4.5.3 Conjunto de Imagens de Animais

Foram coletadas um total de 1232 imagens de animais através de provedores de busca na internet como *google search, pixabay, flickr e shutterstock*. As imagens estão dividas em 11 classes.

Figura 4.8: *Exemplos de imagens coletadas na categoria de animais*

Tabela 4.11: *Conjunto de dados 3 - Animais*

CLASSE	OBJETOS
HOMEM	110
MACACO	116
SERPENTE	110
TUBARÃO	111
CROCODILO	110
CAVALO	115
ELEFANTE	112
BEIJA-FLOR	115
GATO	113
GIRAFA	110
AVESTRUZ	111
TOTAL	1232

As imagens estão em diversas resoluções e com diferentes perspectivas de captura, algumas possuem partes do animal, outras o animal inteiro. A figura 4.8 mostra exemplos de imagens coletadas nessa categoria.

4.5.4 Conjunto de Imagens de CAPTCHA

Foram coletadas imagens através de sites que utilizam o **CAPTCHA** *(Completely Automated Public Turing test to tell Computers and Humans Apart)* como medida de segurança contra ataques automatizados por máquinas.

Para essa categoria foram capturadas imagens de testes CAPTCHA através da captura de tela do computador, e então as imagens de cada um dos testes capturados foram separadas como novos arquivos, totalizando 124 imagens diferentes.

As imagens foram separadas em apenas 2 classes para simulação do teste de "Selecione Um Veículo", cujo objetivo é apenas classificar as imagens como contendo ou não veículos, as mesmas imagens poderiam ser organizadas em outros conjuntos binários (2 classes) para simular outros tipos de situações que ocorrem nesse teste.

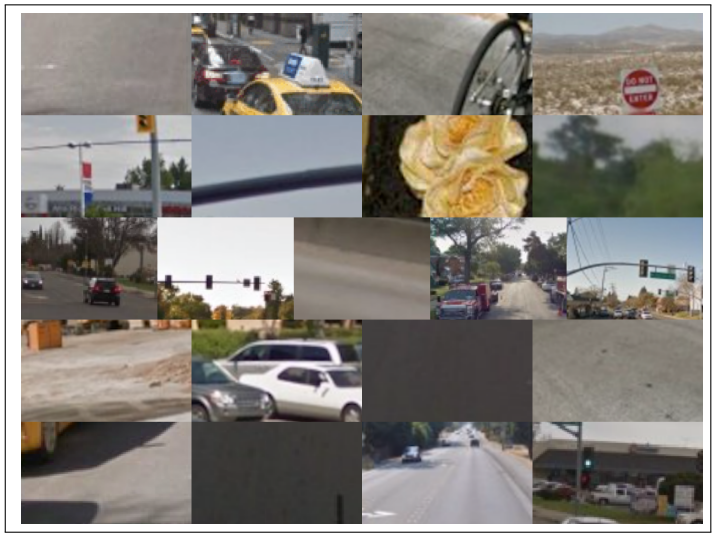


Figura 4.9: *Exemplos de imagens coletadas na categoria de testes CAPTCHA*

Tabela 4.12: *Conjunto de dados 4 - CAPTCHA*

CLASSE	OBJETOS
VÉICULOS	54
NÃO_VEÍCULOS	70
TOTAL	124

As imagens possuem resoluções similares com várias perspectivas de capturas e podem possuir apenas uma cor sólida, apenas um objeto ou uma grande variedade de objetos misturados na mesma imagem. A figura 4.9 mostra exemplos de imagens coletadas nessa categoria.

4.5.5 Conjunto de imagens de faces humanas

Este conjunto de imagens, disponibilizado pela **Universidade de Cambridge**, chamado de *'ORL Database of Faces'*[13], contém imagens da face de 40 indivíduos diferentes, 10 imagens por indivíduo, totalizando 400 imagens.

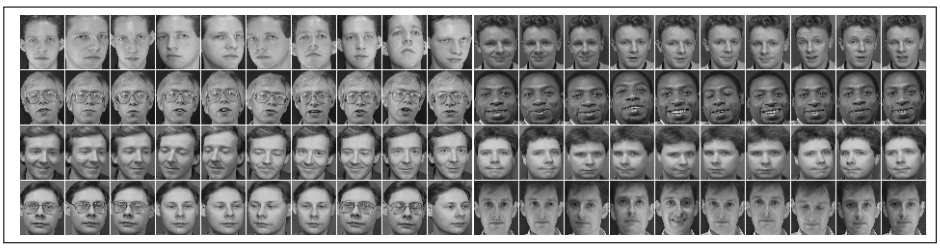


Figura 4.10: *Exemplos de imagens coletadas na categoria de faces humanas*

Tabela 4.13: *Conjunto de dados 5 - 'ORL Database Of Faces'*

CLASSE	OBJETOS
INDIVÍDUO0	10
INDIVÍDUO1	10
...	...
INDIVÍDUO39	10
TOTAL	400

As imagens foram capturadas entre 1992 e 1994, no Laboratório de Pesquisas Olivetti (ORL - Olivetti Research Laboratory) em Cambridge, para testes de reconhecimento facial, estão todas dispostas na mesma resolução com a mesma perspectiva de captura.

Tradução e adaptação de um trecho informativo disponível junto ao conjunto de dados:

"Para alguns dos indivíduos, as suas 10 imagens foram capturadas em momentos diferentes, variando levemente a luz, expressões faciais e detalhes faciais. Todas as imagens foram tiradas em um fundo escuro homogêneo e todos os indivíduos estão encarando a câmera com possivelmente uma pequena variação de movimentos e inclinações laterais da cabeça."

Dados e informações disponíveis em:

https://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html

4.5.6 Conjunto de Pinturas

Foram coletadas, da enciclopédia de artes visuais WikiArt[2], imagens de pinturas de 6 movimentos artísticos diferentes, cada classe (movimento artístico) possui no mínimo 31 amostras e totalizam 255 imagens.

Foram utilizadas pinturas de 6 movimentos de 3 eras diferentes:

- Medieval (Pinturas dos movimentos Bizantino e Gótico);
- Pós Renascença (Pinturas dos movimentos Barroco e Romantismo);
- Moderna (Pinturas dos movimentos Cubismo e Expressionismo).

Figura 4.11: *Exemplos de imagens coletadas na categoria de pinturas*

As imagens variam bastante em resolução, e cada uma contém apenas uma pintura pertencente predominantemente a algum movimento artístico específico, podendo também exibir alguma características de outros movimentos na mesma.

Tabela 4.14: *Conjunto de dados 7 - Movimentos Artísticos*

CLASSE	OBJETOS
BIZANTINO	37
GOTICO	31
BARROCO	49
ROMANTISMO	52
CUBISMO	43
EXPRESSIONISMO	43
TOTAL	255

[2]Endereço para acesso: https://www.wikiart.org/

4.5.7 Pré-processamento das imagens

Munido de uma fonte de imagens, a primeira tarefa do algoritmo para construir o modelo é aplicar um pré-processamento que consiste em uma técnica de compressão com perda para redução de informação e consequentemente maior generalização de valores de pixels. Isso é feito com base na visão humana, que faz algo semelhante ao descartar o excesso de informação (em especial as cores) visual das imagens capturadas pelo olho humano[90, 26]. O pré-processamento definido começa com um simples redimensionamento das imagens para o tamanho 256x256, garantindo que todas imagens contenham a mesma quantidade de pixels, e, posteriormente, uma quantização para 1(um) canal, que pode possuir 4, 8, 16, 32, 64, 128 ou 256.

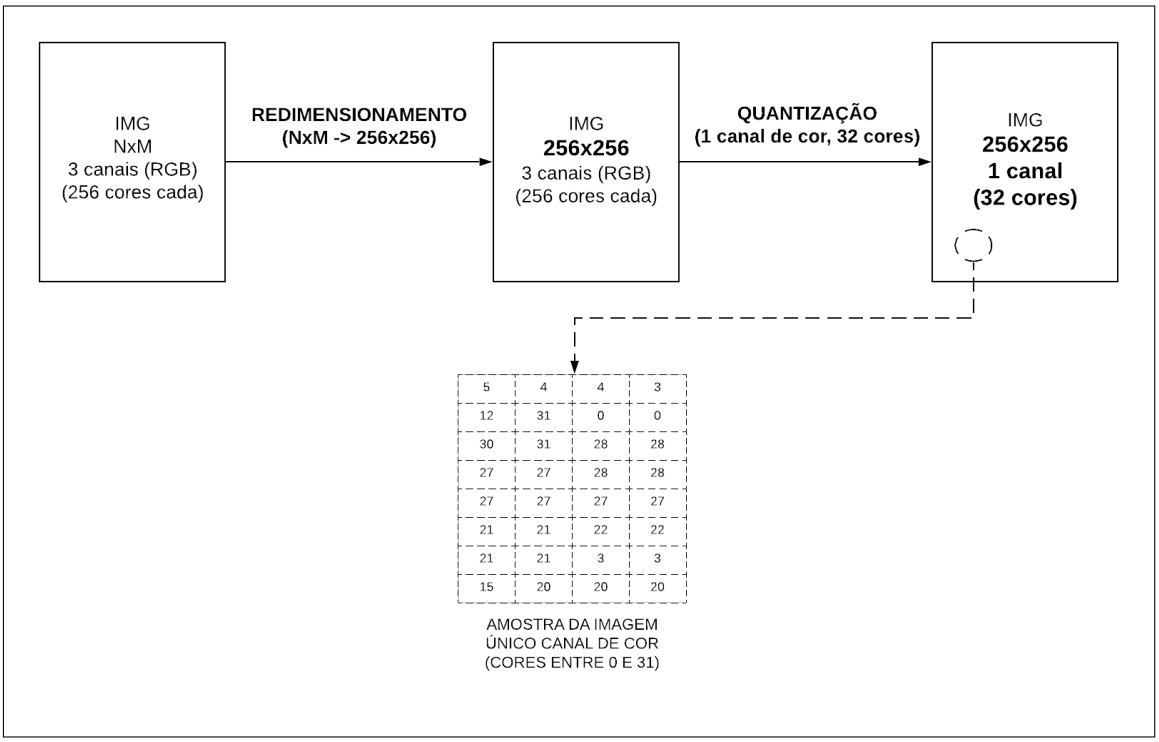


Figura 4.12: *Pré-processamento da imagem*

Para o exemplo com 32 cores, como pode ser visto na figura 4.12, após pré-processamento, tem-se uma única matriz de tamanho 256x256 que possui valores de pixels que variam de 0 até 31. A partir dessa matriz, são realizados processos de granulação para a contagem de ocorrências de cada grânulo. Forma-se, assim, um dicionário que contém a contagem de quantas vezes cada grânulo ocorreu, a partir daí são realizados cálculos probabilísticos para cada um desses grânulos.

Para avaliação da perda de informação ε em uma imagem (X), após o processamento, foi determinada uma expressão baseada na variação da informação original, comparada com a informação após uma reconstrução, depois do processamento realizado

(redimensionamento e quantização), esse cálculo é feito através da seguinte equação:

$$\varepsilon(X) = \frac{1}{n \times 3} \cdot \sum_{j=1}^{n} \left(\sum_{c=1}^{3} \left(\frac{\max(f_{jc}, g_{jc}) - \min(f_{jc}, g_{jc})}{\max(f_{jc}, g_{jc})} \right) \right), \tag{4-1}$$

em que:

- n é o total de número de pixels na imagem original X;
- f é uma matriz de n $\times$ 3 posições e que contém todos os valores de pixels da imagem original X em 3 canais de cores diferentes (RGB);
- g é uma matriz de n $\times$ 3 posições que contém todos os valores de pixels da imagem X em canais de cores diferentes (RGB) após o processamento (redimensionamento e quantização) e reconstrução para a quantidade de cores e resolução original.

CAPÍTULO 5

Experimentos e análise de desempenho

Este capítulo apresenta os resultados dos experimentos realizados nas fontes de dados (FD) descritas no capítulo anterior, além da avaliação de hipóteses H de agrupamento originadas destas fontes por meio do método *General Mining GM*. A determinação dos grânulos e configuração da dependência da informação entre os grânulos nos experimentos é realizada pela configuração dos parâmetros da máquina de Turing (MT) como mostram as tabelas 5.1 e 5.2. As MTs podem ser configuradas a partir de seu processo de compressão pela definição de seu grau de dependência da informação e seus analisadores léxicos e sintáticos. Para os experimentos com objetos de texto, os processos de compressão podem ser configurados conforme a tabela 5.1.

Analisador léxico	**Analisador sintático**	**Dependência da informação**
Grânulo elementar (caractere ASCII)	RLE	Ordem 1
Digrama de caracteres	RLE	Ordem 2
Palavras (LZW)	RLE	Ordem 1

Tabela 5.1: *Parâmetros de configuração de uma MT para objetos do tipo texto*

Em uma MT, o analisador léxico definido como "grânulo elementar"é o parâmetro fundamental de comparação com outras abordagens de compressão.

Grande parte das técnicas de compressão em textos utilizam uma estratégia de representação dos dados *sem perdas* de informação. Muitos experimentos com imagem utilizam uma estratégia de representação dos dados *com perdas* de informação como é o caso do compressor JPG. Para os experimentos com objetos de imagem, os processos de compressão podem ser configurados conforme a tabela 5.2.

Analisador léxico	Analisador sintático	Dependência da informação
Grânulo elementar (Pixel)	RLE	Ordem 1
Grânulos da quantização de 5 bits em imagem com resolução: 256 x256 pixels)	RLE	Ordem 1
Conjunto de Pixels (LZW)	(Sistema L: curva de Hilbert) + RLE	Ordem 1

Tabela 5.2: *Parâmetros de configuração de uma MT para objetos do tipo imagem*

Os experimentos intercalam espaços de agrupamento $E = (SI, R)$ e $ECL = (SD, R)$ utilizando como relação R a NCD. Posteriormente, também, são expostas avaliações a partir dos resultados apresentados.

A figura 5.1 demostra a estrutura do modelo a partir do paradigma dos 4 universos [30]. Cada camada, separada pelas linhas vermelhas da figura 5.1, expressa a relação do modelo proposto, nesta tese, com o paradigma dos 4 universos.

Os termos "acurácia"e "generalização"são utilizados, neste capítulo, como sinônimos de capacidade de generalização do modelo. Para aferição do modelo, avaliação da acurácia é utilizada a métrica G que denota a acurácia da árvore filogenética, derivada da execução do protótipo a partir uma fonte de dados.

Como mostrado no capítulo 4, a função do protótipo é demostrar o fluxo de execução dos experimentos envolvendo o método *GM*, a partir de um computador. Para cada fonte de dados descrita, os resultados são exibidos a partir da métrica G com base na configuração das máquinas de Turing. As próximas seções exibem os resultados para fontes de dados texto e imagem.

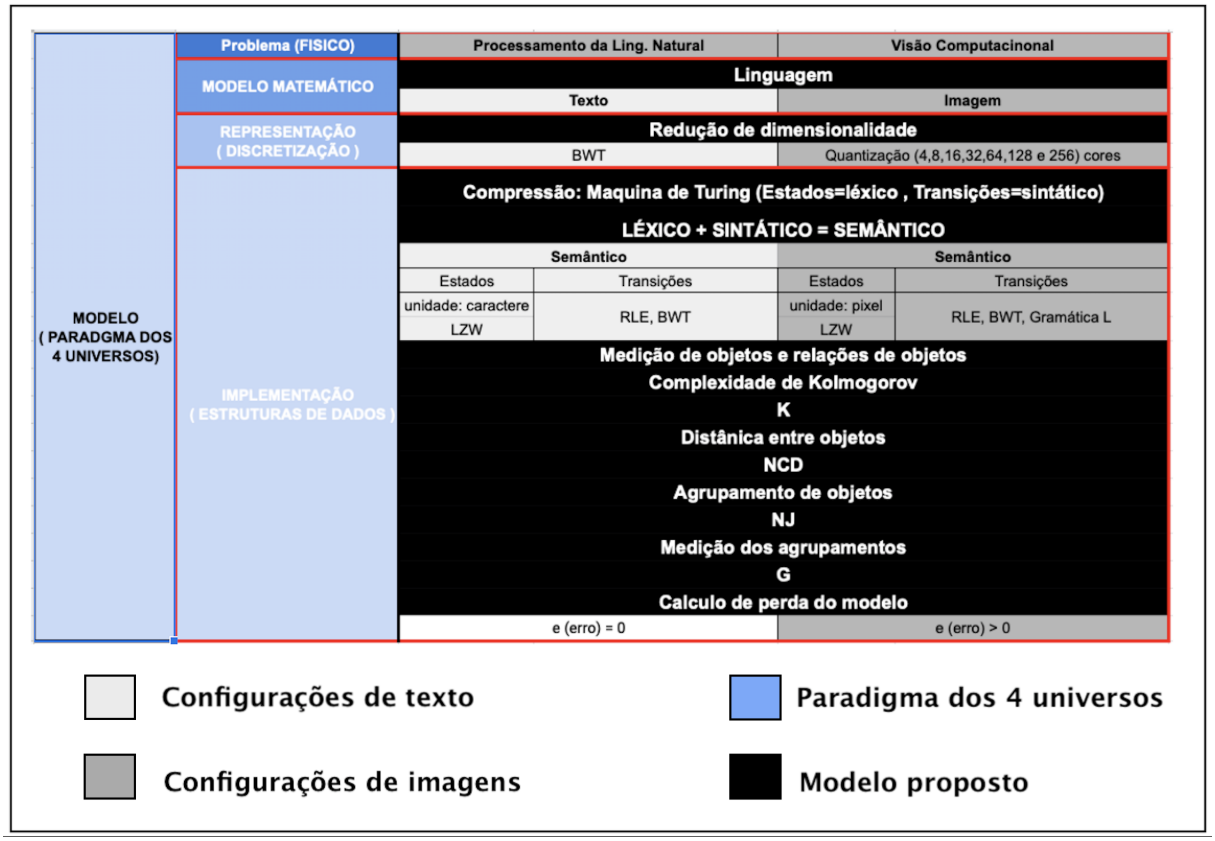


Figura 5.1: *Modelo GM X Paradigma dos 4 Universos*

5.1 Experimentos com objetos comprimidos do tipo texto

Nesta secção, são apresentados os resultados de dois tipos de experimentos texto:

- experimentos para fontes texto (idiomas) com número de classes variadas. Foram utilizadas configurações com 1, 2, 3, 4, 5, 6 e 7 classes, com diferentes quantidades de objetos por classe e diferentes configurações adaptativas de compressão. As estratégias adaptativas de compressão podem ser combinadas;
- experimentos para fontes de texto com semântica (Sinopses de filmes, Abstracts de artigos, Avaliação de Produtos, Texto com Sentimento, Twitter - notícias, Artigos noticiados e Idiomas). Esses experimentos utilizam o analisador léxico do dicionário de palavras do algoritmo adaptativo LZW e o analisador sintático RLE.

5.1.1 Algoritmos adaptativos envolvendo objetos comprimidos do tipo texto

Os resultados serão analisados na execução de 8 (oito) experimentos divididos em 2 (dois) tipos:

- B.1 - experimentos envolvendo objetos de dados não-estruturados no formato texto, comprimidos e mensurados com NCD em agrupamento hierárquico;
- B.2 - experimentos envolvendo objetos de dados não-estruturados no formato texto, comprimidos em agrupamento particionado.

A tabela 5.3 apresenta os resultados dos experimentos obtidos da FD1 utilizando NCD com textos dos idiomas francês e alemão comprimidos por uma MT e compondo um total de 55 objetos. Enquanto que para objetos com grânulos de informação configurados como símbolos, obteve-se 78% de acurácia para a MT Digrama e MT Palavra, as hipóteses H não apresentam nenhum erro na classificação, ou seja, $H = E_x$.

FD1 **Dist. NCD - Gráfico:** $B.1.1$			
	MT		
Informação: (Resolução/Semântica)	Símbolo	Digrama	Palavra
Classes completamente identificadas	1	2	2
Acurácia	78%	100%	100%

Tabela 5.3: *Resultados dos experimentos com a FD1 com NCD.*

Os resultados da Tabela 5.3 podem ser observados graficamente nos agrupamentos hierárquicos da figura 5.2. Observa-se que o resultado utilizando a MT Palavra separa de maneira mais assertiva as duas classes consideradas com todos os objetos agrupados de forma correta. Este resultado se mostrou presente em todos os experimentos envolvendo a MT Palavra realizados nas fontes de dados de número 1 a 8, sendo assim, essa configuração da MT será considerada como limite máximo ou ideal de comparação das hipóteses H.

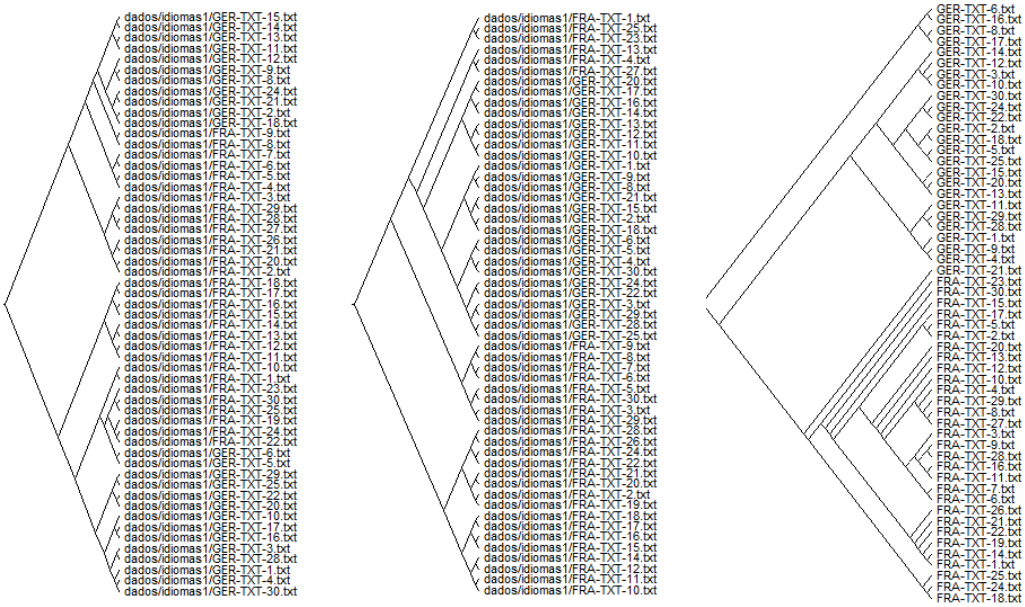


Figura 5.2: *Cladograma de 55 objetos comprimidos por MT Símbolo (Agrup. à esquerda), MT Digrama (Agrup. do meio) e MT Palavra (Agrup. à direita)*

Os resultados apresentados, em agrupamentos particionados, podem ser observados graficamente nas figuras 5.3 e 5.4. Os agrupamentos particionados da FD1 são concebidos com objetos comprimidos a partir de MTs configuradas como símbolos e palavras.

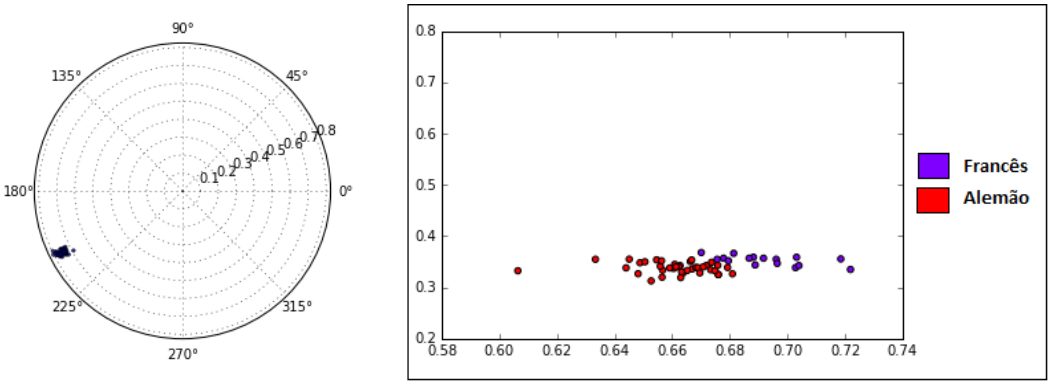


Figura 5.3: *Plano coord. polar de 55 objetos comprimidos por MT Símbolo (plano à esquerda) e plano coord. retangular (plano. à direita)*

Com relação a FD1, as hipóteses H de agrupamento são similares para a MT Símbolo e MT Palavra, porém seus agrupamentos particionados revelam boa capacidade de separação de classes para MT Palavra. Note, também, que para a MT Palavra a margem de separação é maior que na MT Símbolo. Mesmo havendo um objeto (francês) mais afastado, no caso de uma superfície linear de separação, este objeto estaria classificado

como pertencente a classe francês evitando erro de classificação e futuros problemas de generalização em classificadores que utilizarão esta hipótese H de agrupamento.

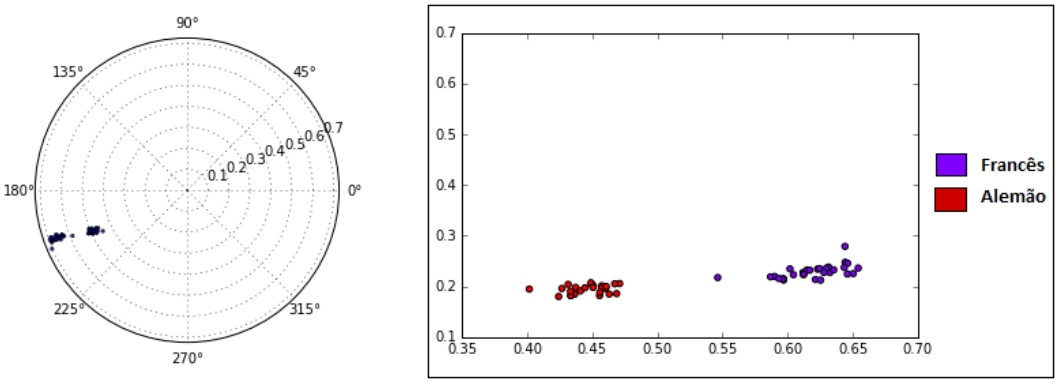


Figura 5.4: *Plano coord. polar de 55 objetos comprimidos por MT Palavra (plano à esquerda) e plano coord. retangular (plano. à direita)*

A tabela 5.4 apresenta os resultados dos experimentos obtidos da FD2 utilizando NCD com textos das línguas francês, alemão, inglês, português e italiano comprimidos por uma MT e compondo um total de 15 objetos. Enquanto que para objetos com grânulos de informação configurados para a MT Símbolo obteve-se 80% de acurácia, para a MT Digrama obteve-se 76%. Hipóteses H com MT Palavra seguem em acurácia de 100% em todos os experimentos como já mencionado.

FD2 **Dist. NCD - Gráfico:** $B.1.2$			
	MT		
Informação: (Resolução/Semântica)	Símbolo	Digrama	Palavra
Classes completamente identificadas	2	3	5
Acurácia	80%	76%	100%

Tabela 5.4: *Resultados dos experimentos com a FD2 com NCD.*

Os resultados da Tabela 5.4 podem ser observados graficamente na figura 5.5.

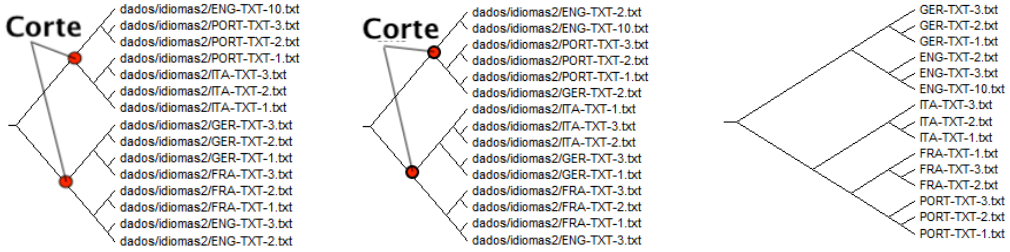


Figura 5.5: *Cladograma de 15 objetos comprimidos por MT Símbolo (Agrup. à esquerda), MT Digrama (Agrup. do meio) e MT Palavra (Agrup. à direita)*

Os agrupamentos particionados da FD2 são concebidos com objetos comprimidos a partir de MTs configuradas como símbolo e palavra. Os resultados apresentados, em agrupamentos particionados da FD2, podem ser observados graficamente nas figuras 5.6 e 5.7.

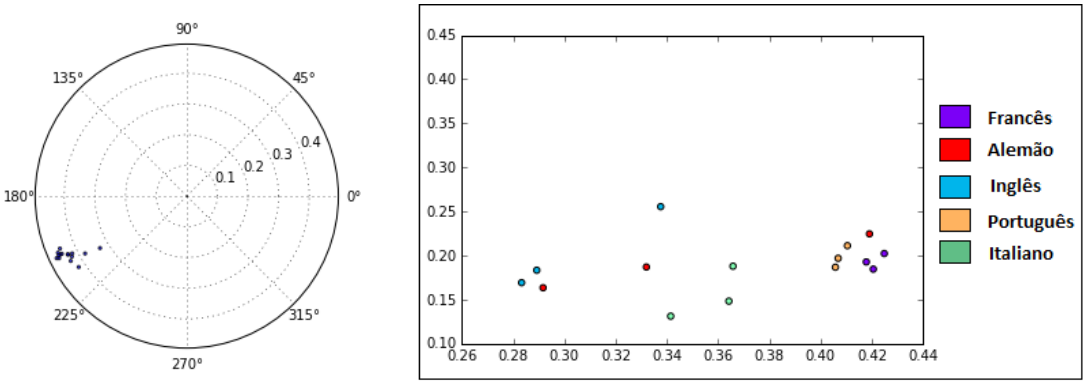


Figura 5.6: *Plano coord. polar de 15 objetos comprimidos por MT Símbolo (plano à esquerda) e plano coord. retangular (plano. à direita)*

Na FD2, a hipótese H de agrupamento para a MT Símbolo possui melhor porcentagem de acurácia que para a MT Palavra. Mesmo o agrupamento particionado da MT Palavra possuindo melhor definição dos grupos, sua representatividade em termos de números de objetos é baixa, em outras palavras, a MT Palavra possui melhor capacidade de separação de classes, mas a quantidade de objetos que representam cada classe é muito baixa.

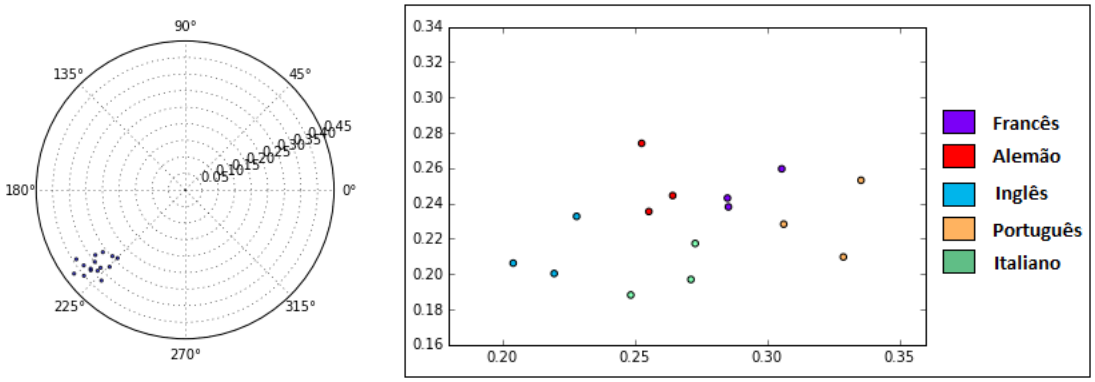


Figura 5.7: *Plano coord. polar de 15 objetos comprimidos por MT Palavra (plano à esquerda) e plano coord. retangular (plano. à direita)*

A tabela 5.5 apresenta os resultados dos experimentos obtidos da FD3 utilizando NCD com textos das línguas francês, alemão e português comprimidos por uma MT e compondo um total de 30 objetos. Enquanto que para objetos com grânulos de informação configurados para a MT Símbolo obteve-se 100% de acurácia, para a MT Digrama obteve-se 78%.

FD3 **Dist. NCD - Gráfico:** *B*.1.3			
	MT		
Informação: (Resolução/Semântica)	Símbolo	Digrama	Palavra
Classes completamente identificadas	3	1	3
Acurácia	100%	78%	100%

Tabela 5.5: *Resultados dos experimentos com a FD3 com NCD.*

Os resultados da Tabela 5.5 podem ser observados graficamente na figura 5.8.

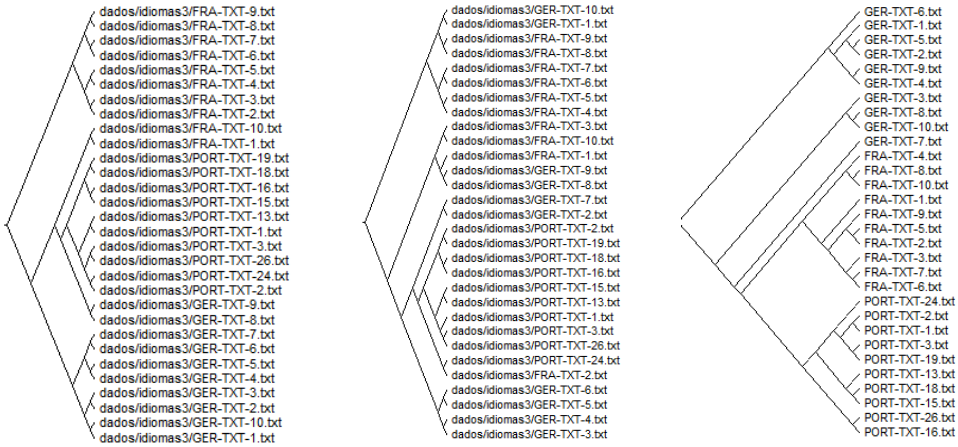


Figura 5.8: *Cladograma de 30 objetos comprimidos por MT Símbolo (Agrup. à esquerda), MT Digrama (Agrup. do meio) e MT Palavra (Agrup. à direita)*

Com base em agrupamentos particionados, os resultados apresentados em relação a FD3 podem ser observados graficamente nas figuras 5.9 e 5.10. Os agrupamentos particionados da FD3 são concebidos com objetos comprimidos a partir de MTs configuradas como símbolo e palavra.

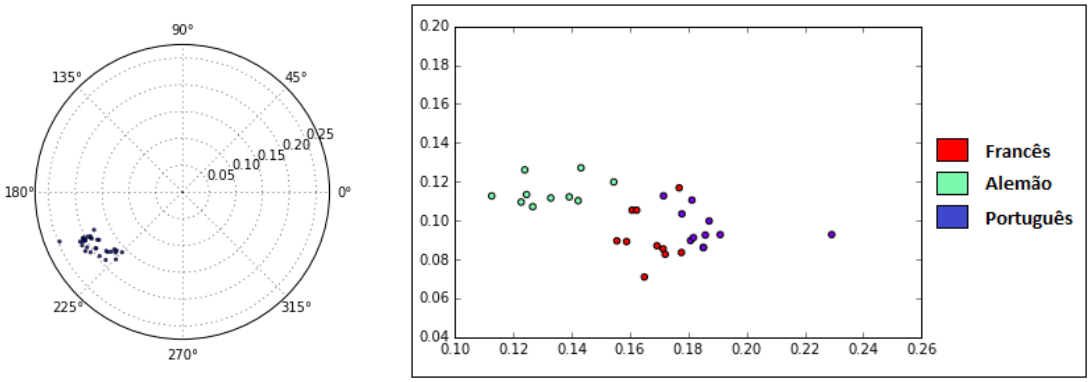


Figura 5.9: *Plano cood.polar de 30 objetos comprimidos por MT RLE (plano à esquerda) e plano coord. retangular (plano. à direita)*

Tanto as MTs configuradas como Símbolo e Palavra apresentaram boas hipóteses H de agrupamento. Os agrupamentos particionados gerados, a partir das duas hipóteses de agrupamento, são linearmente separáveis. Para a concepção de um classificador poderia ser utilizado a hipótese H para a MT Símbolo ou a hipótese H para a MT Palavra.

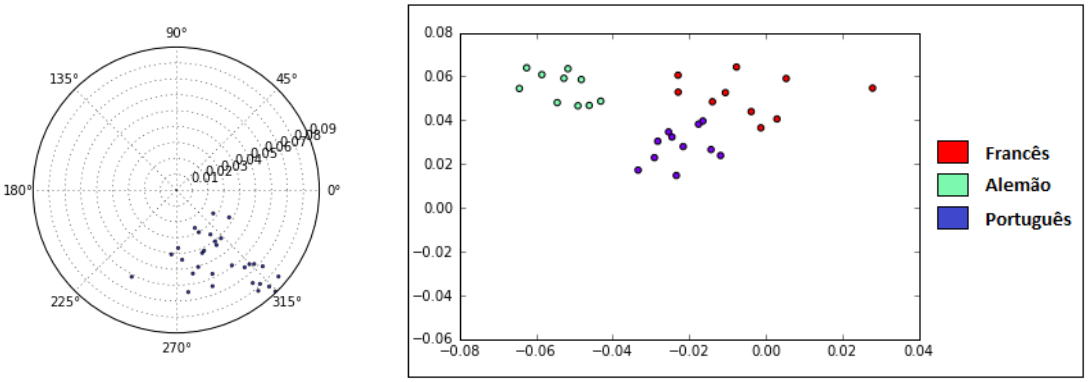


Figura 5.10: *Plano coord. polar de 30 objetos comprimidos por MT Palavra (plano à esquerda) e plano coord. retangular (plano. à direita)*

A tabela 5.6 apresenta os resultados dos experimentos obtidos da FD4 utilizando NCD com textos das línguas francês, alemão, português e italiano comprimidos por uma MT e compondo um total de 38 objetos. Enquanto que para objetos com grânulos de informação configurados para a MT Símbolo obteve-se 65% de acurácia, para a MT Digrama obteve-se 68%.

FD4 **Dist. NCD - Gráfico:** *B*.1.4			
	MT		
Informação: (Resolução/Semântica)	Símbolo	Digrama	Palavra
Classes completamente identificadas	1	0	4
Acurácia	65%	68%	100%

Tabela 5.6: *Resultados dos experimentos com a FD4 com NCD.*

Os resultados da Tabela 5.6 podem ser observados graficamente na figura 5.11.

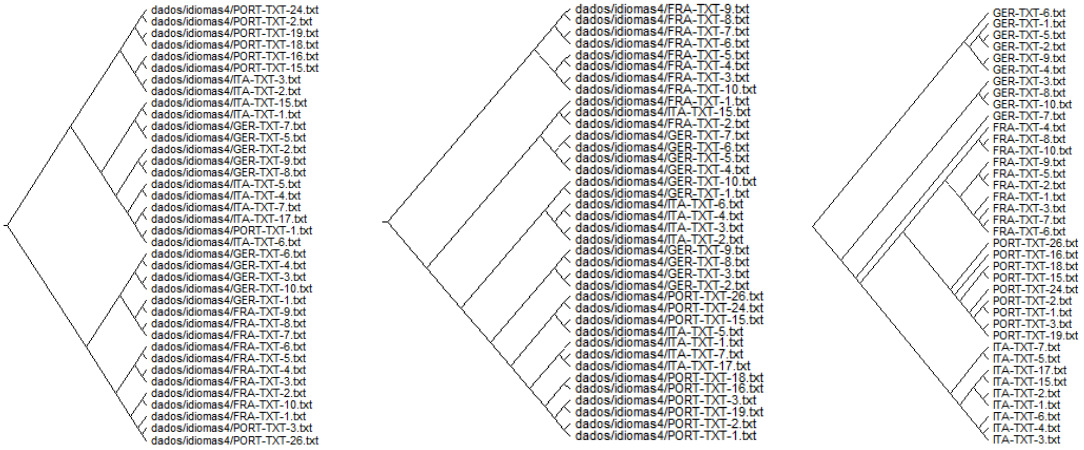


Figura 5.11: *Cladograma de 38 objetos comprimidos por MT Símbolo (Agrup. à esquerda), MT Digrama (Agrup. do meio) e MT Palavra (Agrup. à direita)*

Os agrupamentos particionados da FD4 são concebidos com objetos comprimidos a partir de MTs configuradas como símbolo e palavra. Os resultados apresentados, em agrupamentos particionados da FD4, podem ser observados graficamente nas figuras 5.12 e 5.13.

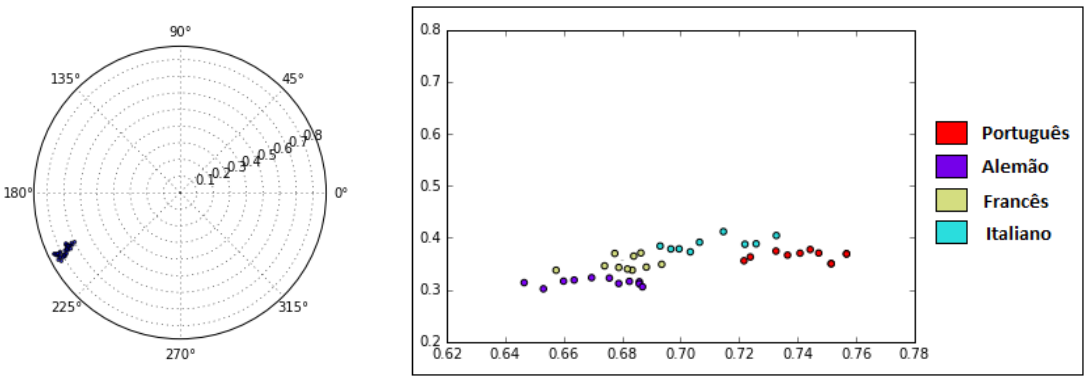


Figura 5.12: *Plano coord. polar de 38 objetos comprimidos por MT Símbolo (plano à esquerda) e plano coord. retangular (plano. à direita)*

Nos experimentos realizados, na FD4, acontece algo semelhante aos experimentos da FD1, porém com a diferença da FD4 possuir mais classes e, então, necessitar de mais fronteiras lineares de separação. As hipóteses H de agrupamento são similares para as MT configuradas com Símbolo e Digrama, porém seus agrupamentos particionados revelam boa capacidade de separação de classes para a MT Palavra.

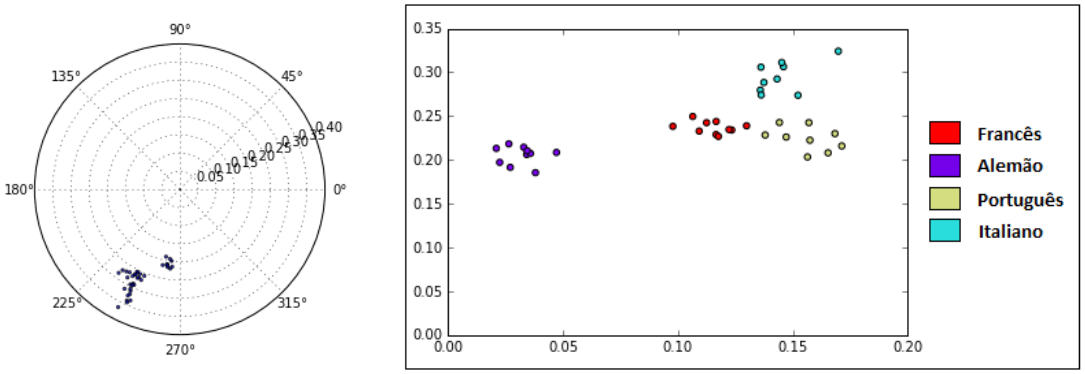


Figura 5.13: *Plano coord. polar de 38 objetos comprimidos por MT Palavra (plano à esquerda) e plano coord. retangular (plano. à direita)*

A tabela 5.7 apresenta os resultados dos experimentos obtidos da FD5 utilizando NCD com textos em português de política brasileira e política internacional, comprimidos por uma MT e compondo um total de 8 objetos. Para objetos com grânulos de informação configurados para a MT com Símbolos e Digrama, obteve-se 87%.

FD5 **Dist. NCD - Gráfico:** *B*.1.5			
	MT		
Informação: (Resolução/Semântica)	Símbolo	Digrama	Palavra
Classes completamente identificadas	1	1	2
Acurácia	87%	87%	100%

Tabela 5.7: *Resultados dos experimentos com a FD5 com NCD.*

Os resultados da Tabela 5.7 podem ser observados graficamente na figura 5.14.

Figura 5.14: *Cladograma de 8 objetos comprimidos por MT Símbolo (Agrup. à esquerda), MT Digrama (Agrup. do meio) e MT Palavra (Agrup. à direita)*

Os resultados apresentados, em agrupamentos particionados da FD5, podem ser observados graficamente nas figuras 5.15 e 5.16. Os agrupamentos particionados da fonte

de dados 5 são concebidos com objetos comprimidos a partir de MTs configuradas como símbolo e palavra.

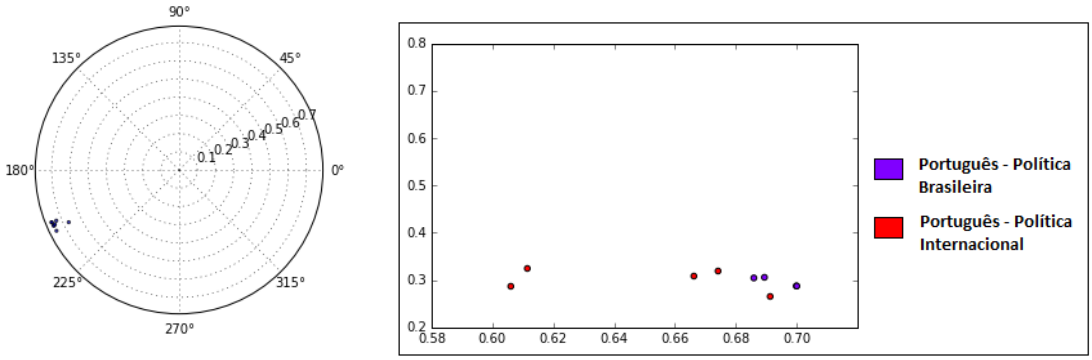


Figura 5.15: *Plano coord. polar de 8 objetos comprimidos por MT Símbolo (plano à esquerda) e plano coord. retangular (plano. à direita)*

Para a FD5, tanto para as MTs com compressão Símbolo como Palavra podem ser utilizadas por classificadores, porém os agrupamentos formados com a hipótese de MT Palavra formam agrupamentos linearmente separáveis, mesmo com poucos objetos que representam cada classe.

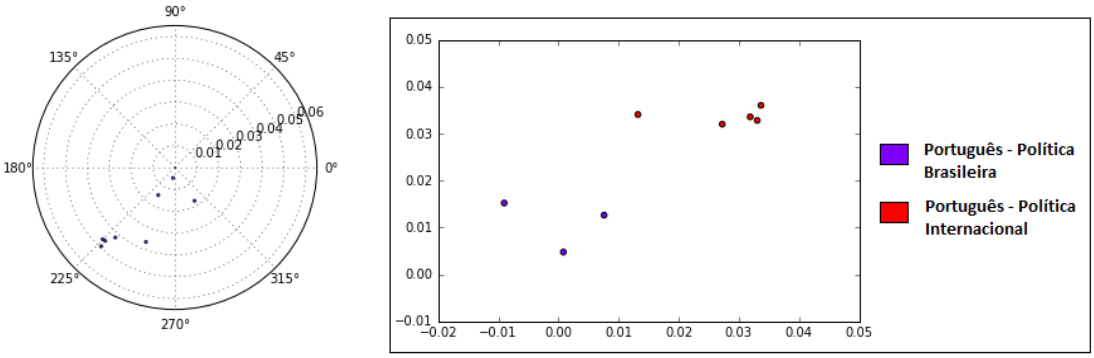


Figura 5.16: *Plano coord. polar de 8 objetos comprimidos por MT. Palavra (plano à esquerda) e plano coord. retangular (plano. à direita)*

A tabela 5.8 apresenta os resultados dos experimentos obtidos da FD6 utilizando NCD com textos em francês, alemão, italiano, e português (subdividido nas classes Biologia, Política Brasileira e Política Internacional) comprimidos por uma MT e compondo um total de 38 objetos. Para objetos com grânulos de informação configurados para a MT Símbolo, obteve-se 59%, enquanto que para a MT Digrama, obteve-se 68%.

FD6 **Dist. NCD - Gráfico:** *B*.1.6			
	MT		
Informação: (Resolução/Semântica)	Símbolo	Digrama	Palavra
Classes completamente identificadas	0	1	4
Subclasses completamente identificadas	0	1	3
Acurácia no agrupamento das classes	59%	68%	100%
Acurácia no agrupamento das subclasses	20%	33%	100%

Tabela 5.8: *Resultados dos experimentos com a FD6 com NCD.*

Os resultados da Tabela 5.8 podem ser observados graficamente na figura 5.17.

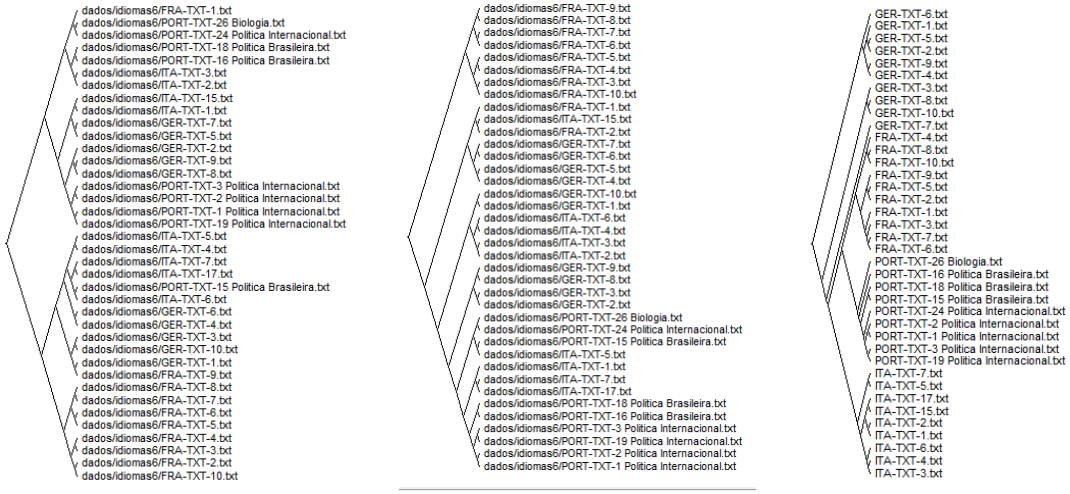


Figura 5.17: *Cladograma de 38 objetos comprimidos por MT Símbolo (Agrup. à esquerda), Digrama (Agrup. do meio) e Palavra (Agrup. à direita)*

Os agrupamentos particionados da FD6 são concebidos com objetos comprimidos a partir de MTs configuradas como símbolo e palavra.Os resultados apresentados em agrupamentos particionados da FD6 podem ser observados graficamente nas figuras 5.18 e 5.19.

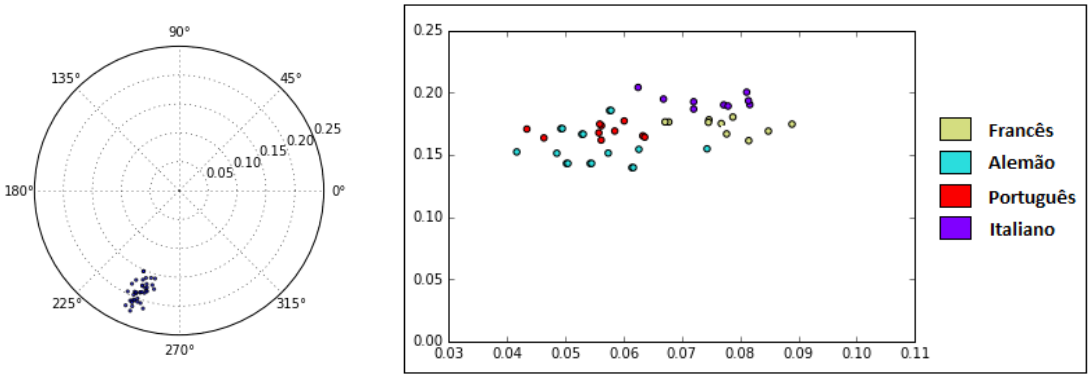


Figura 5.18: *Plano coord. polar de 38 objetos comprimidos por MT Símbolo (plano à esquerda) e plano coord. retangular (plano. à direita)*

Nos experimentos realizados na FD6, as hipóteses que utilizaram as MTs Símbolo e Palavra, apresentaram o fenômeno de pico [87] onde os objetos de todas as classes se misturam num espaço n-dimensional não possibilitando a concepção de fronteiras de decisão tão essenciais para classificadores.

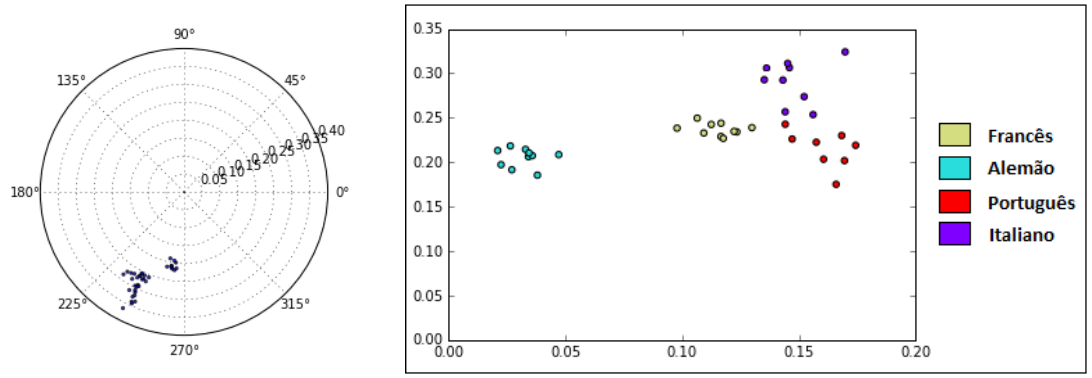


Figura 5.19: *Plano coord. polar de 38 objetos comprimidos por MT Palavra (plano à esquerda) e plano coord. retangular (plano. à direita)*

A tabela 5.9 apresenta os resultados dos experimentos obtidos da FD7 utilizando NCD com textos em francês, alemão, italiano, e português (subdividido nas classes Política Brasileira e Política Internacional) comprimidos por uma MT e compondo um total de 37 objetos. Para objetos com grânulos de informação configurados para a MT Símbolo obteve-se 68% enquanto que para a MT Digrama, obteve-se 82%.

FD7 **Dist. NCD - Gráfico:** *B*.1.7			
	MT		
Informação: (Resolução/Semântica)	Símbolo	Digrama	Palavra
Classes completamente identificadas	1	1	4
Subclasses completamente identificadas	2	0	2
Acurácia no agrupamento das classes	68%	82%	100%
Acurácia no agrupamento das subclasses	100%	63%	100%

Tabela 5.9: *Resultados dos experimentos com a FD7 com NCD.*

Os resultados da tabela 5.9 podem ser observados graficamente na figura 5.20.

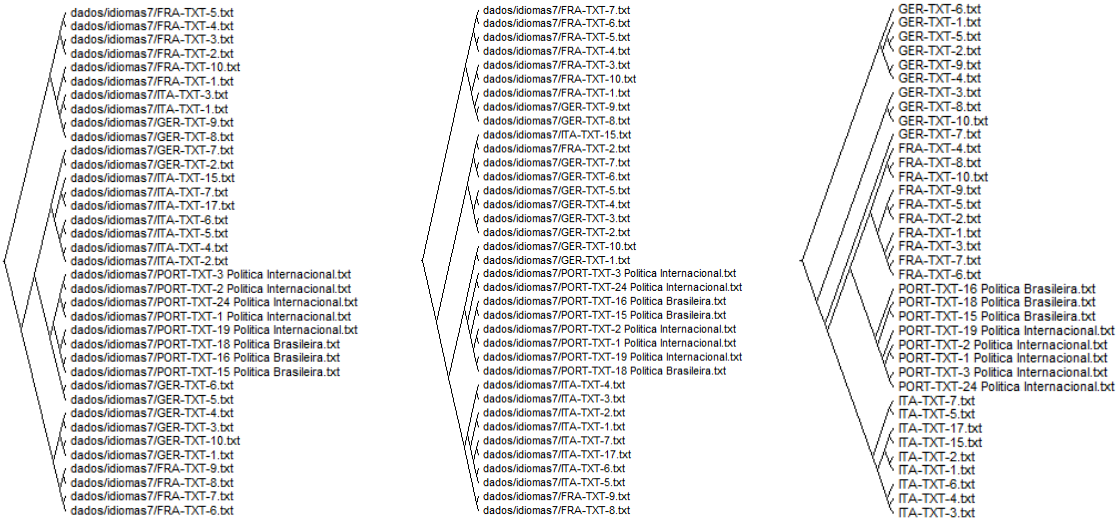


Figura 5.20: *Cladograma de 37 objetos comprimidos por MT Símbolo (Agrup. à esquerda), MT Digrama (Agrup. do meio) e MT Palavra (Agrup. à direita)*

Com base em agrupamentos particionados, os resultados apresentados em relação a FD7 podem ser observados graficamente nas figuras 5.21 e 5.22. Os agrupamentos particionados da FD7 são concebidos com objetos comprimidos a partir de MTs configuradas como símbolo e palavra.

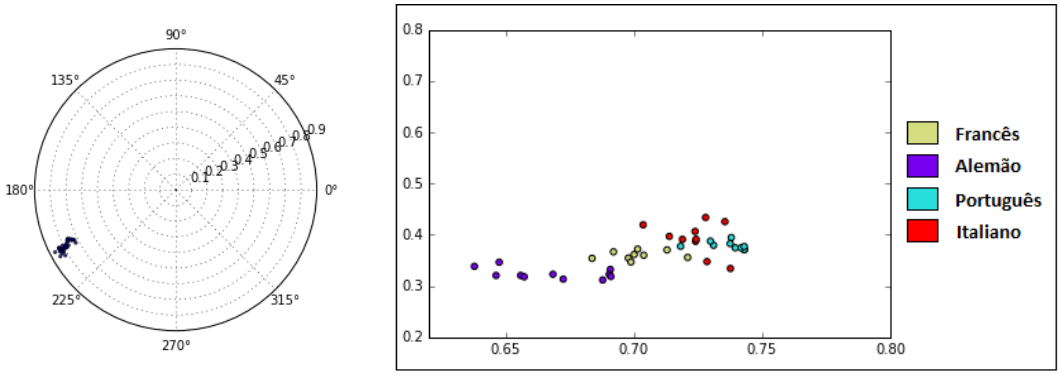


Figura 5.21: *Plano coord. polar de 37 objetos comprimidos por MT Símbolo (plano à esquerda) e plano coord. retangular (plano. à direita)*

Da mesma forma que nos experimentos realizados na FD6, na MT Símbolo, detecta-se o fenômeno de pico, onde os objetos de todas as classes se misturam num espaço n-dimensional não possibilitando a concepção de fronteiras de decisão tão essenciais para classificadores.

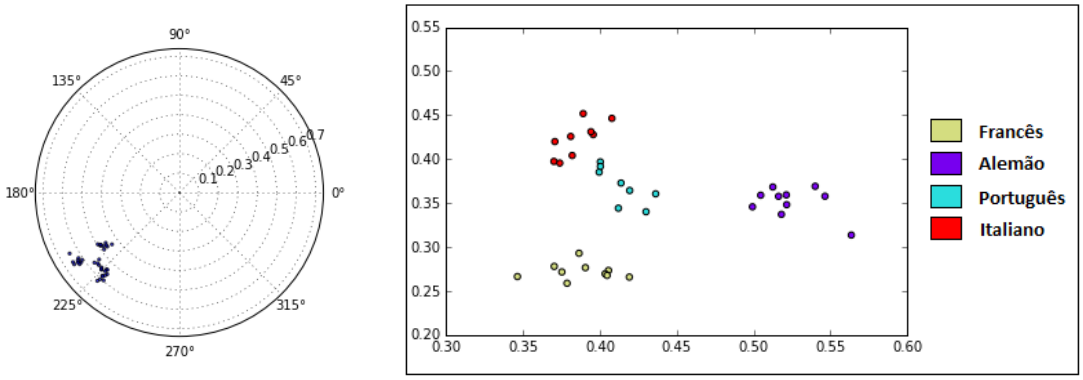


Figura 5.22: *Plano coord. polar de 37 objetos comprimidos por MT Palavra (plano à esquerda) e plano coord. retangular (plano. à direita)*

A tabela 5.10 apresenta os resultados dos experimentos obtidos da FD8 utilizando NCD com textos em francês, alemão, italiano, e português (subdividido nas classes Saúde, Política Brasileira e Política Internacional) comprimidos por uma MT e compondo um total de 39 objetos. Para objetos com grânulos de informação configurados para a MT Símbolo, obteve-se 60% enquanto que para a MT Digrama, obteve-se 100%.

FD8 **Dist. NCD - Gráfico:** *B*.1.8			
	MT		
Informação: (Resolução/Semântica)	Símbolo	Digrama	Palavra
Classes completamente identificadas	1	4	4
Subclasses completamente identificadas	1	1	3
Acurácia no agrupamento das classes	60%	100%	100%
Acurácia no agrupamento das subclasses	26%	75%	100%

Tabela 5.10: *Resultados dos experimentos com a FD8 com NCD.*

Os resultados da tabela 5.10 podem ser observados graficamente na figura 5.23.

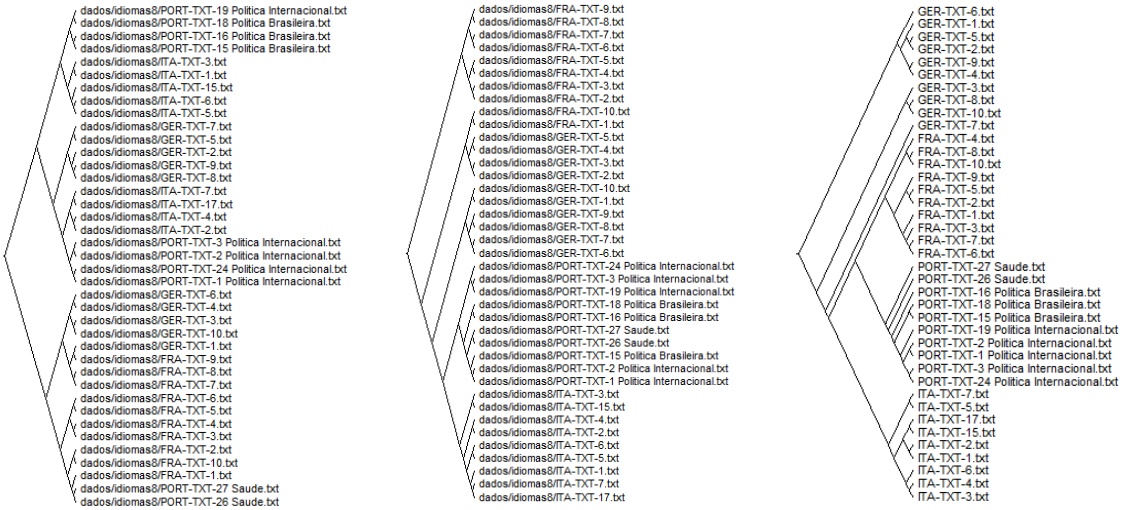


Figura 5.23: *Cladograma de 39 objetos comprimidos por MT Símbolo (Agrup. à esquerda), MT Digrama (Agrup. do meio) e MT Palavra (Agrup. à direita)*

Os agrupamentos particionados da FD8 são concebidos com objetos comprimidos a partir de MTs configuradas como símbolo e palavra. Os resultados apresentados, em agrupamentos particionados da FD8, podem ser observados graficamente nas figuras 5.24 e 5.25.

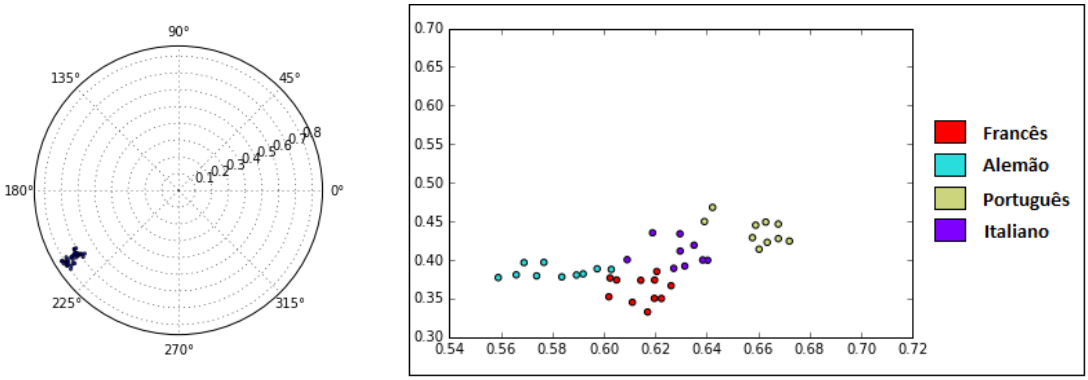


Figura 5.24: *Plano coord. polar de 39 objetos comprimidas por MT Símbolo (plano à esquerda) e plano coord. retangular (plano. à direita)*

Para os experimentos realizados na FD8, as MTs Símbolo e Palavra apresentaram boas hipóteses H de agrupamento. Os agrupamentos particionados gerados, a partir das duas hipóteses de agrupamento, são linearmente separáveis. Para a concepção de um classificador poderia ser utilizado a hipótese H para a MT Símbolo ou Palavra.

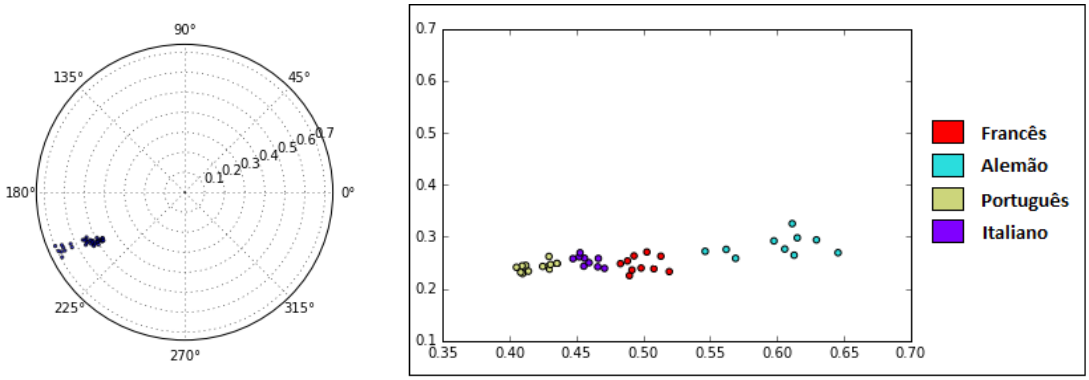


Figura 5.25: *Plano coord. polar de 39 objetos comprimidos por MT Palavra (plano à esquerda) e plano coord. retangular (plano. à direita)*

5.1.2 Resultados empíricos de algoritmos adaptativos para textos

Diante dos experimentos envolvendo algoritmos adaptativos de compressão, são exibidos dois gráficos, que representam a acurácia dos agrupamentos de classes e subclasses, baseados no coeficiente G (capacidade de generalização) em função de cada uma das fontes de dados.

O resultado obtido, nos agrupamentos de classes com 5 (cinco) idiomas distribuídos em 8 (oito) fonte de dados, é ilustrado no gráfico da Figura 5.26

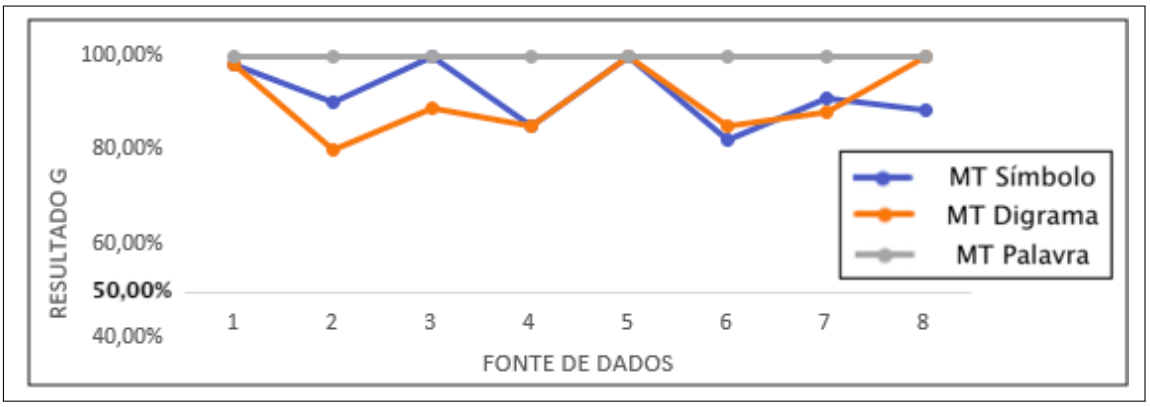


Figura 5.26: *Resultado G do agrupamento de classes (idiomas)*

As fontes de dados possuem diferentes características simulando diversas situações através de fontes de dados com variados números de objetos e diferentes números de classes. O algoritmo MT Palavra obteve uma capacidade de generalização de 100%, em todas as fontes de dados, isto é, obteve uma média de 100% nos resultados e apresentado, assim, um desvio padrão igual a 0. Os algoritmos MT Símbolo e MT Digrama apresentaram algumas variações nos resultados, o MT Símbolo obteve uma média igual a 91,90% e uma variância de 6,77 enquanto o MT Digrama obteve uma média de 90,68% e variância de 7,67. A partir dos experimentos empíricos, detectou-se que o MT Palavra produz agrupamentos com maior valor de generalização, sendo, então, utilizado nos demais experimentos para a fase de granulação.

A Figura 5.27 ilustra os resultados obtidos no agrupamento de subclasses dentro do idioma português, somente as fontes de dados 6, 7 e 8 apresentavam conjuntos de dados com subclasses.

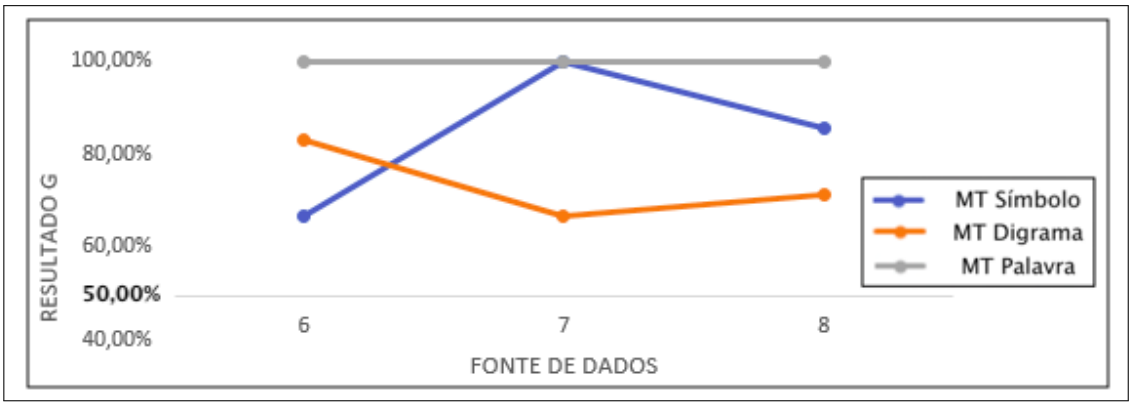


Figura 5.27: *Resultado G do agrupamento de subclasses dentro da classe idiomas*

Após análise gráfica e estatística dos resultados, nota-se que o algoritmo MT Palavra mantém assertividade em 100% apresentando assim, uma média de 100% com

desvio padrão igual 0. O algoritmo MT Símbolo demostra algumas variações e obtém uma média de 84,13% e desvio padrão de 16,72 enquanto o MT Digrama possui uma média de 73,81% e um desvio padrão de 8,58. Nos resultados com agrupamento de subclasses o algoritmo MT Palavra, também, mantém melhores resultados justificando, assim, sua escolha para a realização dos demais experimentos com conjuntos de dados variados.

5.1.3 Resultados para fontes texto com semântica

As fontes de dados de texto de naturezas diferentes são organizadas em tabelas, que dispõem de quantidades diferentes de classes. Os resultados envolvem a avaliação de agrupamentos a partir da análise semântica de textos com base na avalização da métrica G.

A Tabela 5.11 contém dados de sinopses de filmes e apresenta , em cada linha, uma fonte de dados distinta, com um número variado de classes, com o intuito de analisar o comportamento do algoritmo em diversos cenários.

Tabela 5.11: *Bases de dados com sinopses de filmes e divisão por classes*

Num. de CLASSES	**Sinopses de filmes**
2	AÇÃO(9), BIOGRAFIA(7)
3	AÇÃO(9), BIOGRAFIA(7), COMÉDIA(5)
4	AÇÃO(9), BIOGRAFIA(7), COMÉDIA(5), DRAMA(3)
5	AÇÃO(9), BIOGRAFIA(7), COMÉDIA(5), DRAMA(3), HISTÓRIA(4)
6	AÇÃO(9), BIOGRAFIA(7), COMÉDIA(5), DRAMA(3), HISTÓRIA(4), ROMANCE(6)
7	AÇÃO(9), BIOGRAFIA(7), COMÉDIA(5), DRAMA(3), HISTÓRIA(4), ROMANCE(6), GUERRA(6)

A Tabela 5.12 contém dados de abstracts de artigos, descreve 4 (quatro) fontes de dados, simulando um ambiente com as classes todas balanceadas e variando de 2 (duas) a 5 (cinco) classes.

Tabela 5.12: *Bases de dados com abstracts de artigos e divisão por classes*

Num. de CLASSES	**Abstracts de artigos**
2	BIOLOGIA(50), QUÍMICA(50)
3	BIOLOGIA(50), QUÍMICA(50), CIÊNCIA DA COMPUTAÇÃO(50)
4	BIOLOGIA(50), QUÍMICA(50), CIÊNCIA DA COMPUTAÇÃO(50), FÍSICA(50)
5	BIOLOGIA(50), QUÍMICA(50), CIÊNCIA DA COMPUTAÇÃO(50), FÍSICA(50), CIÊNCIAS SOCIAIS(50)

A Tabela 5.13 contém dados de avaliação de produtos e dispõe de uma fonte de dados com 2 (duas) classes balanceadas contendo 100 objetos cada uma dessas classes.

Tabela 5.13: *Bases de dados com avaliação de produtos e divisão por classes*

Num. de CLASSES	**Avaliação de produtos**
2	NEGATIVO(100), POSITIVO(100)

A 5.14 contém 5 fontes de dados de textos contendo sentimentos divididas em 6 classes diferentes. As classes foram balanceadas onde cada uma contém 50 objetos.

Tabela 5.14: *Bases de dados com textos (Twitters)) contendo sentimentos e divisão por classes*

Num. de CLASSES	**Texto com sentimentos**
2	TRISTEZA(50), SURPRESA(50)
3	TRISTEZA(50), SURPRESA(50), ALEGRIA(50)
4	TRISTEZA(50), SURPRESA(50), ALEGRIA(50), DESGOSTO(50)
5	TRISTEZA(50), SURPRESA(50), ALEGRIA(50), DESGOSTO(50), RAIVA(50)
6	TRISTEZA(50), SURPRESA(50), ALEGRIA(50), DESGOSTO(50), RAIVA(50), MEDO(50)

A Tabela 5.15 apresenta dados de twitters distribuídas em 2 (duas) fontes de dados que podem possuir até 3 (três) classes distintas. As fontes de dados são balanceadas contendo aqui 100 objetos por classe.

Tabela 5.15: *Bases de dados com notícias do Twitter e divisão por classes*

Num. de CLASSES	**Twitter - notícias**
2	NEGATIVO(100), NEUTRO(100)
3	NEGATIVO(100), POSITIVO(100), NEUTRO(100)

A Tabela 5.16 contém dados de noticiários com 5 classes distribuídas entre 4 fontes. As fontes de noticiários possuem quantidades variadas de objetos, simulando, assim, um cenário onde os dados não estão balanceados.

Tabela 5.16: *Bases de dados com artigos noticiários e divisão por classes*

Num. de CLASSES	**Artigos - noticiários**
2	ECONOMIA(15), ESPORTE(21)
3	ECONOMIA(15), ESPORTE(21), POLÍTICA(17)
4	ECONOMIA(15), ESPORTE(21), POLÍTICA(17), SAÚDE(27)
5	ECONOMIA(15), ESPORTE(21), POLÍTICA(17), SAÚDE(27), TECNOLOGIA(27)

A Tabela 5.17 representa uma base de dados com 5 (cinco) idiomas separados em 4 (quatro) fontes, as fontes representam um cenário onde os dados não estão balanceados, variando a quantidade de objetos em cada classe.

Tabela 5.17: *Bases de dados com idiomas e divisão por classes*

Num. de CLASSES	**Idiomas**
2	FRANCÊS(30), ALEMÃO(30)
3	FRANCÊS(10), ALEMÃO(10), PORTUGUÊS(10)
4	FRANCÊS(10), ALEMÃO(10), PORTUGUÊS(9), ITALIANO(9)
5	FRANCÊS(3), ALEMÃO(3), PORTUGUÊS(3), ITALIANO(3), INGLÊS(3)

Para os experimentos envolvendo fontes texto de naturezas diferentes, são exibidos gráficos como o da figura 5.28, onde cada ponto representa o valor de G, referente a uma árvore filogenética, em função do número de classes do conjunto de dados que esta

mesma árvore representa. Foram utilizadas 4 (quatro) estratégias diferentes para obtenção dos resultados. Cada estratégia foi implementada por um conjunto de 4 (quatro) baterias.

Na primeira bateria de experimentos, é utilizada a estratégia, que envolve a implementação do algoritmo LZW de forma tradicional no protótipo. Os resultados obtidos são descritos na tabela 5.18 e apresentados na figura 5.28.

Tabela 5.18: *Resultados de G da primeira bateria de testes*

Num. de CLASSES	**2**	**3**	**4**	**5**	**6**	**7**
Sinopses de filmes	50,00%	44,44%	29,17%	25,93%	28,12%	21,62%
Abstracts de artigos	62,24%	48,98%	37,76%	35,51%	-	-
Avaliação de produtos	50,51%	-	-	-	-	-
Texto com sentimentos	54,08%	34,01%	26,53%	22,86%	22,11%	
Twitter - notícias	56,06%	28,96%	-	-	-	-
Artigos noticiários	64,71%	52,00%	48,68%	41,18%	-	-
Idiomas	100,00%	100,00%	100,00%	100,00%	-	-

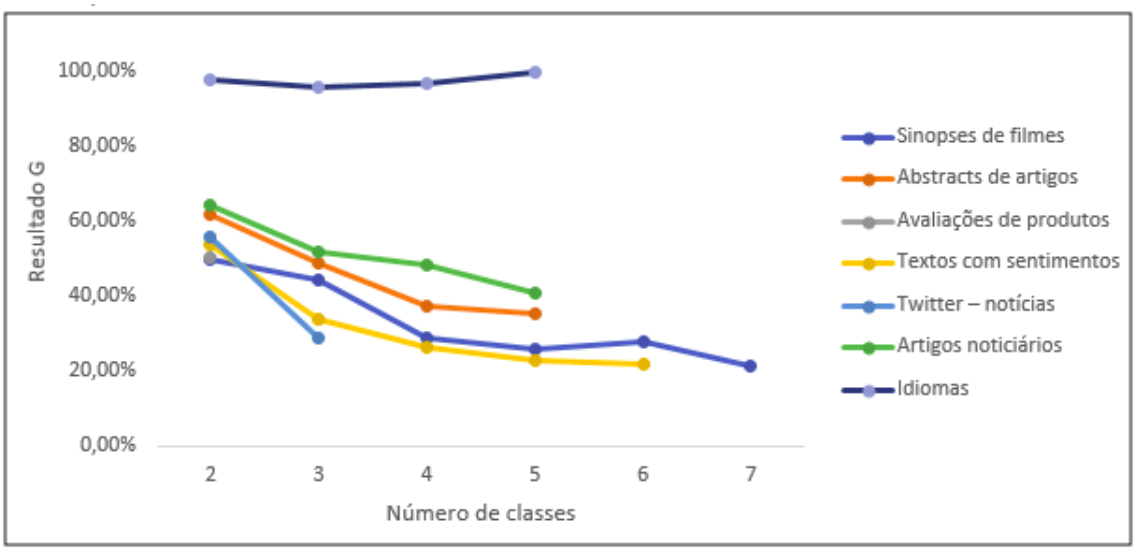


Figura 5.28: *Resultados com LZW tradicional*

Percebe-se que a estratégia de utilizar o algoritmo adaptativo LZW, tem resultados baixos de generalização, sendo que, conforme a quantidade de classes aumenta, a taxa de acerto diminui.

A segunda estratégia adotada é a partir da concepção de um único dicionário para todo o conjunto de documentos. Os resultados, dessa bateria de experimentos, são mostrados na tabela 5.19 e apresentados na figura 5.29.

Tabela 5.19: *Resultados de G da segunda bateria de testes*

Num. de CLASSES	2	3	4	5	6	7
Sinopses de filmes	78,57%	61,11%	37,50%	33,33%	31,25%	16,22%
Abstracts de artigos	74,49%	53,74%	30,61%	31,43%	-	-
Avaliação de produtos	58,08%	-	-	-	-	-
Texto com sentimentos	54,08%	30,61%	21,43%	23,27%	19,05%	
Twitter - notícias	63,13%	34,68%	-	-	-	-
Artigos noticiários	67,65%	58,00%	46,05%	50,98%	-	-
Idiomas	100,00%	100,00%	100,00%	100,00%	-	-

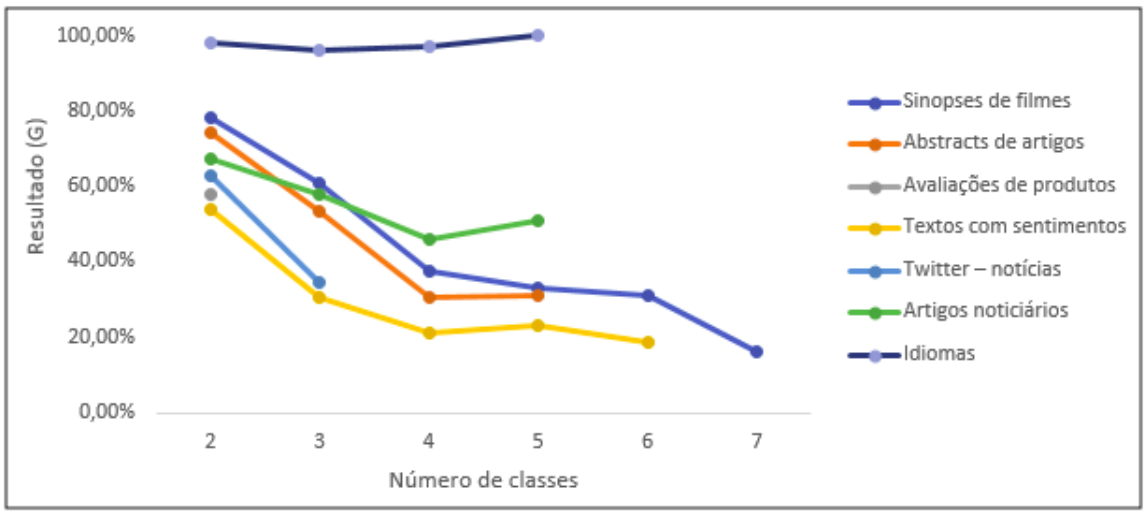


Figura 5.29: *Resultados com LZW utilizando um único dicionário*

Diante dos resultados da bateria 2 de experimentos, nota-se melhora nos resultados, mas conforme aumenta-se a quantidade de classes, os resultados, ainda, caem consideravelmente em termos de generalização. Assim, como na bateria 1, na bateria 2, notou-se que o algoritmo adaptativo tem dificuldade em encontrar e definir padrões em fontes de dados compostas por arquivos com poucas palavras.

A terceira estratégia utiliza um método de pré-processamento textual para remover *stopwords* (palavras vazias) diminuindo, ainda mais, os textos pequenos, mas mantendo somente padrões relevantes para o significado da expressão. A figura 5.30 ilustra os resultados da terceira bateria de experimentos descritos pela tabela 5.20.

Tabela 5.20: *Resultados de G da terceira bateria de testes*

Num. de CLASSES	2	3	4	5	6	7
Sinopses de filmes	42,86%	50,00%	54,17%	51,85%	40,62%	56,76%
Abstracts de artigos	71,43%	57,82%	60,20%	50,20%	-	-
Avaliação de produtos	59,09%	-	-	-	-	-
Texto com sentimentos	50,00%	80,27%	84,69%	83,27%	81,63%	
Twitter - notícias	56,57%	81,82%	-	-	-	-
Artigos noticiários	50,00%	50,00%	39,47%	39,22%	-	-
Idiomas	100,00%	100,00%	100,00%	100,00%	-	-

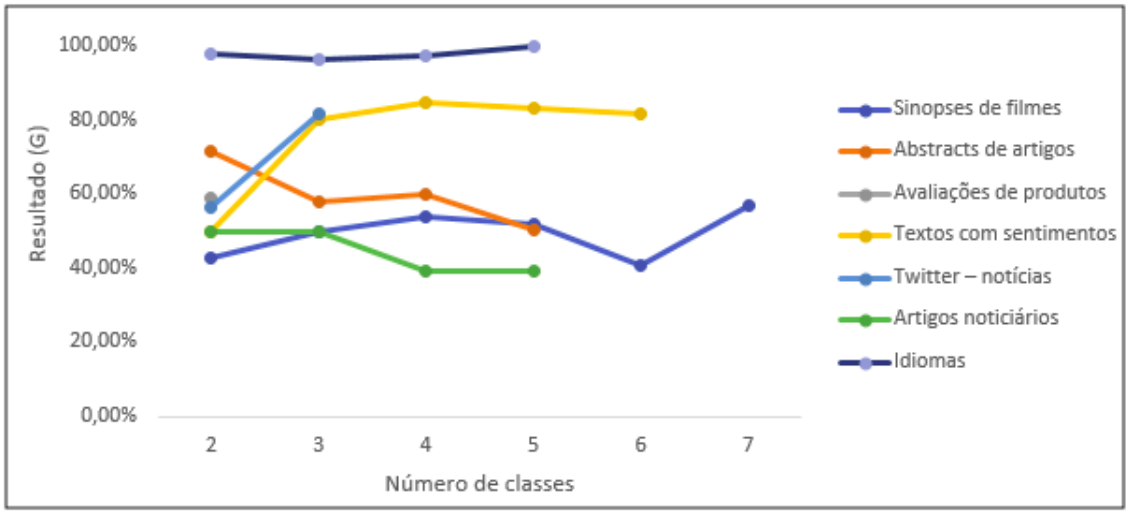


Figura 5.30: *Resultados com LZW aplicando stopwords*

Com essa estratégia, nota-se que, conforme a quantidade de classes aumenta, a taxa de generalização cai suavemente diferente das abordagens em baterias anteriores. Para algumas fontes de dados, houve melhora significativa de generalização com o aumento de classes.

A quarta estratégia possui uma melhora considerável de generalização comparada as estratégias anteriores, implementando no projeto de granulação, tag de negação, combinação com bigramas e aplicando a remoção das *stopwords*. Após executar o algoritmo da quarta estratégia, os resultados obtidos são descritos na tabela 5.21 e apresentados na figura 5.31.

Tabela 5.21: *Resultados de G da quarta bateria de testes*

Num. de CLASSES	2	3	4	5	6	7
Sinopses de filmes	71,43%	66,67%	75,00%	51,85%	43,75%	51,35%
Abstracts de artigos	70,41%	67,35%	74,49%	71,02%	-	-
Avaliação de produtos	61,11%	-	-	-	-	-
Texto com sentimentos	61,22%	81,63%	85,71%	84,49%	82,65%	
Twitter - notícias	60,61%	82,15%	-	-	-	-
Artigos noticiários	52,94%	62,00%	52,63%	60,78%	-	-
Idiomas	100,00%	100,00%	100,00%	100,00%	-	-

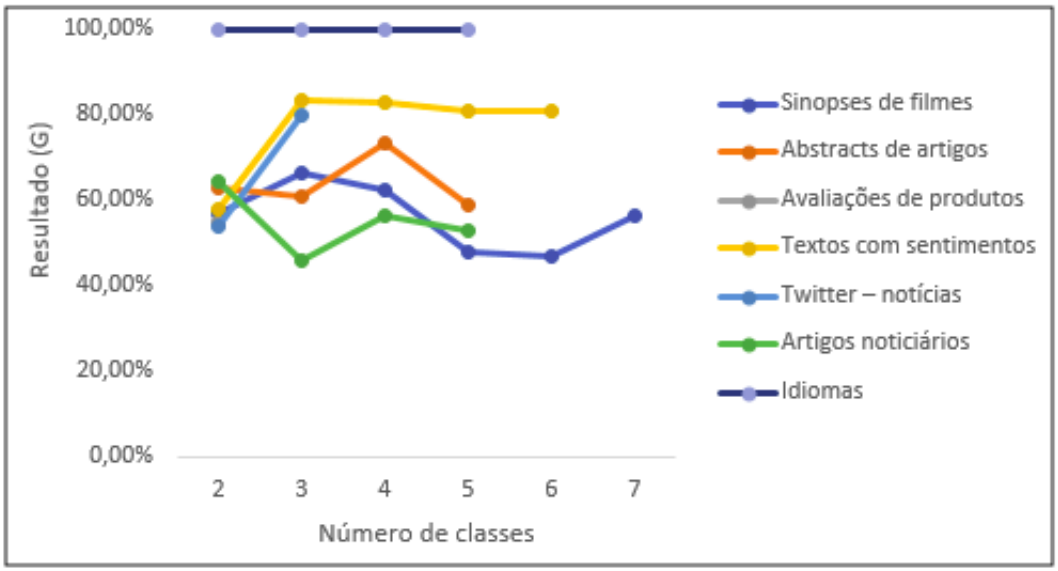


Figura 5.31: *Resultados com LZW aplicando stopwords e frequência*

Na quarta bateria de experimentos, os resultados da terceira bateria se mantiveram e os resultados regulares anteriores obtiveram melhores resultados nesta bateria, desta forma, o aumento da quantidade de classes não deprecia a taxa de generalização, assim como os textos com poucas palavras agora mostram melhor desempenho nos resultados.

5.1.4 Comparações (Benchmark com arquivos de texto)

Para fins comparativos, foi executado o ambiente DAMICORE (aplicação de [20], disponível em sua página do GitHub[0]) para as mesmas bases de dados e a mesma métrica proposta nesta pesquisa. Os resultados são apresentados na figura 5.32, figura 5.33 e a tabela 5.22 com as taxas de generalização para cada teste.

Tabela 5.22: *Resultados de G dos testes DAMICORE*

Num. de CLASSES	2	3	4	5	6	7
Sinopses de filmes	35,71%	33,33%	12,50%	11,11%	15,62%	10,81%
Abstracts de artigos	89,98%	83,67%	81,63%	84,90%	-	-
Avaliação de produtos	59,09%	-	-	-	-	-
Texto com sentimentos	53,06%	39,46%	25,51%	21,63%	22,45%	
Twitter - notícias	59,60%	36,70%	-	-	-	-
Artigos noticiários	88,24%	74,00%	78,95%	76,47%	-	-
Idiomas	100,00%	96,30%	97,06%	100,00%	-	-

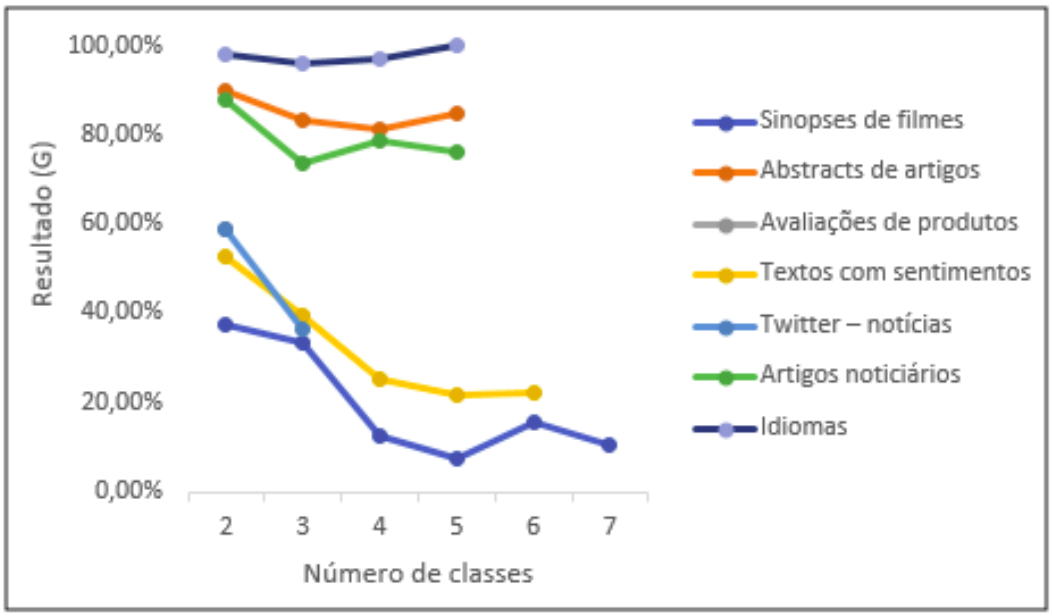
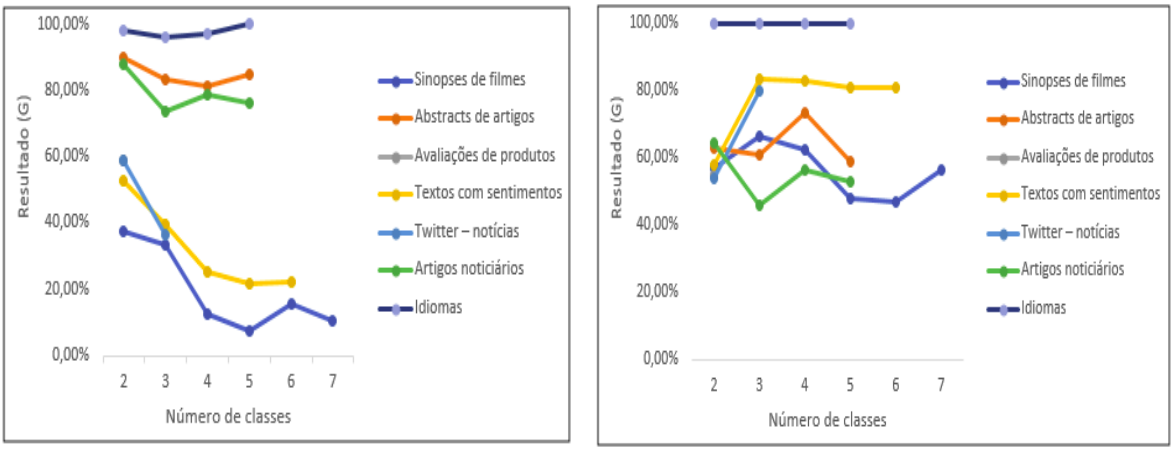


Figura 5.32: *Resultados DAMICORE*

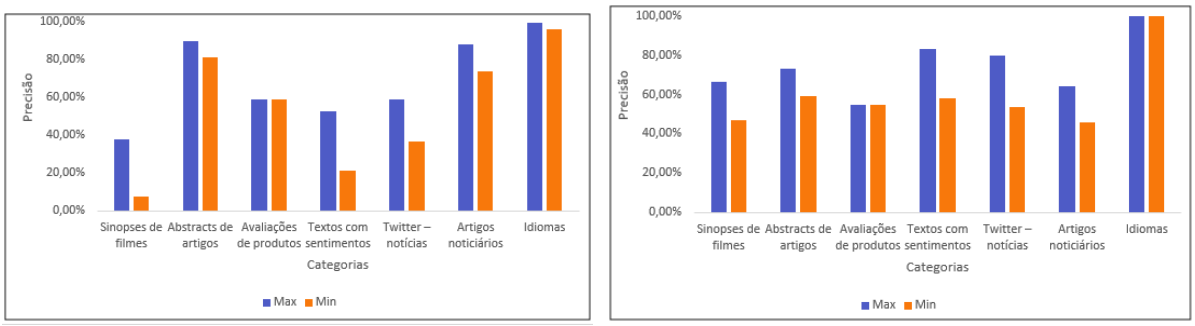
As figuras 5.33 e a tabela 5.34 dispõem de um comparativo entre a solução proposta e o DAMICORE.



(a) *Resultado DAMICORE* (b) *Resultado Solução Proposta*

Figura 5.33: *Comparativo geral DAMICORE X Solução Proposta (Benchmark de textos)*

Nas figuras (a) e (b), são mostrados os limites assintóticos (min - max) entre o índice de generalização G encontrados na abordagem DAMICORE e na proposta desta pesquisa.



(a) *Melhores e piores resultados DAMICORE* (b) *Melhores e piores resultados Solução Proposta*

Figura 5.34: *Comparativo melhores e piores resultados DAMICORE X Solução Proposta*

5.2 Experimentos com objetos comprimidos do tipo imagem

Foram realizados os experimentos com imagens de frutas, plantações, animais, captcha, faces humanas e pinturas. A compressão partiu da utilização de uma MT (léxico=LZW e sintático=Sistema L - curva de Hilbert + RLE (transições)], pois as estratégias adaptativas de compressão podem ser combinadas. Os experimentos foram realizados com o protótipo desenvolvido e disponibilizado nesta pesquisa, e, também, com o protótipo da aplicação de [20] para fins comparativos da seguinte forma:

- primeiramente são apresentados os resultados obtidos do protótipo, desta pesquisa, para imagens com quantização de 32 cores utilizando tabelas que exibem os valores de generalização G de cada fonte de dados;
- são apresentados os resultados empíricos encontrados para todas as fontes de dados;
- posteriormente no mesmo capítulo é realizada uma comparação (Benchmark) dos resultados com as mesmas fontes de dados contidos na aplicação de [20];
- e, finalmente, é calculada a perda de informação do modelo dada por $\varepsilon(X)$, para X sendo uma imagem digital.

5.2.1 Resultados para imagens com quantização de 32 cores

Nesta seção, são apresentados os resultados obtidos com a aplicação proposta nas diferentes categorias de imagens coletadas, os resultados estão descritos da forma em que foram organizados os experimentos, em diferentes fontes de dados (diretórios contendo imagens). O número de cores utilizado, na etapa de quantização, para os resultados aqui descrito foi de 32 cores.

Frutas

Foram utilizadas 102 imagens de frutas cujos experimentos foram divididos em 6 fontes de dados:

Fonte	Classes	G
FONTE1(40)	2: BANANA(20); LARANJA(20)	0.7631
FONTE2(60)	3: BANANA(20); LARANJA(20); MAÇA(20)	0.6842
FONTE3(80)	4: BANANA(20); LARANJA(20); MAÇA(20) E LIMÃO(20)	0.6184
FONTE4(100)	5: BANANA(20); LARANJA(20); MAÇA(20); LIMÃO(20) E MANGA(20)	0.5474
FONTE5(122)	6: BANANA(20); LARANJA(20); MAÇA(20); LIMÃO(20); MANGA(20) E MORANGO(22)	0.5862

Tabela 5.23: *Resultados - Frutas*

Este experimento visa testar a capacidade do algoritmo de identificar diferentes frutas, como pode ser visto na tabela 5.23, as árvores resultantes demonstram uma precisão acima de 58% para todos os experimentos realizados, chegando até 76%, no caso mais trivial, com apenas 2 classes.

Animais

Foram utilizadas 360 imagens de animais cujos experimentos foram divididos em 9 fontes de dados:

Fonte	Classes	G
FONTE11(30)	2: HOMEM(15); MACACO(15)	0.8928
FONTE12(60)	2: HOMEM(30); MACACO(30)	0.8793
FONTE13(120)	4: HOMEM(30); MACACO(30); SERPENTE(30); TUBARÃO(30)	0.6293
FONTE14(240)	8: HOMEM(30); MACACO(30); SERPENTE(30); TUBARÃO(30); CROCODILO(30); CAVALO(30); ELEFANTE(30); BEIJA-FLOR(30)	0.4827
FONTE15(270)	9: HOMEM(30); MACACO(30); SERPENTE(30); TUBARÃO(30); CROCODILO(30); CAVALO(30); ELEFANTE(30); BEIJA-FLOR(30); GATO(30)	0.4636
FONTE16(270)	9: HOMEM(30); MACACO(30); SERPENTE(30); TUBARÃO(30); CROCODILO(30); CAVALO(30); ELEFANTE(30); BEIJA-FLOR(30); GIRAFA(30)	0.4367
FONTE17(270)	9: HOMEM(30); MACACO(30); SERPENTE(30); TUBARÃO(30); CROCODILO(30); CAVALO(30); ELEFANTE(30); BEIJA-FLOR(30); AVESTRUZ(30)	0.4444
FONTE18(300)	10: HOMEM(30); MACACO(30); SERPENTE(30); TUBARÃO(30); CROCODILO(30); CAVALO(30); ELEFANTE(30); BEIJA-FLOR(30); GATO(30); GIRAFA(30)	0.4275
FONTE19(330)	11: HOMEM(30); MACACO(30); SERPENTE(30); TUBARÃO(30); CROCODILO(30); CAVALO(30); ELEFANTE(30); BEIJA-FLOR(30); GATO(30); GIRAFA(30); AVESTRUZ(30)	0.3824

Tabela 5.24: *Resultados - Animais*

Este experimento visa testar a capacidade do algoritmo de identificar diferentes animais. Como pode ser visto na tabela 5.25, as árvores resultantes demonstram uma precisão de apenas 38%, no pior caso, chegando até 89% com apenas 2 classes.

Plantações

Foram utilizadas 360 imagens de plantações cujos experimentos foram divididos em 9 fontes de dados:

Fonte	Classes	G
FONTE21(60)	4: ALGODAO(15); ARROZ(15); GIRASSOL(15); MILHO(15)	0.75
FONTE22(120)	4: ALGODAO(30); ARROZ(30); GIRASSOL(30); MILHO(30)	0.6896
FONTE23(150)	5: ALGODAO(30); ARROZ(30); GIRASSOL(30); MILHO(30); MANDIOCA(30)	0.6758
FONTE24(180)	6: ALGODAO(30); ARROZ(30); CANAVIAL(30); GIRASSOL(30); MANDIOCA(30); MILHO(30)	0.6494
FONTE25(210)	7: ALGODAO(30); ARROZ(30); CANAVIAL(30); GIRASSOL(30); MANDIOCA(30); MILHO(30); MAMÃO(30)	0.6256
FONTE26(240)	8: ALGODAO(30); ARROZ(30); CANAVIAL(30); GIRASSOL(30); MANDIOCA(30); MILHO(30); MAMÃO(30); BANANA(30)	0.612
FONTE27(270)	9: ALGODAO(30); ARROZ(30); CANAVIAL(30); GIRASSOL(30); MANDIOCA(30); MILHO(30); MAMÃO(30); BANANA(30); JABUTICABA(30)	0.5977
FONTE28(300)	10: ALGODAO(30); ARROZ(30); CANAVIAL(30); GIRASSOL(30); MANDIOCA(30); MILHO(30); MAMÃO(30); BANANA(30); JABUTICABA(30); ABACAXI(30)	0.5655
FONTE29(360)	12: ALGODAO(30); ARROZ(30); CANAVIAL(30); GIRASSOL(4); MANDIOCA(30); MILHO(30); MAMÃO(30); BANANA(30); JABUTICABA(30); ABACAXI(30); TOMATE(30); ALFACE(30)	0.5402

Tabela 5.25: *Resultados - Plantações*

Este experimento visa testar a capacidade do algoritmo de identificar diferentes plantações. Como pode ser visto na tabela 5.25, as árvores resultantes demonstram resultados medianos, uma precisão acima de 54% para todos os experimentos realizados, chegando até 75%, no experimento mais trivial, com 4 classes.

CAPTCHA

Foram utilizadas 124 imagens de testes CAPTCHA cujos experimentos foram divididos em 4 fontes de dados:

Fonte	Classes	G
FONTE31(16)	2: VEÍCULO(8); NÃO_VEÍCULO(8)	1
FONTE32(32)	2: VEÍCULO(16); NÃO_VEÍCULO(16)	0.9642
FONTE33(64)	2: VEÍCULO(32); NÃO_VEÍCULO(32)	0.85
FONTE34(124)	2: VEÍCULO(54); NÃO_VEÍCULO(70)	0.8278

Tabela 5.26: *Resultados - CAPTCHA*

Este experimento visa testar a capacidade do algoritmo de identificar imagens que possuam ou não veículos, assim, como é feito nos testes CAPTCHA (Complete Automated Test to tell Computers and Humans Apart). O teste CAPTCHA é uma tarefa, supostamente, exclusiva para humanos. Como pode ser visto na tabela 5.26, os resultados dessa categoria foram muito bons, provavelmente pela baixa quantidade de classes a serem identificadas(2), as árvores resultantes demonstram uma precisão acima de 85% para todos os experimentos realizados, chegando até 100%, no experimento mais trivial, com apenas 16 imagens, 8 contendo veículos e 8 sem veículos, vale destacar , também, o experimento com maior quantidade de imagens (fonte34 com 124 imagens) que teve um resultado de 99%.

Faces Humanas

Foram utilizadas 400 imagens de faces humanas cujos experimentos foram divididos em 8 fontes de dados:

Fonte	Classes	G
FONTE41(20)	2: INDIVIDUO_0(10); INDIVIDUO_1(10)	0.9444
FONTE42(30)	3: INDIVIDUO_0(10); INDIVIDUO_1(10);; INDIVIDUO_2(10)	0.9629
FONTE43(40)	4: INDIVIDUO_0(10); INDIVIDUO_1(10); INDIVIDUO_2(10); INDIVIDUO_3(10)	0.9722
FONTE44(80)	8: INDIVIDUO_0(10); INDIVIDUO_1(10); INDIVIDUO_2(10); INDIVIDUO_3(10); INDIVIDUO_4(10); INDIVIDUO_5(10); INDIVIDUO_6(10); INDIVIDUO_7(10)	0.9861
FONTE45(80)	10: INDIVIDUO_0(10); INDIVIDUO_1(10); INDIVIDUO_2(10); INDIVIDUO_3(10); INDIVIDUO_4(10); INDIVIDUO_5(10); INDIVIDUO_6(10); INDIVIDUO_7(10); INDIVIDUO_8(10); INDIVIDUO_9(10)	0.9888
FONTE46(200)	20: INDIVIDUO_0(10); INDIVIDUO_1(10); INDIVIDUO_2(10); INDIVIDUO_3(10); INDIVIDUO_4(10); INDIVIDUO_5(10); .....; INDIVIDUO 19(10)	0.9666
FONTE47(300)	30: INDIVIDUO_0(10); INDIVIDUO_1(10); INDIVIDUO_2(10); INDIVIDUO_3(10); INDIVIDUO_4(10); INDIVIDUO_5(10) ; .....; INDIVIDUO_29(10)	0.8962
FONTE48(400)	40: INDIVIDUO_0(10); INDIVIDUO_1(10); INDIVIDUO_2(10); INDIVIDUO_3(10); INDIVIDUO_4(10); INDIVIDUO_5(10); .....; INDIVIDUO_39(10)	0.8361

Tabela 5.27: *Resultados - Faces Humanas*

Este experimento visa testar a capacidade do algoritmo de identificar diferentes indivíduos, com possível aplicação, para o reconhecimento facial, por exemplo. Como

pode ser visto na tabela 5.27, os resultados, desse experimento, foram ótimos, as árvores resultantes demonstram uma precisão acima de 83% para todos os experimentos realizados, chegando até 98%, em um dos experimentos, com 10 classes e 80 imagens, vale destacar também, o experimento com maior quantidade de imagens (fonte48 com 400 imagens) que teve um resultado de 83%.

Pinturas

Foram utilizadas 255 imagens de pinturas de diferentes movimentos artísticos cujos experimentos foram divididos em 9 fontes de dados.

Fonte	Classes	G
FONTE61(20)	2: BIZANTINO(37); ROMANTICISMO(52)	0.9885
FONTE62(30)	2: GOTICO(31); BARROCO(49)	0.9615
FONTE63(40)	2: ROMANTISMO(52); CUBISMO(43)	0.9247
FONTE64(80)	2: BARROCO(49); EXPRESSIONISMO(43)	0.988
FONTE65(80)	2: CUBISMO(43); EXPRESSIONISMO(43)	0.94
FONTE66(200)	4: BIZANTINO(37); GOTICO(31); CUBISMO(43); EXPRESSIONISMO(43)	0.88
FONTE67(300)	4: BIZANTINO(37); GOTICO(31); ROMANTISMO(52); BARROCO(49)	0.8848
FONTE68(400)	4: ROMANTISMO(52); BARROCO(49); CUBISMO(43); EXPRESSIONISMO(43)	0.8251
FONTE69(400)	6: ROMANTISMO(52); BARROCO(49); CUBISMO(43); EXPRESSIONISMO(43); BIZANTINO(37); GOTICO(31)	0.7711

Tabela 5.28: *Resultados - Pinturas*

Este experimento visa testar a capacidade do algoritmo de identificar pinturas pertencentes a diferentes movimentos artísticos.

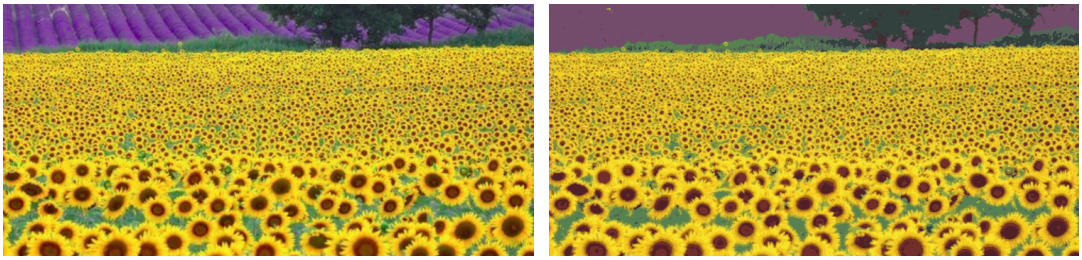
Como pode ser visto na tabela 5.28, os resultados desse experimento foram muito bons, as árvores resultantes demonstram uma precisão acima de 77% para todos os experimentos realizados, chegando até 98%, em alguns dos experimentos, com apenas 2 classes.

5.2.2 Resultados empíricos para imagens

As figuras 5.35 e 5.36 mostram comparações das imagens, que foram utilizadas nos experimentos em seu estado original e no estado após o pré-processamento (redimensionamento e quantização). Percebe-se que o número de cores é suficiente para demonstrar as diferenças entre cores mais relevantes para a percepção da cena, porém em alguns casos (exemplo na figura 5.37) pode-se perder alguns detalhes.

(a) *Imagem original* (b) *Imagem redimensionada e quantizada*

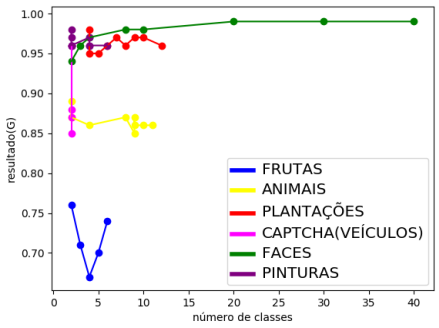
Figura 5.35: *Pintura do momento cubismo (imagem original vs processada)*



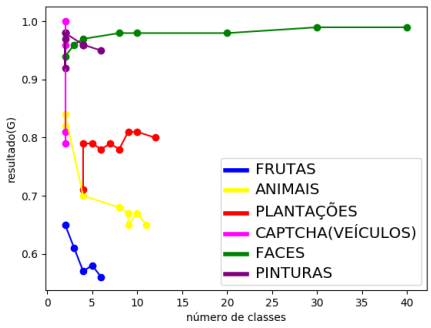
(a) *Imagem original* (b) *Imagem redimensionada e quantizada*

Figura 5.36: *Plantação de girassóis (imagem original vs processada)*

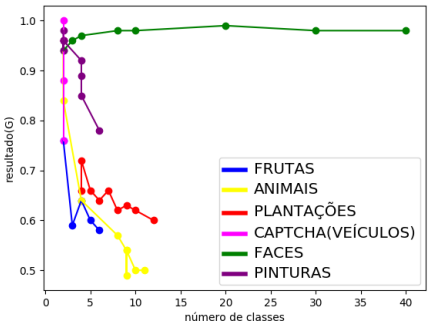
A tabela 5.29 e a figura 5.37 apresentam o valor de generalização G, em função da quantidade de cores na imagem, gerada pelo processo de quantização. A quantização diminui consideravelmente a quantidade excessiva de cores presentes nas imagens com o objetivo de combater a maldição da dimensionalidade.



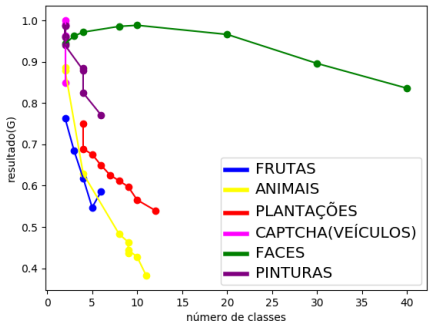
(a) *4 cores*



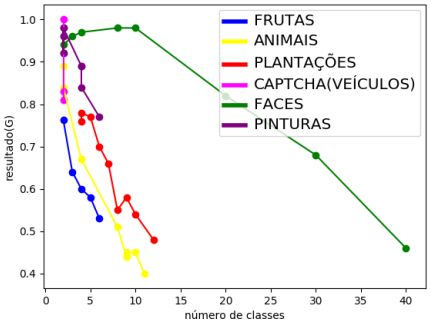
(b) *8 cores*



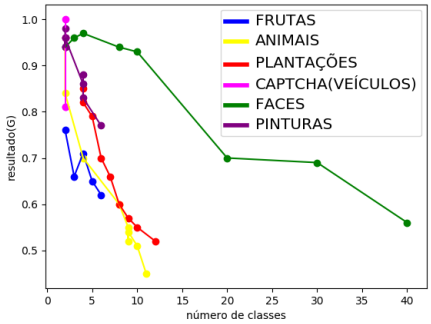
(c) *16 cores*



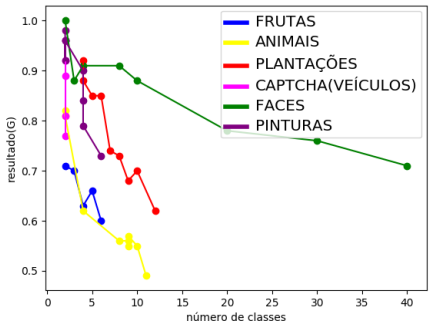
(d) *32 cores*



(e) *64 cores*



(f) *128 cores*



(g) *256 cores*

Figura 5.37: *Comparação dos resultados com diferentes números de cores na etapa de quantização*

Tabela 5.29: *Média de resultados com diferentes números de cores*

Quantidade de Cores	Frutas	Animais	Plantações	CAPTCHA	Faces Humanas	Pinturas	Media Geral
4 cores	0.71	0.86	0.96	0.89	0.97	0.96	0.89
8 cores	0.59	0.7	0.78	0.89	0.97	0.96	0.81
16 cores	0.63	0.62	0.64	0.90	0.97	0.91	0.78
32 cores	0.63	0.55	0.63	0.90	0.94	0.90	0.76
64 cores	0.62	0.56	0.64	0.90	0.84	0.90	0.74
128 cores	0.68	0.63	0.67	0.89	0.83	0.90	0.75
256 cores	0.66	0.61	0.77	0.84	0.85	0.89	0.77

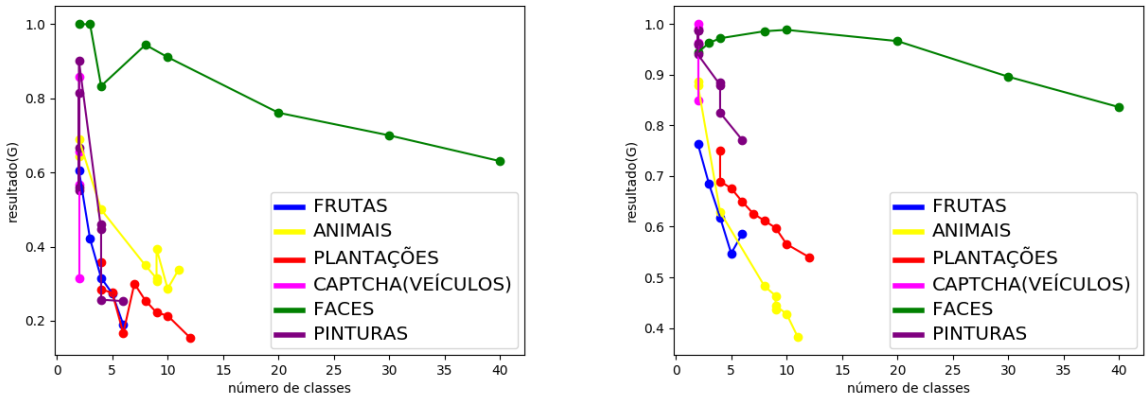
A escolha de 32 cores para exemplos de experimentos e comparações com outras abordagens de compressão, baseia-se de forma empírica em experimentos realizados com a intenção de alcançar boa capacidade G de generalização. O valor **32** se encontra no meio da série ordenada dos valores utilizados para testes (4, 8, 16, **32**, 64, 128, 256), e representa aproximadamente a média dos resultados obtidos nos diferentes experimentos realizados na etapa de quantização.

5.2.3 Comparações (Benchmark com arquivos de imagens)

Os experimentos também foram realizados na aplicação **damicore-py** disponibilizada por [20] (aplicação disponível em sua página do GitHub[0]). A aplicação faz uso do fluxo tradicional da metodologia DAMICORE em experimentos com textos realizados pelo autor.

Um problema, na abordagem tradicional, é que sejam os dados, imagens, textos, áudio ou qualquer outro tipo, a representação dos mesmos é totalmente dependente do algoritmo encapsulado pelo compressor, levando a representações muitas vezes imprecisas, que não representam informações relevantes dos dados para tarefas de classificação.

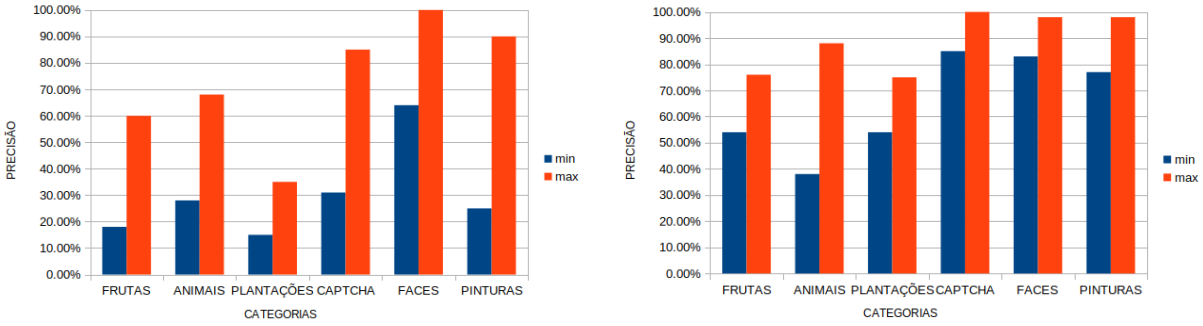
As árvores obtidas com a alteração proposta para o fluxo de operações do framework, realizando o desmembramento da etapa de compressão em uma fase de extração de características (Granulação) e outra de codificação, apresentam resultados de precisão (G) superiores comparados a abordagem DAMICORE, como podem ser vistas nas figuras 5.38 e 5.39.



(a) *Resultados DAMICORE* (b) *Resultados Solução Proposta*

Figura 5.38: *Comparativo DAMICORE X Solução Proposta (Benchmark de imagens)*

A aplicação proposta nesta tese atinge o seu limite inferior na categoria de animais, com uma precisão de apenas 38%, no experimento realizado com 11 classes, enquanto que a aplicação damicorepy chega a 15%, em um experimento com 12 classes, na categoria de plantações (nesse mesmo experimento a aplicação proposta atinge 54%).



(a) *Melhores e piores resultados DAMICORE* (b) *Melhores e piores resultados Solução Proposta*

Figura 5.39: *Comparativo melhores e piores resultados DAMICORE X Solução Proposta*

Percebe-se que, nos dois casos, de acordo com o crescimento do número de classes, a precisão costuma ser prejudicada em ambas aplicações. Porém, também, é perceptível que a aplicação GM possui melhores resultados com o aumento do número de classes, com declínios mais suaves.

5.2.4 Perda de informação do modelo

Nesta seção, são abordados os cálculos envolvendo a perda de informação do modelo com base na médias dos valores de ε (erro), obtidos para cada base de

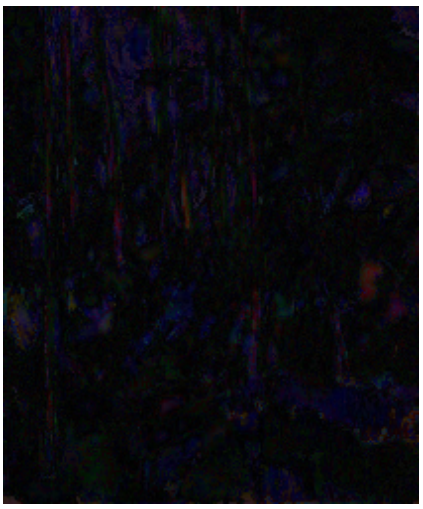
dados utilizando quantização de 32 cores. Os resultados da perdas de informação são sumarizados por categoria. Para as imagens das figuras 5.40 e 5.41, os valores de ε são respectivamente 0.084 (8.4%) e 0.136 (13.6%). As diferenças entre as imagens apresentadas tanto na figuras 5.40 como na figura 5.41 simbolizam a perda de informação do modelo no decorrer do processo de compressão de uma imagem comparado com sua função inversa de descompressão e reconstrução da mesma imagem.

A imagem 5.41 apresenta uma perda significativa de detalhes (variação de informação em relação a imagem original), na porção superior, onde a textura de fundo foi completamente substituída por uma única cor sólida enquanto que na imagem 5.40, de forma menos perceptível, também, ocorre uma perda de gradientes, mas nenhum detalhe é drasticamente eliminado não prejudicando a análise de semântica da imagem.

(a) *Imagem original*

(b) *Imagem redimensionada e quantizada*

(c) *Imagem reconstruída*



(d) *Diferença entre imagem real e reconstruída*

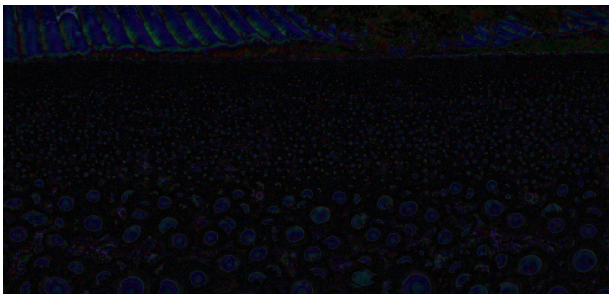
(e) *Diferença entre imagem real e reconstruída (cores invertidas*

Figura 5.40: *Pintura do momento cubismo - Comparações*

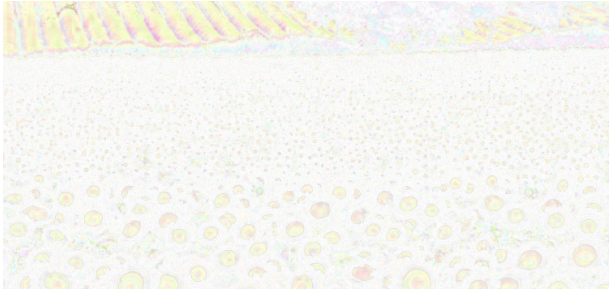
(a) *Imagem original*

(b) *Imagem redimensionada e quantizada*

(c) *Imagem reconstruída*



(d) *Diferença entre imagem real e reconstruída*



(e) *Diferença entre imagem real e reconstruída (cores invertidas*

Figura 5.41: *Plantação de girassóis - Comparações*

Foram calculadas as variações de informação entre as imagens coletadas em seu formato real com relação as imagens após o pré-processamento realizado na etapa de **extração de características**. A tabela 5.30 e o gráfico 5.42 apresentam os resultados das perdas de informação do modelo de exemplo para uma quantização de 32 cores.

Categoria	Média da variação da informação ε
Frutas	0.065
Animais	0.070
Plantações	0.099
CAPTCHA	0.045
Faces Humanas	0.017
Pinturas	0.075
Média Geral	0.061

Tabela 5.30: *Médias da variação da informação*

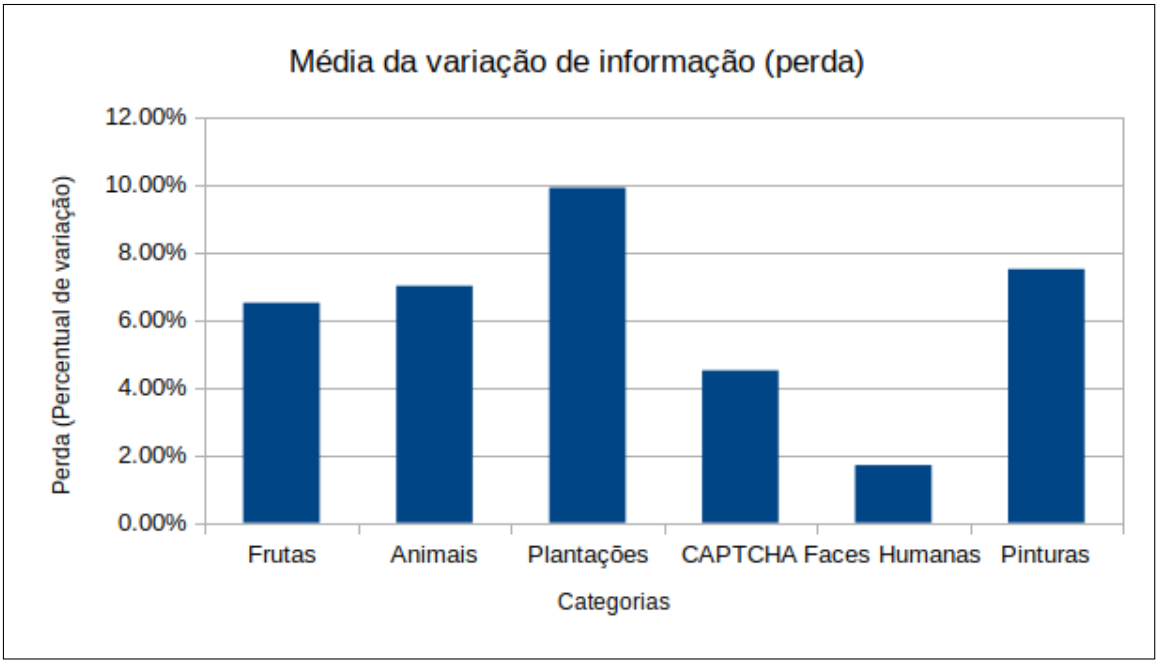


Figura 5.42: *Médias da variação da informação*

A perda média de informação, de todos os experimentos, de imagem para a quantização de 32 cores foi de 6,1%, como mostra a tabela 5.30.

CAPÍTULO 6

Conclusões

Neste capítulo, são discutidas as conclusões da tese a partir da comparação do modelo proposto com alternativas encontradas em trabalhos relacionados. São discutidas, também, as avaliações dos resultados encontrados para os formatos de texto e imagem, além de, conclusões gerais sobre o modelo, experimentos e descrição de alguns trabalhos futuros que possam ser desenvolvidos a partir da proposta.

6.1 Conclusões sobre os experimentos e o modelo

Com relação a análise de todos os experimentos realizados, pode-se destacar:

- Para objetos não-estruturados, do tipo imagem, as técnicas de compressão sem perdas não obtiveram bons resultados mensurados a partir da métrica G. A partir da aplicação de técnicas de compressão, com perdas a capacidade de generalização do modelo de aprendizado para a concepção da hipótese H aumenta.
- O modelo de redução de dimensionalidade implementado, com técnicas de compressão para os agrupamentos particionados, representou os objetos de dados. Isso significa que existe uma configuração de dependência da MT, capaz de gerar agrupamentos, que não caracterizam problemas de generalização para classificadores. A redução de dimensionalidade descrita na tese parte de um espaço n-dimensional que é comprimido, assim, a representação dos objetos, neste espaço, torna-se menos esparsa.
- A granulação é um processo essencial para objetos não-estruturados, pois unidades básicas de informação podem ser identificadas e contabilizadas. A granulação é essencial para a definição no bias da aplicação de reconhecimento de padrões contribuindo, também, com mais flexibilidade para a aplicação de RP em diferentes contextos de informação.
- A escolha do nível de dependência da informação da MT influência a geração de agrupamentos com boas hipóteses H. Nada impede que um nível mais baixo de dependência da informação, seja escolhido para a MT para um projeto de

reconhecimento de padrões e, assim, obter uma boa hipótese de agrupamento. Uma boa abordagem é comparar hipóteses H de várias configurações da MT e, então, escolher o melhor nível de dependência para a MT.

- A técnica GM se mostra eficiente obtendo altos níveis de valores de G para conjuntos de dimensionalidade elevada.
- Para problemas de agrupamento mais complexos como, por exemplo, utilizando objetos de dados de imagens, a técnica adaptativa a partir de dicionários implementada no LZW, de forma isolada, não obteve hipóteses H com valores elevados de G. Torna-se necessário a concepção de um analisador sintático. Um exemplo da utilização de um analisador sintático foi a curva de Hilbert, um tipo de curva fractal, que transfere informações de uma estrutura n-dimensional para uma estrutura unidimensional preservando as propriedades de localidade da estrutura n-dimensional.
- Classificadores podem utilizar uma hipótese H ou um conjunto delas para delimitar fronteiras de decisão que maximizem suas margens de separação de classes.

A técnica de compressão utilizando MT mitiga o problema da dimensionalidade reduzindo a dimensão do espaço de características a partir da criação de metadados que fornecem as informações dos grânulos que compõem os objetos. Sendo assim, o modelo implementa alternativas para a resolução de problemas de generalização da seguinte forma:

- Para objetos de dados comprimidos, a NCD, como relação de equivalência R de um espaço de agrupamento $E = (SI, R)$, proporcionou bons resultados concebendo aproximações da hipótese H próximas da expectativa E_X de agrupamento.
- O modelo de concepção dos agrupamentos proposto na pesquisa utiliza conceitos de compressão de dados através da desigualdade de Kraft. Através desta técnica, pode-se identificar as características presentes no objeto e atribuir maiores pesos para as principais características ou mais importantes. Características, que não têm relevância para o problema ou ruidosas, possuem pesos muito pequenos atribuídos, nesta abordagem, sendo eliminadas da hipótese H de agrupamento evitando, assim, o problema da dimensionalidade.
- O problema de separação de classes por idiomas, é na maioria dos casos separável linearmente. A partir de um conjunto representativo de características, um classificador, com fronteira linear, pode separar os objetos de uma fonte de dados em grupos. Vale a pena lembrar que a arquitetura de um classificador simples combinada a boas características, pode evitar problemas de generalização como a superespecialização e o super-treinamento.
- A partir dos diversos níveis de configuração da MT, nota-se que objetos representados com valores de dependência mais altos, tendem a produzir boas hipóteses H de agrupamento.

- A capacidade de generalização apresentada pelos agrupamentos representados pelos dendrogramas da abordagem de compressão com perdas utilizando a quantização e uma gramática (curva de Hilbert), decaiu, suavemente, em função do aumento do número de classes nos experimentos.

6.2 Conclusões a partir de Trabalhos Relacionados

A metodologia DAMICORE mostra-se uma ferramenta de baixa complexidade computacional capaz de produzir bons resultados em tarefas de agrupamentos e classificação, porém o método apresentou resultados baixos da métrica G [20, 75, 80, 63, 43, 52]. Como possível causa dos resultados negativos, é apontada a dependência da representação dos dados com o algoritmo do compressor utilizado, o que pode levar a representações inadequadas.

O trabalho de [20] realizou a implementação do fluxo de operações DAMICORE em linguagem python. Nessa implementação, explorou-se diferentes formas de utilizar o conjunto de métodos presentes no fluxo de operações DAMICORE, além disso, vale destacar a extensão da metodologia para um número ilimitado de classes e também a extensão da metodologia para tarefas de classificação. O sistema desenvolvido por [20] apresentou bons resultados com o método de classificação kNN (k Nearest Neighbours), com precisão acima de 90% em todos testes expostos. Apesar da aplicação ser capaz de funcionar para qualquer tipo de dado, o autor apresenta apenas testes realizados com textos, mais precisamente quanto à capacidade de classificação de arquivos texto do tipo spam e identificação de arquivos de idiomas. Foram desenvolvidos experimentos com imagens utilizando a aplicação de [20], disponível em sua página do GitHub[0]. Em fontes de dados de imagens, o algoritmo de [20] não foi capaz de produzir bons resultados. A capacidade de generalização dada pela métrica G apresentada pelos agrupamentos representados pelos dendrogramas da abordagem DAMICORE decaiu em função do aumento do número de classes, nos experimentos com objetos do tipo imagem, enquanto na proposta, desta pesquisa, manteve-se constante. O modelo de representação genérico dos compressores (compactadores) não produz uma representação adequada para a detecção de padrões em imagens.

O desmembramento da etapa de compactação em duas etapas: uma etapa responsável pela extração de características da imagem, e outra, pela codificação dessas características produziu melhores resultados em comparação com a aplicação de [20], como pode ser visto e demostrado no capítulo 5 (cinco) Resultados.

[0]Aplicação disponível em: https://github.com/brunokim/damicore-python

6.2.1 Imagens

A técnica proposta, na pesquisa, para a representação e codificação das imagens, demonstrou resultados melhores quando comparado ao método original proposto por [75]. Mesmo o protótipo obtendo bons resultados, também obteve resultados ruins em alguns domínios de experimentos realizados.

Entre os experimentos realizados, com a finalidade de testar a aplicação desenvolvida, destacam-se os experimentos envolvendo imagens de faces humanas, imagens CAPTCHA e imagens pinturas de diferentes movimentos artísticos. Nessas categorias, foram produzidas árvores filogenéticas com uma precisão da métrica G acima de 77% para todos experimentos realizados, demonstrando uma boa capacidade de generalização.

Outros experimentos obtiveram resultados abaixo de 50%, isso, provavelmente, deve-se ao fato de que os conjuntos de imagens de plantações, animais e frutas possuem muita variação de fundo e resolução, além do arranjo dos objetos que se buscava identificar, como, por exemplo: uma foto da classe banana pode ter uma única banana e um fundo branco, já em outra, pode-se ter vários cachos de banana com uma floresta de fundo, o que acabaria gerando representações bem diferentes para cada uma dessas imagens mencionadas.

Conclui-se, com essa pesquisa, que a métrica da distância de compressão é capaz de produzir bons resultados, desde que se esteja fazendo uso de um modelo de representação adequado para o tipo de dado a se tratar. O modelo de representação genérico utilizado nesse trabalho, foi capaz de gerar árvores filogenéticas H próximas da expectativa E_x para alguns conjuntos de dados.

6.3 Trabalhos futuros

Os trabalhos futuros podem ser descritos da seguinte forma:

- realizar a análise de componentes principais de um objeto do tipo imagem a partir da implementação de transformadas como a transformada discreta de cosseno (TDC);
- pesquisar abstrações diferentes da máquina de Turing (MT), como, por exemplo, *cálculo Lambda*.

6.4 Considerações finais

A compressão foi utilizada como processo para a representação e medição da informação dos objetos de dados. Foram mensurados os valores redundantes dos grânulos de informação, correspondentes de cada objeto de dados, para objetos não-estruturados. Etapas posteriores encarregaram-se de relacionar entre si, cada objeto, visando análises

que separaram os diferentes e agregam os semelhantes. A estratégia, para o modelo de representação dos objetos, teve por base a proximidade dos conceitos de compressão com aleatoriedade e incerteza.

Outras métricas em teoria da informação como a informação mútua e a informação condicional, foram utilizadas, na pesquisa, para a análise das relações emergentes entre cada objeto, concepção do contexto de agrupamento e análise dos agrupamentos formados.

Como contribuição, o presente trabalho agregou um modelo flexível de arquitetura da informação capaz de representar e agrupar objetos de dados não-estruturados utilizando conceitos desencapsulados dos processos de compressão, tendo como destaque:

- a parametrização do tipo de codificação através de um projeto de representação dos objetos por unidades semânticas de informação chamados de grânulos;
- a concepção e a análise de relações emergentes entre os objetos de dados derivadas de diversos projetos de granulação e da configuração de dependência da informação de cada objeto.

Todas as características listadas possibilitam flexibilidade, na concepção de sistemas de reconhecimento de padrões, com base na adoção de um modelo que trata fontes de dados de diferentes estruturas, mantendo a capacidade de generalização do modelo.

Em suma, observa-se que os resultados exibidos, nesta tese, demostraram que ajustes realizados nos algoritmos de compressão dos compactadores disponíveis, no mercado, associados a divisão da abordagem de compressão em *Granulação* e *Codificação* geram representações mais próximas das expectativas de contextos semânticos exigidos em cenários e modelos de Reconhecimento de Padrões como a Análise de Agrupamentos e a Aprendizagem não-supervisionada.

Referências Bibliográficas

[1] AGOSTINI, L.; SILVA, I.; BAMPI, S. **Projeto de arquitetura de codificador de entropia para a compressão jpeg de imagens em tons de cinza**. In: *VIII Workshop IBERCHIP*, 2002.

[2] AHUMADA JR, A. J.; PETERSON, H. A. **Luminance-model-based dct quantization for color image compression**. In: *Human vision, visual processing, and digital display III*, volume 1666, p. 365–374. International Society for Optics and Photonics, 1992.

[3] ALENCAR, B.; BARROSO, L. C.; ABREU, J. **Análise multivariada de dados no tratamento da informação espacial: uma abordagem com análise de agrupamentos**. *Sistemas, Cibernética e Informática*, 10(2):7–11, 2013.

[4] AMATO, A.; DI LECCE, V. **A knowledge based approach for a fast image retrieval system**. *Image and Vision Computing*, p. 1466–1480, 2008.

[5] APPEL, A. W. **An efficient program for many-body simulation**. *SIAM Journal on Scientific and Statistical Computing*, 6(1):85–103, 1985.

[6] APRIGIO, P.; PANEK, L. **O teorema de shannon para codificação com ruído**. In: *Proceeding Series of the Brazilian Society of Computational and Applied Mathematics*, volume 6, 2018.

[7] ARAÚJO, F. N. C. D. **Rotulação automática de clusters baseados em análise de filogenias**. Master's thesis, Universidade Federal do Piauí, 2018.

[8] BARDLN, L. **Análise de conteúdo**. Almedina (Edições 70), 1977.

[9] BARTOLI, A.; DE LORENZO, A.; MEDVET, E.; TARLAO, F. **Regex-based entity extraction with active learning and genetic programming**. *SIGAPP Appl. Comput. Rev.*, 16(2):7–15, Aug. 2016.

[10] BENEDETTO, D.; CAGLIOTI, E.; LORETO, V. **Language trees and zipping**. *Physical Review Letters*, p. 4, 2002.

[11] BENNETT, C. H.; GÁCS, P.; LI, M.; VITÁNYI, P. M.; ZUREK, W. H. **Information distance**. *IEEE Transactions on information theory*, 44(4):1407–1423, 1998.

[12] C., R. **Statistical Inference Through Data Compression**. Institute for Logic, Language and Computation (ILLC) publications, Dissertation Series, 1th edition, 2007.

[13] CAMBRIDGE, A. L. **The orl database of faces**, 1994.

[14] CAMPANA, B. J. L.; KEOGH, E. J. **A compression-based distance measure for texture**. *Statistical Analysis and Data Mining*, p. 381–398, 2010.

[15] CAMPANI, C. A. P. **Avaliação da Compressão de Dados e da Qualidade de Imagem em Modelos de Animação Gráfica para Web: uma nova abordagem baseada em Complexidade de Kolmogorov**. UFRGS - Tese, 2005.

[16] CAMPANI, C. A. P.; MENEZES, P. F. B. **Teorias da aleatoriedade**. *Revista de informática teórica e aplicada. Porto Alegre, RS. Vol. 11, n. 2 (dez. 2004)*, p. 86–92, 2004.

[17] CAMPANI, C. A.; MENEZES, P. B. **Aplicação da complexidade de kolmogorov na caracterização e avaliação de modelos computacionais e sistemas complexos**. In: *5th Workshop on Formal Methods*, p. 100–112, 2002.

[18] CASTRO, J. M. D.; MESQUITA, I. **Estudo das implicações do espaço ofensivo nas características do ataque no voleibol masculino de elite**. *Revista Portuguesa de Ciências do Desporto*, 8(1):114–125, 2008.

[19] CERRA, D.; DATCU, M. **A fast compression-based similarity measure with applications to content-based image retrieval**. *Journal of Visual Communication and Image Representation*, 23(2):293–302, 2011.

[20] CESAR, B. K. M. **Estudo e extensão da metodologia DAMICORE para tarefas de classificação**. PhD thesis, Universidade de São Paulo, 2016.

[21] CHUNG, K.-L.; HUANG, Y.-L.; LIU, Y.-W. **Efficient algorithms for coding hilbert curve of arbitrary-sized image and application to window query**. *Information sciences*, 177(10):2130–2151, 2007.

[22] CILIBRASI, R.; VITÁNYI, P. M. **Clustering by compression**. *IEEE Transactions on Information theory*, p. 1523–1545, 2005.

[23] CILIBRASI, R. L.; VITÁNYI, P. M. **A fast quartet tree heuristic for hierarchical clustering**. *Pattern recognition*, 44(3):662–677, 2011.

[24] COHEN, A. R.; VITÁNYI, P. M. **Normalized compression distance of multisets with applications**. *IEEE transactions on pattern analysis and machine intelligence*, 37(8):1602–1614, 2014.

[25] DAVIDSON, R.; SULLIVANT, S. **Distance-based phylogenetic methods around a polytomy**. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, 11(2):325–335, mar 2014.

[26] DE LAS PALMAS DE GRAN CANARIA, U. **Explicación del sistema de compresión jpeg**, 2014.

[27] DIVERIO, T. A.; MENEZES, P. B. **Teoria da Computação–UFRGS: Máquinas Universais e Computabilidade**. Bookman Editora, 2009.

[28] DO ESPÍRITO SANTO, R. **Utilização da análise de componentes principais na compressão de imagens digitais**. *Instituto do Cérebro (INCE), Hospital Israelita Albert Einstein, São Paulo (SP)*, 10:135 – 139, 2012.

[29] EBRAHIM, Y.; AHMED, M.; ABDELSALAM, W.; CHAU, S.-C. **Shape representation and description using the hilbert curve**. *Pattern Recognition Letters*, 30:348–358, 2009.

[30] FALCIDIENO, B.; SPAGNUOLO, M. **A shape abstraction paradigm for modelling geometry and semantics**. In: *Proceedings. Computer Graphics International (Cat. No. 98EX149)*, p. 646–656. IEEE, 1998.

[31] FREI, F. **Introdução À Análise De Agrupamentos**. Editora Unesp, 2006.

[32] GARCIA, E. K.; FELDMAN, S.; GUPTA, M. R.; SRIVASTAVA, S. **Completely lazy learning**. *IEEE Transactions on Knowledge and Data Engineering*, p. 1274–1285, 2010.

[33] HARTLEY, R. **Transmission of information**. *Bell System Technical Journal*, p. 536–560, 1928.

[34] HEIDEMANN, G.; RITTER, H. **Data compression-a generic principle of pattern recognition**. In: *International Conference on Computer Vision and Computer Graphics*, p. 202–212. Springer, 2009.

[35] JAIN, A. K.; DUIN, R. P. W.; JIANCHANG MAO. **Statistical pattern recognition: a review**. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1):4–37, Jan 2000.

[36] JAIN, A. K.; DUIN, R. P. W.; JIANCHANG MAO. **Statistical pattern recognition: a review**. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1):4–37, Jan 2000.

[37] KEOGH, E.; LONARDI, S.; RATANAMAHATANA, C. A.; WEI, L.; LEE, S.-H.; HANDLEY, J. **Compression-based data mining of sequential data**. *Data Mining and Knowledge Discovery*, 14(1):99–129, 2007.

[38] KOLMOGOROV, A. N. **Three approaches to the quantitative definition of information**. *Problemy Peredachi Informatsi*, 1(1):3–1, 1965.

[39] LAM, H. T.; MÖRCHEN, F.; FRADKIN, D.; CALDERS, T. **Mining compressing sequential patterns**. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 7(1):34–52, 2014.

[40] LECUN, Y.; HUANG, F. J.; BOTTOU, L. **Learning methods for generic object recognition with invariance to pose and lighting**. In: *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, volume 2, p. II–104. IEEE, 2004.

[41] LI, M.; CHEN, X.; LI, X.; MA, B.; VITÁNYI, P. M. **The similarity metric**. *IEEE transactions on Information Theory*, 50(12):3250–3264, 2004.

[42] LINDEN, R. **Técnicas de agrupamento**. *Revista de Sistemas de Informação da FSMA*, 4:18–36, 2009.

[43] LOPES, G. R.; DELBEM, A. C. **Classificaç ao de tráfego de ataques em redes de computadores através de técnicas de mineraç ao de dados**, 2019.

[44] LOPES, L. A.; MACHADO, V. P.; RABÊLO, R. A.; FERNANDES, R. A.; LIMA, B. V. **Automatic labelling of clusters of discrete and continuous data with supervised machine learning**. *Knowledge-Based Systems*, p. 231–241, 2016.

[45] LOUDEN, K. C. **Compiladores-Princípios e Práticas**. Cengage Learning Editores, 2004.

[46] MANZINI, G. **An analysis of the burrows—wheeler transform**. *Journal of the ACM (JACM)*, 48(3):407–430, 2001.

[47] MEDEIROS, C.; COSTA, J. A. F. **Uma comparação empírica de métodos de redução de dimensionalidade aplicados a visualização de dados**. *Learning and Nonlinear Models-Revista da Sociedade Brasileira de Redes Neurais (SBRN)*, 6(2):81–110, 2008.

[48] MEILA, M. **Comparing clusterings—an information based distance**. *Journal of multivariate analysis*, 98(5):873–895, 2007.

[49] MENEZES, P. B. **Linguagens Formais e Autômatos: Volume 3 da Série Livros Didáticos Informática UFRGS**. Bookman Editora, 2009.

[50] MENEZES, P. B. **Linguagens formais e autômatos**. Bookman Editora, 6 edition, 2010.

[51] MERIVUORI, T.; ROOS, T. **Some observations on the applicability of normalized compression distance to stemmatology**. In: *Proceedings of 2nd Workshop on Information Theoretic Methods in Science and Engineering*, 2009.

[52] MORO, L. F. D. S.; RODRIGUES, C. L.; ANDRADE, F. R.; DELBEM, A. C.; ISOTANI, S. **Caracterização de alunos em ambientes de ensino online: Estendendo o uso da damicore para minerar dados educacionais.** In: *Anais dos Workshops do Congresso Brasileiro de Informática na Educação*, volume 3, p. 631, 2014.

[53] NELSON, M.; GAILLY, J.-L. **The data compression book**. M & t Books New York, 1996.

[54] NEWMAN, M. E.; GIRVAN, M. **Finding and evaluating community structure in networks**. *Physical review E*, 69(2):1–15, 2004.

[55] NIKVAND, N. **Image information distance analysis and applications**. Master's thesis, University of Waterloo, 2014.

[56] PEREIRA, F. **Estudo das interacções entre evolução e aprendizagem em ambientes de computação evolucionária**. PhD thesis, PhD Thesis, Universidade de Coimbra, 2002.

[57] PIMENTEL, B.; ARRAIS, J. **Implementação de algoritmo de compressão e descompressão de dados para modelo de co-processamento baseado em fpga's**. *Electrônica e Telecomunicações*, 4(2):215–220, 2004.

[58] PINEDA, J. O. D. C. **A entropia segundo claude shannon: o desenvolvimento do conceito fundamental da teoria da informação**. Mestrado em história da ciência, Pontifícia Universidade Católica. São Paulo, 2006.

[59] PINHO, A. J.; FERREIRA, P. J. **Image similarity using the normalized compression distance based on finite context models**. In: *2011 18th IEEE International Conference on Image Processing*, p. 1993–1996, 2011.

[60] PINHO, A. J.; PRATAS, D.; FERREIRA, P. J. **A new compressor for measuring distances among images**. In: *International Conference Image Analysis and Recognition*, p. 30–37. Springer, 2014.

[61] PINHO, A. J.; PRATAS, D.; FERREIRA, P. J. **Authorship attribution using relative compression**. In: *Data Compression Conference Proceedings*, p. 329–338. IEEE, 2016.

[62] PINTO, M. C. **Um algoritmo para comparação sintática de genomas baseado na complexidade condicional de kolmogorov**. Mestrado em ciência da computação, Unicamp, Campinas - SP, 2002.

[63] PINTO, R. S.; DELBEM, A. C.; MONACO, F. J. **Caracterização do perfil de carga a partir de programas binários**. In: *ERAD SP 2017*, Porto Alegre, RS, Brasil, 2017. SBC.

[64] RAO, M. S.; REDDY, B. E. **Comparative analysis of pattern recognition methods: An overview**. *Indian Journal of Computer Science and Engineering (IJCSE)*, 2(3):385–390, 2011.

[65] REGHBATI, H. K. **Special feature an overview of data compression techniques**. *Computer Magazine, IEEE*, 1(4):71–75, 1981.

[66] RENZINI, A.; GREGGIO, L.; RITOSSA, C.; FERRARIO, L. **Why stars inflate to and deflate from red giant dimensions**. *The Astrophysical Journal*, 400:280–303, 1992.

[67] RITA, L. R. V.; DE SOUZA, G. S.; DE OLIVEIRA REOLON, L.; NICOLEIT, E. R. **Compactador de arquivos utilizando algoritmo de huffman**. *Anais SULCOMP*, 6, 2013.

[68] RONQUIST, F.; HUELSENBECK, J. P. **Mrbayes 3: Bayesian phylogenetic inference under mixed models**. *Bioinformatics*, 19(12):1572–1574, 2003.

[69] ROSENBERG, A.; HIRSCHBERG, J. **V-measure: A conditional entropy-based external cluster evaluation measure**. In: *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*, p. 410–418, 2007.

[70] ROUSSEEUW, P. J. **Silhouettes: a graphical aid to the interpretation and validation of cluster analysis**. *Journal of computational and applied mathematics*, 20:53–65, 1987.

[71] SADALAGE, P. J.; FOWLER, M. **NoSQL Essencial: Um guia conciso para o Mundo emergente da persistência poliglota**. Novatec Editora, 2013.

[72] SAITOU, N.; NEI, M. **The neighbor-joining method: a new method for reconstructing phylogenetic trees.** *Molecular biology and evolution*, 4(4):406–425, 1987.

[73] SALOMON, D. **Data compression: the complete reference**. Springer Science & Business Media, 2004.

[74] SANCHES, A. M. **Um estudo de compactação de dados, com a implementação do método de lzw**. *UEMS, Ciência da Computação*, 2001.

[75] SANCHES, A.; CARDOSO, J. M.; DELBEM, A. C. **Identifying merge-beneficial software kernels for hardware implementation**. In: *2011 International Conference on Reconfigurable Computing and FPGAs*, p. 74–79. IEEE, 2011.

[76] SANTOS, J. M.; EMBRECHTS, M. **On the use of the adjusted rand index as a metric for evaluating supervised classification**. In: *International Conference on Artificial Neural Networks*, p. 175–184. Springer, 2009.

[77] SANTOS, Z. T. S. D. **Ensino de entropia: um enfoque histórico e epistemológico**. Doutorado em educação, Universidade Federal do Rio Grande do Norte, 2010.

[78] SAYOOD, K. **Introduction to data compression**. Morgan Kaufmann, 2017.

[79] SHANNON, C. E. **A mathematical theory of communication**. *ACM SIGMOBILE mobile computing and communications review*, 5(1):3–55, 2001.

[80] SILVA, J. J.; SALVINI, R. **Análise de sentimentos de conteúdos textuais de redes sociais por meio de modelos de compressão de dados**. In: *Anais da VI Escola Regional de Informática de Goiás*, p. 107–120. SBC, 2018.

[81] SIPSER, M. **Introduction to the theory of computation**. *SIGACT News*, 27(1):111–129, mar 1996.

[82] STRIMMER, K.; VON HAESELER, A. **Quartet puzzling: a quartet maximum-likelihood method for reconstructing tree topologies**. *Molecular Biology and Evolution*, 13(7):964–969, 1996.

[83] THEODORIDIS, S.; KOUTROUMBAS, K. **Pattern Recognition, Fourth Edition**. Academic Press, 1th edition, 2008.

[84] THEODORIDIS, S.; KOUTROUMBAS, K. **"Nonlinear classifiers", Pattern Recognition**. Academic Press, 2009.

[85] TORRES, P. H. L. **Utilizando Árvores filogenÉticas para a identificaÇÃo de similaridades em pinturas digitalizadas**, 2018.

[86] VALVERDE, M. A. G. **Geração de redes vasculares sintéticas tridimensionais utilizando sistemas de Lindenmayer estocásticos e parametrizados**. PhD thesis, Universidade de São Paulo, 2012.

[87] VERLEYSEN, M.; FRANÇOIS, D. **The curse of dimensionality in data mining and time series prediction**. In: *International Work-Conference on Artificial Neural Networks*, p. 758–770. Springer, 2005.

[88] VIGO, A. **Modeling common outcomes: bias and precision**. *Cadernos de saúde publica*, 22(11):2496–2496, 2006.

[89] VINH, N. X.; EPPS, J.; BAILEY, J. **Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance**. *Journal of Machine Learning Research*, 11(Oct):2837–2854, 2010.

[90] WILDING, J. **Perception: From Sense to Object**. Psychology Library Editions: Perception. Taylor & Francis, 2017.

[91] XU, D.-H.; KURANI, A. S.; FURST, J. D.; RAICU, D. S. **Run-length encoding for volumetric texture**. *Heart*, 27(25), 2004.

[92] YE, J.; JANARDAN, R.; LI, Q. **Gpca: An efficient dimension reduction scheme for image compression and retrieval**. In: *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, p. 354–363, New York, NY, USA, 2004. ACM.