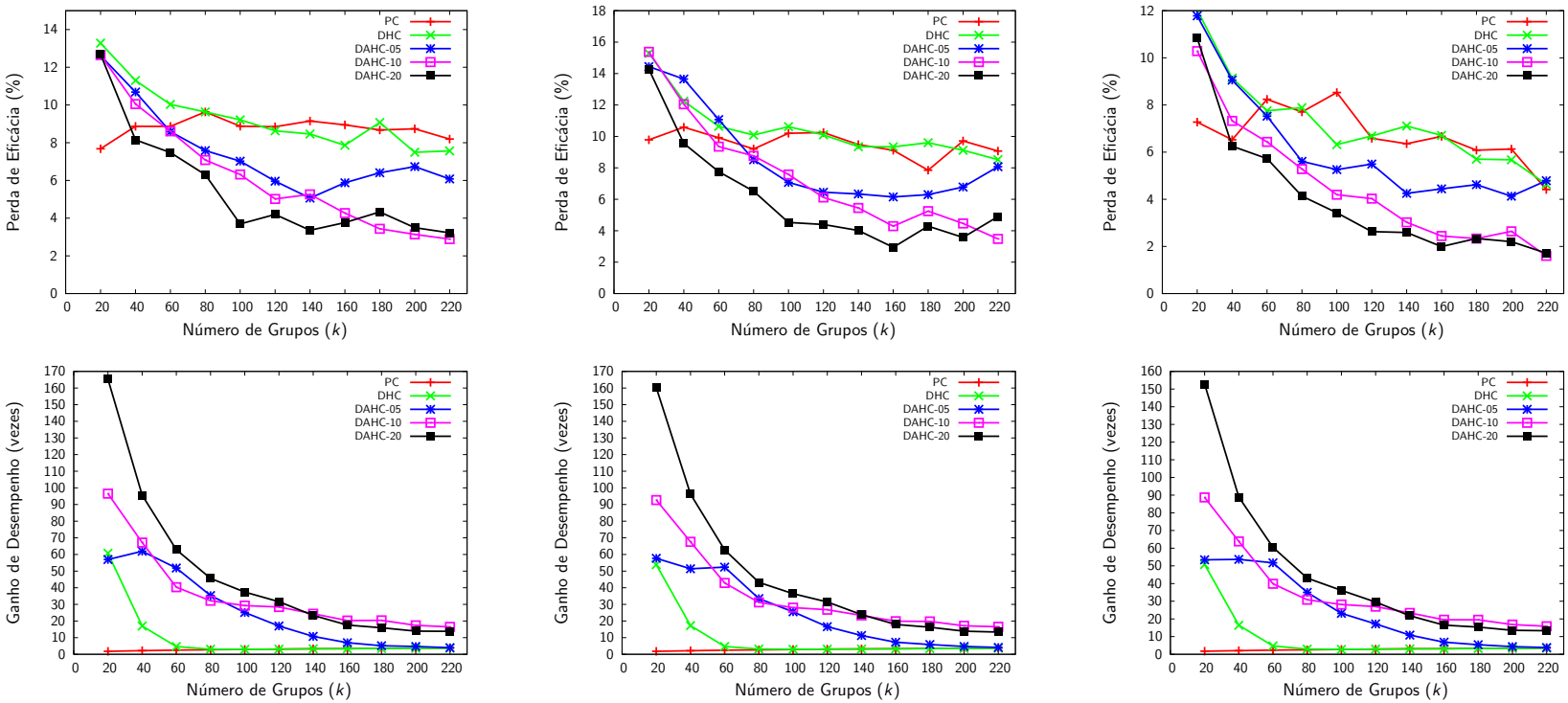


Figura: Perda de eficácia (acima) e ganho de desempenho (abaixo) para o banco *FreeFoto*. Resultados dos descritores GCH, CCV e BIC, são mostrados, respectivamente, da esquerda para direita.

Problemas sendo tratados

1. Suporte a inserção e a remoção de dados
2. Suporte aos diversos tipos de consultas por similaridade
3. Suporte a otimização da estrutura depois de criada

Problemas sendo tratados

1. Suporte a inserção e a remoção de dados
2. Suporte aos diversos tipos de consultas por similaridade
3. Suporte a otimização da estrutura depois de criada

Problemas sendo tratados

1. Suporte a inserção e a remoção de dados
2. Suporte aos diversos tipos de consultas por similaridade
3. Suporte a otimização da estrutura depois de criada

Dúvidas?

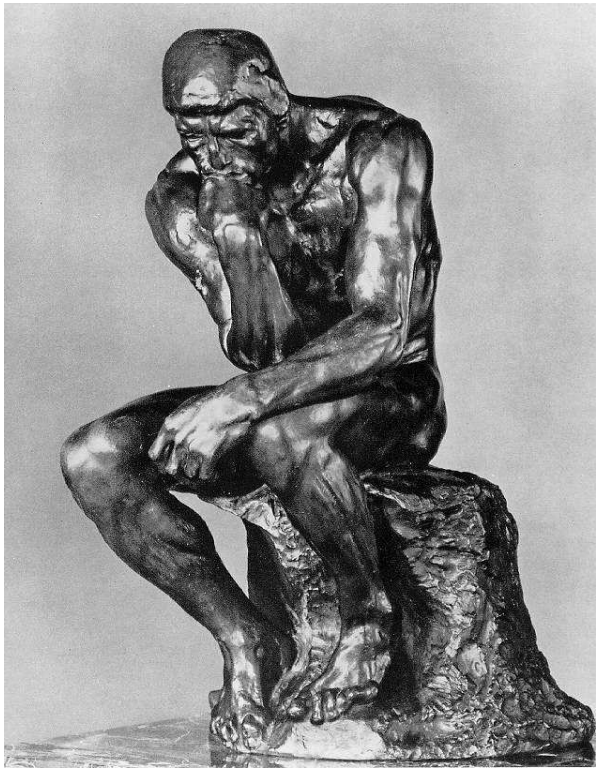


Figura: *O pensador* - Rodin

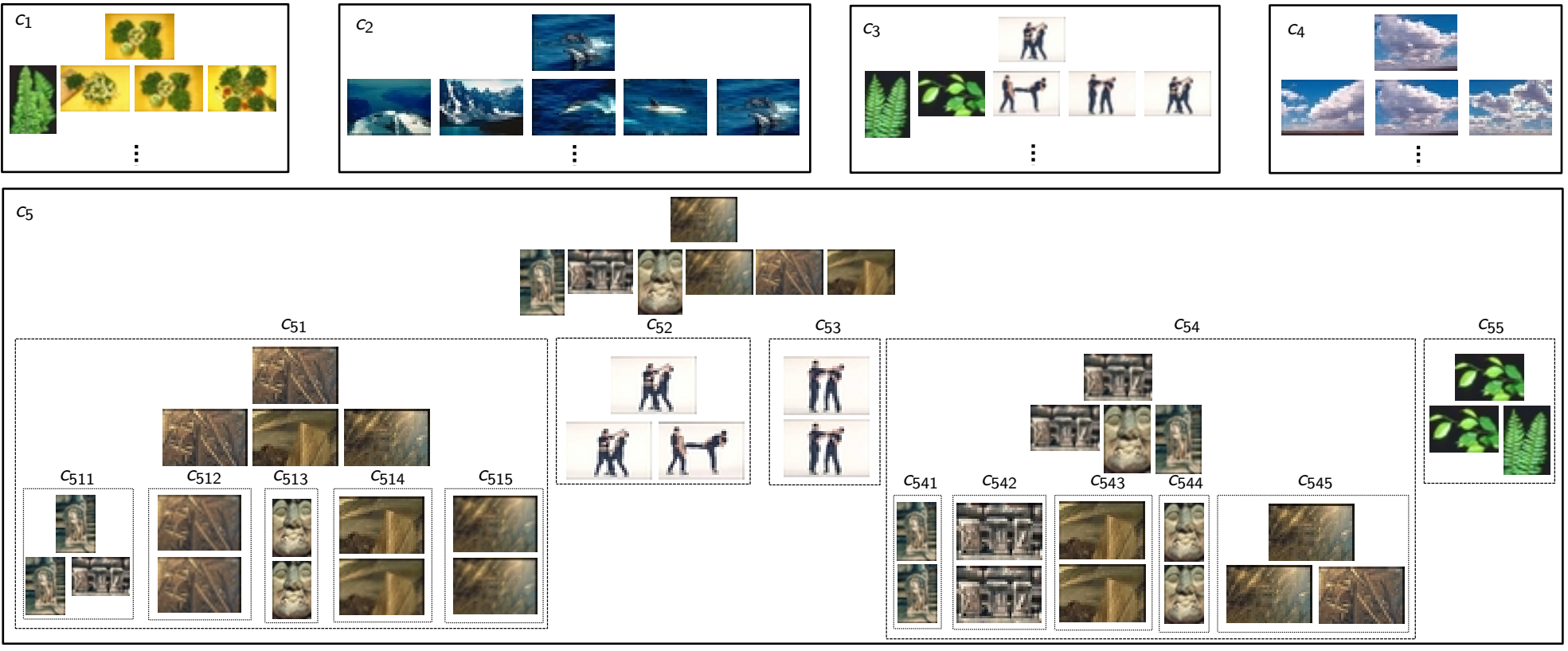


Figura: Exemplo da estrutura hierárquica criada pelo DAHC.

Parada

Locomotivas

Montanhas

Surf

Peixes

Casas

Cervos

Árvores

Cogumelos

Árvores

Praias

Nuvens

Flores

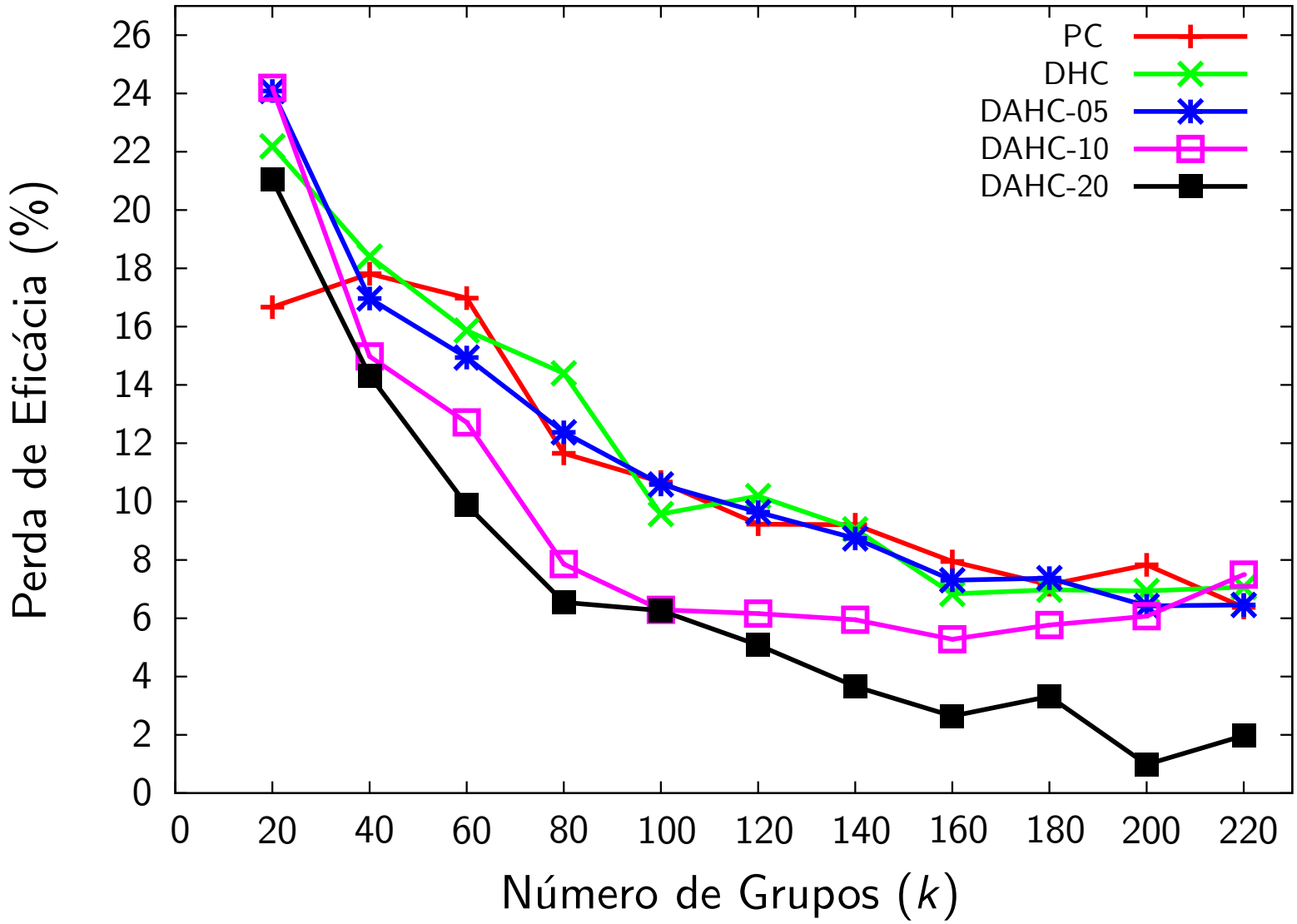
Folhas

Poente

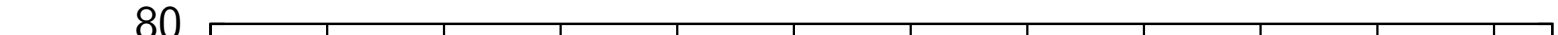
Luar

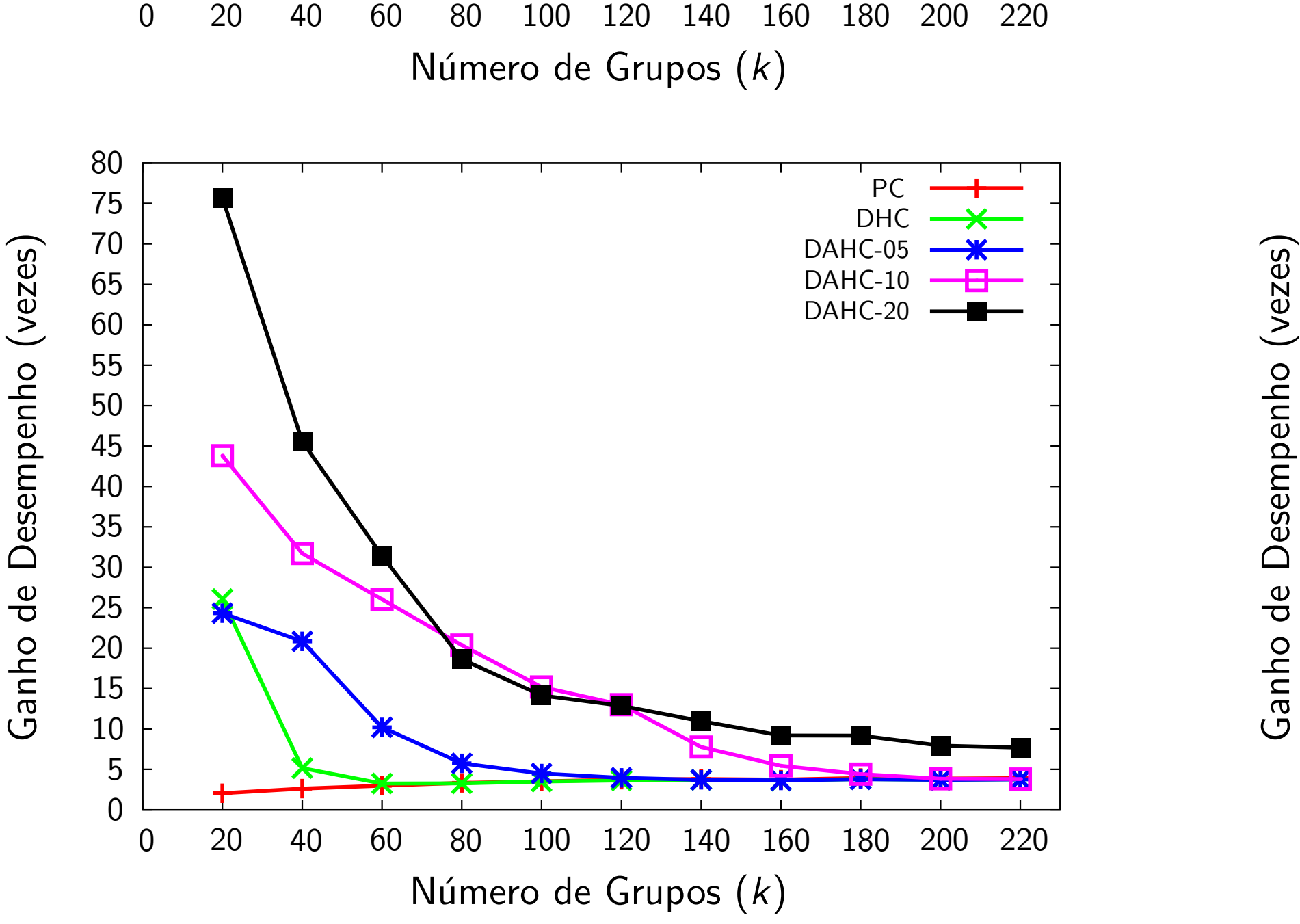
Montanhas

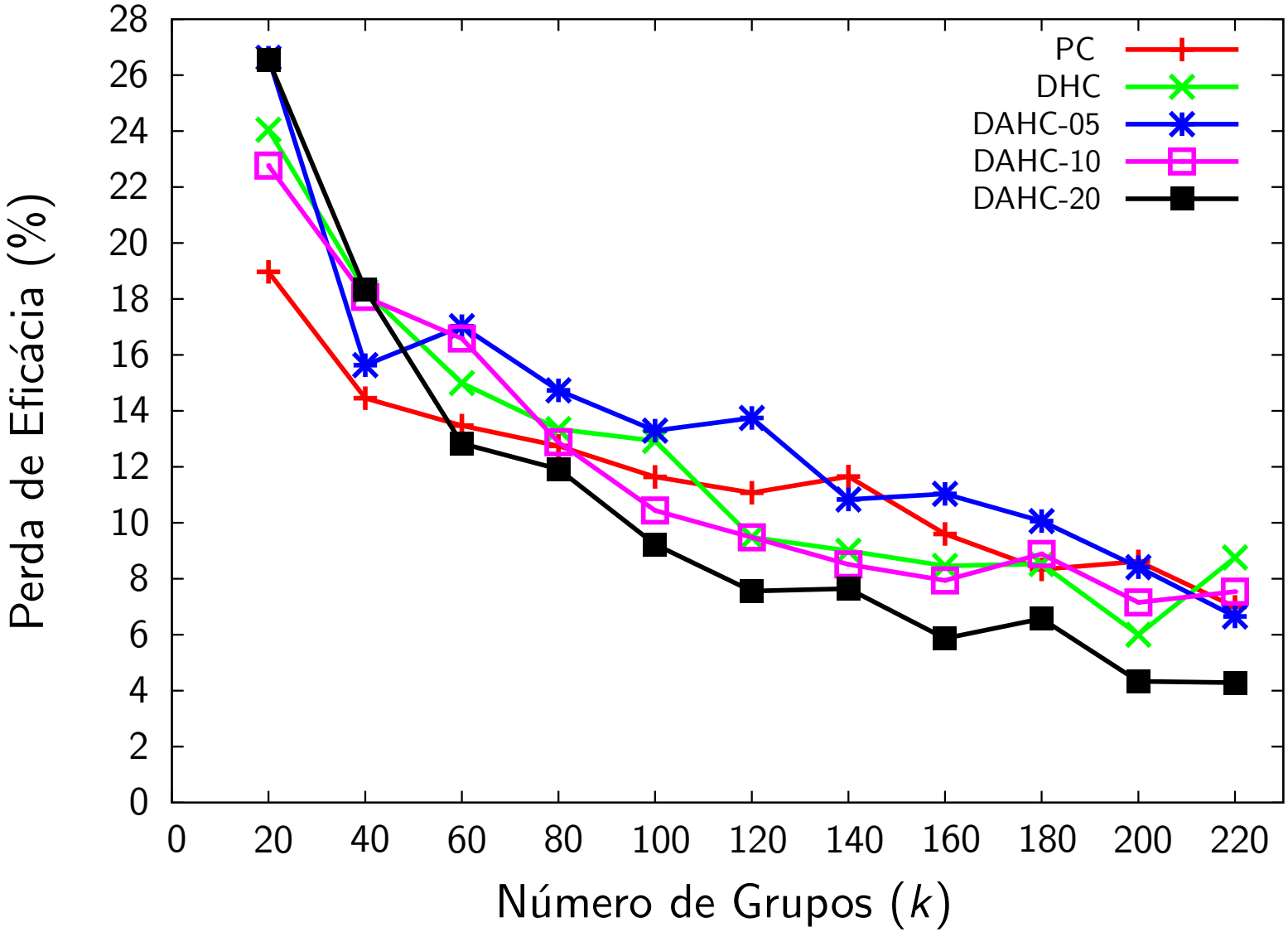
Rios

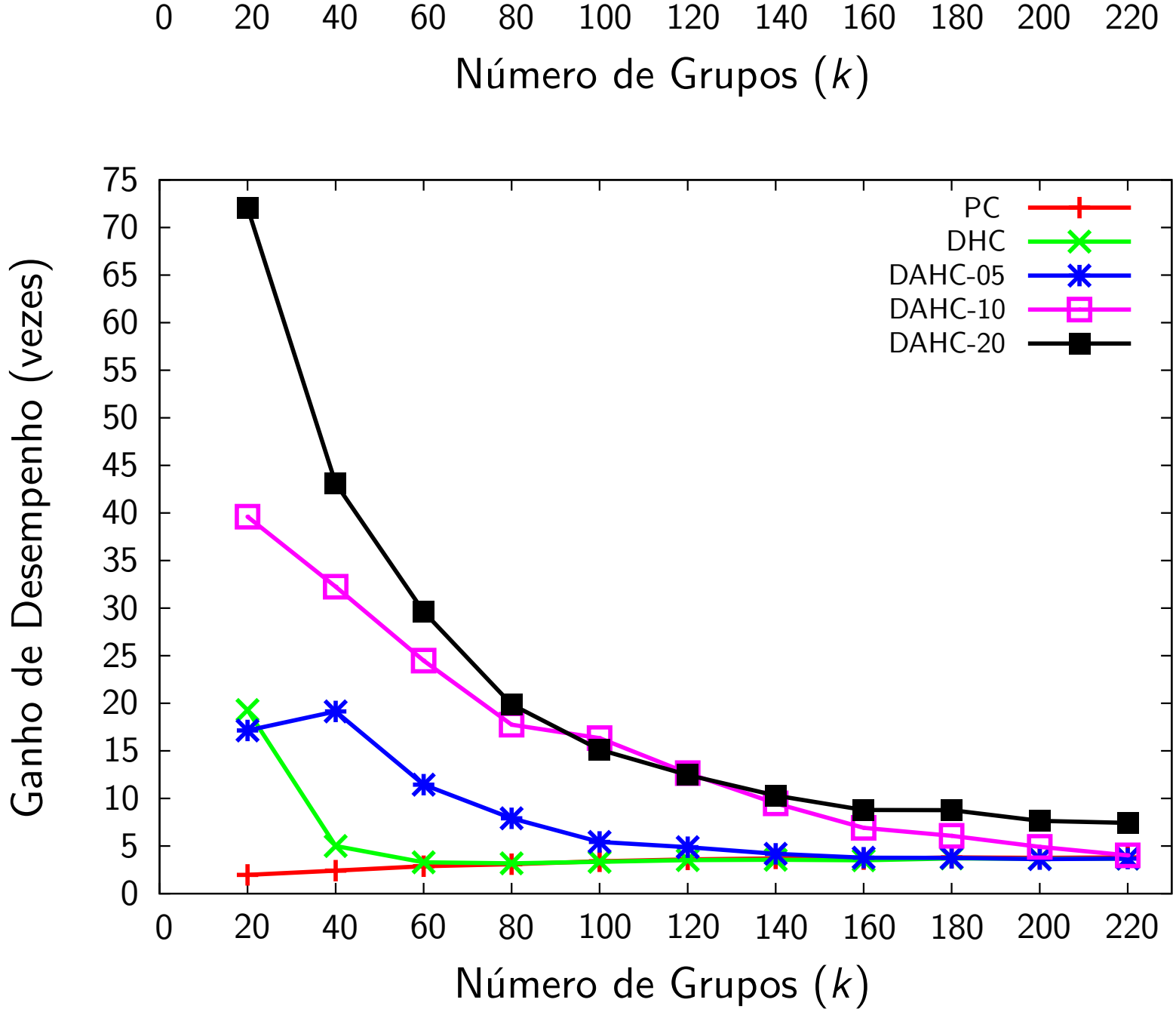

PC
DHC
DAHC-05
DAHC-10
DAHC-20
Perda de Eficácia (%)
Número de Grupos (k)
0
2
4
6
8
10
12
14
16
18
20
22
24
26
0
20
40
60
80
100
120
140
160
180
200
220

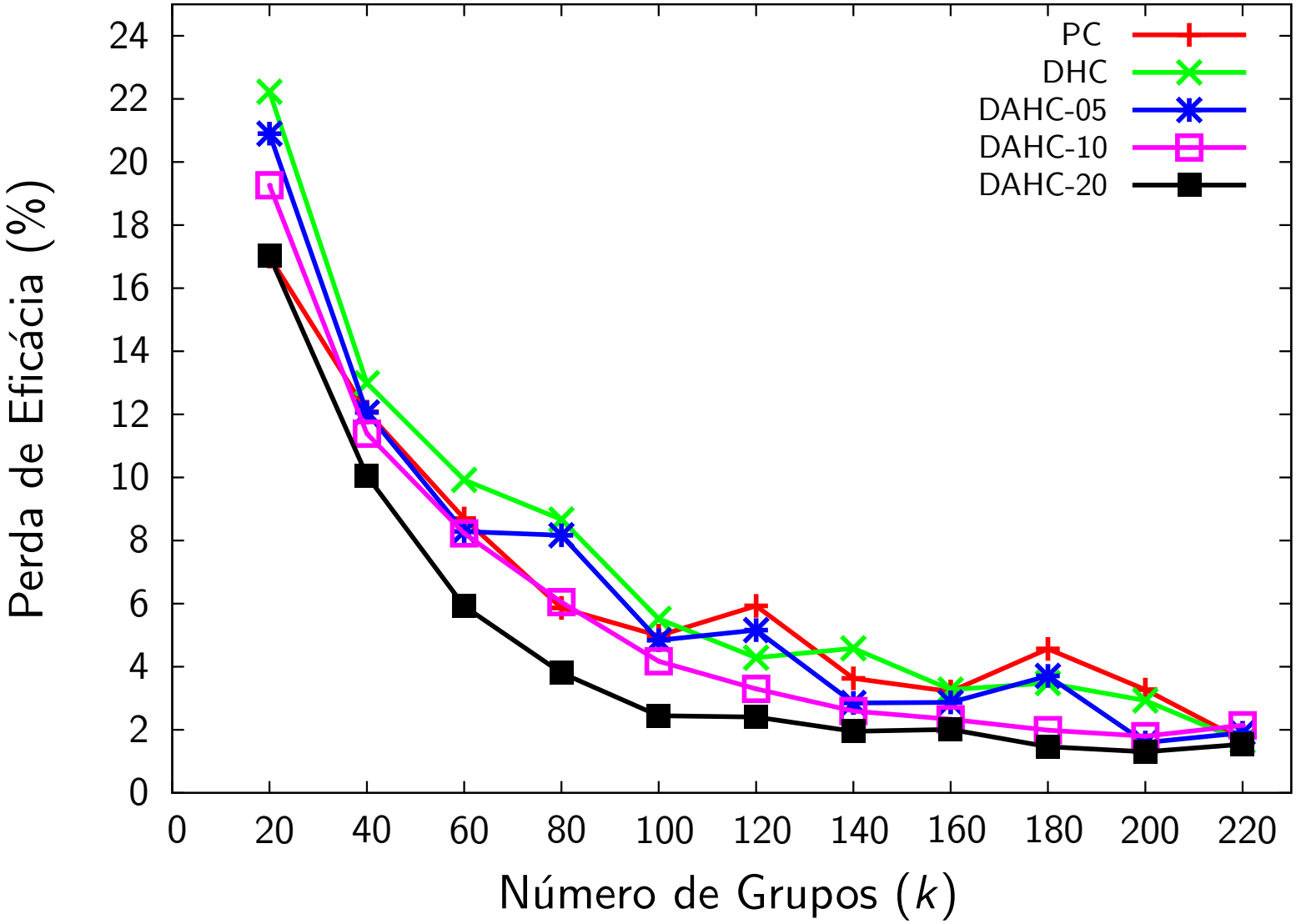

Perda de Eficácia (%)

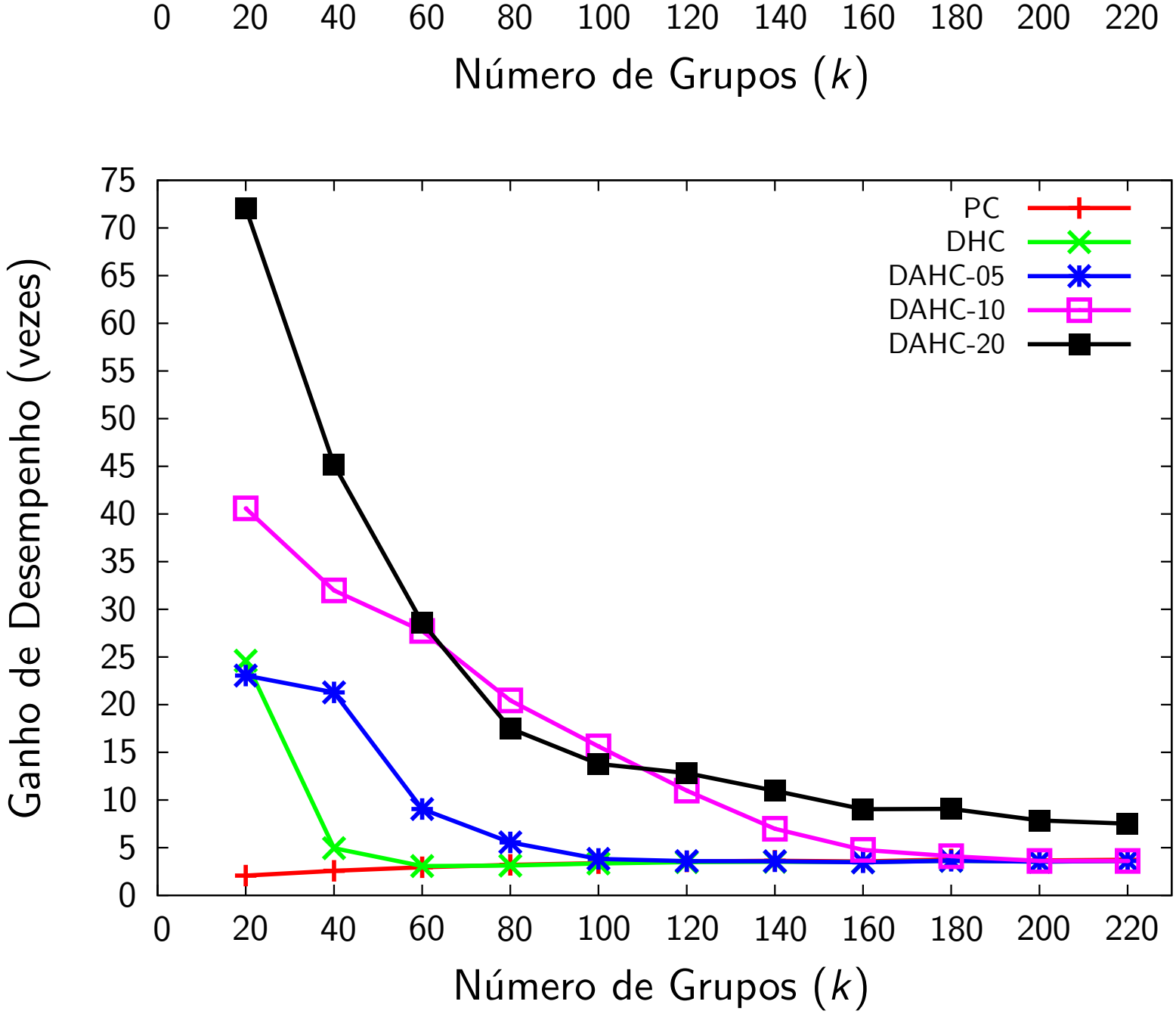

80

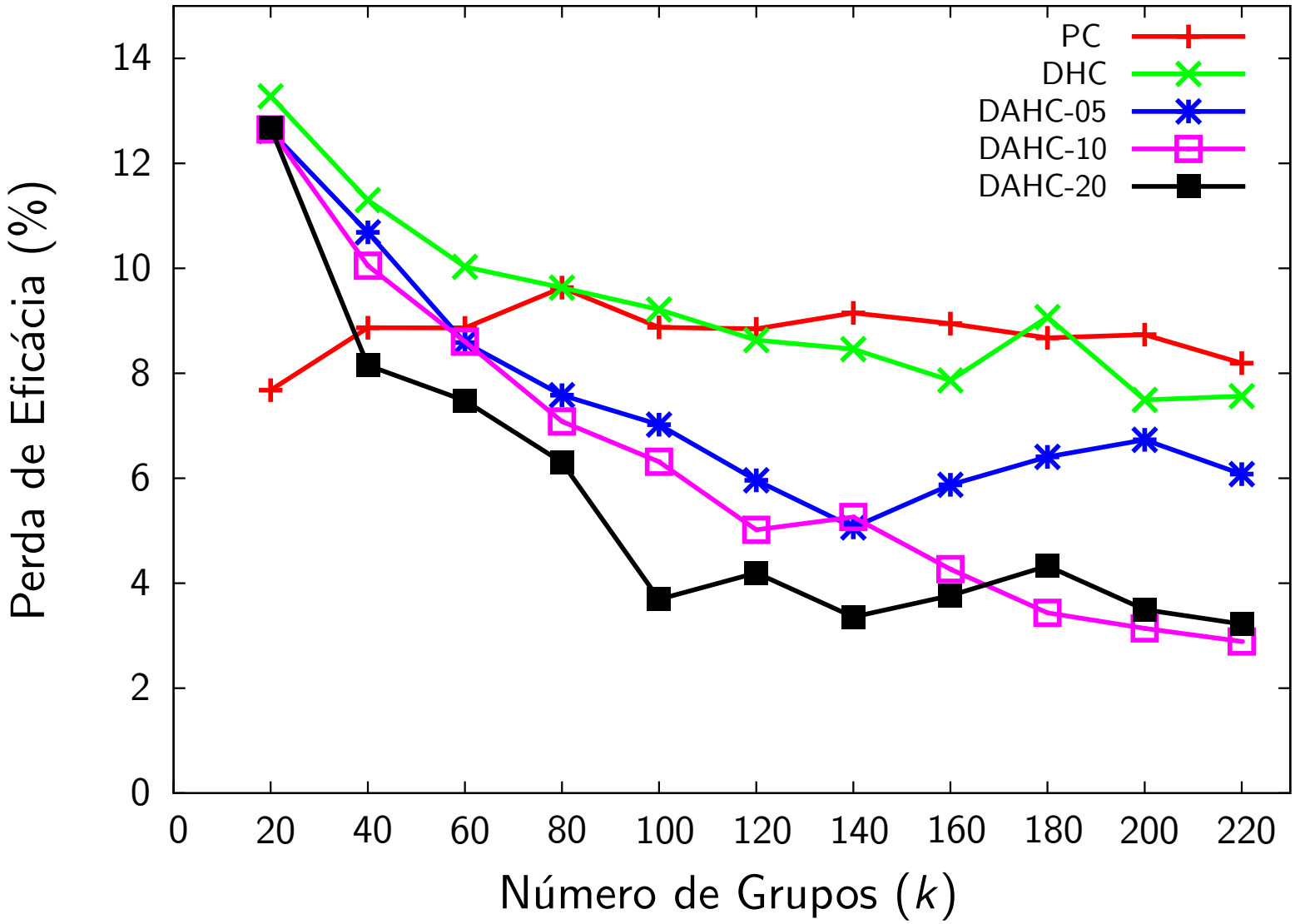

0 20 40 60 80 100 120 140 160 180 200 220
Número de Grupos (k)
Ganho de Desempenho (vezes)
80 75 70 65 60 55 50 45 40 35 30 25 20 15 10 5 0
PC
DHC
DAHC-05
DAHC-10
DAHC-20
0 20 40 60 80 100 120 140 160 180 200 220
Número de Grupos (k)
Ganho de Desempenho (vezes)

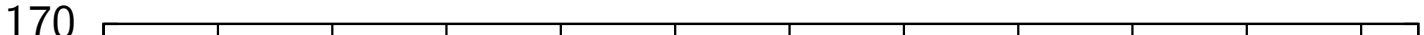

Perda de Eficácia (%)
Número de Grupos (k)
PC
DHC
DAHC-05
DAHC-10
DAHC-20
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
0
20
40
60
80
100
120
140
160
180
200
220


0
20
40
60
80
100
120
140
160
180
200
220
Número de Grupos (k)
Ganho de Desempenho (vezes)
75
70
65
60
55
50
45
40
35
30
25
20
15
10
5
0
PC
DHC
DAHC-05
DAHC-10
DAHC-20
0
20
40
60
80
100
120
140
160
180
200
220
Número de Grupos (k)

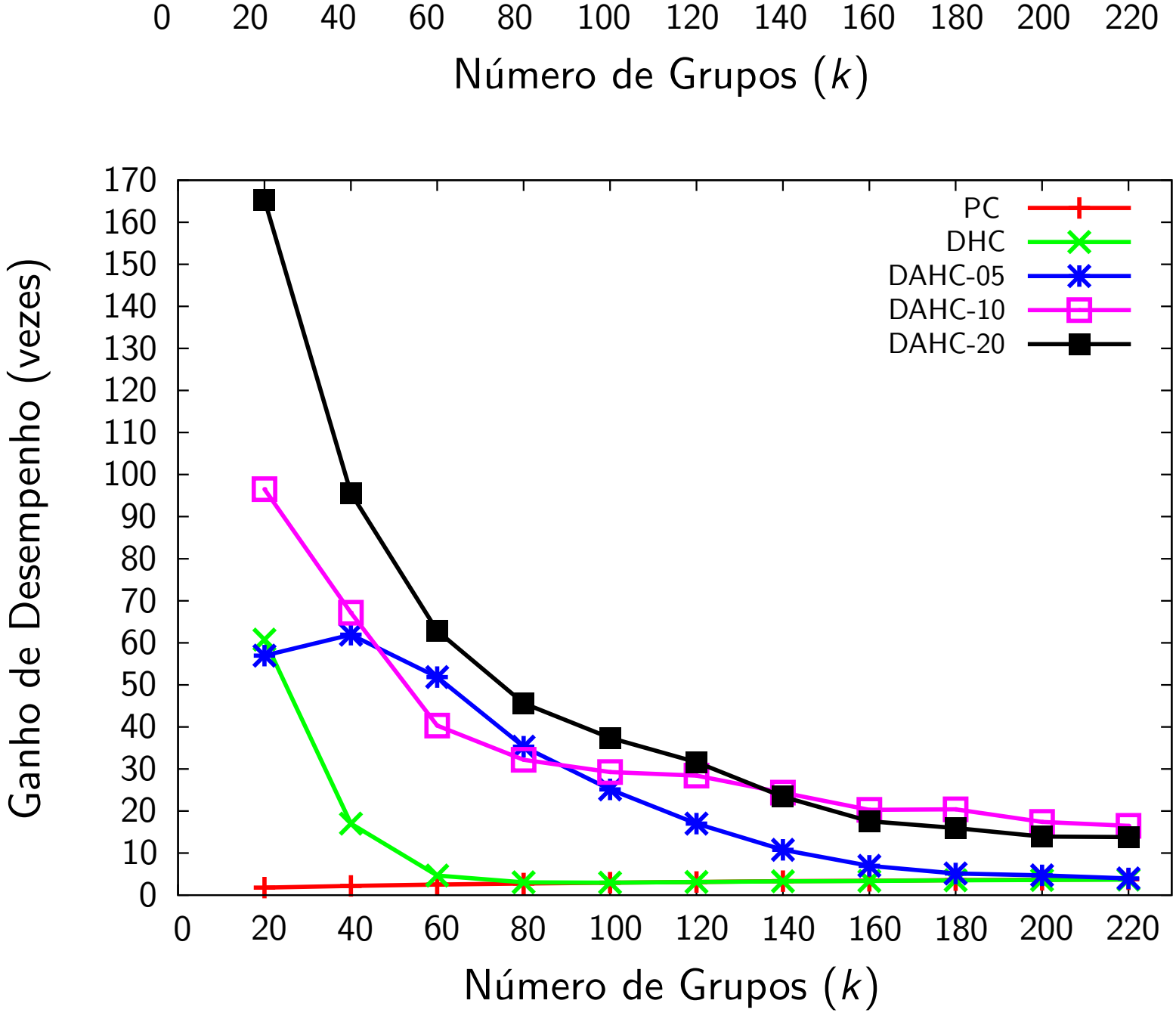

PC
DHC
DAHC-05
DAHC-10
DAHC-20
Perda de Eficácia (%)
Número de Grupos (k)
0
2
4
6
8
10
12
14
16
18
20
22
24
0
20
40
60
80
100
120
140
160
180
200
220

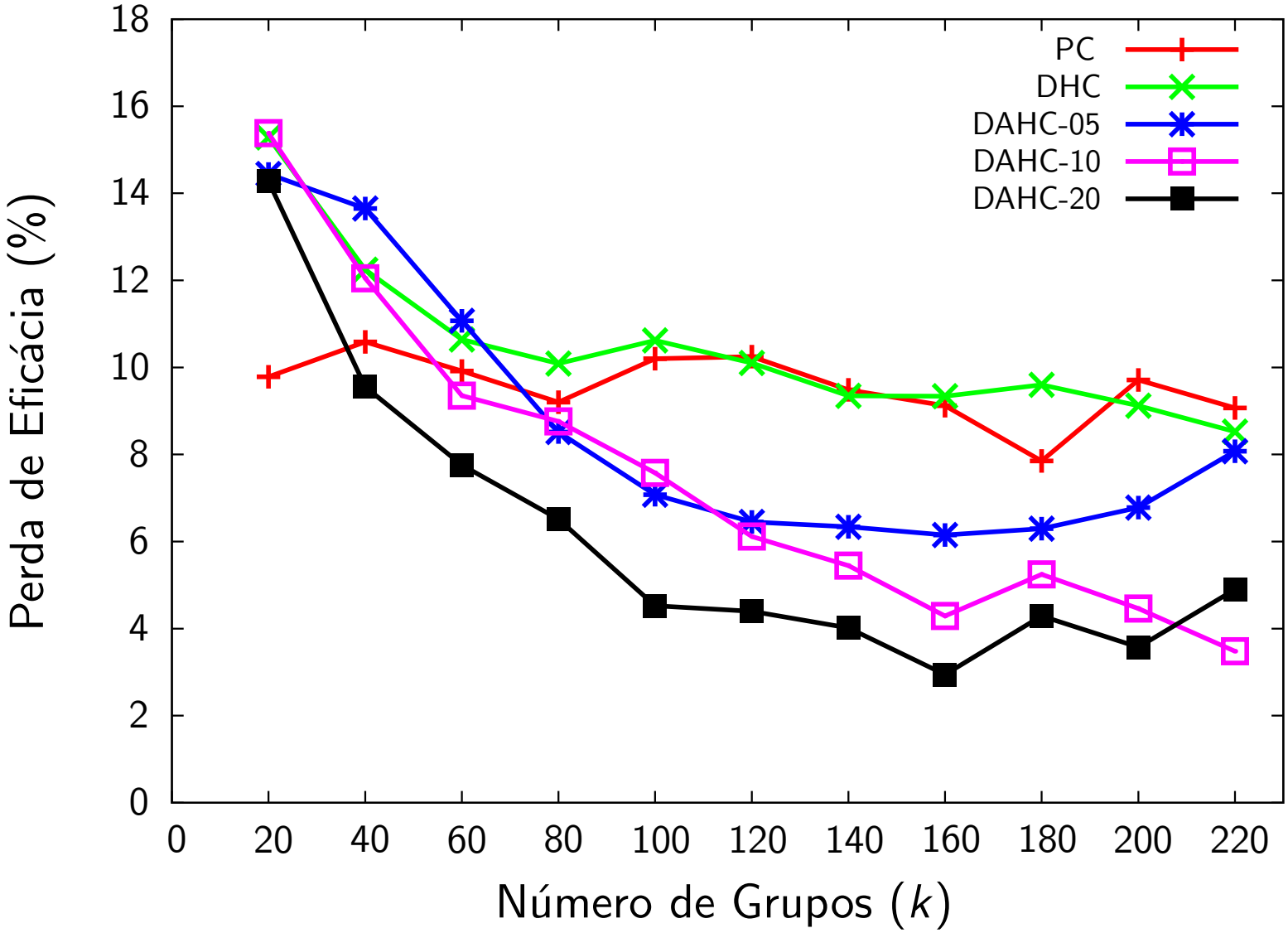

0
20
40
60
80
100
120
140
160
180
200
220
Número de Grupos (k)
Ganho de Desempenho (vezes)
75
70
65
60
55
50
45
40
35
30
25
20
15
10
5
0
PC
DHC
DAHC-05
DAHC-10
DAHC-20
0
20
40
60
80
100
120
140
160
180
200
220
Número de Grupos (k)

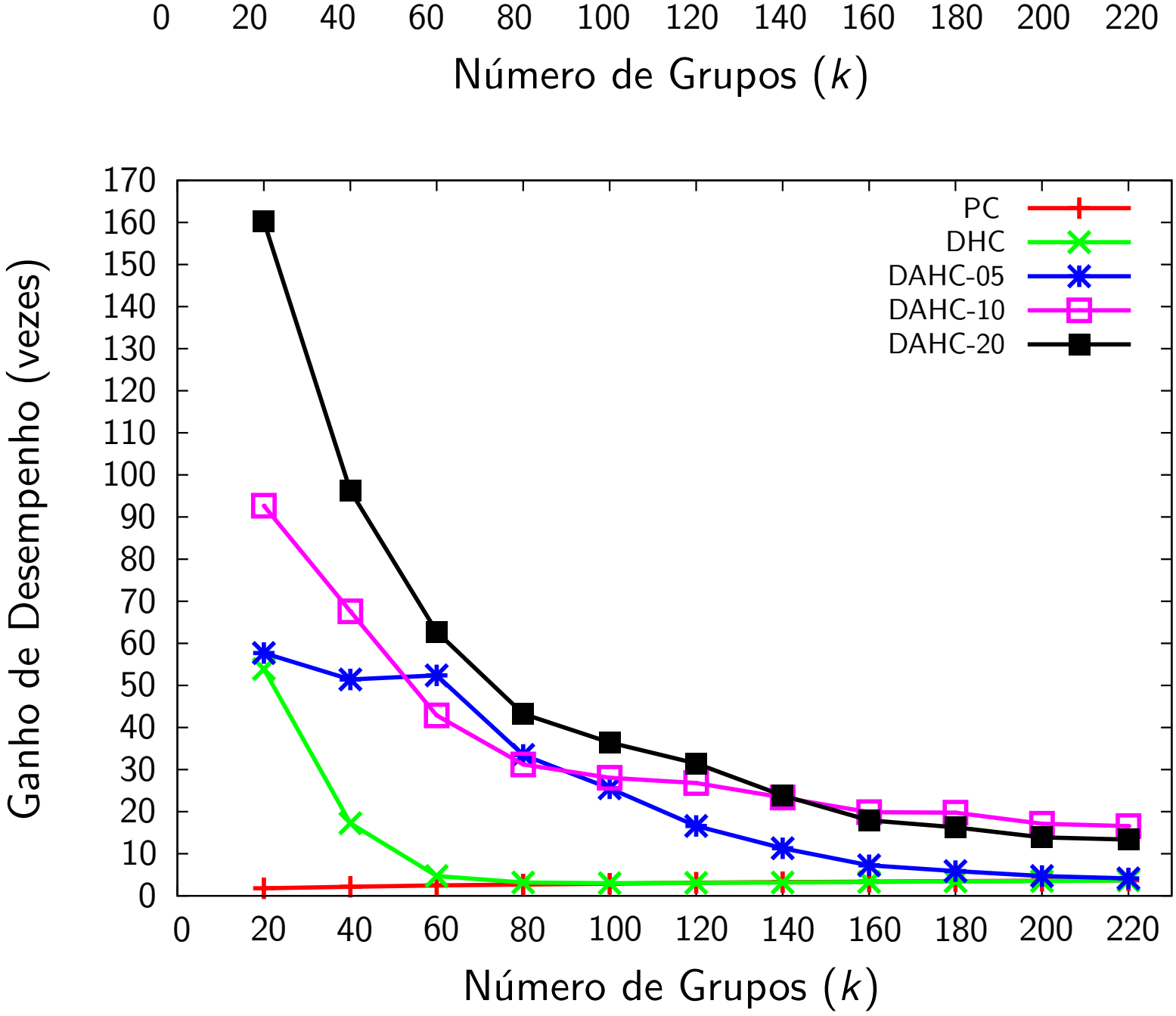

PC
DHC
DAHC-05
DAHC-10
DAHC-20
Perda de Eficácia (%)
Número de Grupos (k)
0
2
4
6
8
10
12
14
0
20
40
60
80
100
120
140
160
180
200
220

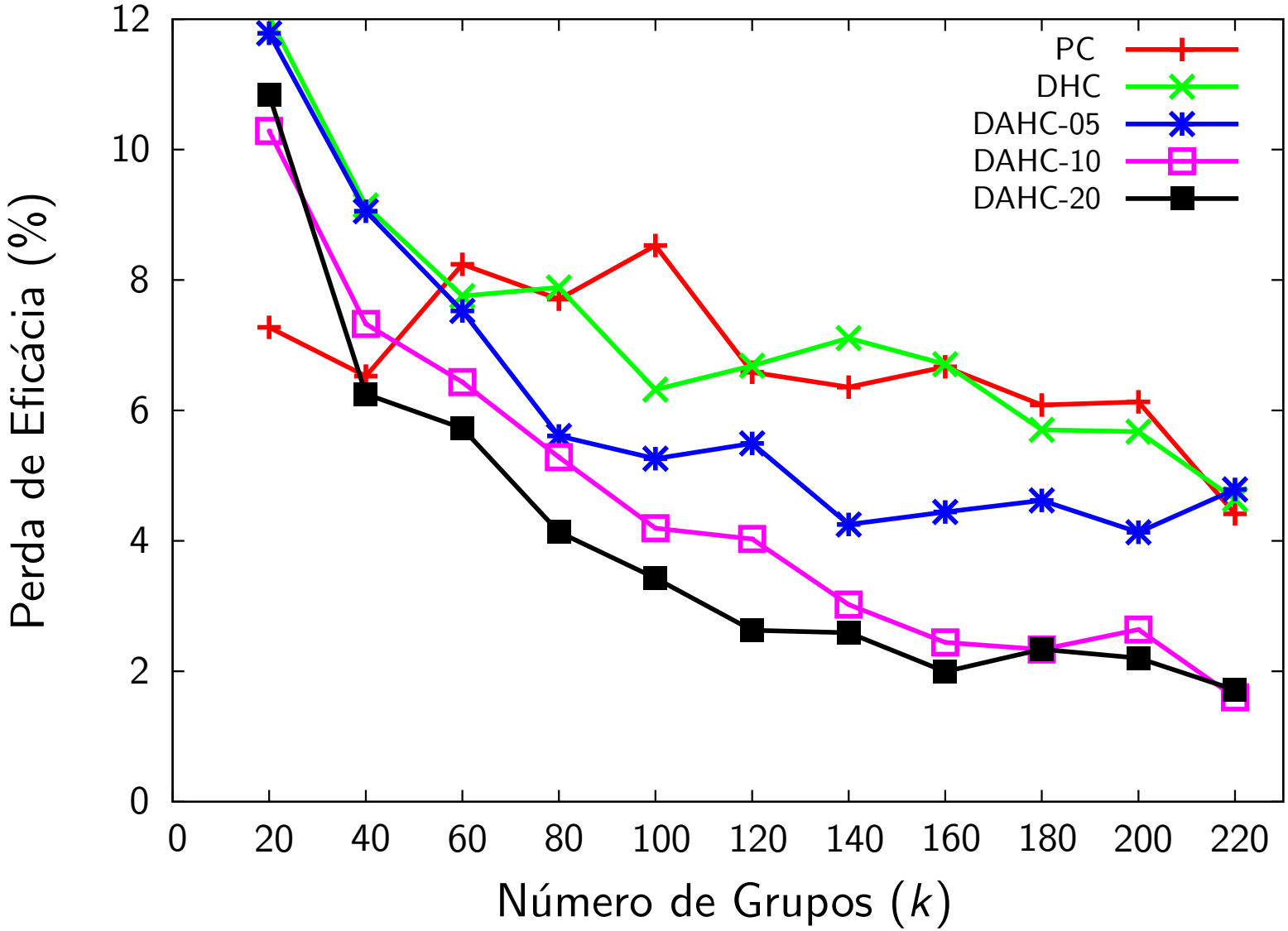

170

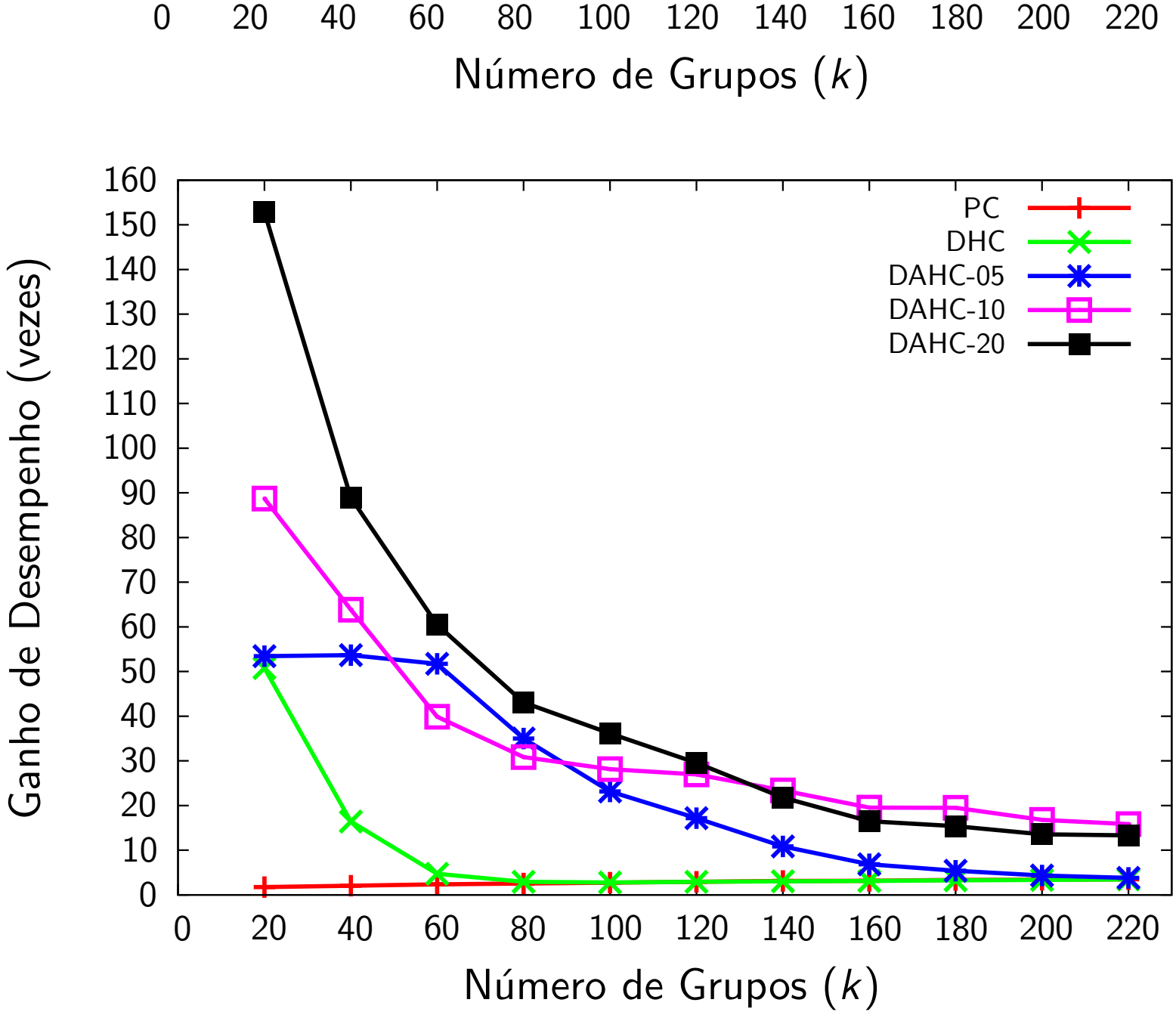

Perda de Eficácia (%)


0 20 40 60 80 100 120 140 160 180 200 220
Número de Grupos (k)
Ganho de Desempenho (vezes)
170 160 150 140 130 120 110 100 90 80 70 60 50 40 30 20 10 0
PC
DHC
DAHC-05
DAHC-10
DAHC-20
0 20 40 60 80 100 120 140 160 180 200 220
Número de Grupos (k)


Perda de Eficácia (%)
Número de Grupos (k)
PC
DHC
DAHC-05
DAHC-10
DAHC-20
0
2
4
6
8
10
12
14
16
18
0
20
40
60
80
100
120
140
160
180
200
220


0 20 40 60 80 100 120 140 160 180 200 220
Número de Grupos (k)
Ganho de Desempenho (vezes)
170
160
150
140
130
120
110
100
90
80
70
60
50
40
30
20
10
0
PC
DHC
DAHC-05
DAHC-10
DAHC-20
0 20 40 60 80 100 120 140 160 180 200 220
Número de Grupos (k)


Perda de Eficácia (%)
Número de Grupos (k)
PC
DHC
DAHC-05
DAHC-10
DAHC-20
0
2
4
6
8
10
12
0
20
40
60
80
100
120
140
160
180
200
220


0 20 40 60 80 100 120 140 160 180 200 220
Número de Grupos (k)
Ganho de Desempenho (vezes)
160
150
140
130
120
110
100
90
80
70
60
50
40
30
20
10
0
PC
DHC
DAHC-05
DAHC-10
DAHC-20
0 20 40 60 80 100 120 140 160 180 200 220
Número de Grupos (k)